WanderGuide: Indoor Map-less Robotic Guide for Exploration by Blind People

Masaki Kuribayashi
Waseda University
Tokyo, Japan
Miraikan - The National Museum of Emerging Science and Innovation
Tokyo, Japan
rugbykuribayashi@toki.waseda.jp

Kohei Uehara
Miraikan - The National Museum of Emerging Science and Innovation
Tokyo, Japan
kouhei.uehara@jst.go.jp

Allan Wang
Miraikan - The National Museum of Emerging Science and Innovation
Tokyo, Japan
allan.wang@jst.go.jp

Shigeo Morishima
Waseda Research Institute for Science and Engineering
Tokyo, Japan
shigeo@waseda.jp

Chieko Asakawa
Miraikan - The National Museum of Emerging Science and Innovation
Tokyo, Japan
chieko.asakawa@jst.go.jp

Figure 1: Five core functionalities of WanderGuide. The system assists users in recreational exploration by explaining the surrounding environment through images obtained from the robot's camera. Users can adjust the level of detail and ask questions about their surroundings. Additionally, the system can guide users to locations they have visited before.

Abstract

Blind people have limited opportunities to explore an environment based on their interests. While existing navigation systems could provide them with surrounding information while navigating, they have limited scalability as they require preparing prebuilt maps. Thus, to develop a map-less robot that assists blind people in exploring, we first conducted a study with ten blind participants at a shopping mall and science museum to investigate the requirements of the system, which revealed the need for three levels of detail to describe the surroundings based on users' preferences. Then, we developed WanderGuide, with functionalities that allow users to adjust the level of detail in descriptions and verbally interact with the system to ask questions about the environment or to go to points of interest. The study with five blind participants revealed that WanderGuide could provide blind people with the enjoyable experience of wandering around without a specific destination in their minds.

CCS Concepts

• **Human-centered computing** → **Accessibility systems and tools**.

Keywords

visual impairment, map-less navigation, recreational exploration

ACM Reference Format:
Masaki Kuribayashi, Kohei Uehara, Allan Wang, Shigeo Morishima, and Chieko Asakawa. 2025. WanderGuide: Indoor Map-less Robotic Guide for Exploration by Blind People. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26-May 1, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3706598.3713788

CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713788

Table 1: Comparison to previous work. Our work explores a unique characteristic that has not been investigated in the past. In the Purpose row, "Navigation" refers to systems primarily designed to guide users to their intended destinations, while "Perception" refers to those focused on understanding the surrounding environment. "Multi-purpose" refers to systems capable of performing various tasks, and "Exploration" refers to those designed for navigating and enjoying facilities, characterized by constantly discovering and changing goals (*e.g.*, window-shopping [35, 37]).

System	Map-less	Purpose	Independent from Human Assistance	Device
NavCog [72]	✗	Navigation	✓	Smartphone
CaBot [24]	✗	Navigation	✓	**Robot**
Tactile Compass [51]	✗	Navigation	✓	Handheld Device
ChitChatGuide [37]	✗	**Exploration**	✓	Smartphone
Kayukawa *et al.* [39]	✗	**Exploration**	✗	**Robot**
Corridor-Walker [46]	✓	Navigation	✓	Smartphone
Snap&Nav [43]	✓	Navigation	✗	Smartphone
PathFinder [45]	✓	Navigation	✗	**Robot**
WorldScribe [13]	✓	Perception	✓	Smartphone
GPT-4o Demo [62]	✓	Perception	✓	Smartphone
MLLM Powered Applications (*e.g.*, Seeing AI [56])	✓	Perception	✓	Smartphone
RSA [1, 7]	✓	Multi-Purpose	✗	Smartphone
WanderGuide	✓	**Exploration**	✓	**Robot**

1 Introduction

Exploration is a fundamental skill that allows one to gain familiarity with novel environments that blind people do not know. Sighted people explore by visually perceiving points of interest (POI) and navigating to desirable destinations. However, blind people face significant challenges in independently exploring new environments [17, 58]. They typically rely on sighted assistants, such as friends or family members, to help them navigate and describe their surroundings. Unfortunately, these assistants are not always readily available, resulting in limited opportunities for blind people to explore independently.

Over the past recent years, various guide systems [40, 44, 54], that are aimed for navigation [49, 72] or exploration [37, 39], have been developed to guide blind people and provide details about surrounding POIs in the environment. These systems typically rely on prebuilt maps and localization infrastructure (*e.g.*, Bluetooth Low Energy (BLE) beacons [14, 30, 41, 59, 72] and ultrawide-bandwidth beacons [53]) that are highly customized to the environments to continually update their current locations and offer turn-by-turn navigation guidance. As access to the prebuilt maps also allows these systems to convey information about nearby POIs while navigating, some systems are specialized in assisting exploration activity [37, 39]. However, only a limited number of guide systems (*e.g.*, InclusiveNavi [30] and BlindSquare [8]) are publicly deployed because configuring and maintaining prebuilt maps and localization infrastructure is expensive, and it is infeasible for them to be deployed in unseen environments. Several systems that do not require maps, as well as remote sighted assistance (RSA) [1, 7, 36], have been developed to guide blind people in various locations [19, 45–47]. However, these systems primarily focus on navigation to target destinations, not exploration, thus providing only navigation-related information to users (*e.g.*, intersections [45, 46] and signs [45]). These systems are also not independent from human assistance. To promote social inclusion and equality for blind people, there is a need to develop a *map-less* guide system that assists blind people in exploring diverse novel locations without relying on prebuilt maps or infrastructure.

To bridge the gaps and address the shortcomings of existing systems, we developed a system with the following characteristics as shown in Table 1: 1. Our system does not rely on prebuilt maps or preinstalled infrastructure. 2. Our system focuses on exploration. 3. Our system does not require supplementary assistance from humans. 4. Our system can automatically guide users physically during exploration. None of the prior systems possess the combination of all these characteristics. Given that the design space for a map-less exploration guide robot remains underexplored, this work aims to investigate and establish the key components of such a system. We begin by selecting a wheeled robot platform as the device. The decision to use a wheeled robot is based on its ability to autonomously guide blind users. It alleviates the challenge of navigation, which is cognitively demanding while learning about the surrounding environment. Additionally, we equip the robot system with the ability to convey real-time information about the surrounding environment to users using natural language, accomplished through a multimodal large language model (MLLM [62]).

Using our prototype system, we employed an iterative process with the direct involvement of target users to develop our system. In the formative study, the participants were asked to follow the robot, which was controlled in a Wizard-of-Oz fashion [68], along predetermined routes while listening to the environment descriptions. The study revealed three groups of user preferences in the

system's descriptions with respect to varying levels of details in the descriptive information received. It also revealed requirements in certain functionalities, such as revisiting locations where the system had mentioned, specifying directions to proceed, and obtaining in-depth information through question-and-answering (Q&A) functionality.

In the second stage, taking the lessons learned from the first study, we present *WanderGuide*, a map-less exploration system for blind people (Fig. 1). Taking into consideration the previously discovered three groups of user preferences, the system offers three modes for describing the surroundings: (1) Detailed description — in-depth information with high granularity, (2) Balanced-Length description — balanced level of information, and (3) Concise description — minimal but essential details for obtaining quick awareness. We also implemented various new features based on the feedback received from the first study, which includes adopting a high-resolution fisheye camera for better perception of the surrounding environment, allowing users to verbally interact to query about the environment and set explored POIs to be navigation destinations, and allowing users to use directional buttons to control the robot for navigation towards the direction of interest. Our system is also fully integrated with the automatic mapping, localization, map-less navigation, and obstacle avoidance functions of the wheeled mobile robot.

Finally, we conducted a main user study with five blind participants, who were asked to freely explore two floors of the science museum. All participants appreciated the experience of wandering freely without a fixed destination, and they expressed their desire to use the system to explore both familiar and unfamiliar areas. Participants also highlighted the need to incorporate recognition of auditory cues from the environment. Additionally, differences in how they interacted with the system were observed: one frequently used buttons to guide the robot towards their areas of interest, one passively followed the robot, and others often asked questions. We also identified a limitation in the system's MLLM when conveying detailed information about the surroundings, such as identifying specific names of objects, which suggests the need for further development in how we input information into the MLLM for exploration purposes.

To the best of our knowledge, our work is the first to investigate the design space of a map-less system for blind people to explore independently. To this end, we made the following contributions.

- We formulated the requirements for the system through a formative study, such as the ability to adjust the level of description based on user preferences and to guide users to previously visited locations of interest, thereby enhancing the exploration experience.
- We developed a full stack map-less exploration system that consists of a waypoint detection algorithm and an MLLM-based perception interaction system on top of an existing navigation guide robot. Additionally, we integrated several functionalities based on the formative study that facilitates the exploration experience.
- We confirmed key design requirements, such as varying the level of descriptions based on user preferences through a usability study. We also gained further insights into users' interaction preferences and into design implications for improving the system, including better recognition of audio cues.

The codes of the system are publicly available in the following link: https://github.com/chestnutforestlabo/WanderGuide.

2 Related Work

2.1 Exploration for Blind People

Previous research has emphasized the importance of exploration for blind people to familiarize themselves with the environment [34] or for enjoying recreational areas where exploration is essential (*e.g.*, museums [39] or shopping malls [35]). The investigation by Engel *et al.* [17] shows that 59.4% of the blind population in the study travels to unfamiliar buildings several times a week, but often cannot explore independently, because they rely on sighted assistants with limited availability. While learning routes and POIs in a building could also be achieved by searching online [17], using interactive maps [60, 64, 71, 79] or applications [22, 31], on-site exploration by blind people is also important because by doing this, they receive rich sensory information and gain a better sense of independence [35]. This overall experience motivates them to explore by themselves [23, 35]. Focusing on on-site exploration, researchers have investigated the information needs of blind people [5, 28, 34, 35, 80], and found that it is essential to include high-level understanding of the environment (*e.g.*, layout information [34]) as well as specific details such as the names of the shops and brands [5]. Additionally, researchers noted that safety during navigation is crucial, as safety concerns can dominate the cognitive load and impede the rich exploration experience [11, 34, 85].

2.2 Assistance Systems for Blind People To Explore

Robotic guide systems have the advantage of addressing the mobility challenges of blind people with their automatic guidance capability. CaBot [24], the first guidance robot that adopted the form of a suitcase, guides users to specified destinations while referring to prebuilt maps or using an object detector to convey surrounding information. Among them, some are specialized in exploration [2, 3, 39]. A robot system by Kayukawa *et al.* [39] allows users to explore by interactively setting destinations on a smartphone and by calling a museum guide to explain the surroundings. However, both systems heavily rely on prebuilt maps and operate in limited locations where the destinations are readily available. Ultimately, our goal is to develop a system that does not require prebuilt maps and enables blind people to explore independently, *i.e.*, without relying on staff assistance within the facility.

Navigation systems for blind people that do not rely on prebuilt maps and infrastructure, *i.e.*, *map-less navigation systems*, have also been proposed in prior research. Besides real-time perception outcomes, these systems primarily depend on externally sourced route information, such as prior route knowledge from blind users [46, 47], routes described by nearby pedestrians [42, 45, 67], and analyzed images of floor maps captured in buildings [43]. For example, PathFinder [45] is a map-less navigation robot system designed to guide blind users to their destinations based on predefined routes.

The system autonomously navigates users by utilizing an intersection detection algorithm [83] and a sign recognition algorithm [45]. These algorithms enable users to determine the correct direction to proceed at key decision points. The system's evaluation found that it is necessary to include functionality that takes users back to their starting location after reaching their destination when navigating unfamiliar buildings. However, these map-less navigation systems are tasked with reaching a specific destination and are not suitable for exploration, as they only focus on providing information related to reaching the destination (*e.g.*, intersections and signs [45]). In an exploration scenario, any information about the environment may prove valuable, such as layout information [34]. In our study, we aim to explore the underexplored design space of map-less guide robots for exploration purposes, such as how the system should describe surroundings and what are the task-specific functionality requirements.

2.3 Autonomy and Control Methods of Assistant Systems

Researchers have emphasized autonomy, *i.e.*, the ability for users to select destinations and routes according to their interests, as an important factor for exploratory activities [37, 38]. To this end, researchers have investigated various control methods based on user inputs [67, 85]. For example, systems with prebuilt maps adopted selecting destinations from a premade list of stores [38, 72], or via conversation [37, 72]. In the case of map-less systems, researchers have adopted feedback-based closed-loop processes to leverage both human inputs and system control. Examples include users specifying proceeding directions at intersections while the robot provides automated guidance to the next intersections [29, 34, 45, 47, 67]. Zhang *et al.* [85] reported that the preferred level of control by blind people may vary depending on context. Therefore, we examine the level of control between users and robots under the novel exploration context.

2.4 Scene Description for Blind People

Knowledge of surrounding information is crucial for blind people to explore [37]. Tools for blind people to understand their surrounding environment have been commercially deployed and the topic remains an ongoing area of research. While researchers have proposed tools using visual captioning [70] and question-answering models [25] to help blind users understand scenes, these often fail to provide accurate descriptions at diverse locations [16]. Alternatively, RSA applications (*e.g.*, Aira [1] and BeMyEyes [7]) have long been a practical aid for providing blind people with surrounding scenes. However, RSA systems are not suitable for our task, as the service quality depends heavily on the sighted assistance provided [36]. RSA services may also not be sustainable for extended use until users feel fully satisfied with their exploration experience. With the emergence of LLMs and MLLMs, scene describing systems (*e.g.*, Seeing AI [56], BeMyAI [6] and GPT4o-demo [62]) have been developed, which enable blind people to understand scenes in diverse scenarios [21, 81]. ChitChatGuide [37] employs LLMs to interpret predetermined maps and deliver exploration-related information during navigation to a specified destination. However, unlike our system, it relies on prebuilt maps and lacks the capability to provide real-time information. MLLM-based systems, such as WorldScribe [13], offer real-time information by analyzing captured images. WorldScribe [13] also adapts the level of description based on user context, such as the speed at which the device is moved. Our WanderGuide similarly provides three levels of descriptions but the selection is adjusted based on individual user preferences rather than situational context. On the system level, the core distinction is that WanderGuide combines MLLM with a navigation robot, allowing users to concentrate fully on the descriptions of novel environments and navigate to interested places. This combination makes WanderGuide particularly well-suited for *exploration while navigating*.

3 System Design Focus

Our goal is to finalize a system that assists blind people in exploring an indoor environment independently. In this section, we describe the key design elements of the system.

3.1 Device

Assistance systems for blind people have been proposed in various devices, such as smartphones [65], handheld haptic devices [15, 51, 75], wearable devices [49], cane-like devices [67] and robots [52]. Each type of device offers unique advantages - Smartphones and handheld haptic devices are portable; Smartphones are also widely used by blind people [55, 57]; Wearable devices free the user's hands [48]; Cane-like devices resemble traditional canes [67]; And robots are able to autonomously guide users [24]. While handheld devices [15, 51, 65, 67, 75] have often been used due to their portability, in exploration scenarios, they require users to point the devices in their directions of interest while navigating around unfamiliar locations and obstacles, which involve high cognitive load. Thus, we chose robots because of their autonomous navigation and obstacle avoidance capabilities. This allows users to concentrate on learning the environment [11, 34, 85]. In particular, we adopted a wheeled robot [24, 78, 85]. While wheeled robots are unable to navigate stairs like quadruped robots [11], blind users often find wheeled robots more suitable due to their silence and stability [78]. Our assumption is that the devices should ensure the users' safety during navigation and allow users to focus on exploration. As a result, the findings in our study can be extended to any similar devices other than wheeled robots.

3.2 Describing Scenes

Previous navigation systems relied on hardcoded information [37, 72] or simple image captioning models [70] to provide scene descriptions. They only convey information related to navigating to destinations. In exploratory tasks, any information and details could be relevant. Therefore, we decided to use MLLM, a foundational model capable of recognizing a variety of objects and describing them in natural language. We injected MLLM into the system to periodically provide comprehensive information about the surroundings to inform blind users during exploration. In this paper, we investigate the appropriate presentation format, such as content types and lengths, and the quality of the responses from MLLMs through our user studies.

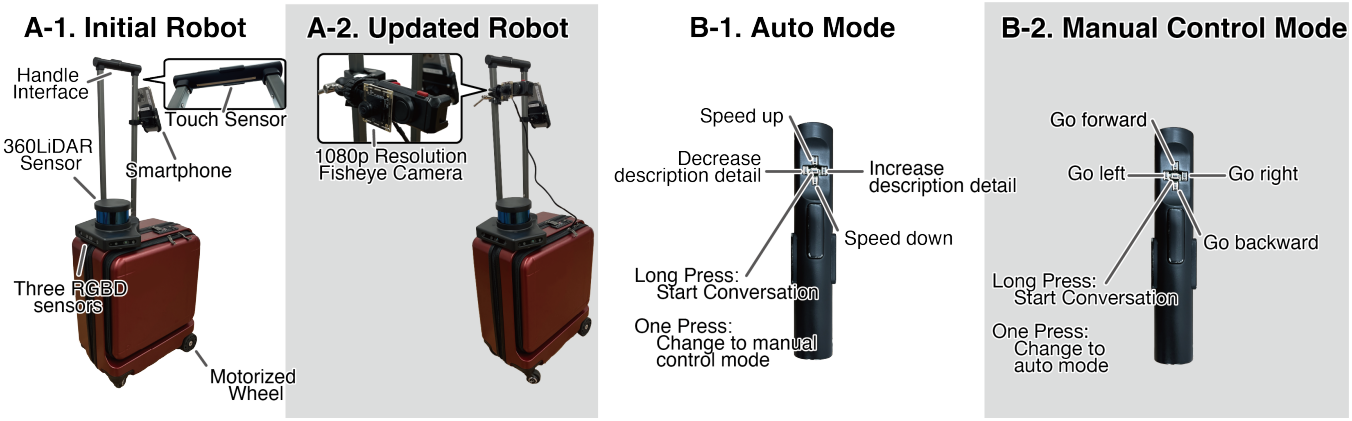

Figure 2: Image of the robot and handle interface used in the study. Panel A-1 shows the robot used in the formative study, while Panel A-2 presents the robot used in the main study. Panels B-1 and B-2 illustrate the mapping of the handle interface buttons' functions, depending on the selected navigation mode.

3.3 Interaction

The ability for users to select destinations and routes according to their interests, often referred to as autonomy, is particularly important for exploration [37, 38]. In our system, to what extent users prefer to take control over the robot (*i.e.*, interaction) remains unknown. Based on the scene descriptions given by the system [37], some blind users may fully embrace letting the robot guide them automatically, while others may prefer to decide which way to go on their own. Additionally, this preference may also be influenced by the robot's descriptions of the scenes. Given that the extent of user preference for autonomy remains unclear, we first conducted the formative study (Sec. 4) to explore the requirement of autonomy based on interaction needs. Then, we conducted a full study (Sec. 6) to evaluate the users' opinions on autonomy in our improved system, which integrated the feedback from the formative study.

4 Formative Study

We first conducted a formative study to investigate the requirements of the system, such as how the system should explain its surroundings and what potential interactions may happen between the robot and the user. To conduct the study, we recruited ten participants through our existing email list. Interestingly, our recruitment emails were shared among blind people, eventually reaching people not on our emailing list. In the recruitment email, we specified that those who are unfamiliar with the experimental location, *i.e.*, even if they have had previous visiting experience, they do not have a clear understanding of the building or know what is there, would be eligible to participate. Tab. 2 shows the demographics of the participants. All studies in this paper have been approved by our institution's review board. Informed consent was read out to all participants in this paper and obtained from them. The study took approximately two hours, and the participants were compensated $20 per hour and reimbursed for their transit costs. Only one participant was present for each session.

4.1 Prototype System

We developed our prototype robot system according to Sec. 3. It was based on an open-source robot platform[1] and could guide users while explaining the surrounding environment. To ensure that the participants experienced the same level of autonomy, we used teleoperation, a Wizrd-of-Oz-based approach [68], to force the robot to be in full-automatic mode when guiding the participants. We adopted a suitcase-shaped wheeled robot for this study. The suitcase's appearance allows blind users to seamlessly blend into their environment, leading to higher social acceptance from users, surrounding pedestrians, and facility managers [38]. As shown in Fig. 2–A-1, the robot has a handle embedded with five buttons, a touch sensor beneath the handle, a 360° Velodyne VLP-16 LiDAR sensor [63] sensor, three RGBD cameras with resolutions of 640×360, one RealSense D455 camera [33] mounted at the front, two RealSense D435 cameras [32] on the left and right, and a pair of motorized wheels in differential drive configuration. Inside the suitcase, it has Ruby R8 powered by an AMD Ryzen R7-4800U CPU [61], and a Jetson Mate featuring multiple Jetson Xavier NX GPUs [76]. The RGBD cameras were attached 0.51 meters above the ground. The touch sensor detects whether or not the user is holding the handle and moves only when it is being held by the user. The cameras combined have a horizontal field of view of approximately 180°. The weight of the robot is approximately 15kg. We set the default speed of the robot to 0.5 meters per second to maintain a balance between a comfortable walking speed and a speed that allows sufficient time to absorb the scene description audios. A smartphone is attached to the suitcase to provide audio feedback through a neck speaker worn by users, connected via Bluetooth.

To convey the surrounding information to the participants, we used GPT-4o [62], a popular MLLM model. We inputted the images from the three RGBD cameras into the MLLM model and asked the model to generate descriptions of the surrounding environment. The robot was designed to describe surrounding information 5-10

[1] https://github.com/CMU-cabot/cabot

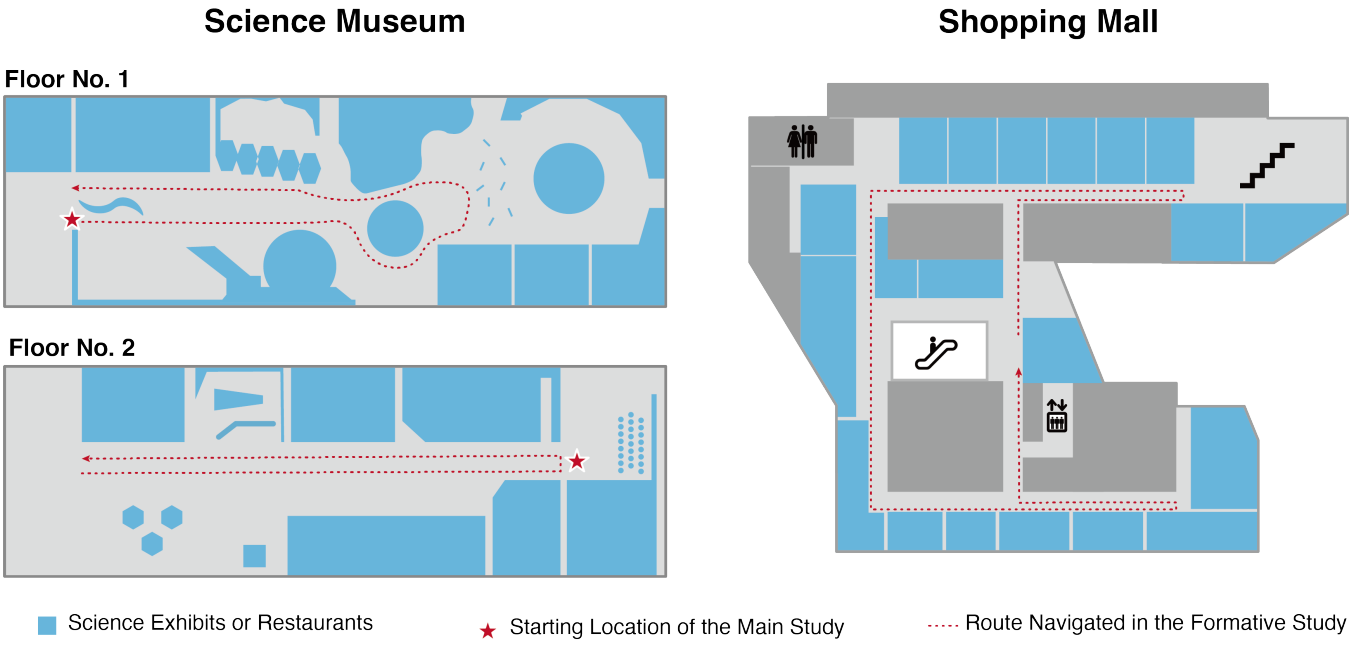

Figure 3: Floor maps of the location of the study. The left panel shows the two floors of the science museum, the fifth floor of Miraikan, which feature exhibits on various topics, such as environmental issues and space exploration. On the right panel is a floor plan of a shopping mall, the fourth floor of Toranomon Hills Station Tower, which includes a variety of restaurants offering different cuisines, including French, Japanese, Chinese, and cafes.

seconds after the end of the previous description every time. We engineered the prompts to ask the MLLM model to first provide a general overview of the scene, followed by specific details on the left, front, and right. We asked the descriptions to include as many objects as possible and incorporate layout information, such as navigable directions and the presence of walls [34]. The processing time and cost to generate a description was 6.087 seconds and $0.00740 on average. We attach the full prompts in Appendix Sec. A.2.

4.2 Experimental Location

To ensure the diversity of the findings we would obtain from this study, we conducted the study in two different locations. We chose to conduct our studies in a science museum and a shopping mall, as these are locations where people typically engage in exploration, and they have been utilized in previous research [2, 3, 37]. A museum is generally a place for learning about exhibits, while a shopping mall often requires exploration both before and during visits to stores. Specifically, we used the fifth floor of Miraikan[2] for the science museum and the fourth floor of Toranomon Hills Station Tower[3] for the shopping mall. The floor map of the science museum is illustrated in the left panel of Fig. 3, which contains two floors, both primarily featuring science exhibits. For the studies, the order of the two floors was counterbalanced. The study in the museum was conducted after business hours, during which customers were absent, but staff were present for their duties. The floor map of the shopping mall is illustrated in the right panel of Fig. 3, a floor that contains several restaurants from various countries. The study in the shopping mall was conducted during regular business hours. As shown in Tab. 2, the study with P01–P05 took place in the science museum, and the study with P06–P10 took place in the shopping mall.

[2]https://www.miraikan.jst.go.jp/en/
[3]https://www.toranomonhills.com/

4.3 Procedure

For each participant, we first conducted a pre-study interview to learn about their experience in exploring buildings, followed by an explanation that the study aimed to gather their opinions on a guide system designed to assist with exploration. Then, participants were given a task to navigate the predetermined route (red arrow of Fig. 3) guided by the robot. Adopting a Wizrd-of-Oz-based approach, an experimenter controlled the robot to navigate along the route and stop when there were nearby pedestrians. During exploration, the robot periodically generated descriptions of the scenes. We show an example of the generated description in Fig. 4. After the exploration, we asked the participants if there were any additional things they wanted to do to partially simulate the potential interaction, such as going to additional places or going around the floor again for more exploration. Finally, we conducted a post-interview session to gather their feedback on the system.

4.4 Result

4.4.1 Interests to Exploration. All participants stated that totally independent exploration is challenging, but they expressed a desire for exploration if a guide system can help them do so. For example:

Table 2: Demographics of participants who attended the formative study. The table reports their gender, age, navigation aid, which they frequently use, frequency of exploration done either independently or with sighted people per year, their experimental location, number of previous visits to the experimental location, and analyzed preference.

	Gender	Age	Aid	Age of Onset	Frequency of Exploration per Year	Experiment Location	Number of Previous Visits	Preference Analysis
P01	F	64	Cane	44	48	Science Museum	1	Exploration-Inclined
P02	M	53	Cane	13	36	Science Museum	0	Destination-Oriented
P03	M	74	Cane	0	1	Science Museum	0	Destination-Oriented
P04	F	54	Cane	0	12	Science Museum	0	Exploration-Inclined
P05	M	56	Cane	52	2	Science Museum	0	Intermediate
P06	M	32	Cane	0	12	Shopping Mall	0	Intermediate
P07	F	55	Cane	52	0	Shopping Mall	1	Exploration-Inclined
P08	M	63	Cane	22	12	Shopping Mall	0	Intermediate
P09	F	78	Guide dog	22	12	Shopping Mall	0	Destination-Oriented
P10	F	49	Cane	3	1	Shopping Mall	0	Exploration-Inclined

C1:*"I don't really explore much. I go out with a specific purpose in mind [...] The reason is that it's just too bothersome. But I do think it would be fun if I did [...] I'm more of an old-timer, so exploration never really caught my interest. It's not that I didn't care at all, but perhaps I've been living this way (not to explore).*[4] (P02)

4.4.2 Positive Feedback and Appreciated Information. Seven participants (P01, P04–P08, and P10) expressed their enjoyment while navigating with the robot, particularly with the provided surrounding descriptions, as described in the following comment: **C2**:*"My first impression was that it was a lot of fun. The reason is, as you just mentioned, unlike the person I usually walk with, the system provided detailed explanations about things like the color of the walls and the signs we saw and even described how the chef was preparing the food. Normally, you might get some of this information from others, but it's rare to get such thorough details. I found myself thinking, "Oh, I see, that's how it looks to sighted people," and I felt there was a lot of new information. In that sense, I really enjoyed it."* (P07)

Participants appreciated a variety of real-time details about their surroundings, notable examples include patterns on the walls, lighting conditions, subjective descriptors such as "beautiful," the presence and actions of nearby people, the existence of signboards, the layout of the environment, and the visibility of a chef in an open kitchen. Additionally, P10, who requested to walk around the floor again, noted that receiving different descriptions of the same location was beneficial, as it gave them a sense of presence: **C3**:*"The system mentioned those things, as well as details about the plants and wall decorations. It's like, you talked about so many different things that it feels like I was actually looking around myself. Honestly, most of the time, I get so occupied with just reaching my destination that I don't notice things around me. [...] The system also mentioned things in the second round of explanations that weren't covered in the first round, which was nice. It conveyed a sense of the ongoing atmosphere and gave a good understanding of the situation at the time."* (P10)

4.4.3 Information Needs. Participants hoped for further polishing of the delivered information about the scenes. Six participants (P01–P03, P06, and P09–P10) felt the information conveyed about the surroundings was too abstract, indicating the need for more specific information: **C4**:*"The system talked about there are just exhibits, or there's information on panels, but I think it would be nice if the system talked about specific titles. There are places where the system talked about them, but there are also places where it did not, so I found myself wondering about that."* (P01) In particular, three participants (P02, P03, and P09) commented that the descriptions neither helped them learn the environment nor make decisions such as determining which shops or exhibits to enjoy: **C5**:*"I expected it to at least tell me the name of the store, but it was disappointing to find out that it didn't do that at all. I really wish there was a system that could provide pinpointed information about what I want to know. Especially in an unfamiliar restaurant area, for example, if I come alone and use the device to enter the premises, it starts running, and then when I think, "Oh, should I have Japanese food today, or maybe tonkatsu?", without such information, I end up just walking around aimlessly."* (P09)

Participants also described specifics about what types of information would be beneficial to include, such as the position of objects given in meters and clock directions, the availability of seats, people on collision paths, identities of surrounding individuals (*e.g.*staff), and specific names of objects. In science museums, participants also wanted to know whether exhibits are touchable. In shopping malls, participants also wanted to learn the store menus and whether there is a spacious area for a guide dog to rest while the user is eating. However, three other participants (P02, P03, and P09) found certain information, such as details about lighting, surrounding people, and wall design, unnecessary.

4.5 Design Considerations

The results of the study affirmed that there are certain appreciations and room for improvement for the exploration robot for blind people. Based on the above results, we derived several requirements for the system, as listed below.

[4]The comments were obtained in the native language where the study was conducted. We translate the comments into English using publicly available LLM to ensure reproducibility. We show the full prompt used for translation in Appendix Sec. A.1.

A. Example at Science Museum

"This is a futuristic exhibition hall that has vibrant displays. To your left, there is a uniquely shaped wooden table and archway. Ahead, you can see a curved blue sofa and a white sign that reads "Entrance." On your right, large colorful panels line the wall, displaying information about the future and health."

B. Example at Shopping Mall

"This is a bright, modern corner of a commercial facility. On the left side, there are tall-backed chairs made of black metal lined up, and beyond them, round tables are arranged. Ahead, a man in a suit is standing, and in the background, there's an electronic menu board, suggesting the presence of a restaurant. To the right, there's an eatery enclosed by warm-colored walls in shades of red and orange, with many metallic chairs and tables, and menu boards are set up."

Figure 4: Examples of descriptions described in the formative study. Panel A shows an example of a description generated at the science museum, and Panel B shows the one generated at a shopping mall.

4.5.1 Vary Detail of Descriptions Based on Preferences and Contexts. We observed three types of preferences: one that enjoyed all the descriptions provided by the system (*Exploration-Inclined*), another that enjoyed the descriptions but preferred to limit certain information (*Intermediate*), and a third group that only wanted information useful for determining where to go (*Destination-Oriented*). In Tab. 2, we show the description preference of each participant. To classify the preferences, we first classified three participants who did not enjoy the description of the system as *Destination-Oriented*. Then, based on the discussion between the authors, we classified the rest as *Intermediate* or *Exploration-Inclined*. Furthermore, the type of information needed varied slightly depending on the experimental location. For instance, participants sought seating information for guide dogs in shopping areas, whereas in the science museum, they were more interested in whether the exhibits were touchable. Given these three types of preferences and context-dependent information needs, we modified the system so that it could adjust the amount and types of information conveyed to each participant.

4.5.2 Add Question and Answer Functionality. There was a clear need for question-and-answer (Q&A) interaction, as seven participants (P02–P05 and P08–P10) noted that they would like the option to ask more detailed questions through conversation. Participants expressed interest in this functionality when they were curious about the system's descriptions. This would allow them to ask more detailed questions about the objects of interest.

4.5.3 Add "Take-Me-There" Functionality. Four participants (P02, P04, P06, and P10) mentioned that they would like to revisit locations they found interesting after walking around the floor. Example situations include deciding to visit a shop, engaging with touchable exhibits, or returning to chairs discovered during the exploration. In unfamiliar locations, where users may lose their sense of direction, participants also expressed the need for a feature that guides them back to their initial location [45].

4.5.4 Vary Speed and Be Able to Stop the Robot. While the majority found the default speed appropriate for listening and understanding the described information, there were requests for customizable speed settings. Eight participants stated that the robot's speed was appropriate for exploring. Two participants (P04 and P06) expressed a preference for a faster speed. P01 additionally wanted to stop when the robot read out the descriptions of interest. In conclusion, users who are *Destination-Oriented* or have already determined the destination through exploration may want to increase the speed, while users who prefer to take time exploring might wish to slow down or stop the robot entirely.

4.5.5 Add Direction Specifying Functionality. Participants expressed a desire for more active engagement by specifying the movement direction themselves. Four participants (P02–P05) mentioned that they wanted more active control over the movement direction based on their interests. Additionally, we extrapolated that instead of simply following the robot, some users may prefer to interactively choose the direction based on the audio description of the surroundings. This could lead to greater autonomy because it would enrich the exploratory experience by aligning the robot's movement with the users' real-time curiosity and needs, creating a more personalized and engaging exploration experience.

5 WanderGuide Implementation

In this section, we provide the implementation of WanderGuide informed by the formative study. Below is a summary of updates made from the implementation of the formative study.

- Attachment of a new fisheye camera for a better view (Sec. 5.1)
- Implementation of a waypoint detection algorithm for realizing autonomous map-less navigation (Sec. 5.2)
- Implementation of three levels of description based on user preferences (Sec. 5.3)
- Implementation of "Take-Me-There" Functionality (Sec. 5.4)
- Implementation of two navigation modes automatic navigation mode and manual control mode (Sec. 5.5.1)
- Implementation of an interface to adjust speed, level of description, and navigation mode (Sec. 5.5.1)
- Implementation of Q&A Functionality (Sec. 5.5.2)

5.1 Hardware Update

One of the notable user feedbacks was the need for more detailed information, such as the names of POIs. However, the cameras in the prototype system were mounted at only 0.51 meters above the ground, had low resolution, and had a limited vertical field of view, making it difficult for the MLLM model to consistently capture details. Thus, as illustrated in Fig. 2–A-2, we attached a fisheye camera with 1080p resolution and a wide field of view to the higher part of the robot.

5.2 Waypoint Detection and Navigation

In order to produce destinations to navigate to for the users, a waypoint detection algorithm (Fig. 5) is necessary to determine navigable points for the robots. As no prebuilt maps were available, we first constructed a cost map, a two-dimensional occupancy grid that assigns costs based on obstacles, and updated it in real-time. We utilized the existing open-source Cartographer package [69], which is a real-time Simultaneous Localization and Mapping (SLAM) algorithm, to generate the cost map. Next, the cost map was skeletonized, and intersection points on the skeleton were identified based on a kernel-based corner detection algorithm [74]. The intersection points, which are typically far from obstacles, were next used to select potential waypoints. To maintain sparsity among waypoints, we applied the DBSCAN clustering algorithm [18] over the intersection points, and selected the centers of the clusters as potential waypoints. In addition, coordinates three meters in front, behind, and to the sides of the robot were also considered potential destinations to address the case where no intersection points were detected through the algorithm. As selecting a waypoint too far may be challenging for the robot to find a suitable path and a waypoint too close would lead to frequent destination changes, we filtered out candidates further than 50 meters and closer than one meter to the robot. After filtering, the final list of candidate waypoints was set.

During navigation, the robot automatically selected its goal from the candidate waypoints. By default, priority was given to waypoints lying in the same forward direction as the robot's initial orientation, where it was placed and activated. If no forward waypoints are available, the robot selects the waypoint with the smallest absolute angle relative to its current orientation. Once a waypoint was chosen from the candidate list, the robot navigated to it by using the onboard open-source navigation algorithm [24]. Once the waypoint was reached, the next waypoint was chosen automatically using the same process. It is important to note that prioritizing the robot's initial orientation was based on the assumption that users can adjust the general direction to proceed, such as starting from the entrance into the building.

5.3 Scene Description Generation

The basic algorithm for scene description generation remains unchanged, but the description was conveyed only when the robot was moving. Also, the MLLM took the overall view image from the fisheye view camera in addition to the three RGB images from the RGBD cameras. According to the results of the formative study, we added three levels of detail in the scene description.

- *Detailed Description* This mode provided rich, immersive descriptions for blind users who wanted to explore their surroundings in detail. The MLLM generated 3-4 sentences (120-240 characters), covering lighting, signs, layout, nearby people, and subjective descriptors like "beautiful" or "modern". The description began with an overview, followed by details of the left, front, and right.
- *Balanced-Length Description* This mode offered clear descriptions for users who preferred concise but informative content. The MLLM generated 2-3 sentences (60-120 characters), focusing on relevant details like signs and layout, while omitting lighting conditions or subjective descriptors. Descriptions covered the left, front, and right, without the overview.
- *Concise Description* This mode provided brief, essential information for users who wanted quick guidance. The MLLM generated 1-2 sentences (less than 60 characters), focusing only on key details needed to navigate, excluding unnecessary information. Descriptions covered the left, front, and right, without the overview.

For MLLM, these three levels were controlled via prompts, which are shown in Appendix Sec. A.3. All prompts shared the following instructions in common: to convey environmental information that assists blind people to explore, to refer to specific details such as genres or the names of objects, to encourage reading any text that helps users explore, to describe spaces for guide dogs to sit in restaurants, to provide information about potential hazards, and to use numbers to indicate the relative positions of surrounding objects. To ensure that the MLLM adhered to the instructions provided in the prompt, we employed a two-stage inference process. First, we instructed MLLM to perform an initial inference, generating a description.

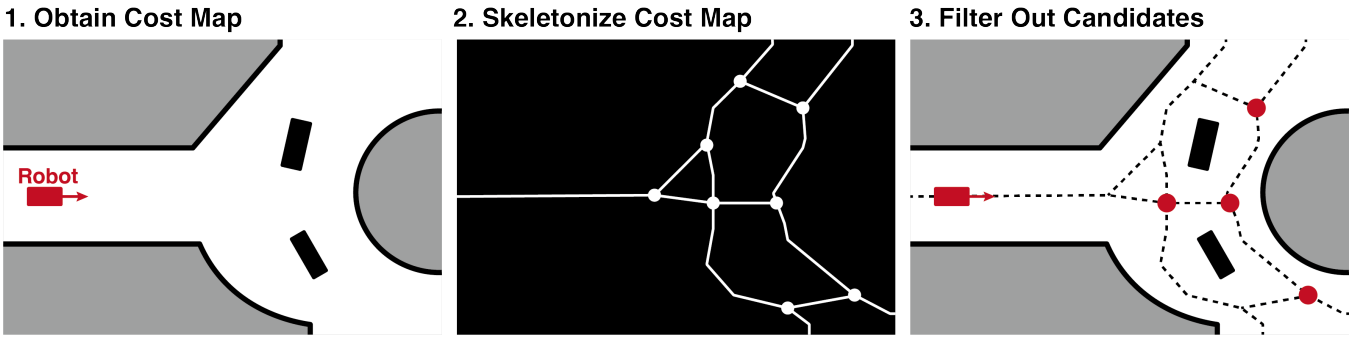

Figure 5: Three steps of the waypoint detection algorithm. Step 1 shows the generated cost map, while Step 2 depicts the skeletonization process of the cost map along with the detection of intersection points. Finally, Step 3 highlights the selected intersection points, which are identified as waypoint candidates.

Then, it self-supervised this generated description to verify if it met the given instructions. Finally, MLLM produced a revised version of the description to be presented to the user. Although this approach resulted in longer inference times, the outputs produced follow complex prompt instructions. The description is read aloud every 5-10 seconds after the previous description has been read out. The processing time and cost to generate a description was 5.78 seconds and $0.00811 for a Detailed Description, 4.75 seconds and $0.00753 for a Balanced-Length Description, and 4.02 seconds and $0.00734 for a Concise Description on average.

5.4 "Take-Me-There" Functionality

Acting on the feedback received from the formative study, we implemented a function that guided users to a destination verbally specified. This feature was typically enabled by the robot's *semantic map* [52, 73, 84]. In our case, we linked the images and generated descriptions, which had been saved as the robot had navigated, to the cost map of the robot. Given a verbal cue from the user (*e.g."I want to go to the blue sofa."*), the system first used our selected MLLM model to extract the name of the target location (*e.g.*, blue sofa). Then, we calculated the embeddings of the target location, all saved captured images, and all saved generated descriptions. We took a dot-product similarity between the extracted target location and the embeddings of images and descriptions to find the closest match. We used pre-trained feature extraction models: a fine-tuned SimCSE [20] model for generating sentence embeddings from text and a pre-trained CLIP [66] model for creating image embeddings. We used models that were trained in the native language where the study was conducted. The coordinate linked to the closest matched image or description would be set as the destination. If the user wanted to go back to the initial location, we used MLLM to detect the user's intent and set the destination to the initial point. We note that a similar functionality, the "Take-Me-Back" functionality, which allows users to return to their initial location, has been implemented in the previous map-less navigation system PathFinder [45]. The "Take-Me-Back" functionality is specifically designed for navigation purposes, as it was motivated by the challenge blind individuals face in returning to their original location after navigating. In contrast, our functionality is tailored for exploration tasks, enabling users to return to any point of interest they identified during their exploration. Ultimately, our functionality encompasses the capabilities of the "Take-Me-Back" feature while extending its application to support exploratory activities.

5.5 Navigation Mode and User Interface

On the high level, we implemented button controls and conversation interaction methods for users to interact with the robot.

5.5.1 Button Controls. We utilized the four directional buttons and the central button on the handle of the suitcase-shaped robot to enable users to control the robot's speed, adjust the level of descriptions, switch between automatic and manual control modes, and specify the direction of movement. The mapping of the buttons is illustrated in Fig. 2–B-1 and B-2. The central button was used for mode changes. The functions of the directional buttons would change depending on the robot's modes: *auto mode*, *manual control mode*, and *conversation mode*. In auto mode, the robot navigated by determining the waypoint automatically. The left and right buttons allowed the user to switch between three levels of description, where the default mode is the balanced-length description mode. The forward and backward buttons were used to adjust the robot's speed. Users can adjust the speed from zero to one meter per second, with increments of 0.05 meters per second. In manual mode, users could specify directions on their own. The robot would instruct the user to press the directional buttons to select the direction to proceed. If there was a suitable waypoint in the specified direction, the robot would inform the users via voice feedback. Otherwise, the robot conveyed that there were no navigable points in the specified direction. In conversation mode, triggered by long-pressing the central button, the robot would pause, and all four directional buttons were disabled until the conversation was ended. Users could manually end the conversation by long-pressing the central button again. The details of the conversation mode are described below.

5.5.2 Conversation. The conversation mode allowed users to give commands or ask questions with verbal input via the smartphone attached to the robot (Fig. 2). When the user inputted their verbal cue, the system used MLLM to classify the user's intent into one of three categories: usage of "Take-Me-There" functionality, usage of Q&A functionality, and direction specification. If the detected intent was direction specification (*e.g., "I want to go to right"*) the

Table 3: Demographics of participants who attended the main study. The table reports their gender, age, navigation aid, which they frequently use, frequency of exploration done either independently or with sighted people per year, their experimental location, and number of previous visits to the experimental location.

	Gender	Age	Aid	Age of Onset	Frequency of Exploration per Year	Experiment Location	Number of Previous Visits
P11	M	59	Cane	29	0	Science Museum	1
P12	F	59	Cane	43	36	Science Museum	1
P13	M	56	Cane	45	1	Science Museum	0
P14	F	60	Cane	45	48	Science Museum	1
P15	M	59	Cane	24	4	Science Museum	0

Table 4: The statistics of duration time and the count of interactions for each mode (Auto, Conversation, and Manual Control). The ratio of the duration time is calculated based on the total duration time of the experiment per participant.

	Auto		Conversation		Manual Control	
	Ratio(%)	Count	Ratio(%)	Count	Ratio(%)	Count
P11	59.77	25	37.52	21	2.70	4
P12	91.66	13	8.16	9	0.18	1
P13	67.88	12	30.56	9	1.56	1
P14	64.86	17	33.94	15	1.20	1
P15	58.53	28	20.03	19	21.44	10

Table 5: Analysis of participants' requests to the system during conversation mode. We defined three types of queries, General Query, Specific Query, and Command Query, and classified each participant request into one of them. The ratios here are simply the count percentages over total counts.

	General Query		Specific Query		Command Query		Total
	Ratio (%)	Count	Ratio (%)	Count	Ratio (%)	Count	
P11	11.43	4	45.71	16	42.86	15	35
P12	30.00	3	10.00	1	60.00	6	10
P13	61.54	8	38.46	5	0.00	0	13
P14	14.29	4	46.43	13	39.29	11	28
P15	8.33	2	25.00	6	66.67	16	24

robot would navigate to the waypoint in the specified direction accordingly. Finally, users could finish the conversation with an ending phrase such as *"Thank you."*

6 Main User Study

This study was conducted to validate WanderGuide and explore further design space. Participants were recruited and compensated similarly to those in the formative study. Similar to the formative study, in the recruitment email, we specified that participants unfamiliar with the experimental location would be eligible to participate. We conducted this study on the same two floors of the science museum. Tab. 3 shows the demographics of the participants. None of the participants from the formative study participated in this study. Similar to the formative study, this study was conducted after business hours.

6.1 Task and Procedure

For each participant, we first conducted a pre-study interview similar to the formative study. Then, the participant joined a 30-minute training session to get familiar with the robot system before the main tasks. For the main tasks, they were asked to freely explore the floor for 20 minutes using the system from a fixed starting location, as illustrated in Fig. 3. The ordering of the floors was counterbalanced to mitigate the order effect. After the main tasks, we conducted a post-study interview to ask several seven-point Likert scale questions (1: Strongly Disagree, 4: Neutral, and 7: Strongly Agree) that measure their self-evaluated exploration performance, Raw Task Load Index (TLX) [10] to measure the task workload, and system usability scale (SUS) [9] to evaluate the usability of the system. Finally, we asked open-ended questions to gather comments on the system. Below, we report the results of the study.

6.2 Analysis of Participants Activity During The Task

We report the statistics of each participant's activity during the task by referring to the system's log and the video captured during the tasks. Tab. 4 shows the analysis of their time spent on the three modes as specified in Sec. 5.5.1. We noticed that the activation quantity and duration of each mode varied significantly among participants. P11, P13, P14, and P15 frequently used the conversation mode. Notably, P11 spent nearly 40% of the total time engaging in conversation with the robot. In contrast, P12 barely used the conversation mode and relied on the auto mode for 90% of the total time. P15 was the only participant who actively used the manual control mode.

6.3 Analysis of Requests from Participants During Within The Conversation Mode

In Tab. 5, we further report the statistics of requests from participants within the conversation mode. Note that the total count of conversations in Tab. 5 is bigger than the conversation mode counts in Tab. 4, as multiple turns of conversation could happen in one conversation mode interaction. We classify each verbal request into three categories.

General Query Request general information in the surrounding area or in a particular direction.

Table 6: Error analysis of outputs from MLLM. We classified the errors into five categories and counted the number of them. Note that a single response could contain multiple errors, so the sum of errors does not match the total output of MLLM.

	Wrong Character Recognition	Wrong Object Recognition	Nonexistent Objects and Texts	Misunderstanding User Input	Inaccurate User Input	No Error	Total output
Scene Description	31	6	11	-	-	117	164
Q&A Response	9	6	15	5	1	21	53

Table 7: The statistics of the usage of each description level. Usage statistics for each description level are calculated by normalizing the duration of each level with respect to the total duration of the experiment.

	Concise	Balanced-Length	Detailed
P11	0.10%	76.46%	23.43%
P12	15.22%	29.88%	54.91%
P13	0.19%	49.14%	50.66%
P14	0.15%	87.94%	11.91%
P15	0.14%	99.86%	0.00%

Specific Query Request detailed information about a specific object in the environment.

Command Query Issue command to guide to destination, triggering "Take-Me-There" functionality or direction specification via conversation.

Overall, we discovered that although our system constantly provided environmental descriptions in auto mode, users still preferred to ask for general information about their surroundings or in a specific direction in conversation mode. For example, P13 predominantly made General Queries (61.54%). Users also had diverse preferences when using our system. Some users such as P11 (45.71%), P13 (38.46%) and P14 (46.43%) were interested in learning the specifics of POIs, reflecting the takeaways obtained in Sec. 4.4.3. Some users such as P11 (42.86%), P12 (60.00%), P14 (39.29%), and P15 (66.67%) favored using conversation mode to instruct the robot to guide them to their destinations. In particular, by referencing Tab. 4, we can see that P11, P12, and P14 preferred conversation mode over manual control mode to issue commands. This validates the extrapolated idea in Sec. 4.5.5.

6.4 Error Analysis of Scene Description and Q&A Responses

In Tab. 6, we report the accuracy of MLLM responses both during auto and conversation modes. We manually analyzed the text output generated by MLLM and compared it with the logs of the images saved. We classified and counted the errors made by MLLM into six categories.

Wrong Character Recognition Misrecognition of text, such as misreading signs.

Wrong Object Recognition Misidentification of objects in the scene.

Nonexistent Objects and Texts Mistakenly recognizing objects or text that are not present. Note that this differs from the previous two categories, where some similar objects or text were actually present.

Misunderstanding User Input Misinterpreting a user's question in conversation mode, such as providing an environmental description when asked to read text from a panel.

Inaccurate User Input Errors made when the user asked about objects or text that were not present.

No Error Accurate responses with no errors.

When multiple errors occur in a single sentence, errors of the same type are grouped together and counted as one. Errors of different types are counted separately. For instance, if there are multiple text recognition errors in a single sentence, they are counted as one text recognition error. If a sentence contains both text recognition errors and object recognition errors, each is counted separately as one text recognition error and one object recognition error. Thus, note that the total number of errors may not match the total number of outputs.

The results showed that 28.6% of the outputs contained some form of error during scene descriptions whereas 60.3% of conversation mode outputs had errors. This difference is likely because users in conversation mode often asked for more detailed explanations, which led MLLM to attempt more complex responses and, as a result, made more mistakes. This was particularly evident in the *Nonexistent Objects and Texts* category, which accounted for only 0.07% of errors during scene descriptions but significantly higher at 28.3% in conversation mode. This means that MLLM often generated descriptions of objects or text that did not exist in the environment when asked for more detailed information. Character recognition errors were common in both modes, likely due to MLLM's limitation in reading distant text. In a general sense, instead of complete failures, MLLM often partially misread the text or misidentified objects with similar-looking ones (*e.g.*, mistaking a tall table for a reception desk). Nevertheless, over 70% of responses in the auto mode were accurate, demonstrating the overall usefulness of the system.

6.5 Analysis of Usage of Each Description Level

In Tab. 7, we report the statistics of how much time participants spend their time using each description level. The result shows that there were three types of usage during the study. P15 only used Balanced-Length mode, P11 and P14 used Balanced-Length mode most of the time while sometimes using Detailed mode, and P12 and P13 used Detailed mode most of the time.

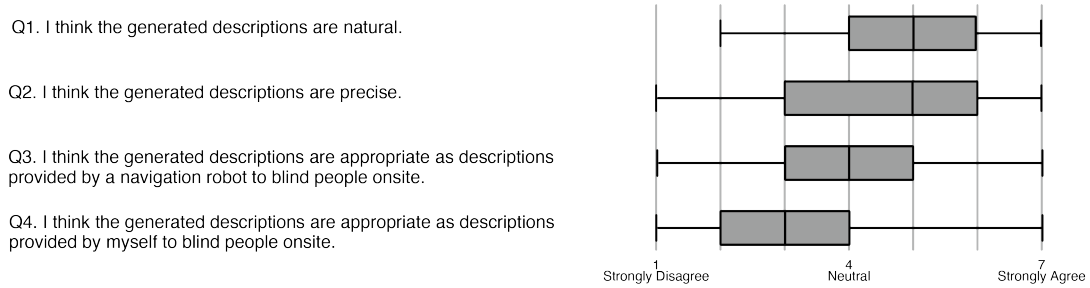

Figure 6: Box plot of evaluation with human experts in seven-point Likert points.

Table 8: Rating to seven-point Likert score questions (1: strongly disagree; 4: neutral; 7: strongly agree).

	P11	P12	P13	P14	P15	Median
Q1. I was able to explore the facility.	4	6	4	6	6	6
Q2. I was able to enjoy the exploration.	4	7	4	6	6	6
Q3. I was able to gain an interest in the things around me.	4	6	6	6	6	6
Q4. The interface of the system was easy to understand.	5	6	6	5	5	5
Q5. I want to explore where I am familiar with this system.	7	7	6	7	6	7
Q6. I want to explore where I am unfamiliar with this system.	7	6	6	7	6	6

Table 9: Scores for the Raw TLX provided by each participant. Lower total scores indicate a lower workload. Each item is scored on a scale from 1 to 10, where 1 represents a lower level, and 10 represents a higher level of Mental Demand, Physical Demand, Temporal Demand, Effort, and Frustration. For Performance, 1 indicates good performance, and 10 indicates poor performance.

	P11	P12	P13	P14	P15	Median
Mental Demand	2	2	5	3	2	2
Physical Demand	2	2	3	2	6	2
Temporal Demand	2	3	1	5	2	2
Performance	7	2	5	7	5	5
Effort	3	2	5	6	6	5
Frustration	8	4	1	3	7	4
Total Score	24	15	20	26	28	

6.6 Scene Description Quality Evaluation

Finally, to analyze the quality of the MLLM-generated scene descriptions from the human expert perspective, we conducted a survey with human museum guides and asked them to evaluate using a seven-point Likert scale. The participants were presented with images captured by the robot, each accompanied by its corresponding generated description, and were asked to evaluate the descriptions in a survey, as shown in Fig. 6. The survey was conducted in a counterbalanced manner to mitigate potential biases. During the main study, 164 descriptions were generated, and we randomly sampled half (82) of the total descriptions for evaluation. The randomly sampled descriptions contain mixed levels of detail. Each description is evaluated by three to four participants. In total, 56 museum guides participated in the evaluation, with each randomly assessing five descriptions. There were 32 males and 20 females, and four participants did not report their gender. Their average age was 39.6 years, with an average of 5.9 years of experience as a museum guide. On seven-point Likert scale items, the median self-reported familiarity with museums was 5.0, and the familiarity with LLMs was 4.0 (1: very unfamiliar, 4: neutral, and 7: very familiar). Our analysis revealed that the experts generally perceived the generated descriptions as somewhat natural (Q1) and precise in describing an image (Q2) as shown by their median of five. Meanwhile, they found the generated descriptions less suitable as image descriptions for blind people (Q3) and as onsite descriptions provided by experts for blind people (Q4).

6.7 Usability and Workload Evaluation

In Tab. 8, we report the results of seven-point Likert items. For Likert items, a median score of five or higher indicates that participants generally responded positively. The total SUS for P11 to P15 were 72.5, 80, 90, 82.5, and 77.5, respectively, showing acceptable usability of all being above 70 [4]. The total Raw TLX scores for P11 to P15 were 24, 15, 20, 26, and 28, respectively. We show the distribution of Raw TLX scores in Tab. 9. Raw TLX [10], a simplified version of NASA TLX [26], is known to have a high correlation with NASA TLX, and the total NASA-TLX scores for people with special needs typically ranged from 26 to 48 in previous research [27]. Overall, our total Raw TLX scores may suggest that participants did not experience a significant load during the task. We also observed that the median value for mental, physical, and temporal demand was relatively lower, scoring 2. This is likely due to the robot navigating them, allowing participants to explore without being burdened by these demands. Nonetheless, a relatively higher median value was observed for Performance, Effort, and Frustration, indicating that

some users experienced a lack of satisfaction with the exploration experience provided by the system.

6.8 Qualitative Analysis

6.8.1 Positive Feedback. All participants expressed their appreciation for the experience of wandering around a building to explore without specific destinations in mind with the help of our system: **C6**:*" When the camera explains things it recognizes, like how bright the room is or what the floor looks like, or what objects are placed where, I found myself nodding in agreement multiple times, like, "Oh, so this is how it looks." I remember when I first held the suitcase robot, I deeply empathized with guide dog users. I thought, "Oh, so this is what it's like to have a guide dog." However, since I can't take care of a guide dog, I've given up on that option. And now, with this navigation system that explains various situations, it's exactly what I need. It's not just about setting a destination and getting there but feeling the freedom to explore spontaneously. For example, the ability to roam a large shopping mall freely and explore on a whim feels like true freedom to me. Instead of pre-planning every move or relying on a guide, I could simply grab my suitcase and decide to venture out spontaneously."* (P12)

The same participant, P12, who had been to the facility previously, noted that they still had new discoveries with the system: **C7**:*" I've been to this museum before, but when the guide explained things to me back then, it was more like a general explanation about the atmosphere and such. But earlier with the system, there was a very detailed explanation that came out of the suitcase. Like, about how bright sunlight comes [...] There were things I didn't know that made me learn new stuff, even though I thought I knew about the facility."* (P12) Also, P12 and P14 noted the feeling of relief not relying on sighted assistance: **C8**:*" I don't think there has ever been a system that explains your surroundings while walking. [...] When walking with other people, I often find myself feeling a sense of obligation. I worry that they're putting in extra effort to describe things because I can't see. And then I feel like I have to respond to them since they're trying so hard—which can be exhausting. But with this system, I feel I can go strolling by myself."* (P14)

Participants also noted the functionality to go to an aforementioned destination and Q&A functionality particularly useful: **C9**:*"(The "Take-Me-There" functionality is) I think it's wonderful. After all, spatial awareness is difficult, so going back to landmarks is very important. If it is accurate, I think it's great because it can be extremely helpful for spatial cognition."* (P11) and **C10**:*"When engaging in a conversation, not knowing what kind of response you'll get, the feeling of unease and excitement that's both a plus and a minus, I think. But I found it really great that you can still ask questions. So even if the response you get doesn't answer your question, or even if it's just "I don't know," the fact that you can at least ask is important."* (P14)

6.8.2 Adjusting Detail of Description. When we discussed their preference in the level of detail of descriptions, all participants described that it would rather depend on the scenario they are in: **C11**:*"It might depend on the location, but I know I can get detailed information in Q&A functionality. So, for familiar places, the Balanced-Length mode might be fine. However, there are parts where I'd want the Detailed Description mode for unfamiliar places. For example, switching between modes could be useful, like having Detailed Description mode first for explanations about the room's brightness and how easy it is to walk around. "* (P12)

6.8.3 Comments to Improve the System. Participants suggested various improvements to the system. One particular suggestion was to incorporate functionality for the robot to understand sounds. As the experiment location was a science museum, various exhibits emitted sounds. P13 noted that they would like to inquire about the sound sources, which were not supported by the system: **C12**:*"We are extremely sensitive to sounds, and it becomes a point of interest. At a place like the exhibition hall we're visiting this time, various sounds are coming from all directions. This prompts questions like, "What's happening at that sound over there?" Therefore, it would be advantageous if we could ask specific questions like, "What's that sound coming from the right?" "* (P13)

Also, four participants (P11-P13 and P15) found the descriptions from the system still insufficient to explore, as described in the following comments: **C13**:*"The place we did the task this time was quite out of the ordinary. Even if you were walking around with my family, I think they would also have difficulty explaining it. Therefore, I felt it might still be somewhat challenging for machines to handle this kind of thing. However, I did feel it was good that I got a sense of what was there. But when it comes to the actual detailed explanations, it was not there [...]"* (P15)

6.8.4 Specification of Proceeding Direction. While we introduced all functionality to participants within the training session, we observed that only P15 used the functionality to specify which way to proceed via a button or conversation. P15 tended to use the functionality when P15 was interested in a specific object: **C14**:*"It seems that when I was told, "There's something on the right," I tried to approach toward it because I wanted to get closer when I used something like that."* (P15)

7 Discussion

7.1 Experience of Using WanderGuide

WanderGuide provided participants with the experience of exploring unfamiliar indoor environments without a specific destination in their minds, mimicking the spontaneous wandering experience of sighted people (C6). Participants expressed a sense of confidence when using the system, noting that it allowed them to navigate independently without relying on traditional tools like white canes (C6). As described by C13 and the ratings of 4 from P11 and P13 to Q1 and Q2 in Tab. 8, there still exists the limitation of being unable to describe specific information. Thus, there is a need for further research on how to appropriately convey surrounding information to blind people. Still, the system's ability to deliver real-time descriptions of objects, walls, and spatial layouts enabled participants to form an imagination (C6) of their surroundings, sparking their desire to use the system in familiar and unfamiliar environments (Tab. 8 Q5 and Q6). In short, WanderGuide has the potential to provide users with an experience similar to that of navigating with sighted assistants to explore the environment, but the users can explore independently. We believe this research opens a new frontier to the concept of *map-less exploration* guide system for blind people.

7.2 Scene Description by MLLM

Our survey in Sec. 6.6 revealed that descriptions by MLLM were rated high for their naturalness and suitability for general image description but were not for actual descriptions to be provided to blind people by sighted experts. This may be because the style and content of the generated descriptions differ from those typically provided to blind people during live interactions. For example, museum guides often focus on explaining notable objects or visible exhibits, complementing their descriptions with additional knowledge about the exhibits. In contrast, the generated description often lacked concrete explanation about exhibits and shops, such as their names (C4, C5, and C13). This problem may be more prominent because the study was conducted in a science museum, where each exhibit contains detailed information that is not visually apparent but needs to be explained. On the other hand, from participant feedback, participants noted that MLLM-generated descriptions are comprehensive (C2, C3, C7, and C8), and provide them with enjoyment (C2) and imagination of vision perception (C3). They noted that MLLM provided them with information that they usually do not get from sighted assistants, leading to new discoveries (C7). The descriptions provided by MLLM additionally allow blind people to tune in without hesitation and the need to rely on sighted people (C8). These results indicate that evaluation from sighted experts may be stricter than that from blind people. Nonetheless, these results suggest that MLLM for blind people's exploration could be further enhanced by providing more specific information about surrounding shops or exhibits, potentially inferring details when necessary.

7.3 Personal Preferences

The studies revealed distinct preferences among participants regarding the levels of detail in the descriptions (Sec. 5.3) and interaction modes (Sec. 5.5.1). From the formative study, participants were divided into three preference groups, highlighting users' diverse information needs regarding exploration and goal-oriented navigation (Sec. 4.5.1). Differences in preferences were mainly attributed to personality traits, because participants who were "Destination-Oriented" (Tab. 2), or were mostly concerned with reaching destinations, mentioned they did not enjoy the detailed explanation of the system and preferred short, concise information. For example, one early blinded participant mentioned that exploration did not interest him, as he had barely done it in his daily life (C1). On the other hand, some participants enjoyed imagining the scenes conveyed by the system. Congenital users commented that the descriptions felt as if they were actually seeing the surroundings, while acquired users likened it to their recalled experiences when they could still see. Interestingly, those who particularly enjoyed the system and were "Exploration-Inclined" were all female, while the Intermediate group, who enjoyed exploration but wanted more control over the information provided, consisted mainly of male participants. We note that "Destination-Oriented" users expressed dissatisfaction with the system because they felt the scene description capabilities of the MLLM did not meet their expectations for exploration. Therefore, if the system was improved and was able to convey more concrete information, they might express different opinions.

In the main study, further differences regarding how users interacted with the system were observed. Firstly, we observed that participants adjusted the system's levels of description, demonstrating our design aligns with their needs, which were based on three types of preferences identified in the formative study (Sec 6.5). The variation in the portions used for each mode further underscores the need for configurable descriptions. Also, how they used the conversation mode varied. Three participants frequently asked questions to the system to gather information about their surroundings (Sec. 6.2), while P15 preferred having more manual control over the robot's navigation. Meanwhile, P12 favored the auto mode, where the robot guided them with minimal intervention. These observations highlight the need to consider customizing to various dimensions of personal preference, from description details to user autonomy, for future development.

7.4 Design Implications and Future Development Directions

Two key design implications were observed in our studies. First, allowing the users to control the level of detail in the scene descriptions emerged as one of the most important design requirements. The system may benefit from further *personalization* by users verbally describing their personal information needs as in previous research [37]. Second, participants expressed the need for audio-based recognition capabilities, especially in environments where sound is an integral part of the experience, such as museums (C12). The ability to answer questions about sounds and potentially guide users to the sounds' sources would enhance their exploration experience.

On the development side, the primary challenge encountered throughout the two studies was the system's inability to provide detailed information that participants required, particularly regarding the identification of POI-related objects, as described in Sec. 7.2. We attempted to address this by upgrading the robot's hardware, *i.e.*, adding a 1080p resolution fisheye camera to a much higher position. Still, participants found the descriptions lacking in detail and conveyed information somewhat vague, as partially shown by the ratings of 4 from P11 and P13 to Q1 Tab. 8. We deduce that this was because the captured images sometimes did not contain useful information, such as the names of certain objects, or because the MLLM failed to accurately identify the useful information. As a possible improvement, the robot could utilize history images by selecting the image with the best view to generate descriptions. Also, the robot could utilize, other modalities, such as colored point clouds by fusing camera images with the LiDAR sensor to provide three-dimentional sensor details to MLLM [50, 82]. In conclusion, the MLLM module is the bottleneck of our system's technological development. Similar system development efforts in the future should allocate the most resources to tackling this technological challenge. Still, the issue may be gradually solved as MLLM is the current core area actively developed by researchers.

Another significant challenge in development we encountered is the challenge of running map-less navigation algorithms in diverse novel environments, which requires extensive development. Incorporating vision modalities [12], which we did not use in this study, could potentially enhance the robot's navigation capabilities.

Achieving this, however, demands human-level object and layout recognition and real-time processing speed, where further research is required.

7.5 Limitation and Future Work

We were unable to examine user preferences over the long term, as participants in our study interacted with the system only for a short duration (20-40 minutes in the formative study and 70 minutes in the main study). Only a small portion of the reliance on concise descriptions may be due to the study's design limiting participants' time to explore. The time constraints may have led users to act on the cost-effective information acquisition. However, if the system is used regularly, users may encounter more situations where they prefer to use the concise mode, as indicated by C11. Also, their preferences might change as they become more adept at utilizing it as a tool to query information, which the MLLM is particularly proficient at. Thus, future research should investigate the effects of long-term use of the system.

We conducted two studies in two indoor locations. To capture more diverse needs, future studies should also explore the system's performance in more diverse environments. This may reveal various additional information needs. The usage of the wheeled robot, while beneficial in guiding blind users because it is silent [78], remains a constraint when navigating stairs or uneven terrain. This limitation, however, could be alleviated through user collaboration, such as assisting the robot in getting onto elevators or slightly lifting the robot over small steps. Thus, future research should investigate the system devices's capabilities in different environments, as well as how these robots can address physical limitations by interacting with users. Finally, for the main study, we were unable to conduct it in crowd environments with bystanders potentially obstructing the cameras, because the primary study was conducted in the science museum outside of regular operational hours. Handling crowded environments with robots, even when prebuilt maps are used, remains a significant challenge in the field of robotics [77]. Therefore, in future work, we aim to address the usability limitations of our system in such scenarios by integrating novel algorithms designed to manage crowded environments [77].

The MLLM often made mistakes or referred to non-existent objects, with these errors being particularly noticeable in its responses within the Q&A functionality (Sec. 6.4). The most common misrecognitions involved either partially reading the text or confusing objects with similar-looking ones. However, the performance of the MLLM is not the primary focus of our research. To ensure users receive the most accurate information possible, we will continue updating the MLLM used in the system. Also, some of the image inputs provided to the MLLM may have been affected by motion blur, potentially leading to a degradation in the quality of the generated descriptions. This issue could be addressed by using cameras that are more resistant to motion blur or by implementing algorithms that detect motion blur and select alternative frames for processing.

Recruitment was conducted through our institution's email list, which includes many participants from previous studies. We acknowledge that these participants may have exhibited a positive bias toward our study, as they had expectations regarding the development of the robot system. Furthermore, we obtained valuable insights from five participants, and involving more participants might have provided additional perspectives. Given the difficulty of recruiting many blind participants, we chose to iterate the study with five participants in each study, rather than conducting a single study with a larger group.

8 Conclusion

Towards realizing a scalable map-less guide system that assists blind people in exploring, we developed WanderGuide, a robotic guide system designed to provide real-time descriptions of surroundings and to offer conversation functionalities that allow users to specify their destinations or ask questions. The formative study with ten blind participants revealed that there are three types of preferences over the levels of details of the descriptions generated by the system. In a subsequent main study with five blind participants, all of them expressed appreciation for the experience of wandering freely without a fixed destination, as well as a desire to use the system for exploring both familiar and unfamiliar areas. The study further revealed that including audio recognition would be the immediate next step for developing our system. It also revealed that customizing to diverse user preferences is important and that MLLM is the key bottleneck of the technology development of our system. We hope this research contributes to the potential deployment of robotic guide systems in general use cases, enabling blind users to explore independently.

Acknowledgments

We would like to thank all the participants in our user study. We are also deeply thankful to Mori Building Co., Ltd. for providing the experimental location. Finally, we thank all members of Miraikan, including Hironobu Takagi and Hiromi Kurokawa, and the Consortium for Advanced Assistive Mobility Platform for their support. This work was supported by JSPS KAKENHI (JP23KJ2048).

References

[1] Aira. 2024. Aira. Retrieved in July 29, 2024 from https://aira.io/.
[2] Saki Asakawa, João Guerreiro, Dragan Ahmetovic, Kris M. Kitani, and Chieko Asakawa. 2018. The Present and Future of Museum Accessibility for People with Visual Impairments. In *ASSETS'18*. ACM, New York, NY, USA, 382–384.
[3] Saki Asakawa, João Guerreiro, Daisuke Sato, Hironobu Takagi, Dragan Ahmetovic, Desi Gonzalez, Kris M. Kitani, and Chieko Asakawa. 2019. An Independent and Interactive Museum Experience for Blind People. In *W4A'19*. ACM, New York, NY, USA.
[4] Aaron Bangor, Philip Kortum, and James Miller. 2009. Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale. *Journal of usability studies* 4, 3 (2009), 114–123.
[5] Nikola Banovic, Rachel L Franz, Khai N Truong, Jennifer Mankoff, and Anind K Dey. 2013. Uncovering Information Needs for Independent Spatial Learning for Users Who Are Visually Impaired. In *ASSETS'13*. ACM, New York, NY, USA, 1–8.
[6] BeMyAI. 2024. Introducing Be My AI. Retrieved in July 29, 2024 from https://www.bemyeyes.com/blog/introducing-be-my-ai.
[7] BeMyEyes. 2024. BeMyEyes. Retrieved in July 29, 2024 from https://www.bemyeyes.com/.
[8] BlindSquare. 2024. BlindSquare. Retrieved in November 25, 2024 from https://www.blindsquare.com/.
[9] John Brooke et al. 1996. SUS-A Quick and Dirty Usability Scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.
[10] James C Byers, AC Bittner, and Susan G Hill. 1989. Traditional and Raw Task Load Index (TLX) Correlations: Are Paired Comparisons Necessary. *Advances in industrial ergonomics and safety* 1 (1989), 481–485.
[11] Shaojun Cai, Ashwin Ram, Zhengtai Gou, Mohd Alqama Wasim Shaikh, Yu-An Chen, Yingjia Wan, Kotaro Hara, Shengdong Zhao, and David Hsu. 2024. Navigating Real-World Challenges: A Quadruped Robot Guiding System for

Visually Impaired People in Diverse Environments. In *CHI'24*. ACM, New York, NY, USA, 1–18.

[12] Matthew Chang, Theophile Gervet, Mukul Khanna, Sriram Yenamandra, Dhruv Shah, So Yeon Min, Kavit Shah, Chris Paxton, Saurabh Gupta, Dhruv Batra, Roozbeh Mottaghi, Jitendra Malik, and Devendra Singh Chaplot. 2024. GOAT: GO to Any Thing. In *Proceedings of Robotics: Science and Systems*. Delft, Netherlands.

[13] Ruei-Che Chang, Yuxuan Liu, and Anhong Guo. 2024. WorldScribe: Towards Context-Aware Live Visual Descriptions. In *UIST'24*. ACM, New York, NY, USA, 1–18.

[14] Hsuan-Eng Chen, Yi-Ying Lin, Chien-Hsing Chen, and I-Fang Wang. 2015. BlindNavi: A Navigation App for the Visually Impaired Smartphone User. In *CHI EA'15*. ACM, New York, NY, USA, 19–24.

[15] Jean-Philippe Choiniere and Clement Gosselin. 2016. Development and Experimental Validation of a Haptic Compass Based on Asymmetric Torque Stimuli. *IEEE transactions on haptics* 10, 1 (2016), 29–39.

[16] Khadidja Delloul and Slimane Larabi. 2022. Image Captioning State-of-the-Art: Is It Enough for the Guidance of Visually Impaired in an Environment?. In *CSA'22*. Springer, New York, NY, USA, 385–394.

[17] Christin Engel, Karin Müller, Angela Constantinescu, Claudia Loitsch, Vanessa Petrausch, Gerhard Weber, and Rainer Stiefelhagen. 2020. Travelling More Independently: A Requirements Analysis for Accessible Journeys to Unknown Buildings for People with Visual Impairments. In *ASSETS'20*. ACM, New York, NY, USA.

[18] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *KDD'96*. AAAI Press, 226–231.

[19] Navid Fallah, Ilias Apostolopoulos, Kostas Bekris, and Eelke Folmer. 2012. The User as a Sensor: Navigating Users with Visual Impairments in Indoor Spaces Using Tactile Landmarks. In *CHI'12*. ACM, New York, NY, USA, 425–432.

[20] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *EMNLP'21*. ACL, Online and Punta Cana, Dominican Republic.

[21] Ricardo E Gonzalez Penuela, Jazmin Collins, Cynthia Bennett, and Shiri Azenkot. 2024. Investigating Use Cases of AI-Powered Scene Description Applications for Blind and Low Vision People. In *CHI'24*. ACM, New York, NY, USA, 1–21.

[22] João Guerreiro, Dragan Ahmetovic, Kris M Kitani, and Chieko Asakawa. 2017. Virtual Navigation for Blind People: Building Sequential Representations of the Real-world. In *ASSETS'17*. ACM, New York, NY, USA, 280–289.

[23] João Guerreiro, Dragan Ahmetovic, Daisuke Sato, Kris Kitani, and Chieko Asakawa. 2019. Airport Accessibility and Navigation Assistance for People with Visual Impairments. In *CHI'19*. ACM, New York, NY, USA, 1–14.

[24] João Guerreiro, Daisuke Sato, Saki Asakawa, Huixu Dong, Kris M Kitani, and Chieko Asakawa. 2019. CaBot: Designing and Evaluating an Autonomous Navigation Robot for Blind People. In *ASSETS'19*. ACM, New York, NY, USA, 68–82.

[25] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz Grand Challenge: Answering Visual Questions From Blind People. In *CVPR'18*. IEEE, Piscataway, NJ, USA, 3608–3617.

[26] Sandra G Hart. 2006. NASA-task Load Index (NASA-TLX); 20 Years Later. In *Proceedings of the human factors and ergonomics society annual meeting*. SAGE Publications, Los Angeles, CA, 904–908.

[27] Morten Hertzum. 2021. Reference Values and Subscale Patterns for the Task Load Index (TLX): a Meta-analytic Review. *Ergonomics* 64, 7 (2021), 869–878.

[28] Karst MP Hoogsteen, Sarit Szpiro, Gabriel Kreiman, and Eli Peli. 2022. Beyond the Cane: Describing Urban Scenes to Blind People for Mobility Tasks. *TACCESS* 15, 3 (2022), 1–29.

[29] Hochul Hwang, Tim Xia, Ibrahima Keita, Ken Suzuki, Joydeep Biswas, Sunghoon I. Lee, and Donghyun Kim. 2022. System Configuration and Navigation of a Guide Dog Robot: Toward Animal Guide Dog-Level Guiding Work.

[30] IncNavi. 2024. IncNavi. Retrieved in September 8, 2024 from https://www.nihonbashi-tokyo.jp/inclusive_navi/.

[31] Gesu India, Mohit Jain, Pallav Karya, Nirmalendu Diwakar, and Manohar Swaminathan. 2021. VStroll: An Audio-based Virtual Exploration to Encourage Walking Among People with Vision Impairments. In *ASSETS'21*. ACM, New York, NY, USA, 1–13.

[32] Intel. 2024. Intel® RealSense™ Depth Camera D435. Retrieved in November 25, 2024 from https://www.intel.com/content/www/us/en/products/sku/128255/intel-realsense-depth-camera-d435/specifications.html.

[33] Intel. 2024. Intel® RealSense™ Depth Camera D455. Retrieved in November 25, 2024 from https://www.intelrealsense.com/depth-camera-d455/.

[34] Gaurav Jain, Yuanyang Teng, Dong Heon Cho, Yunhao Xing, Maryam Aziz, and Brian A Smith. 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 (2023), 1–28.

[35] Rie Kamikubo, Hernisa Kacorri, and Chieko Asakawa. 2024. " We Are at the Mercy of Others' Opinion": Supporting Blind People in Recreational Window Shopping with AI-infused Technology. In *W4A'24*. ACM, New York, NY, USA.

[36] Rie Kamikubo, Naoya Kato, Keita Higuchi, Ryo Yonetani, and Yoichi Sato. 2020. Support Strategies for Remote Guides in Assisting People with Visual Impairments for Effective Indoor Navigation. In *CHI'20*. ACM, New York, NY, USA, 1–12.

[37] Yuka Kaniwa, Masaki Kuribayashi, Seita Kayukawa, Daisuke Sato, Hironobu Takagi, Chieko Asakawa, and Shigeo Morishima. 2024. ChitChatGuide: Enabling Exploration in a Shopping Mall for People with Visual Impairments Through Conversational Interaction Using Large Language Models. *Proceedings of the ACM on Human-Computer Interaction* MHCI (2024).

[38] Seita Kayukawa, Daisuke Sato, Masayuki Murata, Tatsuya Ishihara, Akihiro Kosugi, Hironobu Takagi, Shigeo Morishima, and Chieko Asakawa. 2022. How Users, Facility Managers, and Bystanders Perceive and Accept a Navigation Robot for Visually Impaired People in Public Buildings. In *RO-MAN'22*. IEEE, Piscataway, NJ, USA.

[39] Seita Kayukawa, Daisuke Sato, Masayuki Murata, Tatsuya Ishihara, Hironobu Takagi, Shigeo Morishima, and Chieko Asakawa. 2023. Enhancing Blind Visitor's Autonomy in a Science Museum Using an Autonomous Navigation Robot. In *CHI'23*. ACM, New York, NY, USA, 1–14.

[40] Sulaiman Khan, Shah Nazir, and Habib Ullah Khan. 2021. Analysis of Navigation Assistants for Blind and Visually Impaired People: A Systematic Review. *IEEE Access* 9 (2021), 26712–26734.

[41] Jee-Eun Kim, Masahiro Bessho, Shinsuke Kobayashi, Noboru Koshizuka, and Ken Sakamura. 2016. Navigating Visually Impaired Travelers in a Large Train Station Using Smartphone and Bluetooth Low Energy. In *SAC'19*. ACM, New York, NY, USA, 604–611.

[42] J. Taery Kim, Wenhao Yu, Yash Kothari, Jie Tan, Greg Turk, and Sehoon Ha. 2023. Transforming a Quadruped Into a Guide Robot for the Visually Impaired: Formalizing Wayfinding, Interaction Modeling, and Safety Mechanism.

[43] Masaya Kubota, Masaki Kuribayashi, Seita Kayukawa, Hironobu Takagi, Chieko Asakawa, and Shigeo Morishima. 2024. Snap&Nav: Smartphone-based Indoor Navigation System For Blind People Via Floor Map Analysis and Intersection Detection. *Proceedings of the ACM on Human-Computer Interaction* MHCI (2024).

[44] Bineeth Kuriakose, Raju Shrestha, and Frode Eika Sandnes. 2020. Tools and Technologies for Blind and Visually Impaired Navigation Support: A Review. *IETE Technical Review* 0, 0 (2020), 1–16.

[45] Masaki Kuribayashi, Tatsuya Ishihara, Daisuke Sato, Jayakorn Vongkulbhisal, Karnik Ram, Seita Kayukawa, Hironobu Takagi, Shigeo Morishima, and Chieko Asakawa. 2023. PathFinder: Designing a Map-less Navigation System for Blind People in Unfamiliar Buildings. In *CHI'23*. ACM, New York, NY, USA, 1–16.

[46] Masaki Kuribayashi, Seita Kayukawa, Jayakorn Vongkulbhisal, Chieko Asakawa, Daisuke Sato, Hironobu Takagi, and Shigeo Morishima. 2022. Corridor-Walker: Mobile Indoor Walking Assistance for Blind People to Avoid Obstacles and Recognize Intersections. *Proceedings of the ACM on Human-Computer Interaction* 6, MHCI (2022), 1–22.

[47] Gerard Lacey and Shane MacNamara. 2000. Context-aware Shared Control of a Robot Mobility Aid for the Elderly Blind. *The International Journal of Robotics Research* 19, 11 (2000), 1054–1065.

[48] Young Hoon Lee and Gerard Medioni. 2014. Wearable RGBD Indoor Navigation System for the Blind. In *ECCV'14*. Springer, New York, NY, USA, 493–508.

[49] Bing Li, J Pablo Munoz, Xuejian Rong, Jizhong Xiao, Yingli Tian, and Aries Arditi. 2016. ISANA: Wearable Context-aware Indoor Assistive Navigation with Obstacle Avoidance for the Blind. In *ECCV'16*. Springer, New York, NY, USA, 448–462.

[50] Dingning Liu, Xiaoshui Huang, Yuenan Hou, Zhihui Wang, Zhenfei Yin, Yongshun Gong, Peng Gao, and Wanli Ouyang. 2024. Uni3d-llm: Unifying Point Cloud Perception, Generation and Editing with Large Language Models.

[51] Guanhong Liu, Tianyu Yu, Chun Yu, Haiqing Xu, Shuchang Xu, Ciyuan Yang, Feng Wang, Haipeng Mi, and Yuanchun Shi. 2021. Tactile Compass: Enabling Visually Impaired People to Follow a Path with Continuous Directional Feedback. In *CHI'21*. ACM, New York, NY, USA, 1–13.

[52] Shuijing Liu, Aamir Hasan, Kaiwen Hong, Runxuan Wang, Peixin Chang, Zachary Mizrachi, Justin Lin, D Livingston McPherson, Wendy A Rogers, and Katherine Driggs-Campbell. 2024. DRAGON: A Dialogue-based Robot for Assistive Navigation with Visual Language Grounding. *RA-L* (2024).

[53] Chen-Lung Lu, Zi-Yan Liu, Jui-Te Huang, Ching-I Huang, Bo-Hui Wang, Yi Chen, Nien-Hsin Wu, Hsueh-Cheng Wang, Laura Giarré, and Pei-Yi Kuo. 2021. Assistive Navigation Using Deep Reinforcement Learning Guiding Robot With UWB/Voice Beacons and Semantic Feedbacks for Blind and Visually Impaired People. *Frontiers in Robotics and AI* 8 (2021).

[54] Kanak Manjari, Madhushi Verma, and Gaurav Singal. 2020. A survey on Assistive Technology for visually impaired. *Internet of Things* 11 (2020), 100188. https://doi.org/10.1016/j.iot.2020.100188

[55] Natalina Martiniello, Werner Eisenbarth, Christine Lehane, Aaron Johnson, and Walter Wittich. 2022. Exploring the Use of Smartphones and Tablets Among People with Visual Impairments: Are Mainstream Devices Replacing the Use of Traditional Visual Aids? *Assistive Technology* 34, 1 (2022), 34–45.

[56] Microsoft. 2024. Seeing AI. Retrieved in July 29, 2024 from https://www.microsoft.com/en-us/ai/seeing-ai.

[57] John Morris and James Mueller. 2014. Blind and Deaf Consumer Preferences for Android and IOS Smartphones. In *Inclusive designing*. Springer, New York, NY, USA, 69–79.
[58] Karin Müller, Christin Engel, Claudia Loitsch, Rainer Stiefelhagen, and Gerhard Weber. 2022. Traveling More Independently: A Study on the Diverse Needs and Challenges of People with Visual or Mobility Impairments in Unfamiliar Indoor Environments. *TACCESS* 15, 2 (2022), 1–44.
[59] Masayuki Murata, Dragan Ahmetovic, Daisuke Sato, Hironobu Takagi, Kris M Kitani, and Chieko Asakawa. 2018. Smartphone-based Indoor Localization for Blind Navigation Across Building Complexes. In *PerCom'18*. IEEE, Piscataway, NJ, USA, 1–10.
[60] Ruth G Nagassa, Matthew Butler, Leona Holloway, Cagatay Goncu, and Kim Marriott. 2023. 3D Building Plans: Supporting Navigation by People Who Are Blind or Have Low Vision in Multi-Storey Buildings. In *CHI'23*. ACM, New York, NY, USA, 1–19.
[61] NUC. 2024. Ruby R8 – AMD Ryzen R7-4800U. Retrieved in November 25, 2024 from https://simplynuc.co.uk/wp-content/uploads/briefs/SimplyNUCProductBrief-CBM1r8RB.pdf.
[62] OpenAI. 2024. Hello GPT-4o | OpenAI. Retrieved in September 8, 2024 from https://openai.com/index/hello-gpt-4o/.
[63] Ouster. 2024. VLP 16. Retrieved in November 25, 2024 from https://ouster.com/products/hardware/vlp-16.
[64] Benjamin Poppinga, Charlotte Magnusson, Martin Pielot, and Kirsten Rassmus-Gröhn. 2011. TouchOver Map: Audio-tactile Exploration of Interactive Maps. In *MobileHCI'11*. ACM, New York, NY, USA, 545–550.
[65] Giorgio Presti, Dragan Ahmetovic, Mattia Ducci, Cristian Bernareggi, Luca Ludovico, Adriano Baratè, Federico Avanzini, and Sergio Mascetti. 2019. WatchOut: Obstacle Sonification for People with Visual Impairment or Blindness. In *ASSETS'19*. ACM, New York, NY, USA, 402–413.
[66] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International conference on machine learning*. PMLR, 8748–8763.
[67] Vinitha Ranganeni, Mike Sinclair, Eyal Ofek, Amos Miller, Jonathan Campbell, Andrey Kolobov, and Edward Cutrell. 2023. Exploring Levels of Control for a Navigation Assistant for Blind Travelers. In *HRI'23*. ACM, New York, NY, USA, 4–12.
[68] Laurel D Riek. 2012. Wizard of Oz Studies in HRI: a Systematic Review and New Reporting Guidelines. *Journal of Human-Robot Interaction* 1, 1 (2012), 119–136.
[69] Cartographer ROS. 2024. Cartographer ROS Integration. Retrieved in November 25, 2024 from https://google-cartographer-ros.readthedocs.io/en/latest/.
[70] Manaswi Saha, Alexander J Fiannaca, Melanie Kneisel, Edward Cutrell, and Meredith Ringel Morris. 2019. Closing the Gap: Designing for the Last-few-meters Wayfinding Problem for People with Visual Impairments. In *ASSETS'19*. ACM, New York, NY, USA, 222–235.
[71] Elen Sargsyan, Bernard Oriola, Marc JM Macé, Marcos Serrano, and Christophe Jouffrais. 2023. 3D Printed Interactive Multi-Storey Model for People with Visual Impairments. In *CHI'23*. ACM, New York, NY, USA, 1–15.
[72] Daisuke Sato, Uran Oh, João Guerreiro, Dragan Ahmetovic, Kakuya Naito, Hironobu Takagi, Kris M Kitani, and Chieko Asakawa. 2019. NavCog3 in the Wild: Large-scale Blind Indoor Navigation Assistant with Semantic Features. *TACCESS* 12, 3 (2019), 14.
[73] Nur Muhammad Mahi Shafiullah, Chris Paxton, Lerrel Pinto, Soumith Chintala, and Arthur Szlam. 2022. Clip-fields: Weakly Supervised Semantic Fields for Robotic Memory.
[74] Pierre Soille et al. 1999. *Morphological Image Analysis: Principles and Applications*. Springer, New York, NY, USA.
[75] Adam J Spiers and Aaron M Dollar. 2016. Outdoor Pedestrian Navigation Assistance with a Shape-changing Haptic Interface and Comparison with a Vibrotactile Device. In *HAPTICS'16*. IEEE, Piscataway, NJ, USA, 34–40.
[76] Seeed Studio. 2024. Jetson Mate Getting Started. Retrieved in November 25, 2024 from https://wiki.seeedstudio.com/Jetson-Mate/.
[77] Allan Wang, Christoforos Mavrogiannis, and Aaron Steinfeld. 2022. Group-based Motion Prediction for Navigation in Crowded Environments. In *CoRL'22*. 871–882.
[78] Luyao Wang, Qihe Chen, Yan Zhang, Ziang Li, Tingmin Yan, Fan Wang, Guyue Zhou, and Jiangtao Gong. 2022. Can Quadruped Navigation Robots Be Used as Guide Dogs?
[79] Xiyue Wang, Seita Kayukawa, Hironobu Takagi, and Chieko Asakawa. 2022. BentoMuseum: 3D and Layered Interactive Museum Map for Blind Visitors. In *ASSETS'22*. ACM, New York, NY, USA, 1–14.
[80] Michele A Williams, Caroline Galbraith, Shaun K Kane, and Amy Hurst. 2014. "just Let the Cane Hit It" How the Blind and Sighted See Navigation Differently. In *ASSETS'14*. ACM, New York, NY, USA, 217–224.
[81] Jingyi Xie, Rui Yu, He Zhang, Sooyeon Lee, Syed Masum Billah, and John M Carroll. 2024. Emerging Practices for Large Multimodal Model (LMM) Assistance for People with Visual Impairments: Implications for Design.
[82] Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. 2023. Pointllm: Empowering Large Language Models to Understand Point Clouds.
[83] Fan Yang, Dung-Han Lee, John Keller, and Sebastian Scherer. 2021. Graph-based Topological Exploration Planning in Large-scale 3d Environments. In *ICRA'21*. IEEE, Piscataway, NJ, USA, 12730–12736.
[84] Naoki Yokoyama, Sehoon Ha, Dhruv Batra, Jiuguang Wang, and Bernadette Bucher. 2024. VLFM: Vision-Language Frontier Maps for Zero-Shot Semantic Navigation. In *ICRA'24*. IEEE, Piscataway, NJ, USA, 42–48.
[85] Yan Zhang, Ziang Li, Haole Guo, Luyao Wang, Qihe Chen, Wenjie Jiang, Mingming Fan, Guyue Zhou, and Jiangtao Gong. 2023. "I Am the Follower, Also the Boss": Exploring Different Levels of Autonomy and Machine Forms of Guiding Robots for the Visually Impaired. In *CHI'23*. ACM, New York, NY, USA, 1–22.

A Appendix: Prompts to MLLM

In this section, we list full prompts to MLLM and LLM, which were used in this paper.

A.1 Prompt Used For Translating Native Language to English

As the research was conducted in a country where English is not spoken, we used the below prompt to translate any data obtained in the native language throughout the paper. Note that the authors manually refined the output to keep the nuances of the original language. This prompt was also used to translate the prompt engineered in the native language, which was fed into the MLLM for generating scene descriptions.

```
Please translate the given <Native Language> to English
    . Make sure to keep the nuances and context of the
     original text.
<Native Language>:  Text written in native Language
English:
```

A.2 Prompt Used In The Formative Study

Below is the prompt used to generate descriptions in the formative study.

```
# Instructions
Please describe the image.
The text you generate will be read directly to visually
     impaired individuals. Make sure your description
    is engaging so that visually impaired individuals
    can enjoy listening to it.
To describe the image, you must follow the rules
    outlined below.

## Rules you must strictly follow to comply with the
    instructions

### Rules on what you should do
1. Since visually impaired individuals will listen
    while walking, provide a description in one
    cohesive sentence. Please describe as many objects
     and their details as possible.
2. Generate the description in 1 to 4 sentences in
    total.
3. If necessary, first describe the overall layout or
    the general view of the location.
4. After that, identify the objects located on the left
    , in front, and on the right of the image, and
    explain the information required to understand the
     scene.
5. Always describe the scene in the following order:
    overall view, left side, front, right side.
6. When describing, use a tone similar to a guide for
    the visually impaired, such as "On the right,
    there is..."
```

```
7. If there is a store, make sure to include
     information about what the store offers (for
     example, the type of cuisine if it is a restaurant
     ). Also, include a description of the store's
     atmosphere (e.g., bright, calm).
8. Only describe objects that are clearly visible.
     Include descriptions of distinctive objects.
9. Create a description that is enjoyable to listen to
     and allows the listener to learn about their
     surroundings.

### Rules on what you should not do
10. Avoid unnatural words for the listener, such as "
     the image" or "viewpoint."
11. You do not need to include common and unremarkable
     objects (e.g., tables and chairs in a restaurant)
     in the description.
12. If there is nothing to describe in a particular
     direction (e.g., there is nothing on the right),
     you do not need to mention that direction.
13. Do not describe the floor, ceiling, shadows,
     distant unclear objects, or the brightness or
     darkness of the lighting.

## Response Format
If you generate a good description that follows the
     above rules, you will receive a tip.
Please respond in JSON format.
Include the image description under the "description"
     key.
Start your response with ```json\n{ to indicate the
     beginning of the JSON.

Here is an example of a response:
```json
{
"description": "<description of the surroundings>",
}
```
```

A.3 Prompts Used In The Formative Study

This section provides prompts used in the main study.

A.3.1 The Prompt for Generating Detailed Description. Below is the prompt used to generate a detailed description.

```
# Instructions
Please describe the image.
You are given three images that provide a view of your
     left, right, and front, as well as a view from a
     fisheye camera that captures the overall view from
      a high point of view.
The text you generate will be read directly to visually
      impaired individuals.
When writing the description, please aim to make it
     appealing so that it creates an enjoyable
     experience for the listener.
The most important thing is to provide detailed and
     specific information so that the listener can feel
      as if they are actually at the scene.
Being specific means describing the category or name of
      objects, their condition, and the role they play.
For example, a description like "circular wooden object
     " is not specific, but "a circular wooden table
     with YYY written on the nearby guide" is specific.
Similarly, "iron exhibit" is vague, while "a tall, iron
      exhibit, possibly XXX" is specific.
When describing the image, you must follow the rules
     below.

## Rules that must be followed to comply with the
     instructions
1. The description must be something that a visually
     impaired person can listen to while walking.
     Provide a coherent description in one block of
     text. Explain as many objects and their details as
      possible.
2. Keep the description to 3-4 sentences at most
     (120-240 characters).
3. Use polite language (honorifics).
4. Identify and describe objects located in the overall
      scene, to the left, front, and right that are
     necessary to understand the scene.
5. Only describe clearly visible objects. Include
     distinctive objects in your description.
6. Always describe the overall scene first, followed by
      objects on the left, in front, and then on the
     right.
7. Strive to include the following information:
    - Details about the building's interior and
        decoration.
    - Information about the layout of the building (
        such as whether the front is open, where walls
         are, and the directions one can go).
    - Information on the surrounding brightness and the
         amount of light coming through windows.
    - Information about people in the surroundings,
        their actions, clothing colors, and whether
        they are staff or customers.
    - If it's a store, provide information on whether
        the entrance is open like a terrace and
        whether guide dogs can wait there.
    - Include information on visible stores or exhibits
        . Be sure to mention their category (e.g., the
         type of food if it's a restaurant, or what
        kind of place the exhibit is). If possible,
        include the name of the place. For exhibits,
        state whether they are interactive or for
        viewing only.
    - When describing objects, be specific (mention the
         category or name). For example, if there's a
        counter, specify if it looks like a cafe
        counter.
    - Mention people walking toward the front if there'
        s a risk of collision.
    - Use numbers when explaining object positions (e.g
        ., "5 meters to the right").
    - If there is a sign or guidepost, describe what it
         is and read out the text written on it.
    - Read out visible text.
    - Use adjectives like futuristic, stylish, modern,
        or classic to make the exploration more
        enjoyable and to help the listener visualize
        the scene.
8. Do not use unnatural words for the listener like "
     image," "viewpoint," or "overall."
9. If there's nothing to describe in a certain
     direction (e.g., nothing on the right side), do
     not describe that direction.
10. Do not summarize or conclude with a description of
     the overall direction or scene when finishing the
     explanation.
11. Do not describe anything not visible in the image.
     Do not lie or hallucinate details.

## Response Format
If you follow the rules above, you will receive a tip.
If you ignore the rules, you will be penalized and have
      to pay a fine.
Please do your best to comply with these instructions.

Respond in JSON format.
First, include the initial description of the image
     under the "initial_description" key.
Next, include points for improvement under the "
     improve_thoughts" key.
Finally, include the revised image description under
     the "description" key.
Start your response with `json\n{`.
Here is an example response:
```

```
```json
{
"initial_description": "<initial description>",
"improve_thoughts": "<points for improvement>",
"description": "<revised description>",
}
```
```

A.3.2 The Prompt for Generating Balanced-Length Description. Below is the prompt used to generate a balanced-length description.

```
# Instructions
Please describe the image.
You are given three images that provide a view of your
     left, right, and front, as well as a view from a
     fisheye camera that captures the overall view from
      a high point of view.
The text you generate will be read directly to visually
      impaired individuals.
Keep the description concise, but aim to make it
     appealing and enjoyable for the listener.
The most important thing is to provide detailed and
     specific information so that the listener can feel
      as if they are actually at the scene.
Being specific means describing the category or name of
      objects, their condition, and the role they play.
For example, a description like "circular wooden object
     " is not specific, but "a circular wooden table
     with YYY written on the nearby guide" is specific.
Similarly, "iron exhibit" is vague, while "a tall, iron
      exhibit, possibly XXX" is specific.
When describing the image, you must follow the rules
     below.

## Rules that must be followed to comply with the
     instructions
1. The description must be something that a visually
     impaired person can listen to while walking.
     Provide a coherent description in one block of
     text. Explain as many objects and their details as
      possible.
2. Keep the description to 2-3 sentences at most
     (60-120 characters).
3. Use polite language (honorifics).
4. Identify and describe objects located to the left,
     front, and right that are necessary to understand
     the scene.
5. Only describe clearly visible objects. Include
     distinctive objects in your description.
6. Always describe objects in the following order: left
     , front, and right.
7. Strive to include the following information:
    - If it's a store, provide information on whether
         the entrance is open like a terrace and
         whether guide dogs can wait there.
    - Include information on visible stores or exhibits
         . Be sure to mention their category (e.g., the
          type of food if it's a restaurant, or what
         kind of place the exhibit is). If possible,
         include the name of the place. For exhibits,
         state whether they are interactive or for
         viewing only.
    - When describing objects, be specific (mention the
          category or name). For example, if there's a
         counter, specify if it looks like a cafe
         counter.
    - Mention people walking toward the front if there'
         s a risk of collision.
    - Use numbers when explaining object positions (e.g
         ., "5 meters to the right...").
    - If there is a sign or guidepost, describe what it
          is and read out the text written on it.
    - Read out visible text.
8. Do not use unnatural words for the listener like "
     image," "viewpoint," or "overall."
9. If there's nothing to describe in a certain
     direction (e.g., nothing on the right side), do
     not describe that direction.
10. Do not summarize or conclude with a description of
     the overall direction or scene when finishing the
     explanation.
11. Do not describe objects if you cannot provide
     specific information about them.
12. Do not include information about people in the
     surroundings unless there is a risk of collision.
13. Do not include information about the amount of
     light or brightness in the surroundings.
15. Do not use subjective adjectives like futuristic,
     stylish, modern, or classic.
16. Do not describe anything not visible in the image.
     Do not lie or hallucinate details.

## Response Format
If you follow the rules above, you will receive a tip.
If you ignore the rules, you will be penalized and have
      to pay a fine.
Please do your best to comply with these instructions.

Respond in JSON format.
First, include the initial description of the image
     under the "initial_description" key.
Next, include points for improvement under the "
     improve_thoughts" key.
Finally, include the revised image description under
     the "description" key.
Start your response with `json\n{`.
Here is an example response:

```json
{
"initial_description": "<initial description>",
"improve_thoughts": "<points for improvement>",
"description": "<revised description>",
}
```
```

A.3.3 The Prompt for Generating Concise Description. Below is the prompt used to generate a concise description.

```
# Instructions
Please describe the image.
You are given three images that provide a view of your
     left, right, and front, as well as a view from a
     fisheye camera that captures the overall view from
      a high point of view.
The text you generate will be read directly to visually
      impaired individuals.
The description should be concise and minimal, allowing
      the listener to quickly understand their
     surroundings.
Visually impaired individuals are listening to the
     image description to locate their destination.
The most important thing is to provide detailed and
     specific information so that the listener can feel
      as if they are actually at the scene.
Being specific means describing the category or name of
      objects, their condition, and the role they play.
For example, a description like "circular wooden object
     " is not specific, but "a circular wooden table
     with YYY written on the nearby guide" is specific.
Similarly, "iron exhibit" is vague, while "a tall, iron
      exhibit, possibly XXX" is specific.
When describing the image, you must follow the rules
     below.

## Rules that must be followed to comply with the
     instructions
1. The description must be something that a visually
     impaired person can listen to while walking.
     Provide a coherent description in one block of
     text.
```

```
2. Keep the description to 1-2 sentences at most (0-60
     characters).
3. Use polite language (honorifics).
4. Identify and describe objects located to the left,
     front, and right that are necessary to understand
     the scene.
5. Only describe clearly visible objects. Include
     distinctive objects in your description.
6. Always describe objects in the following order: left
     , front, and right.
7. Strive to include the following information:
    - If it's a store, provide information on whether
         the entrance is open like a terrace and
         whether guide dogs can wait there.
    - Include information on visible stores or exhibits
         . Be sure to mention their category (e.g., the
          type of food if it's a restaurant, or what
         kind of place the exhibit is). If possible,
         include the name of the place. For exhibits,
         state whether they are interactive or for
         viewing only.
    - Use numbers when explaining object positions (e.g
         ., "5 meters to the right...").
    - If there is a sign or guidepost, describe what it
          is and read out the text written on it.
    - Read out visible text.
8. Only convey specific information.
9. Keep the description short, direct, and concise.
10. Do not use unnatural words for the listener like "
     image," "viewpoint," or "overall."
11. If there's nothing to describe in a certain
     direction (e.g., nothing on the right side), do
     not describe that direction.
12. Do not summarize or conclude with a description of
     the overall direction or scene at the beginning or
      end of the explanation.
13. Do not describe decorations. Simply convey what is
     there and provide specific information.
14. To keep the length minimal, do not include
     subjective adjectives.
15. Do not include unnecessary information that does
     not help the listener locate their destination (e.
     g., details about furniture such as chairs or
     tables).
16. Do not describe objects if you cannot provide
     specific information about them.
17. Do not include information about people in the
     surroundings unless there is a risk of collision.
18. Do not include information about the amount of
     light or brightness in the surroundings.
19. Do not describe anything not visible in the image.
     Do not lie or hallucinate details.

## Response Format
If you follow the rules above, you will receive a tip.
If you ignore the rules, you will be penalized and have
      to pay a fine.
Please do your best to comply with these instructions.

Respond in JSON format.
First, include the initial description of the image
     under the "initial_description" key.
Next, include points for improvement under the "
     improve_thoughts" key.
Finally, include the revised image description under
     the "description" key.
Start your response with `json\n{`.
Here is an example response:

```json
{
"initial_description": "<initial description>",
"improve_thoughts": "<points for improvement>",
"description": "<revised description>",
}
```
```