

Collar-VLM: A VLM-Powered Collar Headset Supporting Daily Assistance for the Visually Impaired

Kaixin Chen, Yongzhi Huang, Usman Saleh Toro, Chang Huang, Yueling Che,
Kaishun Wu, *Fellow, IEEE*, and Lu Wang, *Senior Member, IEEE*

Abstract—Wearable assistive devices (WADs) for visually impaired people enable hands-free interaction and support richer usage scenarios than smartphone-based solutions. Following a user-centered design philosophy, we develop Collar-VLM, a collar-style headset prototype that integrates binocular cameras, a high-capacity battery, smartphone-side stereo depth perception, and a cloud-edge computing architecture into a discreet and ergonomic 3D-printed form factor. To support daily assistive functions, the software layer of Collar-VLM is powered by a vision-language model (VLM), which jointly interprets users' natural voice commands and camera-captured environmental images for intent understanding and semantic perception. The system maintains hierarchical long- and short-term memory on the smartphone, returns compact structured outputs, and performs depth-based spatial computation and proximity monitoring locally. In addition, parallel perception, bounded VLM outputs, memory summarization, and streaming speech feedback are introduced to reduce latency and enhance real-time interaction. Technical evaluation shows that Collar-VLM improves scene description, object and text recognition, and indoor mobility assistance over prompt-based VLM baselines. In an outdoor wild study, all participants successfully completed the route using Collar-VLM without remote human guidance. Compared with state-of-the-art glasses-based devices, Collar-VLM further provides longer battery life, reduced self-consciousness, better wearing comfort, and more stable camera views.

Index Terms—User-centered Design, Assistive Devices, Wearable.

1 INTRODUCTION

ACCORDING to the World Health Organization, there are about 285 million visually impaired (VI) people (VIP) around the world [1], [2]. Due to visual impairment, VIP experience significant or complete loss of visual perception, which substantially impacts their ability to interact with their environment and quality of life [3], [4]. Many common tasks in daily life pose significant challenges for them, such as cooking in the kitchen, navigating through pedestrians on the streets, shopping in supermarkets, and ordering food in restaurants.

To help VIP overcome these challenges and improve their quality of life, researchers have proposed a variety of assistive technologies (ATs). The most accessible ATs are those based on smartphone apps. For example, AI

applications like Envision AI [5] and Seeing AI [6] provide essential functions such as object recognition, color identification, and currency detection, thereby offering significant support to VIP. Recent advances in large multimodal language models have further expanded their adaptability across scenarios [7]. On the other hand, services such as Aira [8], and BeMyEyes [9] connect VIP with remote sighted assistants via smartphone cameras. It avoids the limitations of purely technological ATs by providing natural conversations, switchable contexts, and trustworthy companion experiences [10], [11]. However, smartphone-based solutions have drawbacks, such as occupying the user's hands and unstable camera view [12], [13]. For example, holding a mobile phone will hinder users from cooking [14], shopping [15], walking [15], participating in sports [16], and artistic creation [17]. In particular, using a handheld phone can be tiring in prolonged scenarios such as reading, public transportation, and social activities [18].

Recently, some wearable assistive devices (WADs) were introduced, to overcome the limitations of smartphone usage scenarios and the inconvenience of usage. For instance, Horizon Smart Glasses [19] connect VIP with trained professional agents. Envision Glasses [7], resembling Google Glass, allow VIPs to recognize paper currency and seek human assistance via video calls. OrCam [20] can recognize text and convert it into speech.

However, recent research indicates that these WADs, i.e., smart glasses, also face adoption hurdles due to inappropriate or uncomfortable form factors. For instance, Horizon Smart Glasses and Envision Glasses have been deemed as

- YL. Che, and L. Wang are with the College of Computer Science and Software Engineering, Shenzhen University, China. E-mail: wanglu@szu.edu.cn, wangqiqi2023@email.szu.edu.cn, and yuelingche@szu.edu.cn.
- KX. Chen, YZ. Huang, U. S. Toro, C. Huang, and KS. Wu are with the Hong Kong University of Science and Technology (Guangzhou), China. E-mail: kchen977@connect.hkust-gz.edu.cn, yhuang849@connect.hkust-gz.edu.cn, usmantoro@hkust-gz.edu.cn, chuang932@connect.hkust-gz.edu.cn, and wu@szu.edu.cn.

The research was supported in part by Guangdong Provincial Key Lab of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things (No.2023B1212010007), the China NSFC Grant (No.62372307, No.62472366, No.U2001207), 111 Center (No.D25008), Guangdong NSF (No. 2024A1515011691), Shenzhen Science and Technology Program (No. RCYX20231211090129039), Shenzhen Science and Technology Foundation (No.JCYJ20230808105906014, ZDSYS20190902092853047), the Project of DEGP (No.2024GCZX003, No.2023KCXTD042).
(Corresponding author: Lu Wang.)

peculiar and can induce self-consciousness among users [21], [22]. OrCam [20] attaches a rectangular camera module magnetically to the right side of glasses, but its unbalanced weight distribution has been found to cause discomfort in wearing, particularly around the nose and ears [23].

Thus, a primary challenge is to redesign a discreet WAD form, which not only needs to meet the demands of multi-scenario usability but also takes into account self-awareness and comfort. Low self-awareness of VIP is positively correlated with device adoption [24], and users are unlikely to adopt or discontinue the use of peculiar assistive devices [25]. Moreover, WAD should also be capable of allowing users to comfortably wear it for extended periods to provide assistance and companionship in long-term scenarios.

In response, we co-designed with VIP and developed a 3D-printed collar headset system (Figure 1). This design combines functional assistive capabilities with socially discreet and ergonomic wearability, suitable for prolonged wear and anytime use. Beyond hardware design, Collar-VLM is powered by a VLM-based cloud-edge software pipeline that jointly interprets users' natural voice commands and camera-captured environmental images for intent understanding and semantic perception. The smartphone maintains hierarchical context memory, performs stereo-depth-based spatial computation, and monitors near obstacles locally, while the cloud VLM returns compact task-specific structured outputs. Parallel perception, bounded VLM generation, memory summarization, and streaming speech feedback further improve real-time interaction.



Fig. 1: Collar-VLM collar headset prototype, depicted in (1) inactive mode, (2) frontal view, and (3) side view.

Our main contributions are as follows:

(1) Through formative research, we gained insights into the usage context, challenges, and design requirements of AT usage. This led to the ideation phase with VI participants as the protagonists, inspiring them to conceptualize potential wearable forms, including collar headset, regular glasses, and sunglasses form.

(2) We developed three prototypes and elaborated on the design considerations in detail, including hardware design, material and dimensions of shells, battery capacity, and weight management. Plus, we tested three different prototype forms and gained feedback from VI participants to refine the most promising collar headset design, resulting in the iterative version called Collar-VLM.

(3) We designed a VLM-based cloud-edge assistive software architecture that supports multimodal perception, hierarchical memory, compact structured outputs, stereo-depth-based spatial grounding, mobility safeguards, and latency-reduction mechanisms.

(4) We conducted technical evaluations and an outdoor wild study. The technical evaluation shows that Collar-VLM

improves scene description, object and text recognition, and indoor mobility assistance over prompt-based VLM baselines. The wild study further shows that Collar-VLM can provide assistance without remote human guidance, while reducing user self-consciousness and improving wearing comfort compared with XiaoQ Glasses.

2 BACKGROUND AND RELATED WORK

2.1 Smartphone-based ATs for VIP

Smartphones are widely used for developing ATs for VIP due to their ubiquity among this population. Applications such as Envision AI [5], LookTel Recognizer [26], and Seeing AI [6] provide essential functions like object recognition, color identification, and currency detection, thereby offering significant support to VIP. Additionally, the HCI community has developed various smartphone-based applications for different tasks, such as obstacle avoidance [27], queuing [28], navigation [29], and shopping [30]. Although smartphones have demonstrated promising results, a major drawback of their use is that they occupy the user's hands. More importantly, maintaining hand stability is fatiguing, and their long-term usability has rarely been evaluated.

2.2 Design of WADs for VIP

The forms of WADs for VIP proposed in current research include glasses, bone conduction earphones, rings, wristbands, gloves, and smartphones hanging on the chest. Glasses are a mainstream form. For example, OrCam [20] is a monocular camera glasses that can recognize words and faces. It magnetically attaches a strip-shaped camera module to the right side of the glasses and plays audio through the micro speaker above the ear. However, it was found that the unbalanced weight caused the nose and ears to wear uncomfortably [23]. Intoer [31], [32] is an intelligent eyeglass that can identify obstacles. It integrates RealSense R200 depth camera [33] and bone conduction earphones, and forms a sunglasses appearance. However, after adding cameras, hardware, and batteries, the appearance of the auxiliary glasses has undergone a significant form change (increased glasses thickness or thickened legs), which can easily attract the attention of bystanders [15], [22]. Some strange forms (such as Google Glasses, Horizon Smart Glasses [19], and Envision Glasses [7]) may be considered clumsy and rude by most people and abandoned [21], [22]. Horus [34] and AiSee [15] are bone conduction earphones integrated with cameras, which can identify product categories and text information. However, their cameras are easily obscured by their heads. Compared with the nose and ears of the head, the neck is more load-bearing [35]. Wearing a smartphone on the neck is a common wearable solution. Mobile phones worn on the chest are often used for face recognition and target recognition [36]–[38]. GIST [39] is a depth camera worn on the chest, allowing VIP to use different gestures to access the visual information (color and depth) of the space in front of them. However, hanging the mobile phone or camera on the neck with a single rope is easy to shake from side to side, and the image will be blurred [36], [40]. LiSee [41] adopts an unobtrusive neckband-headphone form for object-reaching assistance,

but focuses on a single task rather than multi-task assistance across daily scenarios. FingerReader [42], [43] is a ring with a camera to help with text recognition. DLWV2 [44] is a wrist strap with a monocular camera and five micro vibration motors to help VIP grasp target objects. GFR [45] is a watch with a camera, which can realize face recognition. However, the way of holding a mobile phone or wearing a camera on the hand requires VIP to maintain the orientation and stability of the hand to ensure the clarity of the image, which in fact is also a limitation on hand activity, and may cause unacceptable busy hands [42], [43], [46]. On the other hand, the voice feedback from the hand is easy to be heard by people nearby and causes unnecessary attention. ALVU [47] is a belt that can identify obstacles, integrating time of flight sensors and feedback motors. However, the waist vision is difficult to use for other tasks, such as identifying the target object on the desktop.

In summary, current WADs for VIP are primarily designed with task-oriented approaches, with few wearable devices considering multiple usage scenarios, and experimenting in social environments. In contrast, our aim is to design an easily adoptable WAD that can be used across various social settings, overcoming the main drawbacks of current WADs, including their weird appearance and poor comfort.

2.3 VLM-Based Visual Assistance for VIP

Recent advances in vision-language models (VLMs) have enabled a broader range of visual-assistance functions within a unified interaction framework [48]. The @BENCH benchmark [49] evaluates VLMs on tasks such as environment understanding, visual question answering, text recognition, and spatial perception, demonstrating their potential to address diverse visual-assistance needs of VIP. Prompt engineering further allows a VLM to flexibly adapt to different assistive scenarios. Existing studies have developed accessibility-oriented prompts for scene description [50], visual question answering [51], and local wayfinding assistance [52]. OpenGlass [53] and Vinci [54] further demonstrate the feasibility of deploying VLM-based visual assistants on wearable or portable devices.

While VLMs provide flexible support for everyday visual assistance, mobility guidance additionally requires accurate and reliable perception of the surrounding environment [55]. Collar-VLM therefore incorporates stereo depth sensing to provide more accurate spatial perception and support more reliable mobility guidance.

3 FORMATIVE STUDY

In order to gain a better understanding of the usage context, challenges, and design requirements of WADs for VIP, we conducted semi-structured interviews [56] with participants who utilize WADs.

3.1 Participants

We recruited VIP with experience using ATs from local organizations of the China Disabled Persons' Federation (CDPF), blind massage parlors, and the internet. When recruiting participants, we informed them of the purpose

of our interviews: to understand their usage of ATs and to facilitate the subsequent design of a WAD that is easy to adopt. In the end, we interviewed 78 VIP with varied careers (aged 21-62, including 42 males and 36 females, with an average age of 37.8 years [SD=12.1]).

3.2 Procedure

All interviews were conducted orally, with 28 participants interviewed at blind massage shop, 35 interviews conducted online, and the remaining interviews conducted at the participants' homes. We first collected their personal information, and then conducted interviews around themes of ATs usage frequency, usage scenarios, challenges encountered with existing ATs, and design requirements. All interviews were recorded and lasted 30-45 minutes. They were paid 20 RMB after completing the interview.

3.3 Results

After transcribing the recordings and organizing the notes, we carried out a statistical analysis.

We found that all participants had used smartphone-based ATs. Plus, 23.3% of participants had used WADs, e.g., XiaoQ Glasses [57]). Their most common usage scenarios were in unfamiliar social environments, such as traveling indoors and outdoors (92.8%), describing scenes (86.5%), and performing other miscellaneous tasks like shopping and ordering food in restaurants. 87.2% of participants claimed to use ATs every day, however, 72.6% complained about the inconvenience of handhelding smartphones and mentioned that adjusting and stabilizing the camera was annoying (68.0%). Meanwhile, 77.2% of participants had a positive attitude towards using wearable. Unfortunately, although participants' usage scenarios involved social environments, only 14.1% of smart glasses owners wore them when going out. Key factors included: odd or conspicuous appearance (73.5%), discomfort due to excessive weight or imbalance (61.8%), short battery life (48.7%), poor sound quality of bone conduction headphones (29.5%), and obstructive wired connections (21.3%). In addition to addressing existing smart glasses issues, they also mentioned some design elements, including not obstructing environmental sounds (58.1%), lightweight (53%), wireless connectivity (43%), and a sufficiently large field of view (FOV) (28.5%).

Overall, participants affirmed the wide range of uses and usefulness of smart glasses WADs, but they believed that existing smart glasses had design issues that hindered their adoption. Participants' positive attitudes towards WADs and a range of usage issues motivated us to co-design a cautious and comfortable WAD to facilitate participants' long-term use in social environments.

4 SYSTEM ARCHITECTURE

To achieve a lightweight, low-cost, and low-power wearable, we adopted a cross-platform cloud-edge computing architecture (Figure 2) [15], [42], [46], [58], [59], comprising a wearable client, a mainstream smartphone, and cloud VLM services. Fully utilizing smartphones, which are nearly ubiquitous among the visually impaired [60], can reduce the cost, power consumption, and weight of wearable. The wearable

TABLE 1: Detailed demographic information of 14 VI participants.

Participant	Gender	Age	Vision Condition	Diagnosis	Age of Vision Loss	Education	Occupation
P1	F	24	Blind	Glaucoma	0	University	Undergraduate student
P2	F	36	Low vision Visual acuity=0.1	Retinitis pigmentosa	16	High school	Beautician
P3	M	45	Low vision Visual acuity=0.2	Optic Atrophy	12	High school	Blind masseur
P4	M	33	Blind	Glaucoma	9	University	Accessibility engineer
P5	F	58	Blind	Cataract	25	Primary school	Unemployed
P6	M	49	Blind	Birth blind	0	Primary school	Blind masseur
P7	M	27	Low vision Visual acuity=0.2	Glaucoma	7	University	Music practitioners
P8	F	39	Low vision Visual acuity=0.1	Retinitis pigmentosa	9	Junior high school	Blind masseur
P9	F	43	Low vision Visual acuity=0.1	Cataract	13	Technical school	Blind masseur
P10	M	58	Blind	Cataract	14	Primary school	Unemployed
P11	F	23	Low vision Visual acuity=0.2	Glaucoma	21	High school	Undergraduate student
P12	M	32	Blind	Birth blind	0	Primary school	Blind masseur
P13	F	34	Blind	Improper medication	19	High school	Anchor
P14	M	37	Blind	Conjunctivitis	15	Primary school	Blind masseur

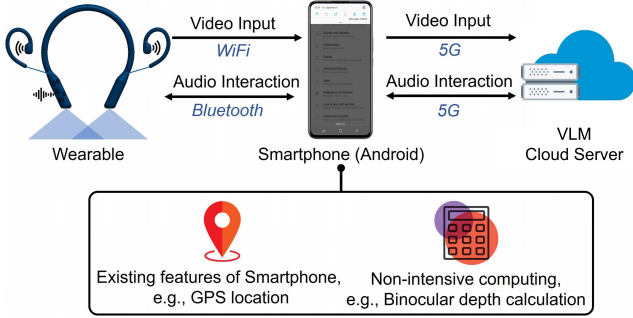


Fig. 2: A cross-end architecture combining wearable aids with existing mainstream devices (such as smartphones) and the cloud.

only needs to integrate essential camera modules, voice-interaction modules, and batteries to form a wearable interface. Due to limitations in computing power and battery life, wearable handles tasks with low computational and energy demands, such as video capture, transmission, and voice interaction. Mainstream devices, connected to wearables via WiFi and Bluetooth and to cloud VLM services via 5G, manage tasks that require more memory and computation (e.g., depth calculations) and run existing smartphone functions (e.g., navigation, emergency calling).

5 SYSTEM HARDWARE DESIGN

5.1 Form Factor Ideation

In this section, VI participants explore and design preliminary forms of potential WADs through brainstorming.

5.1.1 Participants

Fourteen VI participants in Table 1 (24-58 years old, including 7 men and 7 women, with an average age of

32.3 years [SD=15.4]) who participated in the previous formative study and had prior experience using WADs use experience participated in this ideation study. We informed them that our purpose was to work with them to design a user-friendly WAD. To ensure diverse perspectives, we intentionally selected participants from various age groups, educational backgrounds, and professions.

5.1.2 Procedure

Firstly, we presented the results of the formative study to help participants draw design insights from a broader population of VIP. They were also informed about the system's architecture (Section 4) to better integrate technology and design needs.

Then, we guided participants through a three-step brainstorming process using the Lotus Flower Method [61], which helps explore a central theme and generate more ideas by extending this theme. Through verbal communication, participants were first asked, "What are the contexts in which WADs is used?" In the second step, participants selected 2-3 usage contexts and answered important questions about the wearing position of the wearable [62], considering the conspicuousness, weight-bearing capacity, camera stability, and camera FOV in different parts of the body. Building on prior ergonomic research, we briefed participants about eight anatomically optimal wearable placements (e.g., waist and neck) known to provide both comfort and discretion [63]. In the third step, participants were encouraged to use their imagination to propose wearable forms while considering the wearing position. We suggested that the expanded design approach (adding new hardware to existing mainstream devices [?], [64]–[66]) is a commonly used method to ensure that the device is non-stigmatizing and not weird. Participants were instructed to focus primarily on the form factor, camera and earphone placement, and main weight distribution. Moreover, participants were requested to provide design considerations that support their

concepts. When researchers had differing understandings of participants' descriptions, participants were asked for further clarification.

After completing a 30-50 minute brainstorming session, each participant was compensated 50 RMB. Five researchers assisted with the ideation process, with each participant going through an independent session. Finally, five researchers summarized the wearable form proposed by participants using sketches and tables.

5.1.3 Results

Scenario Context: Guiding participants through brainstorming helped them recall various scenarios for using ATs. Participants reported various real-life challenges they encountered (each providing 3–7 examples), such as finding toys at home, navigating to train ticket gates, ordering from a menu at a restaurant, and describing unfamiliar scenes. Overall, the challenges they described encompassed a wide range of scenarios, including both indoor and outdoor environments, with a particular emphasis on mobility-related situations.

Consideration of Wearing Position: We found that the face (12/14), collar area (12/14), and ribcage (7/14) were the top three preferred wearing positions. All participants chose chest or higher positions to maximize their FOV for capturing daily activities. For example, P4 noted, *"Upper body placement allows a better view to see people we are talking to, items on a table, or things we are holding."* Twelve participants emphasized that the device should not be worn on the torso due to concerns about maintaining a stable view. P5 remarked, *"If worn on the hand or arm, the camera would shake, making it hard for the remote assistant to see clearly."* Eight participants suggested ear proximity for discreet sound feedback and privacy protection. P9 said, *"Considering the integration of the wearable, it should be near the ear to play sounds."*

Wearable Form and Design Considerations: In step three, participants proposed four primary form factors: (1) collar headset form (7 participants), (2) regular glasses form (4 participants), (3) sunglasses form (2 participants), and (4) bone conduction headset form (1 participant). Their proposed concept is shown in Figure 4(top), with expected weight concentration (pink regions). We excluded the bone conduction headset form due to its evident camera occlusion issues and limited endorsement. VI participants' concepts for this solution were very similar, but there were discrepancies regarding the weight distribution of the sunglasses form factor. One participant advocated for front-loaded weight distribution to accommodate additional hardware components, while another preferred side-balanced weighting to reduce nasal pressure. Thus, a compromise configuration was shown in Figure 4(top).

Their proposals were supported by the following thematic analysis results of design considerations. First, they all favored wearable forms that subtly extended from existing equipment to avoid obtrusiveness, with P1 stating: *"I reject a camera strapped to my head as it attracts stares; it should blend with daily accessories."* Second, ergonomic and practical advantages significantly influenced proposals: seven participants preferred collar headset form for its comfort (P11: *"I wore an assistive glasses device before, but the weight on my nose*

was too much"), five highlighted the collar's superior space for battery integration (P7: *"The collar's larger space allows for bigger batteries and extended runtime."*). Also, ten participants mentioned the importance of a lightweight design to reduce wearing pressure. Third, accessibility considerations were crucial. Five participants stressed the importance of preserving residual vision (P11: *"Side-mounted hardware maintains my light perception"*), while six emphasized unobstructed camera views to ensure reliable functionality (P4: *"Glasses or collars work, but bone conduction headsets might block the camera's view"*). Seven participants also noted that the necessity of wide-angle lenses for adequate visual information capture. This consideration serves two distinct purposes: it help reduce fatigue from having to carefully turn the head with the glasses forms (three participants), and (2) compensate for the limited camera mobility inherent in collar headset form (four participants). Nine prioritized maintaining environmental sound awareness (P8: *"Hearing traffic is non-negotiable for street navigation"*).

5.2 Prototype Implementation

Combining technical feasibility with VI participants' design proposals, this section presents a detailed implementation of three promising prototypes. Note that, the deployment strictly adheres to the VI participants' specified criteria regarding shells, camera and headphone positioning, weight distribution, and design considerations.

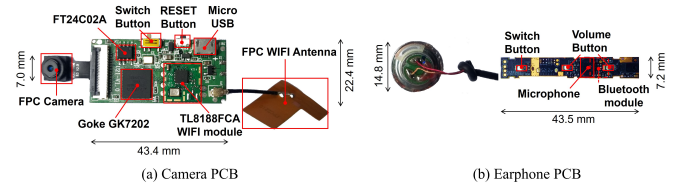


Fig. 3: WiFi camera PCB and Bluetooth headphone PCB.

5.2.1 Electrical Design

We considered excellent performance and ease of assembly when designing the following hardware:

Wide angle binocular camera: A camera is essential for capturing the VIP's surrounding environment for visual assistance. However, research has found that agents struggle to estimate distances from monocular video streams. Therefore, it is necessary to provide depth information to supplement monocular cameras [13], which can be obtained through binocular camera ranging [67], TOF sensing [68], structured light coding [69], or depth learning [70]. To achieve rapid response and reduce energy consumption, we opt not to choose additional sensors and complex computations, but rather opt for a binocular camera. Calculating depth information using methods based on binocular disparity and feature point matching can reduce computational overhead. To obtain a wider FOV, we select a 100° camera.

WiFi camera PCB and Bluetooth earphone PCB: As shown in Figure 3(a), we designed WiFi camera PCB that includes two buttons for setup and power control. It uses a TL8188FCA WiFi module supporting 802.11b/g/n with a 150M transmission rate and a 2.4G FPC antenna. Initial

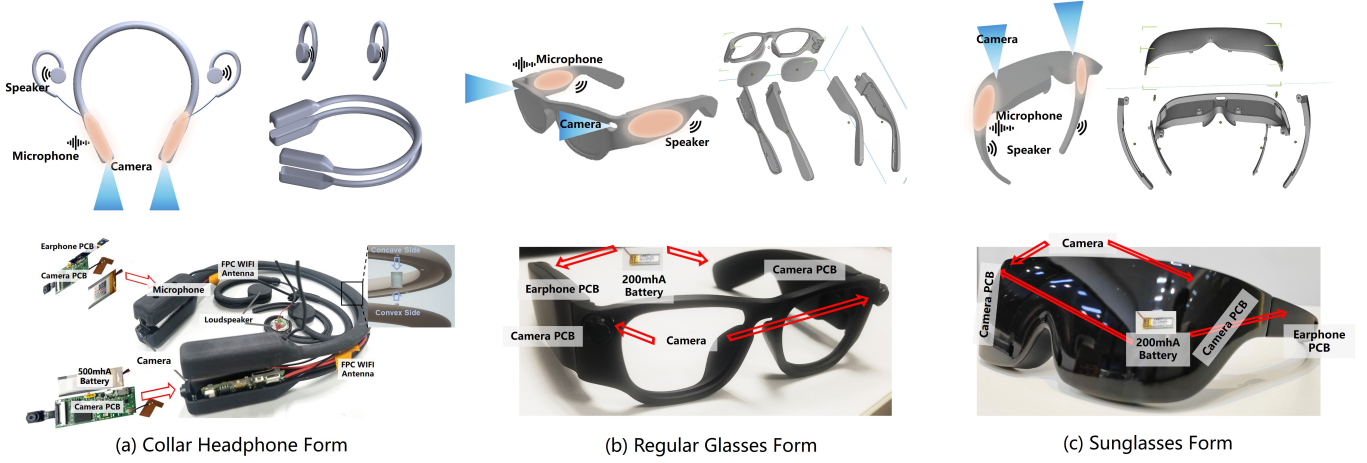


Fig. 4: VI participants' ideation (top) and implementation details (bottom) for three wearable form factors.

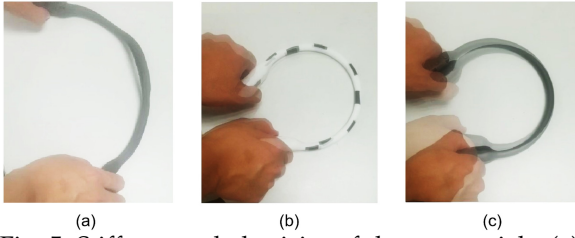


Fig. 5: Stiffness and elasticity of three materials. (a) Rubber. (b) Nylon. (c) Resin.

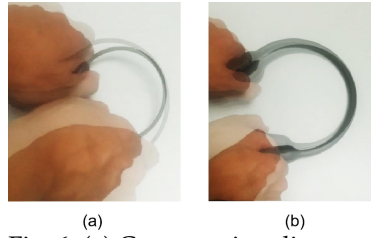


Fig. 6: (a) Cross-section diameter: 4mm, thickness: 1.2mm; (b) Ø: 9.5mm, thickness: 1.4mm.

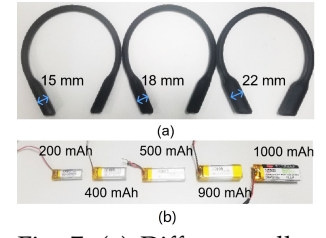


Fig. 7: (a) Different collar end cavity sizes' collars. (b) Batteries of different capacities.

occasional instability in video transmission was resolved by upgrading from an Anyka AK3918EN080V200 to a Goke GK7202 processor. Video is streamed to mobile phones at 30 FPS via Real-Time Streaming Protocol (RTSP). The push-pull button can simultaneously control the power switch of both modules. The Bluetooth earphone PCB has three buttons for power and volume control, a microphone, and a Bluetooth module, with wired speakers. Figure 3(b) shows layouts of the electronic components for one pair of elongated PCBs. Note that, we designed PCBs of different sizes to facilitate subsequent assembly on three prototypes.

Battery and other interfaces: We use two elongated polymer lithium batteries to power the WiFi camera PCB and the Bluetooth earphone PCB, respectively. Five battery capacities (Figure 7) are available for selection. The WiFi camera PCB also provides a Micro-USB charging interface and indicator lights for debugging purposes.

5.2.2 Shells Fabrication and Assembly

We used 3ds Max to model three types of forms and the Stereolithography Appearance process to print the shells.

Collar Headset Form: As shown in Figure 4, the collar is designed with a smooth arc for ergonomic fit around a cylindrical neck, with wider spaces at two ends to accommodate the WiFi camera PCB and batteries. The collar's end features a round hollow for camera assembly. We use an ear-hung design, positioning the speaker in front of the ear without blocking the ear canal, and create a rectangular hollow near the ear for sound spread. This ensures sound remains undistorted without being heard by nearby people. As shown in

Figure 5, we compared three materials (rubber, nylon, resin) with thermophysiological and skin sensorial wear comfort [71]. Resin (acrylonitrile-butadiene-styrene) was chosen for its strong stiffness keeping the baseline distance between the two cameras constant and for good elasticity to tolerate the stretching during putting on/taking off. We measured the neck and ear dimensions of one participant (P14) to design the sizes of the collar and headphones. The diameter of the collar is set to 135mm, which is larger than P14's neck width of 122mm, to ensure a loose and comfortable fit. The dimensions of the headphones are designed to be 64mm in height and 39mm in width for a snug fit. As shown in Figure 6, we set the shell thickness to 1.4 mm and the collar cross-sectional diameter to 9.5 mm, balancing strength and resilience for a 68 mm pull-on/take-off. As shown in Figure 7, we designed different-sized collar end cavities and assembled batteries of different capacities (200 mAh, 400 mAh, 500 mAh, 900 mAh, and 1000 mAh). We asked participants to wear and choose. We found that the participants prefer to choose a thinner shape (18 mm wide collar end) with a larger battery capacity (500 mAh), which maintains a larger battery capacity while ensuring a lower wearing pressure (70 g) and an unobtrusive appearance. The total weight of 70 g is significantly below the 169 g neck load capacity threshold [35].

Regular Glasses Form: We started the design with an ergonomic standard eyeglass frame. The prototype is made of ABS, with a shell thickness of 1.4 mm. Two camera PCBs and a earphone PCB are deployed on the sides of the glasses' legs. An opening for sound propagation was made

on the glasses' legs near the ear. The cameras are positioned at the circular openings at the ends of the glasses frame, rather than in the center, to reduce oddness. To reduce the weight on the nose and avoid significant shape changes, we discreetly placed two 200mAh batteries at the end of the legs. A total weight of approximately 73 g complies with the 75 g bearing capacity limit for ear and nose bridge support established in [66]. Figure 4 (bottom) shows the assembled appearance.

Sunglasses Form: As shown in Figure 4, we designed the front of the glasses with a single-piece cavity, where two camera PCBs are strategically mounted beneath transparent sunglass lenses for optimal space utilization. The earphone PCB is placed on the side of the glasses for easy volume adjustment. Two 200mAh batteries are deployed on the side to reduce the weight on the nose. The opening that allows the speaker sound to play is also located above the ear. The main part of the shell is made of ABS with a thickness of 1.4mm. The back of the glasses is designed to be made of transparent resin, while the positioning of the camera PCBs is optimized to avoid obstructing the remaining vision of individuals with visual impairments. This prototype weigh approximately 78 grams, close to meeting the weight tolerance for both the nose and ears [66].

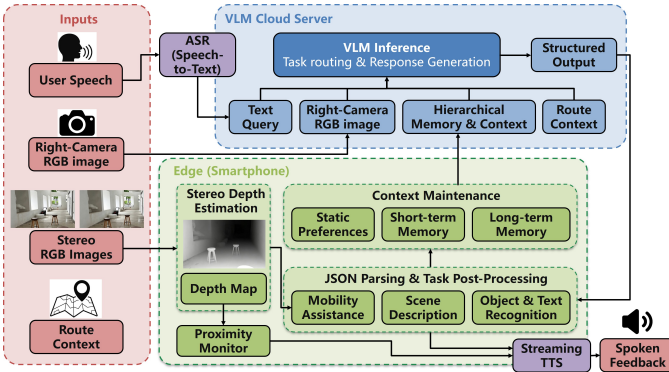


Fig. 8: Cloud-edge processing pipeline of Collar-VLM.

6 SYSTEM SOFTWARE DESIGN

6.1 Cloud-Edge Processing Pipeline

Collar-VLM combines cloud-based vision-language model (VLM) understanding with on-device spatial perception to support three representative assistive tasks: mobility assistance, scene description, and object and text recognition based on the formative study in Section 3.

As shown in Figure 8, user speech is transcribed into text, while the right-camera RGB image, text query, interaction context, and a fixed task-aware prompt are provided to the cloud VLM. Navigation context is additionally supplied for mobility assistance. In parallel, stereo depth is computed locally on the smartphone and is never uploaded to the cloud. The structured VLM response is parsed on-device and converted into concise spoken feedback through streaming TTS. For mobility assistance, stereo depth and route context further support local walking guidance and proximity warnings.

6.2 Parallel Multimodal Perception

Speech recognition and stereo depth estimation run in parallel. For the i -th interaction, the ASR module $\mathcal{A}$ converts user speech a_i into a text query:

$$q_i = \mathcal{A}(a_i), \quad (1)$$

where q_i denotes the recognized request.

The right-camera RGB frame is denoted by I_i and directly used as the VLM visual input. Meanwhile, the smartphone applies Semi-Global Matching (SGM) [41], [72] to the stereo images:

$$D_i = \mathcal{S}(I_i^L, I_i^R), \quad (2)$$

where D_i is the depth map aligned with I_i . Depth data remain on-device.

6.3 Hierarchical Context and Memory

Collar-VLM maintains hierarchical context without unbounded prompt growth:

$$M_i = \text{Concat}(C, S_i, R_i), \quad (3)$$

where C stores static user preferences, S_i denotes short-term session memory, and R_i contains retrieved long-term records.

In our implementation, S_i contains a rolling summary and the three most recent dialogue turns, with earlier turns merged into the summary. Long-term memory stores at most 50 compact records. For each query, local BM25 retrieval [73] selects up to three relevant records:

$$R_i = \text{TopK}_{\text{BM25}}(q_i, \mathcal{L}, 3), \quad (4)$$

where $\mathcal{L}$ denotes the long-term memory store. Memory is updated in the background after each interaction, while raw images, depth maps, and audio are not retained.

6.4 Structured VLM Interaction

Collar-VLM uses a fixed system prompt Π defining supported tasks, task-specific instructions, evidence constraints, and the JSON output format. The VLM processes the RGB image I_i , text query q_i , interaction context M_i , and navigation context N_i :

$$P_i = \mathcal{V}_\theta(I_i, q_i, M_i, N_i; \Pi), \quad (5)$$

where N_i contains the current route instruction and upcoming maneuver for mobility-assistance tasks, and $N_i = \emptyset$ otherwise. Stereo depth D_i is not provided to the VLM.

Task routing and response generation are completed within a single VLM call, without a separate task classifier. The output follows a compact JSON schema:

$$P_i = \{\text{task}_i, \text{data}_i\}, \quad (6)$$

where task_i identifies the active task and data_i contains scene semantics, ordered text, or a local waypoint coordinate. The final spoken response is constructed locally from data_i .

6.5 Task-Specific Assistance

The three tasks share the same VLM pipeline but differ in their outputs and on-device post-processing.

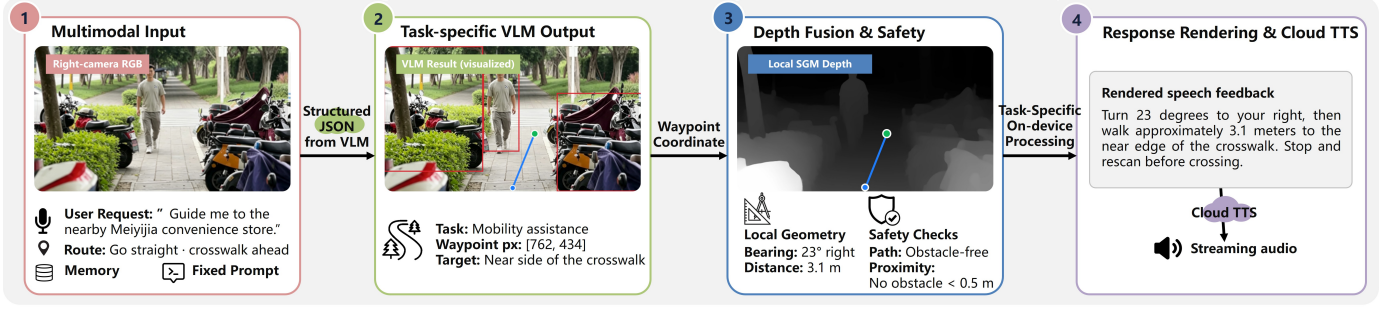


Fig. 9: Deployment-time execution workflow of Collar-VLM, illustrated with a mobility-assistance example.

6.5.1 Scene Description

The VLM summarizes the visible environment or answers scene-specific questions, focusing on relevant objects, activities, and spatial relationships. The result is passed directly to the response renderer.

6.5.2 object and text recognition

The VLM identifies task-relevant objects or products and extracts visible words and numbers in reading order. For field-specific requests, such as brand, date, price, or volume, it returns only the requested information. Long text is split into ordered segments for progressive speech playback.

6.5.3 Mobility Assistance

As shown in Figure 9, during mobility assistance, N_i provides the current route instruction and upcoming maneuver. Based on N_i and the current RGB image, the VLM selects a visible local waypoint consistent with the planned route and returns its pixel coordinate:

$$\mathbf{p}_i = (u_i, v_i). \quad (7)$$

Valid depth measurements within a neighborhood $\mathcal{N}(\mathbf{p}_i)$ are averaged to estimate the waypoint depth:

$$z_i = \text{Mean} \{ D_i(u, v) \mid (u, v) \in \mathcal{N}(\mathbf{p}_i), D_i(u, v) \text{ is valid} \}. \quad (8)$$

Using focal length f_x and principal point c_x , the waypoint's horizontal position, bearing, and distance are computed as

$$x_i = \frac{(u_i - c_x)z_i}{f_x}, \quad (9)$$

$$\alpha_i = \text{atan2}(x_i, z_i), \quad d_i = \sqrt{x_i^2 + z_i^2}, \quad (10)$$

where α_i is the bearing relative to the wearable's forward axis and d_i is the physical distance to the waypoint.

Before issuing guidance, the smartphone checks depth along the movement direction toward $\mathbf{p}_i$. If the path is clear, a heading adjustment is issued when

$$|\alpha_i| > \tau_\alpha, \quad \tau_\alpha = 20^\circ. \quad (11)$$

The system then combines turning direction and waypoint distance into a concise instruction; otherwise, it directly provides a forward-movement instruction. If the intended path is blocked, the movement instruction is suppressed and a proximity warning is issued. After each local movement,

new RGB and stereo frames are captured and N_i is updated. The updated route context and RGB image are then used to select the next waypoint, forming a closed loop until the next maneuver or destination.

6.6 Response Rendering

Let t_i denote the task identified by task_i . The structured VLM output is rendered into concise text:

$$r_i = \mathcal{R}_{t_i}(P_i, G_i), \quad (12)$$

where

$$G_i = \begin{cases} (N_i, \alpha_i, d_i), & \text{mobility assistance,} \\ \emptyset, & \text{otherwise.} \end{cases} \quad (13)$$

The resulting text is delivered through streaming TTS:

$$y_i = \mathcal{T}(r_i), \quad (14)$$

where y_i denotes the audio feedback.

6.7 Response Acceleration and Mobility Safeguards

Collar-VLM reduces latency through parallel processing, compact context, bounded output, and streaming audio feedback. ASR and stereo depth estimation run in parallel, while VLM inference can proceed concurrently with local depth computation. Streaming TTS and memory updates also run concurrently after response generation.

Hierarchical memory limits prompt growth, while the fixed JSON schema and output-token limit constrain generation to task-relevant content.

During mobility assistance, an independent on-device proximity monitor examines a 30° horizontal sector centered on the wearable's forward axis. For horizontal pixel coordinate u ,

$$\beta(u) = \arctan\left(\frac{u - c_x}{f_x}\right), \quad (15)$$

and the monitored region is

$$\Omega_f = \{(u, v) \mid |\beta(u)| \leq 15^\circ\}. \quad (16)$$

A near-depth measurement is considered reliable when at least 15 pixels in its 5×5 neighborhood are within 0.5 m (Figure 9). Once such a locally consistent region is detected, a high-priority proximity warning is issued. This suppresses isolated stereo-matching outliers without smoothing the scene's natural depth variation.

6.8 Implementation Details

We implement the cloud VLM using Qwen3-VL-Flash-2025-10-15 through the DashScope API, with temperature set to 0 and a maximum output length of 192 tokens. Speech is transcribed using the Paraformer real-time ASR API, while CosyVoice-v2 provides streaming TTS.

As shown in Figure 9, the fixed system prompt contains common routing instructions and task-specific sub-instructions for scene description, object and text recognition, and mobility assistance. For mobility assistance, navigation context N_i is added as a dynamic input while the system prompt remains fixed. Task routing and response generation are completed within a single VLM call. The prompt requires only the predefined `task` and `data` fields and prevents the VLM from estimating metric distance or making depth-based proximity judgments. These quantities are computed locally from stereo depth. Full prompt templates, JSON schemas, parsing rules, and deployment settings are provided in the supplementary material.

7 PILOT STUDY

This pilot study evaluated whether the deployed prototypes met participants' expectations in daily tasks. It was conducted before integrating the VLM-based software pipeline. Therefore, a human agent was used to isolate form-factor and wearability feedback from AI-related errors.

7.1 Prototype Apparatus and Setup

All studies used the three prototypes we developed (Figure 4), with a Samsung S9 phone as the edge device. Participants were informed how to use the system and how it connects to the phone. A 38-year-old male with 6 years of experience in remote assistance for VI individuals acted as an agent. He was aware of these three tasks in advance. He provided remote assistance to VI participants using the right-side camera images and depth images of the prototype. Note that, we used human agents instead of AI to minimize the frustrations that AI can sometimes cause, allowing participants to focus more on the wearability aspect.



(a) Scenario 1

(b) Scenario 2

(c) Scenario 3

Fig. 10: Experimental setup: Three scenarios mentioned by participants in the formative study that they typically encounter when using ATs.

7.2 Participants, Tasks, and Procedure

All participants listed in Table 1 were invited to test prototypes in this study. However, P10 did not participate in the study since he was not in the region. Our experiment

adhered to the Declaration of Helsinki [74], and all participants provided informed consent.

We defined three tasks based on problem scenarios encountered by users in formative research and considering participant safety. (1) Passing through corridors while avoiding pedestrians. In a corridor 2 meters wide and 60 meters long, researchers stood at random positions and blocked the path of VI individuals. Participants need to avoid pedestrians while crossing corridors. (2) Familiarizing with four scenes in unfamiliar environments. Participants were taken by researchers to four different scenes near the laboratory (elevator, conference room, classroom, and car interior downstairs). Participants needed to identify these scenes. (3) Ordering from a menu. Participants were asked to identify a menu containing 12 dishes in a meeting room and order three. Researchers acted as waiters and recorded the ordered dishes. Each VI participant was required to use all three prototypes to perform each task, with the order of device use and task settings randomized to minimize memory effects. Before performing the tasks, they were advised to feel the user experience when using these three prototypes. Participants continuously completed these three tasks. After finishing all tasks, they were asked to select the wearable form they were most likely to continue using and to provide feedback and suggestions for improvement. They were paid 60 RMB after completing all the tasks. The study lasted 70–100 minutes.

7.3 Results

All participants successfully completed all tasks using different prototypes, and they all reported that the prototypes met their expectations. Ten participants chose the collar headset, while three participants chose the regular glasses. The primary reason for selecting the collar headset was its comfort during prolonged wear (mentioned 9 times). The second reason was that they appeared less odd (6 times). The reasons for choosing the glasses were their alignment with participants' aesthetic preferences (mentioned 3 times) and comfortable wearing experience (2 times). Moreover, the sunglasses form was not chosen by the participants due to the excessive weight on the nose (8 times) and the bulky design (5 times).

Since more VI participants preferred the collar headset over the regular glasses, we believe the collar headset form has greater potential and focus on improving the collar headset. Suggestions for improving the appearance of the regular glasses include adding lenses to make them look less unusual (mentioned by P5 and P12) and an adjustable nose bridge to fit different nose shapes (P4 and P7).

Participants who chose the collar headset form provided the following feedback.

Camera Orientation: Six participants mentioned that the camera orientation was not directly facing forward, sometimes requiring manual adjustments to slightly tilt it toward the right. We also observed this phenomenon in the videos, even though we used a lens with a large FOV of 100°.

Comfort: Participants also considered the current prototype's skin sensorial wear comfort could be improved. For example, six participants mentioned that the end of the collar was too square and did not conform well to the skin,

causing neck discomfort. Additionally, five participants mentioned that the cylindrical part of the external earpiece could not be attached to the skin, easily protruding outward. Furthermore, two participants mentioned the issue of the prototype's shell not being smooth enough.

Camera Stability: Four female participants complained that the collar frame was too loose, causing it to shake while walking, and were reminded by remote agents to maintain a more stable gait. This may be an oversight because we only measured the neck size of a male participant (Section 5.2.2), as female participants may have smaller necks.

8 PROTOTYPE ITERATION

The pilot study raises several important issues regarding the collar headset prototype. Thus, we do prototype iteration in this section to solve these issues and improve the prototype.

8.1 Modifying Camera Orientation

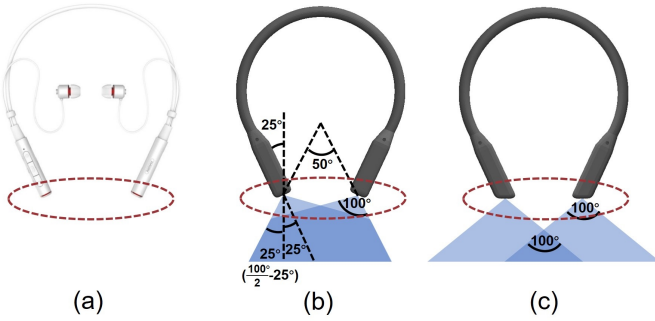


Fig. 11: We modified the end of the collar so that the vision of the left and right cameras are parallel to the front and obtain a wide enough FOV.

The shape of the end of the current collar (Figure 11(b)) is very similar to that of the ordinary collar headset (Figure 11(a)). However, because the fields of view of both cameras are inward-facing (Figure 11(b)), it results in the right camera pointing towards the left front, which is not conducive for remote agents to provide assistance. Moreover, we found that the FOV of the cameras was also reduced. Therefore, as shown in Figure 11(c), we modified the end of the collar so that the visual axes of the left and right cameras are parallel to the forward direction.

8.2 Improve Shape of Collar and External Earpiece

Based on user feedback, we have improved the originally square-shaped collar end into a more rounded form, smoothing out the corner positions and providing a smoother transition in curvature. Furthermore, recognizing that the tragus protrudes outward, and the cylindrical speaker portion of the external earpiece tends to overlap with the tragus, we have modified the cylindrical speakers into a teardrop shape to avoid interference with the tragus. The revised collar and external headphones are depicted in Figure 12 (b), featuring a smoother overall curvature for enhanced aesthetics. In addition, we polished the prototype with acetone to make it even smoother, and the final prototype is shown in Figure 1.

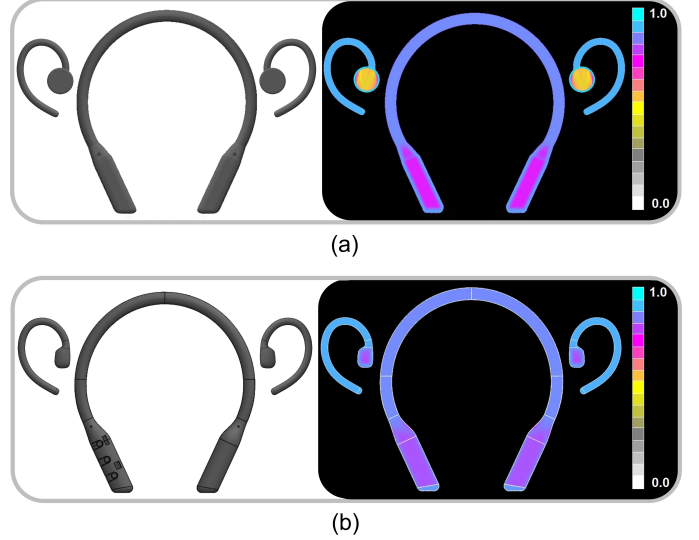


Fig. 12: We modified the shape of the collar and the external earpiece and made their curvature smoother.

8.3 Personalized Collar Size

We measured the sizes of all participants' necks and ears. Figure 13 is the measured results. We observed that male necks had significantly greater width and depth compared to females, while the difference in ear size was not significant. Therefore, we designed a variety of collar sizes, with widths and depths ranging from 95 mm to 128 mm in 3 mm increments, resulting in twelve different sizes. This allows users to choose a collar that fits their neck, ensuring camera stability. Note that, customizing collar sizes is not difficult in industrial production, and the fitting process for VIP takes about 5 minutes.

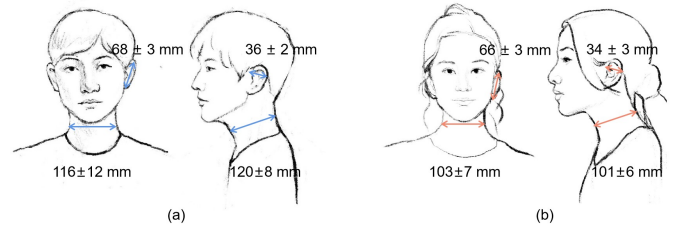


Fig. 13: (a) Men's neck size and ear size. (b) Women's neck size and ear size.

8.4 Adding Camera Switch Prompt

Considering the social impact in outdoor environments, bystanders may want to know whether the camera is on [75]–[78]. We adopted the constructive suggestion of the study [79], that is, to adopt a design with obvious meaning and an opaque occlusion. As shown in Figure 14, we designed an opaque slideable switch that, when turned on, the camera resembles an open eye, and the curved "eyebrows" convey a friendly meaning. When the switch is off, the camera is completely blocked to eliminate privacy concerns.

9 TECHNICAL EVALUATION

After integrating the VLM-based software pipeline, we conducted a technical evaluation to assess Collar-VLM's



Fig. 14: (a) There is a curved “eyebrows” when the camera is turned on. (b) The camera is completely blocked when the camera is turned off.

task performance, latency, power consumption, and module contributions.

9.1 Participants

We recruited 8 new participants (4 male, 4 female; 5 blind, 3 low-vision), aged 22–47 with vision loss onset from 0 to 23 years. All used assistive technologies before, signed informed consent, and received 30 RMB compensation post-experiment. Given the exploratory nature of the technical evaluation, the results are interpreted as user-centered evidence of system performance rather than population-level estimates.

9.2 Apparatus

Participants wore our revised Collar-VLM prototype paired with a Samsung Galaxy S9. We compared Collar-VLM with three software baselines under the same camera input, ASR, TTS, and VLM backend: Vanilla VLM, a prompt baseline inspired by OpenGlass [53], and a prompt baseline inspired by Vinci [54]. These baselines were implemented as prompting variants rather than full reimplementations of the original systems, allowing us to isolate the effect of Collar-VLM’s structured output, memory, and depth-based spatial grounding. Each participant completed all three tasks on every system. We adopted a counterbalanced design with distinct, difficulty-matched scenes, products, and routes to eliminate memory bias.

9.3 Procedure

Scene tests covered eight indoor environments, requiring environment classification and queries about key objects or spatial relations. Product trials used packaged daily goods; participants identified items and requested attributes such as brand or capacity. Mobility tasks adopted fixed 10–50 m, 2 m-wide routes with two obstacle types: pedestrian blockages and static objects, with 1–3 obstacles on each visible path. A brief pre-test tutorial familiarized users with voice commands and workflows. A researcher accompanied participants during mobility trials only for safety, without task guidance. The phone logged all module latencies, task durations and obstacle warnings.

9.4 Evaluation Metrics

For scene description, we report scene type accuracy and key element spatial matching accuracy, measuring the system’s ability to classify environments and answer spatial and object queries. For object and text recognition, we use

object identity accuracy and requested field extraction accuracy to quantify attribute retrieval performance. For indoor mobility, core metrics include task completion rate, traversal time, and obstacle warning count.

9.5 Quantitative Comparison with Baselines

Figure 15 compares Collar-VLM with baseline methods across scene understanding, object-text recognition, and indoor navigation. Collar-VLM achieves the best overall performance by combining VLM reasoning with on-device spatial perception.

For scene understanding, Collar-VLM improves spatial element matching accuracy from 78% to 93% compared with vanilla VLM, enabling better alignment between semantic descriptions and physical layouts. For object and text recognition, task-specific VLM outputs and post-processing achieve 89% accuracy in target text-field extraction.

The largest gains are observed in indoor mobility. Collar-VLM increases task completion rate from 56% to 88% ($1.57\times$ higher than vanilla VLM), while reducing navigation time by 31% and obstacle warning frequency by 66%. These results demonstrate that combining cloud-based semantic reasoning with on-device spatial perception enables reliable assistive interaction in real-world scenarios.

9.6 Latency Analysis

Figure 16 compares latency across methods. Collar-VLM achieves lower VLM latency and shorter output sequences across all three tasks, mainly because the structured response constrains generation to task-required fields. The dominant latency components are ASR, cloud VLM inference, and initial TTS synthesis. RGB upload and smartphone-side SGM ranging are executed in parallel with ASR, and their measured latencies, 33 ms and 104 ms respectively, are both shorter than ASR latency. Therefore, the latency of this parallel front-end stage is determined by ASR, and RGB upload and SGM ranging are not shown as independent critical-path components in Figure 16. Other local operations are negligible, including BM25 retrieval over the 50-record local memory store (0.45 ms) and JSON parsing (<0.01 ms).

9.7 Power Consumption

We measured the prototype’s power consumption using an IT9100 power meter [80]. The WiFi camera PCB and Bluetooth headphone PCB consumed 1123.2 mW and 47.2 mW during continuous operation, respectively. Therefore, the total hardware power consumption was 2293.6 mW for two camera PCBs and one Bluetooth audio PCB. During startup, the average power consumption was 360.3 mW, and startup took approximately 5 s. Powered by two 500 mAh batteries, the current prototype can operate continuously for about 2 hours. This duration is sufficient to support a long journey. Note that by replacing the batteries with ones of higher density, the battery life can be improved.

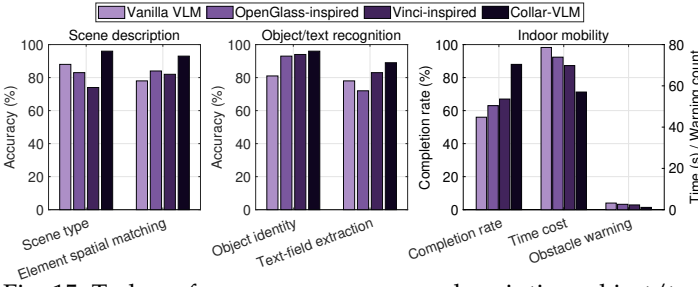


Fig. 15: Task performance across scene description, object/text recognition, and indoor mobility.

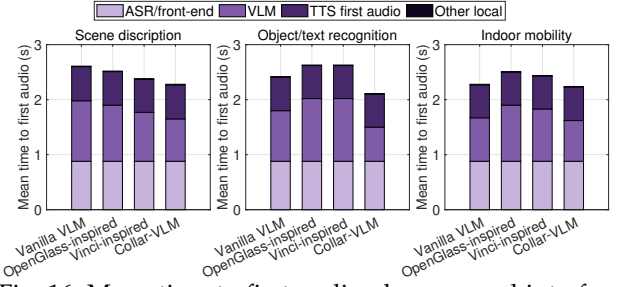


Fig. 16: Mean time to first audio, decomposed into front-end, VLM, TTS, and local processing latency.

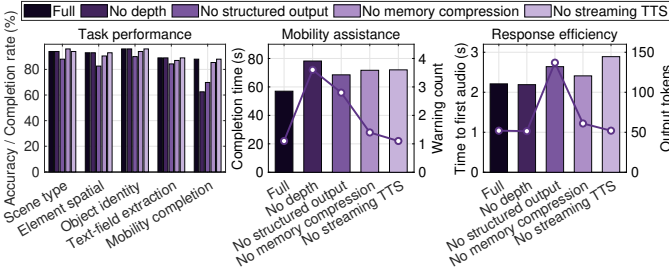


Fig. 17: Ablation study on task performance, mobility-assistance outcomes, and response efficiency.

9.8 Ablation Study

We ablate four key components of Collar-VLM: depth-based spatial perception, structured VLM output, memory compression, and streaming TTS. Results are shown in Figure 17.

Removing depth perception has limited impact on scene understanding but reduces mobility completion from 88.0% to 62.5%, while increasing completion time from 57.0 s to 78.2 s and warning frequency from 1.1 to 3.6, demonstrating the importance of depth for spatial grounding and safe navigation.

Without structured output, object-text recognition drops from 89.0% to 84.3% and output tokens increase from 52 to 137, showing that constrained formats reduce generation ambiguity. Removing memory compression increases tokens from 52 to 61 and latency from 2.2 s to 2.4 s, confirming the benefit of compact context management. Disabling streaming TTS keeps task accuracy unchanged but increases first-audio latency from 2.2 s to 2.9 s.

10 WILD STUDY

We conducted user studies in outdoor environments with higher demands for camera FOV and stability. Plus, since the scenarios in the pilot study were controlled, we aimed to investigate participants' self-awareness and long-term wearing comfort regarding the prototype in social environments, as these are two of the most important design objectives.

10.1 Participants, Apparatus and Setup

The same eight participants from Section 9.1 joined the wild study, and each received compensation of 50 RMB upon completing all assigned tasks. As international products such as Horizon Smart Glasses [19] and Envision Glasses [7] are not available domestically, we compared our prototype with XiaoQ Glasses [57], a representative commercially

available WAD in China. It consists of glasses with a 60° FOV camera in the upper right corner of the frame connected to a host with a battery. Collar-VLM generated all guidance through the proposed VLM-based pipeline without remote human assistance. XiaoQ Glasses [57] were used in their standard remote-assistance mode, where the same trained sighted agent from Section 7.1 provided guidance through the camera stream. The agent knew the task and route in advance. XiaoQ Glasses provide remote-agent voice feedback within 0.7 s after an assistance request. Since the two devices differ in form factor, camera placement, field of view, depth sensing capability, and audio interaction interfaces, this experiment provides an end-to-end overall comparison of wearables instead of a controlled test focusing on a single hardware variable. While performing the task, a researcher follows participants to ensure their safety, while another researcher records videos for observational analysis.



Fig. 18: Experiment setup illustration (on the left is a VI participant wearing xiaoQ Glasses, and on the right is a VI participant wearing Collar-VLM).

10.2 Tasks and Procedure

Participants first choose a collar size that fits their neck, making sure it is slightly attached to the neck and not easily swayed. Participants descended the stairs from the second-floor laboratory between 3:00 PM and 5:00 PM and walked to the nearest supermarket to purchase a bottle of water. Along the way, they passed a traffic light junction and then passed the office area, where many people were working. The distance of the journey was approximately 500 meters. They used our iterated prototype (Collar-VLM) in one direction and commercial smart glasses WAD (XiaoQ Glasses [57]) in the other direction, with the order of using these two methods balanced. We recorded the time each participant spent completing the tasks using these two wearables, and

the number of triggered proximity warnings. Afterward, participants completed a 9-point Likert scale survey regarding self-awareness and comfort. We conducted a statistical analysis of the collected scale scores, calculated the mean and standard deviation, and performed the analysis using the Wilcoxon signed-rank test [81]. The study lasted 60–110 minutes.

10.3 Results

Given the limited number of participants, these results provide initial insights into real-world user experiences with Collar-VLM rather than population-level conclusions.

10.3.1 Route Completion, Time, and Proximity Warnings

All participants successfully reached the supermarket and finished their shopping. We calculated average travel time by excluding the duration spent shopping inside the store. Participants spent an average of 1037 s (S.D. = 245 s) traveling with Collar-VLM, compared to 1219 s (S.D. = 382 s) with XiaoQ Glasses. The difference in travel time was not statistically significant ($p > 0.05$). Even so, six out of eight walked faster using Collar-VLM rather than XiaoQ Glasses. This outcome suggests the practical potential of VLM-powered guidance in real route-level outdoor assistance.

It is worth noting that XiaoQ Glasses lack proximity warning functions entirely. In contrast, Collar-VLM generated an average of 3.4 proximity alerts per route (S.D. = 1.2), mostly appearing in crowded, narrow spaces including office corridors and traffic light intersections. This proves our system can actively remind users of potential hazards along the way without hindering their travel progress.

The faster travel speed reported by most participants can be largely attributed to better camera stability. The camera integrated into Collar-VLM stays steady no matter how users rotate their heads. In contrast, participants struggled to keep their heads straight and stable while wearing the facial smart glasses. They would naturally turn their heads toward surrounding sound sources when crossing roads, which constantly disrupted the camera view. None of the participants mentioned wobbly camera footage or the need to readjust the camera angle, which validates the effectiveness of our two rounds of prototype updates optimizing camera layout and offering adjustable, personalized collar sizes.

Overall, participants agreed that the proximity warnings helped them detect nearby pedestrians and physical barriers. Still, several participants commented that automated voice prompts felt less reassuring than real human communication when facing uncertain surroundings. They proposed that future assistive systems should optimize warning timing, voice tone, and hazard descriptions, so users can clearly tell whether the alert is triggered by passersby, walls, or other obstacles nearby. We additionally observed that low-vision participants finished the journey more quickly than fully blind participants, an effect likely brought by their remaining visual ability.

10.3.2 Camera Field and Stability

Figure 19 shows the camera FOV of a VI participant wearing the Collar-VLM and XiaoQ. Collar-VLM offers a significantly larger FOV, allowing for the visibility of a person's

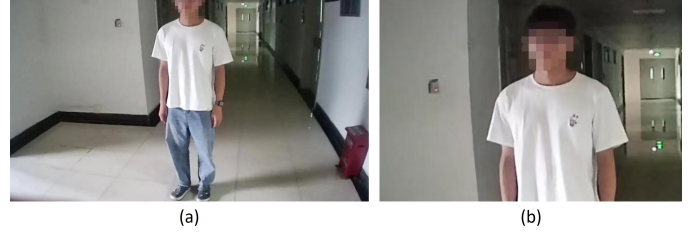


Fig. 19: Field of view (FOV) of (a) Collar-VLM and (b) XiaoQ.

entire body at a distance of approximately 1.5 meters, as well as the ground and a wider environment. We conducted an objective assessment based on recorded camera videos, using the well-known No-Reference Image Quality Assessment model BRISQUE [82] as a measure of image blur. This model closely aligns with human subjective quality evaluations and is easily accessible. We calculated the blur level of each frame captured by the camera during the task performed by the 8 participants in Section 10 and averaged the results. The findings show that our Collar-VLM prototype scored 32.6 (SD = 8.1), while the XiaoQ Glasses scored 51.5 (SD = 12.7), with higher scores indicating more blur and lower image quality.

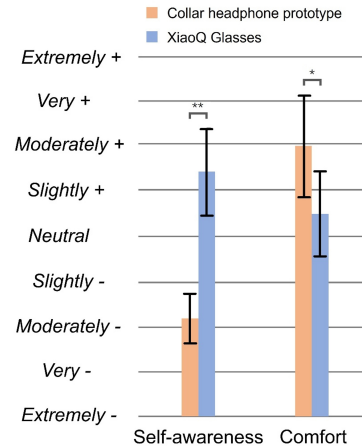


Fig. 20: 9-score Likert Scale, 1 represents "Extremely hard to feel self-aware or extremely uncomfortable" and 9 represents "Extremely easy to feel self-aware or extremely comfortable". The orange and blue bar represent wearing Collar-VLM and XiaoQ Glasses, respectively. Error bars represent standard deviation. Self-awareness ratings are as follows: Collar-VLM (M = 2.2, S.D. = 1.1), XiaoQ Glasses (M = 5.4, S.D. = 1.9). Comfort ratings are: Collar-VLM (M = 6.0, S.D. = 2.3), XiaoQ Glasses (M = 4.5, S.D. = 1.8). Statistical significance is denoted as ***P-value<0.001, **P-value<0.01, and *P-value<0.05.

10.3.3 Self-awareness

As shown in Figure 20, nearly all participants found Collar-VLM to be less likely to cause self-awareness and more inconspicuous. The difference is significant ($p < 0.01$). Four participants felt that smart glasses were more likely to attract attention because they were worn on the face, which will increase their self-awareness. "I think it's hard not to be noticed when the glasses are on the face," one participant said.

Three participants thought the glasses form was weird, that is, relatively thick legs and protruding cameras are still uncommon, which puts a certain psychological burden and implication on the participants. For example, a participant said: "I know of glasses with thick legs and large frames, which seems to be a fashion. But the legs of the glasses are still too thick and it has a protruding camera." Another participant explained: "I still prefer the inconspicuous headphone, which looks like normal collar headset."

Moreover, five participants mentioned that perceiving the wearable device as too large or heavy could trigger strong self-awareness. For instance, one female participant stated, "When wearing this glasses, I can clearly feel the large frames, you know? It's like you can distinctly feel the glasses pressing against your face." Interestingly, one participant mentioned that the glasses often slid down, requiring him to push them up multiple times, which also led to self-consciousness. In contrast, Collar-VLM, with its comfortable and stable design, is less likely to trigger feelings of self-consciousness compared to other headphone styles.

Additionally, three participants criticized the wired connection of the XiaoQ Glasses as uneasy. One participant said, "I wanted to put the host in my pocket, but when I did, the glasses' wire would often tangle around my hand, which annoyed me. I had to hold the heavy host in my hand, and it made me feel like passers-by were looking at me with strange eyes."

10.3.4 Comfort

As shown in Figure 20, participants believed that headsets were more comfortable and the difference was statistically significant. Seven participants found Collar-VLM to be less stressful to wear. A participant stated, "I find wearing headsets more comfortable; my ears hurt a bit when wearing glasses." Another participant said, "The collar headset seemed lightweight and I didn't feel neck or shoulder pain after wearing it for so long." We have reason to believe that the neck can bear more weight and is a good comfortable position to wear.

Four participants were accustomed to wearing collar headset but not glasses. This habit also suggests that collar headset form are more likely to be accepted and adopted. For instance, a participant mentioned, "I'm used to wearing headsets, but I don't often wear glasses. My eyes are not sensitive to light, so I don't need to cover them." Another participant described his usual scenario of using collar headset, "I usually use collar headset to listen to music, so when I wore your collar headset device, I felt very natural and comfortable, just like wearing my own headsets."

Three participants noted that the overall smooth design of Collar-VLM felt comfortable against the skin. For instance, one participant remarked, "The smooth collar headset feel great, and even after wearing them for so long, it's still hard to feel their presence."

However, three participants noted the issue with our prototype heating up over time. Using an infrared thermometer, we traced the heat mainly to the microprocessor of the WiFi camera PCB. The maximum measured surface temperature was 37°C, close to the reported thermal-comfort boundary of wearable devices (36–38°C) [83]–[85], but still below the 43°C long-duration contact safety threshold [86]. Thus, reducing heat from the WiFi camera PCB remains important for improving long-term comfort.

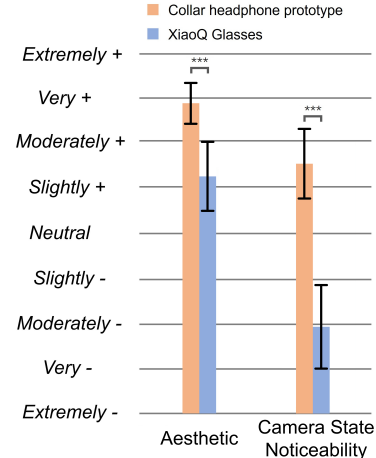


Fig. 21: 9-score Likert Scale, 1 represents "Extremely hard to perceive as aesthetically pleasing or to notice camera status" and 9 represents "to perceive as aesthetically pleasing or to notice camera status". The orange and blue bar represent wearing Collar-VLM and XiaoQ Glasses, respectively. Error bars represent standard deviation. Aesthetic ratings are as follows: Collar-VLM (M = 7.8, S.D = 0.9), XiaoQ Glasses (M = 6.3, S.D = 1.5). Camera status awareness ratings are: Collar-VLM (M = 6.5, S.D = 1.6), XiaoQ Glasses (M = 2.9, S.D = 1.8). Statistical significance is denoted as ***P-value<0.001, **P-value<0.01, and *P-value<0.05.

10.4 Aesthetics and Comprehensibility of Camera State

To evaluate bystanders' perceptions of the aesthetic appeal of Collar-VLM and camera switch prompt comprehensibility, we enlisted twelve sighted participants (average age 24, 4 males) as bystanders for assessment. Bystanders stood in the middle of the corridor in scenario 1 of the pilot study, waiting for a VI participant wearing Collar-VLM (with camera switch on) and XiaoQ Glasses to pass by. Moreover, in scenario 3, bystanders sat in chairs on the side and observed the two devices when the VI participant was not performing a task, specifically when the camera switch was in the off position. They were asked to complete a 9-point Likert scale to rate each device's aesthetics and the visibility of the camera switch prompt comprehensibility.

As shown in Figure 21, the aesthetic ratings of Collar-VLM were significantly higher than those of the XiaoQ Glasses. For example, one of the sighted participants stated, "The headsets looked more like a regular headset and are easily hidden under the black collar. The legs of the glasses are too thick and uncoordinated." Moreover, the camera switch prompt designed on Collar-VLM was considered noticeable. For instance, a sighted participant reported, "I could see the camera cover open and knew the camera might be on." Another sighted participant mentioned, "The XiaoQ Glasses itself have no indication; I could only guess whether the visually impaired was using the device based on their actions." The results indicate that we proposed a promising design for a camera status indicator.

11 CONCLUSION

In this paper, we focus on the exploratory redesign of WAD through a human-centered design process. We first conducted formative research, and then used ideation workshops to inspire VI participants to come up with concepts that best matched their expectations, leading to ideas that included a novel collar headset. We then designed easy-to-assemble hardware and mechanical structure, including 3D-printed shells, material and battery selection, size design, and assembly, resulting in three meticulous system prototypes. Following a test study involving 13 participants, we gathered user feedback and iterated on the collar headset, incorporating modifications such as adjusting camera orientation, improving the shape of the collar and external earpiece, and personalizing collar size. Collar-VLM further integrates a VLM-based cloud-edge software pipeline with smartphone-side stereo depth perception, hierarchical context memory, compact structured outputs, local spatial computation, and mobility safeguards. Technical evaluation shows that Collar-VLM improves scene description, object and text recognition, and indoor mobility assistance over prompt-based VLM baselines. Finally, a wild study shows the effectiveness of Collar-VLM, which was preferred over XiaoQ Glasses due to lower user self-consciousness and enhanced comfort.

REFERENCES

- [1] D. Pascolini and S. P. Mariotti, "Global estimates of visual impairment: 2010," *British Journal of Ophthalmology*, vol. 96, no. 5, pp. 614–618, 2012.
- [2] S. Xu, C. Yang, W. Ge, C. Yu, and Y. Shi, "Virtual paving: Rendering a smooth path for people with visual impairment through vibrotactile and audio feedback," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 3, pp. 1–25, 2020.
- [3] M. Khorrami-Nejad, A. Sarabandi, M.-R. Akbari, and F. Askarizadeh, "The impact of visual impairment on quality of life," *Medical hypothesis, discovery and innovation in ophthalmology*, vol. 5, no. 3, p. 96, 2016.
- [4] D. Dakopoulos and N. G. Bourbakis, "Wearable obstacle avoidance electronic travel aids for blind: a survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, no. 1, pp. 25–35, 2009.
- [5] "Envision ai," <https://apps.apple.com/us/app/envision-ai/id1268632314>, 2025.
- [6] "Seeing ai," <https://www.microsoft.com/en-us/ai/seeing-a->, 2025.
- [7] "Envision glasses," <https://aira.io/envision/>, 2025.
- [8] "Aira," <https://aira.io/>, 2025.
- [9] "Bemyeyes," <https://www.bemyeyes.com/>, 2025.
- [10] B. Gamage, T.-T. Do, N. S. C. Price, A. Lowery, and K. Marriott, "What do blind and low-vision people really want from assistive smart devices? comparison of the literature with a focus study," in *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility*, 2023, pp. 1–21.
- [11] S. Lee, M. Reddie, C.-H. Tsai, J. Beck, M. B. Rosson, and J. M. Carroll, "The emerging professional practice of remote sighted assistance for people with visual impairments," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–12.
- [12] S. Lee, R. Yu, J. Xie, S. M. Billah, and J. M. Carroll, "Opportunities for human-ai collaboration in remote sighted assistance," in *27th International Conference on Intelligent User Interfaces*, 2022, pp. 63–78.
- [13] J. Xie, M. Reddie, S. Lee, S. M. Billah, Z. Zhou, C.-H. Tsai, and J. M. Carroll, "Iterative design and prototyping of computer vision mediated remote sighted assistance," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 29, no. 4, pp. 1–40, 2022.
- [14] F. M. Li, J. Dorst, P. Cederberg, and P. Carrington, "Non-visual cooking: Exploring practices and challenges of meal preparation by people with visual impairments," in *The 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, 2021, pp. 1–11.
- [15] R. Boldu, D. J. Matthies, H. Zhang, and S. Nanayakkara, "Aisee: an assistive wearable device to support visually impaired grocery shoppers," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 4, pp. 1–25, 2020.
- [16] G. Carretti, M. Manetti, and M. Marini, "Physical activity and sport practice to improve balance control of visually impaired individuals: a narrative review with future perspectives," *Frontiers in Sports and Active Living*, vol. 5, p. 1260942, 2023.
- [17] G. Luo and S. Pundlik, "Usage patterns of head-mounted vision assistance app as compared to handheld video magnifier," *Displays*, vol. 75, p. 102303, 2022.
- [18] S. Pundlik, P. Shivshanker, and G. Luo, "Impact of apps as assistive devices for visually impaired persons," *Annual Review of Vision Science*, vol. 9, no. 1, pp. 111–130, 2023.
- [19] "Horizon smart glasses," <https://www.brentonlacombe.com/aira>, 2025.
- [20] "Orcam," <https://www.orcam.com/>, 2025.
- [21] H. P. Profita, "Designing wearable computing technology for acceptability and accessibility," *ACM SIGACCESS Accessibility and Computing*, no. 114, pp. 44–48, 2016.
- [22] H. Profita, R. Albaghli, L. Findlater, P. Jaeger, and S. K. Kane, "The at effect: how disability affects the perceived social acceptability of head-mounted display use," in *proceedings of the 2016 CHI conference on human factors in computing systems*, 2016, pp. 4884–4895.
- [23] Y. M. Kim, S. Bahn, and M. H. Yun, "Wearing comfort and perceived heaviness of smart glasses," *Human Factors and Ergonomics in Manufacturing & Service Industries*, vol. 31, no. 5, pp. 484–495, 2021.
- [24] K. MP Hoogsteen, S. A. Osinga, B. LPA Steenbekkers, and S. FA Szpiro, "Functionality versus inconspicuousness: Attitudes of people with low vision towards ost smart glasses," in *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, 2020, pp. 1–4.
- [25] A. D. P. Dos Santos, A. L. M. Ferrari, F. O. Medola, and F. E. Sandnes, "Aesthetics and the perceived stigma of assistive technology for visual impairment," *Disability and Rehabilitation: Assistive Technology*, vol. 17, no. 2, pp. 152–158, 2022.
- [26] "Looktel recognizer," <https://http://www.looktel.com/recognizer>, 2025.
- [27] S. Kayukawa, K. Higuchi, J. Guerreiro, S. Morishima, Y. Sato, K. Kitani, and C. Asakawa, "Bbeep: A sonic collision avoidance system for blind travellers and nearby pedestrians," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–12.
- [28] M. Kuribayashi, S. Kayukawa, H. Takagi, C. Asakawa, and S. Morishima, "Linechaser: a smartphone-based navigation system for blind people to stand in lines," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–13.
- [29] J. Guerreiro, D. Ahmetovic, D. Sato, K. Kitani, and C. Asakawa, "Airport accessibility and navigation assistance for people with visual impairments," in *Proceedings of the 2019 CHI conference on human factors in computing systems*, 2019, pp. 1–14.
- [30] Y. Kaniwa, M. Kuribayashi, S. Kayukawa, D. Sato, H. Takagi, C. Asakawa, and S. Morishima, "Chitchatguide: Conversational interaction using large language models for assisting people with visual impairments to explore a shopping mall," *Proceedings of the ACM on Human-Computer Interaction*, vol. 8, no. MHCI, pp. 1–25, 2024.
- [31] "Introer," <http://krvision.cn/>, 2025.
- [32] K. Yang, L. M. Bergasa, E. Romera, R. Cheng, T. Chen, and K. Wang, "Unifying terrain awareness through real-time semantic segmentation," in *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2018, pp. 1033–1038.
- [33] L. Keselman, J. Iselin Woodfill, A. Grunnet-Jepsen, and A. Bhowmik, "Intel realsense stereoscopic depth cameras," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 1–10.
- [34] "Horus," <https://horus.tech>, 2025.
- [35] H. Park, J. Pei, M. Shi, Q. Xu, and J. Fan, "Designing wearable computing devices for improved comfort and user acceptance," *Ergonomics*, vol. 62, no. 11, pp. 1474–1484, 2019.

- [36] M. Kianpisheh, F. M. Li, and K. N. Truong, "Face recognition assistant for people with visual impairments," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 3, pp. 1–24, 2019.
- [37] K. Kramer, D. Hedin, and D. Rolkosky, "Smartphone based face recognition tool for the blind," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. IEEE, 2010, pp. 4538–4541.
- [38] Y. Kajiwaru and H. Kimura, "Object identification and safe route recommendation based on human flow for the visually impaired," *Sensors*, vol. 19, no. 24, p. 5343, 2019.
- [39] V. Khambadkar and E. Folmer, "Gist: a gestural interface for remote nonvisual spatial perception," in *Proceedings of the 26th annual ACM symposium on User interface software and technology*, 2013, pp. 301–310.
- [40] M. Nakajima and S. Haruyama, "Indoor navigation system for visually impaired people using visible light communication and compensated geomagnetic sensing," in *2012 1st IEEE International Conference on Communications in China (ICCC)*. IEEE, 2012, pp. 524–529.
- [41] K. Chen, Y. Huang, Y. Chen, H. Zhong, L. Lin, L. Wang, and K. Wu, "Lisee: A headphone that provides all-day assistance for blind and low-vision users to reach surrounding objects," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 3, pp. 1–30, 2022.
- [42] R. Shilkrot, J. Huber, W. Meng Ee, P. Maes, and S. C. Nanayakkara, "Fingerreader: a wearable device to explore printed text on the go," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015, pp. 2363–2372.
- [43] R. Boldu, A. Dancu, D. J. Matthies, T. Buddhika, S. Siriwardhana, and S. Nanayakkara, "Fingerreader2.0: Designing and evaluating a wearable finger-worn camera to assist people with visual impairments while shopping," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 3, pp. 1–19, 2018.
- [44] M.-L. Shih, Y.-C. Chen, C.-Y. Tung, C. Sun, C.-J. Cheng, L. Chan, S. Varadarajan, and M. Sun, "Dlrv2: A deep learning-based wearable vision-system with vibrotactile-feedback for visually impaired people to reach objects," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1–9.
- [45] L. d. Sousa Britto Neto, V. R. M. L. Maie, F. L. Koch, M. C. C. Baranauskas, A. d. Rezende Rocha, and S. K. Goldenstein, "A wearable face recognition system built into a smartwatch and the blind and low vision users," in *International Conference on Enterprise Information Systems*. Springer, 2015, pp. 515–528.
- [46] S. Kayukawa, T. Ishihara, H. Takagi, S. Morishima, and C. Asakawa, "Guiding blind pedestrians in public spaces by understanding walking behavior of nearby pedestrians," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 3, pp. 1–22, 2020.
- [47] R. K. Katzschmann, B. Araki, and D. Rus, "Safe local navigation for visually impaired users with a time-of-flight and haptic feedback device," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 3, pp. 583–593, 2018.
- [48] L. Wang, Y. Li, J. He, Y. Che, K. Wu, X. Qi, and K. Chen, "Meet-sumaid: A mobile human-ai collaborative meeting summarization system," *IEEE Transactions on Mobile Computing*, 2026.
- [49] X. Jiang, J. Zheng, R. Liu, J. Li, J. Zhang, S. Matthiesen, and R. Stiefelhausen, "@ bench: Benchmarking vision-language models for human-centered assistive technology," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025, pp. 3934–3943.
- [50] S. S. Baghel, Y. P. S. Rathore, A. Pradhan, S. Jena, A. Bhavsar, A. Shukla, and P. Goyal, "Towards blind and low-vision accessibility of lightweight vlms and custom llm-evals," in *Proceedings of the 1st Workshop on Multimodal Models for Low-Resource Contexts and Social Impact (MMLoSo 2025)*, 2025, pp. 86–94.
- [51] M. Huh, F. Xu, Y.-H. Peng, C. Chen, D. Gurari, E. Choi, and A. Pavel, "Long-form answers to visual questions from blind and low vision people," in *Workshop on Demographic Diversity in Computer Vision@ CVPR 2025*, 2024.
- [52] J. He, S. Pundlik, and G. Luo, "Using chatgpt to generate walking directions—could it potentially help visually impaired with micronavigation?" *Translational vision science & technology*, vol. 14, no. 5, pp. 28–28, 2025.
- [53] M. Li and Y. Yao, "Openglass: A sensing-computing split architecture for local mllm-driven real-time visual assistance," in *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, 2026, pp. 829–839.
- [54] Y. Huang, J. Xu, B. Pei, L. Yang, M. Zhang, Y. He, G. Chen, X. Chen, Y. Wang, Z. Nie et al., "Vinci: A real-time smart assistant based on egocentric vision-language model for portable devices," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 9, no. 3, pp. 1–33, 2025.
- [55] J. Zeng, K. Chen, R. Wang, Y. Li, M. Fan, K. Wu, X. Qi, and L. Wang, "Contractmind: Trust-calibration interaction design for ai contract review tools," *International Journal of Human-Computer Studies*, vol. 196, p. 103411, 2025.
- [56] B. Still and K. Crane, *Fundamentals of user-centered design: a practical approach*. CRC press, 2017.
- [57] "Xiaoq glasses," http://canjiren.china.com.cn/2019-10/23/content_40930479.html, 2025.
- [58] Y. H. Lee and G. Medioni, "Rgb-d camera based wearable navigation system for the visually impaired," *Computer vision and Image understanding*, vol. 149, pp. 3–20, 2016.
- [59] K. Chen, J. Xiang, W. Tan, K. Chen, Y. Luo, C. Ma, K. Wu, and L. Wang, "Pit: A novel toothbrush providing real-time and robust plaque indication during brushing," in *Proceedings of the 23rd Annual International Conference on Mobile Systems, Applications and Services*, 2025, pp. 15–27.
- [60] H. L. Tan, T. Aplin, T. McAuliffe, and H. Gullo, "An exploration of smartphone use by, and support for people with vision impairment: a scoping review," *Disability and Rehabilitation: Assistive Technology*, vol. 19, no. 2, pp. 407–432, 2024.
- [61] S. Tatsuno, "Created in japan: From imitators to world-class innovators," (No Title), 1990.
- [62] C. Zeagler, "Where to wear it: functional, technical, and social considerations in on-body location for wearable technology 20 years of designing for wearability," in *Proceedings of the 2017 ACM International Symposium on Wearable Computers*, 2017, pp. 150–157.
- [63] F. Gemperle, C. Kasabach, J. Stivoric, M. Bauer, and R. Martin, "Design for wearability," in *digest of papers. Second international symposium on wearable computers (cat. No. 98EX215)*. IEEE, 1998, pp. 116–122.
- [64] J. Rekimoto, "Gesturewrist and gesturepad: Unobtrusive wearable interaction devices," in *Proceedings Fifth International Symposium on Wearable Computers*. IEEE, 2001, pp. 21–27.
- [65] K. Shinohara and J. O. Wobbrock, "In the shadow of misperception: assistive technology use and social interactions," in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2011, pp. 705–714.
- [66] A. Olwal, K. Balke, D. Votintcev, T. Starner, P. Conn, B. Chinh, and B. Corda, "Wearable subtitles: Augmenting spoken communication with lightweight eyewear for all-day captioning," in *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, 2020, pp. 1108–1120.
- [67] K. Mühlmann, D. Maier, J. Hesser, and R. Männer, "Calculating dense disparity maps from color stereo images, an efficient implementation," *International Journal of Computer Vision*, vol. 47, no. 1, pp. 79–88, 2002.
- [68] R. Nair, K. Ruhl, F. Lenzen, S. Meister, H. Schäfer, C. S. Garbe, M. Eisemann, M. Magnor, and D. Kondermann, "A survey on time-of-flight stereo fusion," *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications*, pp. 105–127, 2013.
- [69] J. Han, L. Shao, D. Xu, and J. Shotton, "Enhanced computer vision with microsoft kinect sensor: A review," *IEEE transactions on cybernetics*, vol. 43, no. 5, pp. 1318–1334, 2013.
- [70] J. Li, R. Klein, and A. Yao, "A two-streamed network for estimating fine-scaled depth maps from single rgb images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3372–3380.
- [71] V. T. Bartels, "Physiological comfort of biofunctional textiles," *Bio-functional textiles and the skin*, vol. 33, pp. 51–66, 2006.
- [72] H. Hirschmüller, "Accurate and efficient stereo processing by semi-global matching and mutual information," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2. IEEE, 2005, pp. 807–814.
- [73] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: Bm25 and beyond," *Foundations and trends® in information retrieval*, vol. 4, no. 1–2, pp. 1–174, 2009.

- [74] W. M. Association *et al.*, "World medical association declaration of helsinki: ethical principles for medical research involving human subjects," *Jama*, vol. 310, no. 20, pp. 2191–2194, 2013.
- [75] K. Lee, D. Sato, S. Asakawa, H. Kacorri, and C. Asakawa, "Pedestrian detection with wearable cameras for the blind: A two-way perspective," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–12.
- [76] B. Ens, T. Grossman, F. Anderson, J. Matejka, and G. Fitzmaurice, "Candid interaction: Revealing hidden mobile and wearable computing activities," in *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*, 2015, pp. 467–476.
- [77] J. Jung and M. Philipose, "Courteous glass," in *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing: Adjunct publication*, 2014, pp. 1307–1312.
- [78] D. Pohl and C. F. de Tejada Quemada, "See what i see: Concepts to improve the social acceptance of hmds," in *2016 IEEE Virtual Reality (VR)*. IEEE, 2016, pp. 267–268.
- [79] M. Koelle, K. Wolf, and S. Boll, "Beyond led status lights-design requirements of privacy notices for body-worn cameras," in *Proceedings of the Twelfth International Conference on Tangible, Embedded, and Embodied Interaction*, 2018, pp. 177–187.
- [80] "IT9100," <https://www.itech.sh/en/product/power-meter/IT9100.html>, 2025.
- [81] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics bulletin*, vol. 1, no. 6, pp. 80–83, 1945.
- [82] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on image processing*, vol. 21, no. 12, pp. 4695–4708, 2012.
- [83] M. A. Rupp, "Evaluating the impact of vr headset microclimate temperature on perceptions of thermal discomfort and forehead temperature," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 66, no. 1. SAGE Publications Sage CA: Los Angeles, CA, 2022, pp. 1285–1288.
- [84] K. Chen, U. S. Toro, J. Lin, C. Huang, J. Xiang, L. Wang, H. Cui, L. Zhu, and K. Wu, "Bladdersense: A wearable ultrasound system for continuous bladder monitoring in real-world use," in *Proceedings of the 24th Annual International Conference on Mobile Systems, Applications and Services*, 2026, pp. 28–40.
- [85] M. A. Rupp, "Is it getting hot in here? the effects of vr headset microclimate temperature on perceived thermal discomfort, vr sickness, and skin temperature," *Applied Ergonomics*, vol. 114, p. 104128, 2024.
- [86] International Organization for Standardization, "Iso 13732-1:2006 ergonomics of the thermal environment – methods for the assessment of human responses to contact with surfaces – part 1: Hot surfaces," <https://www.iso.org/standard/43558.html>, 2006, international Standard.



Kaixin Chen received the MS degree from the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China, in 2024. He is currently working toward the PhD degree with the Internet of Things Thrust of Information Hub, Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China. He has authored or coauthored papers in premier conferences, such as ACM MobiCom, ACM UbiComp, and ACM MobiSys. His research interests include mobile

computing and Internet of Things (IoT).



Yongzhi Huang is currently a Ph.D. student at HKUSTGZ, with a focus on advanced technologies in smart sensing, mobile computing, and IoT. His research includes next-generation health sensing, HCI, and wireless networks. He has published papers at top conferences including MobiCom, MobiSys, UbiComp, CHI, etc. He is also a prolific inventor with 12 pending patents in China, reflecting his contributions to the field of smart technologies and IoT.



Usman Saleh Toro is a Postdoctoral researcher at the DSA thrust of Hong Kong University of Science and Technology (Guangzhou). He obtained his PhD from Shenzhen University in 2024. His research interests include low-power wireless sensing for the Internet of Things (IoT). In particular, he focuses on practical applications of optical and acoustic sensing. Currently, he has published 5 first-author papers in top-tier journals, such as IEEE TGCN, IEEE NETWORK, and IEEE TVT.



Chang Huang received the MPhil degree from DSA Thrust, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China, in 2024. He is currently working toward the PhD degree with the Internet of Things Thrust of Information Hub, Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China. He has authored papers in premier conferences, such as ACM MobiCom, AAAI and ACM MM. His research interests include Underwater Internet of

Things and Artificial Intelligence.



Yue Ling Che (S'11-M'15) received the B.Eng. and the M.Eng. degrees from the University of Electronic Science and Technology of China in 2006 and 2009, respectively, and the Ph.D. degree from the Nanyang Technological University, Singapore, in 2014, all in electrical engineering. From 2014 to 2016, she was a postdoc research fellow in Engineering Systems and Design Pillar, Singapore University of Technology and Design. She is now an Associate Professor in College of Computer Science and Software Engineering with the Shenzhen University. Her research interests include integrated sensing and communication systems, AI-enabled wireless networks, UAV-enabled mobile communications, wireless energy transfer, and stochastic modeling and optimization methods.



Kaishun Wu (Fellow, IEEE) is the Associate Vice President for Research of the Hong Kong University of Science and Technology (Guangzhou). He is also a full professor of the DSA & IoT Thrust Area under the Information Hub. He received his Ph.D. degree in computer science and engineering from HKUST in 2011. He is an active researcher with more than 200 papers published on major international academic journals and conferences, as well as more than 100 invention patents, including 9 from the

USA. He received the 2012 Hong Kong Young Scientist Award, the 2014 Hong Kong ICT awards: Best Innovation and 2014 IEEE ComSoc Asia-Pacific Outstanding Young Researcher Award. He is an IET Fellow.



Lu Wang (Senior Member, IEEE) is currently an associate professor in College of Computer Science and Software Engineering, Shenzhen University, China. His research interests focus on wireless communications and mobile computing.