2025 IEEE/SICE International Symposium on System Integration (SII)
January 21-24, 2025, Munich, Germany

A VLM-Drone System for Indoor Navigation Assistance with Semantic Reasoning for the Visually Impaired

Zezhong Zhang, Chenyu Hu, Sunwoh Lye, and Chen Lv

Abstract— **Reduced vision significantly impacts the daily lives of people with visual impairments (PVI), often posing challenges in navigation and spatial awareness. To enhance the semantic reasoning capabilities of assistive technologies, we have developed a guidance system that integrates large vision-language models (VLMs) with a collision-avoidance drone. This system provides navigational assistance in indoor environments by interpreting semantic wayfinding signs. At the software level, we propose a hierarchical cross-prompt VLM (HCP-VLM) structure that leverages both Claude 3.5 Sonnet and ChatGPT 4o[1]. This structure improves the reasoning accuracy of semantic wayfinding signs to 76.73%, outperforming the standalone accuracies of Claude (74.73%) and ChatGPT (66.35%). A specialized wayfinding sign dataset was developed to fine-tune and evaluate the VLM. At the hardware level, an ultralight dual-modal Time of Flight (TOF) Laser-Camera module was integrated into the drone to detect obstacles, track users, and identify signs. Additionally, a vibration module was designed to communicate orientation and mobility information to users. The system's performance was evaluated in unfamiliar office buildings with two blindfolded sighted subjects, both of whom successfully located their target rooms with assistance from the system. To further drive innovation, we have released the dataset and code for public access[2], aiming to inspire advancements in intelligent assistive technologies.**

Index Terms—**VLM-Drone Guidance, System Integration, Assistive Robots**

I. Introduction

Around sixteen percent of the world population lives with some visual impairment, according to WHO's report [1]. As the number of PVI continues to rise, there is an increasingly pressing need for effective assistive technologies to enhance their independence and overall well-being [2].

Over the years, various assistive technologies have been developed, ranging from white canes [3] to navigation-assisting wheeled and quadruped robots [4], [5]. These technologies can be categorized into three types based on user engagement: passive, active, and semi-active. Passive assistive devices, such as the white cane, provide obstacle information and alerts but require active user exploration to be effective [3]. In contrast, active assistive technologies, including wheeled and quadruped guide robots [4], [6], employ high-resolution sensors like Lidar or depth cameras for autonomous navigation, but often neglect personal preferences and other sensory inputs like hearing of PVI [7].

Semi-active assistive technologies offer a balanced approach by enabling human-machine collaboration. These devices not only assist in navigating and avoiding obstacles but also allow PVI to actively engage with their environment, respecting their intention and subjective preferences, which aligns with a more user-centric philosophy [3], [8].

Many assistive technologies have focused on navigation for PVI in obstacle avoidance and movement orientation [9], [10]. However, situations requiring complex reasoning based on semantic information have not been thoroughly studied, such as office buildings, shopping malls, or libraries with diverse way-finding and target signs, as illustrated in Fig. 1. These signs play a critical role in guiding pedestrians to their desired destinations. The research goal is to design a VLM-Drone guidance system to navigate PVI to their desired targets in such environments.

The VLM-Drone guidance system we developed is categorized as a semi-active assistive technology. In this system, VLM is employed for advanced reasoning on complex semantic signs, while the drone actively navigates around obstacles and locates targets, subsequently transmitting orientation and mobility information to the user. The user then makes the final movement decisions based on this information. The system respects the user intentions by closely tracking and following their movements, intervening only when potential collisions are detected.

In essence, the VLM-Drone guidance system serves as a knowledgeable companion, providing navigational recommendations to PVI while respecting their preference, rather than acting as an authoritarian figure imposing its decisions. Overall, our work contributes to the following two aspects:

- The VLM is first integrated into the robotics system to provide semantic reasoning navigation assistance. A novel hierarchical cross-prompt VLM structure is proposed to improve the reasoning accuracy. A corresponding way-finding sign dataset is developed and open-sourced for fine-tuning and evaluating VLM.
- The autonomous drone is first applied as a guidance assistant, which avoids obstacles, identifies targets and track users. The vibration interface instructs users to move accordingly.

This work was supported in part by the Agency for Science, Technology and Research (A*STAR), Singapore, under the MTC Individual Research Grant (M22K2c0079), the ANR-NRF Joint Grant (No.NRF2021-NRF-ANR003 HM Science), and the Ministry of Education (MOE), Singapore, under the Tier 2 Grant (MOE-T2EP50222-0002).

All authors are with School of Mechanical and Aerospace Engineering, Nanyang Technological University, 50 Nanyang Ave, 639798, Singapore. email: {zezhong002, huch0008}@e.ntu.edu.sg, {mswlye,lyuchen}@ntu.edu.sg

[1]Hereinafter, Claude 3.5 Sonnet and ChatGPT 4o are referred as 'Claude' and 'ChatGPT,' respectively.

[2]https://github.com/AlexanderZ3/VLM-Drone-Guidance-System

II. Related Works

People with visual impairments require careful consideration and attention. Over time, human caregivers have developed and refined many effective principles through experience [11]. These guidelines emphasize several key principles, including:

- *DO use words such as "straight ahead," "turn left," "on your right."*
- *DO give specific directions like, "The desk is five feet to your right," as opposed to saying, "The desk is over there."*
- *DON'T insist upon trying to help if your offer of assistance is declined.*

Caregivers are encouraged to use clear language and respect autonomy, thereby effectively supporting PVI while preserving their dignity and independence. However, despite being the most effective form of assistance, the high cost and extensive training required make professional caregiver services inaccessible to many individuals with visual impairments[9].

Compared to human caregivers, active assistive robots have been developed to assist PVI in navigation by relying heavily on high-resolution sensors, such as Lidar or depth cameras, for accurate mapping and path planning [12], [13]. For instance, quadruped guidance robots are designed to mimic guide dogs [5], [6]. Similarly, wheeled guidance robots, such as the suitcase-shaped design by Guerreiro et al. [4], aid in indoor navigation. Seita et al. propose an assistive suitcase system with ultrasonic sensors and sonic feedback to detect obstacles and alert both the user and nearby pedestrians through auditory cues [14]. However, the dependence on advanced sensors in these systems increases weight, cost, and power consumption, which may constrain their practical utility.

Daisuke et al.[15] developed a smartphone-based indoor navigation assistant utilizing Bluetooth Low Energy, but its dependence on pre-installed beacons and detailed floor maps limits its practicality in unstructured and unmapped environments. Slade et al.[3] introduced a lightweight guidance cane equipped with a comprehensive sensor suite for SLAM and user steering, yet it faces challenges due to its limited semantic reasoning capabilities, which are essential for interpreting way-finding signs and gate numbers-critical elements often relied upon by sighted pedestrians [16].

Some pioneering works have explored the application of aerial robots to guide PVI. Studies such as those by [17], [18] investigate the use of drones for PVI navigation, with manually controlled by pilots. The auditory cues produced by drones can alert nearby individuals to the presence of PVI, thereby increasing their visibility and raising public awareness [18], [19]. Moreover, the compact size and portability of drones make them suitable for navigating various terrains, underscoring their significant potential in guidance applications for the visually impaired.

III. Design of VLM-Drone Guidance System

A. Problem Definition

The system is designed to assist PVI in locating specified targets within environments rich in signage information. During navigation, the drone instructs users to stop, turn left, right, or move forward as appropriate when encountering obstacles or detecting way-finding signs. Upon reaching the target, the system alerts the user to stop.

Fig. 1: Way-finding sign samples in the customized datasets.

Fig. 1 showcases scenarios randomly sampled from our collected dataset to demonstrate its complexity. It is important to recognize that way-finding signs may contain implicit semantic information. For instance, if the target is room "310" and the sign in the upper middle sub-image of Fig. 1 shows "302 to 315" with a left arrow, the guidance drone must interpret that the target room is within this range and provide navigational guidance as "turn left".

B. System Hardware and Framework Design

1) Hardware Design: The VLM-Drone guidance system is a typical human-robot collaboration system.

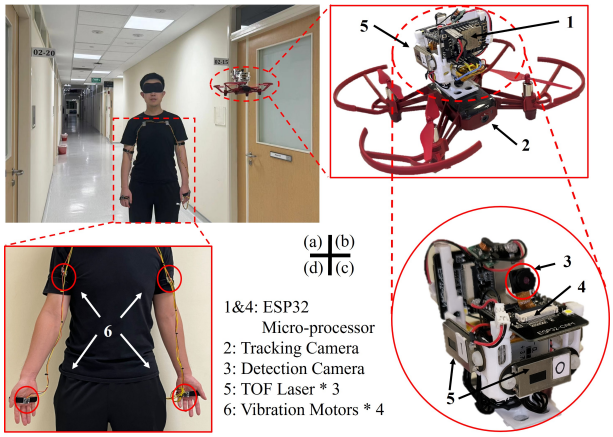

Fig. 2: **Overview of the guidance system:** (a) The system primarily serves PVI in indoor buildings with rich semantic information. (b) The flight platform and sensors. (c) The TOF Laser-Camera module. (d) Four vibration motors, conveying orientation and mobility information to users.

The TOF Laser-Camera, the core perception module carried by the flight platform (Fig. 2 (b)), comprises three TOF sensors, a forward-facing detection camera, and two ESP32 micro-controllers. The TOF lasers (No. 5) scan for obstacles in the front, left, and right, while the detection camera (No. 3) captures forward-facing images used for semantic reasoning.

The ESP32 micro-controllers handle the pre-processing of TOF data (No. 1) and image data (No. 4), which are then wirelessly transmitted to the ground computing station.

The DJI Tello drone (Fig. 2 (b)) serves as the flight platform. Its onboard camera (No. 2) functions as a tracking camera, enabling the drone to effectively follow the user. Given the limited take-off payload of Tello, the TOF Laser-Camera module is optimized to weigh only 35 grams.

TABLE I: Signals of Motors Vibration Interface

	Weak & Long	Strong & Short	Strong & Long
Motor 1	Turn Left	Move Left	Arrive Destination
Motor 2	Move Forward	Stop	
Motor 3	Lower Speed	Normal Speed	
Motor 4	Turn Right	Move Right	

The user-worn vibration interface (Fig. 2 (d)) is also lightweight, weighing around 100 grams. Each motor has two vibration modes: "Weak & Long" indicates a low-amplitude, long-duration vibration (700 ms), while "Strong & Short" represents a high-amplitude, short-duration vibration (300 ms). A strong, long-duration vibration from all motors indicates arriving destination. The combination of the four motors conveys various signals to users, as illustrated in Table I.

2) System Framework: The framework presented in Fig. 3 illustrates the integration of large VLMs into the system, enhancing its environment understanding and enabling collision-free guidance. The software workflow is organized into three parallel processes.

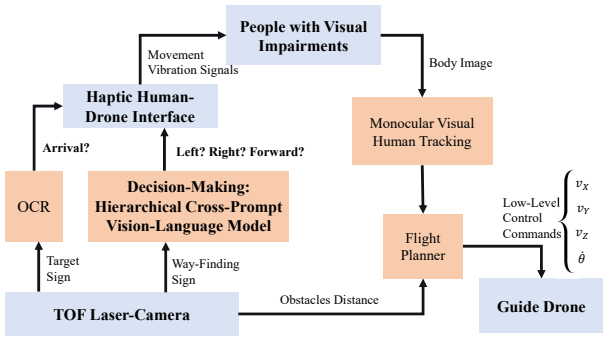

Fig. 3: Overview of system framework. The Hierachical Cross-Prompt VLM (HCP-VLM) module plays critical roles to understand way-finding signs in the real world.

In the first process, the TOF Laser-Camera module actively detects obstacles and the surrounding environment. When way-finding signs are detected using YOLOv8n [20], the proposed HCP-VLM module is triggered, extracting and reasoning the semantic meaning of the signs to provide instructions to the PVI: turn left, turn right, or move forward. Simultaneously, the Optical Character Recognition (OCR) module is activated to verify if the target destination has been identified. The system continuously monitors obstacles and, if the detected distance falls below the safety threshold, alerts the PVI to take evasive action-either by moving left, right, or stopping. Vibration motors facilitate effective communication between the PVI and the guidance system.

In the second process, the detected obstacle distance is transmitted to the flight planner, enabling the drone to execute collision avoidance algorithms. In the third process, the Monocular Visual Human Tracking module analyzes the human body image and generates flight commands to ensure the drone follows the PVI unless a collision is imminent. Finally, the flight planner integrates the obstacle data from the sensors (second process) and tracking commands from the PVI (third process) to achieve collision-free tracking.

IV. Methodologies

A. Dataset

To the best of our knowledge, no existing datasets are designed to address the tasks outlined in Section III-A. We develop a dataset consisting of 320 way-finding sign images and 1,184 image-text Visual Question Answering (VQA) pairs for VLM fine-tuning and testing. The dataset can be augmented, as each image contains multiple signs, enabling the generation of diverse VQA input pairs by varying the target destination in the questions. Although the focus is on indoor navigation assistance, the dataset also includes outdoor way-finding signs to enhance training and evaluation, emphasizing the semantic content. The dataset has been open-sourced and will be continuously expanded to support future research.

B. HCP-VLM Structure Design

The VLM is crucial for decision-making of the guidance system since it can reason the semantic meaning. This subsection evaluates the reasoning accuracy of four VLMs and proposes the HCP-VLM structure based on the two best-performing models, ChatGPT and Claude.

1) Accuracy comparison of single VLM for way-finding sign reasoning: First, we evaluated the zero-shot performance of four VLMs-BLIP [21], BLIP2 [22], ChatGPT [23], and Claude [24] -using our test dataset, which includes **735** VQA samples. Each sample consisted of an image with a way-finding sign and a consistent question: "What is the direction of the TARGET?", where "TARGET" refers to the destination. The expected output was a textual navigation command (turn left, turn right, or move forward).

Furthermore, we fine-tuned BLIP and BLIP2 on this dataset and reassessed their performance post-fine-tuning.

The results in Table II show that zero-shot Claude and ChatGPT significantly outperform other models in terms of visual recognition and logical reasoning accuracy, even exceeding the fine-tuned models. The improvements observed in BLIP and BLIP2 demonstrate that fine-tuning with our dataset can effectively enhance their performance. However, considering the computational demands of running such multi-modal VLMs, even relatively lightweight models like BLIP require high-end GPUs, making them less feasible for deployment on resource-constrained devices. Therefore, using APIs such as ChatGPT and Claude is a more practical solution for our system.

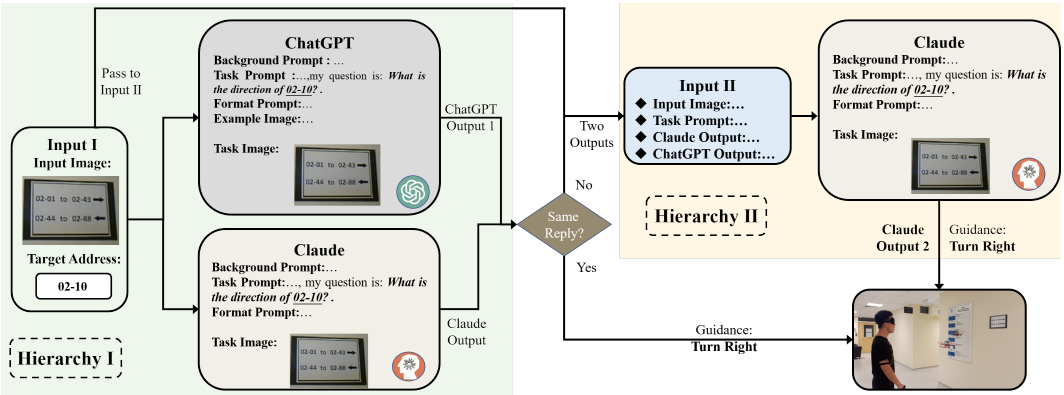

Fig. 4: Architecture of Hierarchical Cross-Prompt VLM. In Hierarchy I, input I, containing the target image and target address, is fed into ChatGPT and Claude simultaneously. Their built-in prompts convert Input I into specific questions and output the answers through their VLM functions. Compare the two outputs, and if the answers are consistent, the output (Turn Right) is adopted as guidance instruction. Otherwise, the two different outputs are fused with Input I to generate Input II, which is a new input to the Claude, and the guidance instruction (Turn Right) is output by Claude in Hierarchy II .

TABLE II: Way-finding sign reasoning comparison

Model	BLIP		BLIP2		ChatGPT	Claude
	ZS	FT	ZS	FT		
Accuracy(%)	37.29	59.11	32.20	59.32	**66.35**	**74.73**
Time(s)	0.21	0.21	4.04	13.87	3.82	2.08

'ZS' represents 'Zero-Shot', and 'FT' represents 'Fine-Tuned'.

2) Prompt Design: According to guidelines from OpenAI [23] and Anthropic [24], prompt engineering is crucial for output quality, whether tasks involve pure language processing or VQA [25]. Our empirical findings also confirm that variations in text prompts can significantly impact results, even with identical input images.

We propose a prompt structure "**Background-Task-Format**" -which we found highly effective. The Background should clearly define the VLM's role and outline the task's key details. The Task prompt must be concise to prevent misinterpretation, and the desired output format should be explicitly defined to facilitate subsequent operations. Fig. 5 illustrates the detailed prompt design.

A bit of difference is designed: for ChatGPT, restricting the output ensures concise answers like "Left" or "Right." For Claude, overly restrictive output formats can reduce inference accuracy, so Claude generates a directive sentence instead of a concise word.

*3) **HCP-VLM Structure**:* To further enhance the reasoning accuracy of ChatGPT and Claude, we propose a hierarchical cross-prompt Vision-Language Model structure, as depicted in Fig. 4. Furthermore, the term "**hierarchical**" refers to the overall reasoning process consisting of two hierarchies. In Hierarchy I, the target image and target address are simultaneously fed into ChatGPT and Claude, respectively. If the outputs from these two models are consistent, they are accepted as the final decision-making results. However, if the outputs differ, the program progresses to Hierarchy II, where the output from ChatGPT and Claude are treated as an additional input and fed into the Claude model along with the original target image and target address. This process, referred to as "**cross-prompt**," allows for a more

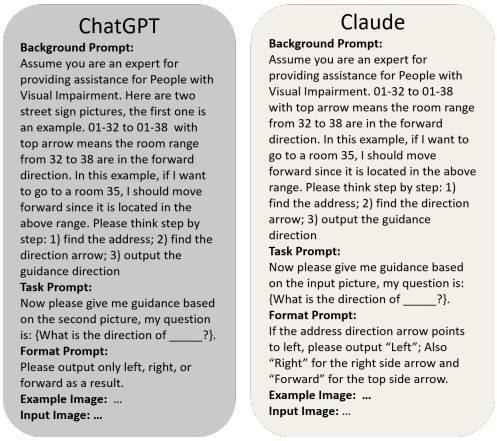

Fig. 5: The prompt design of ChatGPT and Claude.

refined reasoning approach. This operation is equipped with the potential to make full use of the reasoning and understanding capabilities of two large vision language models. The HCP-VLM structure increases the reasoning accuracy from 74.73% (best single VLM performance, Claude) to **76.73%**.

C. Others

Human Tracking By utilizing the renowned human pose detection machine learning package, MediaPipe [26], the frontal keypoints of the human body can be extracted, forming a bounding box encompassing the upper body, as shown in Fig. 7 (e). The position of the bounding box's center point relative to the center of the entire image is then used as input for a proportional-integral-derivative (PID) controller, which adjusts the drone's position to maintain a consistent frontal view of the users while following their movement. The key control parameters for each degree of freedom of the drone's movement speed are as follows:

- Left-Right: the horizontal difference of bounding box center and image center;

- Up-Down: the vertical difference of bounding box center and image center;
- Forward-Backward: the area of bounding box center and image center;
- Yaw angle: the ratio of bounding box height and width.

A right turn human tracking case is shown in Fig. 7 (e), and the corresponding speed control value is calculated in real-time.

Collision Avoidance As shown in Fig. 6, the user and guidance drone are surrounded by a convex rectangular border. d_L represents the minimal distance of the rectangular border to an obstacle on the left side; d_R represents the right minimal distance; d_F represents the front minimal distance. d_B is the alerting boundary threshold. If the distance to obstacles along any direction is lower than d_B, the flight planner will execute a large penalty and push drones away from obstacles. The vibration motors also alert the user to move in the opposite direction. Fig. 6 (a) and (b) describe cases for forward movement and turn in junction, respectively.

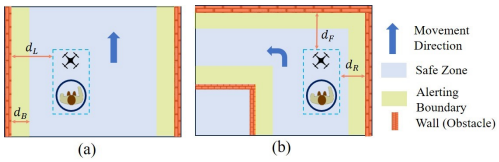

Fig. 6: Schematic of Obstacle Avoidance.

We defined the speed control function as:

$$v_i = \begin{cases} -\omega_i \cdot (d_i - d_B)^2 & \text{if } d_i \leq d_B \\ 0 & \text{if } d_i > d_B \end{cases}, \quad i = R, L, F. \qquad (1)$$

where v_i, ω_i and d_i are the flight speed of the drone, weight of speed, and obstacle distance along corresponding directions, respectively; R, L, F represents right, left and front direction, respevtively. d_B is the safe threshold.

V. Experiments and Results

A. Experiment Setup

"cross-prompt" Designs: Based on our way-finding signs dataset, as illustrated in Section IV-A, we compared the overall accuracy performance of HCP-VLM when the Hierarchy II in Fig. 4 is chosen as ChatGPT or Claude.

Solid Experiment: The experiment was conducted in an office building with multiple rooms, each identified by a unique door sign, and way-finding signs placed at intersections. To validate the effectiveness of the proposed VLM-Drone system, two sighted participants, wearing eye masks, were asked to walk three distinct routes with varying starting and ending points. Without prior knowledge of the building layout, the participants relied solely on the VLM-Drone system to locate the target room. Notably, sighted individuals wearing eye masks may find the task more challenging than visually impaired persons, as they are not accustomed to darkness and have not undergone mobility or orientation training [27].

Hardware Platform: The sensors and flight platform are illustrated in Section III-B. A laptop serves as the ground computing station for processing TOF and image data, equipped with an Intel Core i7-9750H CPU @ 2.60GHz and an NVIDIA GeForce GTX 1660Ti GPU.

TABLE III: Accuracy of the Proposed HCP-VLM

	Hierarchy I	Hierarchy II
Proportion of Each Hierarchy	434/735 = 59.29%	301/735 = 40.95%
Accuracy of Each Hierarchy	402/434 = 92.63%	162/301 = 53.82%
Overall Accuracy of HCP-VLM	402/735 + 162/735 = **76.73%**	

B. Results and Discussions

1) Overall Accuracy of HCP-VLM: The accuracy results of way-finding sign reasoning tasks of the proposed HCP-VLM, Claude is chosen as the Hierarchy II, are listed in Table. III. Totally, 735 various text-image pairs are randomly sampled from our dataset. 59.29% (434 cases) reasoning situations obtained same results from ChatGPT and Claude at the Hierarchy I, and in these 434 cases, 92.63% are correct decisions; the rest 40.95% (301 cases) reasoning situations differs at the Hierarchy I, then they are further processed at the Hierarchy II. 53.82% (162 cases) of secondary processing are correct reasoning. The overall accuracy of HCP-VLM is **76.73%**, exhibiting a 2% improvement in accuracy compared to individual Claude and a 10.38% improvement over individual ChatGPT.

Meanwhile, we also evaluated the situation when the ChatGPT is applied as Hierarchy II VLM. In the total 301 cases processed at Hierarchy II stage, only 94 cases are identified correctly, leading to an overall accuracy of 67.48%, which is also higher than the individual performance of ChatGPT. Therefore, results confirmed that "cross-prompt" mechanism can enhance way-finding signs reasoning capabilities. Furthermore, the optimal HCP-VLM structure is "**ChatGPT+Claude→Claude**", as illustrated in the Fig. 4.

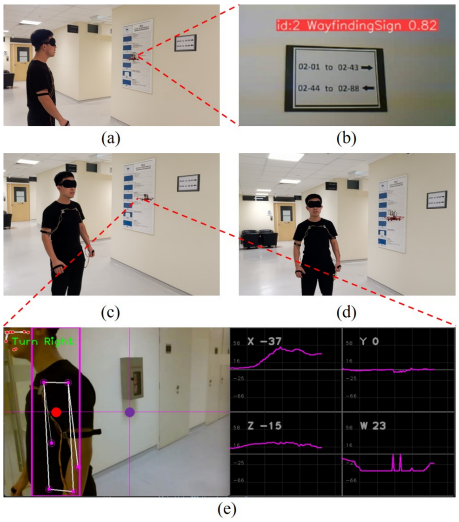

Fig. 7: Figures of one typical scenario for way-finding sign reasoning and human body tracking.

*2) **Experiment Results**:* **Human Tracking and Way-Finding Sign Detection:** Fig. 7 illustrates a right-turn scenario. Initially (Fig. 7 (a)), the room "02-10" is selected as the target destination, and the drone detects a way-finding sign (Fig. 7 (b)). The HCP-VLM decision module processes the image and identifies that room "02-10" is on the right. A "Turn Right" signal is then sent to the user via the vibration device. The user initiates the right turn, as shown in Fig. 7 (c) and (d). The drone immediately tracks the user using a real-time tracking algorithm, as demonstrated in Fig. 7 (e), the perspective of the onboard tracking camera.

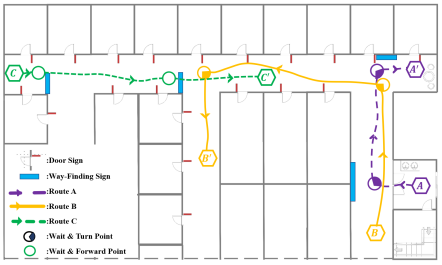

Fig. 8: The routes used in the guidance tasks. A) 8 meters, 2 turns, B) 28 meters, 2 turns, C) 20 meters, 2 intersections.

Route Map: Fig. 8 demonstrates the experiment routes used in our guidance tasks. The experiment subjects accomplished turn and collision avoidance under the assistance of VLM-Drone guidance sytsem, and find all target rooms in three tasks.

VI. CONCLUSIONS

This study developed a VLM-Drone guidance system to provide effective navigational assistance for people with visual impairments. To the best of our knowledge, this is the first study to VLMs for interpreting semantic information from wayfinding signs and to utilize autonomous drones as guidance assistants. Besides, we proposed a novel HCP-VLM framework, which significantly enhances reasoning accuracy for semantic navigation tasks. The effectiveness of the integrated system was validated through experiments.

REFERENCES

[1] World Health Organization, "Blindness and vision impairment," Available: https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment, 2018, accessed: 2019-09-03.

[2] K. Manjari, M. Verma, and G. Singal, "A survey on assistive technology for visually impaired," *Internet of Things*, vol. 11, p. 100188, 2020.

[3] P. Slade, A. Tambe, and M. J. Kochenderfer, "Multimodal sensing and intuitive steering assistance improve navigation and mobility for people with impaired vision," *Science robotics*, vol. 6, no. 59, p. eabg6594, 2021.

[4] J. Guerreiro, D. Sato, S. Asakawa, H. Dong, K. M. Kitani, and C. Asakawa, "Cabot: Designing and evaluating an autonomous navigation robot for blind people," in *Proceedings of the 21st international ACM SIGACCESS conference on computers and accessibility*, 2019, pp. 68–82.

[5] Y. Chen, Z. Xu, Z. Jian, G. Tang, L. Yang, A. Xiao, X. Wang, and B. Liang, "Quadruped guidance robot for the visually impaired: A comfort-based approach," in *Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 12 078–12 084.

[6] A. Xiao, W. Tong, L. Yang, J. Zeng, Z. Li, and K. Sreenath, "Robotic guide dog: Leading a human with leash-guided hybrid physical interaction," in *Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 11 470–11 476.

[7] W. Niemeyer and I. Starlinger, "Do the blind hear better? investigations on auditory processing in congenital or early acquired blindness ii. central functions," *Audiology*, vol. 20, no. 6, pp. 510–515, 1981.

[8] A. Bhowmick and S. M. Hazarika, "An insight into assistive technology for the visually impaired and blind people: state-of-the-art and future trends," *Journal on Multimodal User Interfaces*, vol. 11, pp. 149–172, 2017.

[9] G. I. Okolo, T. Althobaiti, and N. Ramzan, "Assistive systems for visually impaired persons: challenges and opportunities for navigation assistance," *Sensors*, vol. 24, no. 11, p. 3572, 2024.

[10] W. Elmannai and K. Elleithy, "Sensor-based assistive devices for visually-impaired people: Current status, challenges, and future directions," *Sensors*, vol. 17, no. 3, p. 565, 2017.

[11] Wisconsin Department of Health Services, "Do's and don'ts for interacting with a person who has a visual impairment," 2024, accessed: 2024-08-14. [Online]. Available: https://www.dhs.wisconsin.gov/obvi/adjustment/dos-donts.htm

[12] H. Fernandes, P. Costa, V. Filipe, H. Paredes, and J. Barroso, "A review of assistive spatial orientation and navigation technologies for the visually impaired," *Universal Access in the Information Society*, vol. 18, pp. 155–168, 2019.

[13] V. Isazade, "Advancement in navigation technologies and their potential for the visually impaired: a comprehensive review," *Spatial information research*, vol. 31, no. 5, pp. 547–558, 2023.

[14] 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.

[15] D. Sato, U. Oh, K. Naito, H. Takagi, K. Kitani, and C. Asakawa, "Navcog3: An evaluation of a smartphone-based blind indoor navigation assistant with semantic features in a large-scale environment," in *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*, 2017, pp. 270–279.

[16] F. E.-z. El-taher, A. Taha, J. Courtney, and S. Mckeever, "A systematic review of urban navigation systems for visually impaired people," *Sensors*, vol. 21, no. 9, p. 3103, 2021.

[17] M. Al Zayer, S. Tregillus, J. Bhandari, D. Feil-Seifer, and E. Folmer, "Exploring the use of a drone to guide blind runners," in *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility*, 2016, pp. 263–264.

[18] M. Avila, M. Funk, and N. Henze, "Dronenavigator: Using drones for navigating visually impaired persons," in *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*, 2015, pp. 327–328.

[19] X. Zhang, Z. Pan, Z. Song, Y. Zhang, W. Li, and S. Ding, "The aerial guide dog: a low-cognitive-load indoor electronic travel aid for visually impaired individuals," *Sensors*, vol. 24, no. 1, p. 297, 2024.

[20] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLO," Jan. 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[21] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proceedings of the International conference on machine learning*. PMLR, 2022, pp. 12 888–12 900.

[22] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proceedings of the International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.

[23] OpenAI. (2023) Gpt-4o. [Online]. Available: https://platform.openai.com/docs/models/gpt-4o

[24] Anthropic. (2023) Anthropic console. [Online]. Available: https://console.anthropic.com/dashboard

[25] J. Wang, Z. Liu, L. Zhao, Z. Wu, C. Ma, S. Yu, H. Dai, Q. Yang, Y. Liu, S. Zhang *et al.*, "Review of large vision models and visual prompt engineering," *Meta-Radiology*, p. 100047, 2023.

[26] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann, "Mediapipe: A framework for building perception pipelines," 2019.

[27] R. G. Long and E. Hill, "Establishing and maintaining orientation for mobility," *Foundations of orientation and mobility*, vol. 1, p. 45, 1997.