2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

SA-VLM v2: Useful, Comprehensive, and Concise Guidance for Guide-Dog Robots Assisting the Visually Impaired

Woo-han Yun[1,2], Jaeho Shin[2], Beom-Su Seo[1], Jaehong Kim[1], and ByungOk Han*[1,2]

***Abstract*— The development of guide dog robots is expected to enhance the mobility and safety of visually impaired individuals outdoors. To assist these users in real-world navigation, walking guidance should be useful, comprehensive, and concise so that instructions are both actionable and easy to follow. While recent VLMs show promising capabilities in scene understanding, existing approaches do not address the effective delivery of guidance for visually impaired users. In this work, we propose SA-VLMv2 (Space-Aware VLM), a model designed to generate useful, comprehensive, and concise walking guidance based on ego-centric scenes and target destinations. To this end, we first derived four canonical templates for walking guidance through user evaluation with professional guide dog trainers across diverse images, providing insights into preferred guidance formats. We then collected, manually annotated, curated a dataset of 19,945 samples aligned with these templates and trained SA-VLMv2 from the open-sourced VLM, Qwen2.5VL. Experimental results show that SA-VLMv2 outperforms state-of-the-art proprietary MLLMs (Claude 3.5 Sonnet, Gemini 2.5, GPT-4o) and the open-sourced pretrained VLM (Qwen2.5VL) in both holistic and factor-wise evaluations. SA-VLMv2 generated more concise yet informative guidance while achieving higher scores across multiple evaluation factors.**

I. INTRODUCTION

Guide dog robots have attracted sustained attention as a powerful means to improve independent mobility and safety for people who are blind or have low vision, combining advances in robotics, computer vision, and human–robot interaction [1][2][3][4]. Building on these advances, recent research shows that such robots can perceive their surroundings, plan paths, and communicate with users through natural language, enabled by progress in vision–language models (VLMs)[5][6][7][8]. This integration of perception and language has opened new possibilities for providing context-rich navigation support, allowing robots to describe spatial relationships and interpret dynamic environments with a level of detail and flexibility not achievable with traditional perception alone [9]. These developments highlight the potential for robotic companions to provide context-aware walking guidance across diverse environments and everyday situations, making the development of more effective guidance strategies an important next step for the field.

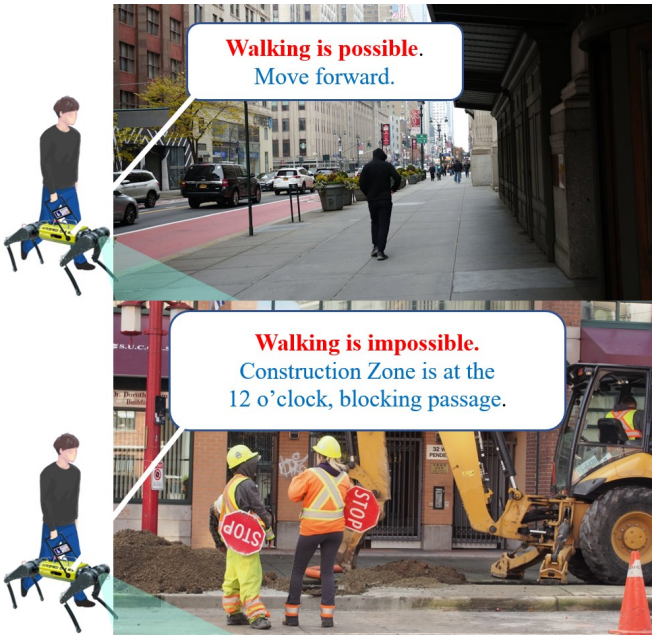


Fig. 1. We introduce a method that, given an image and a specified goal position, identifies the walking situation and generates responses according to predefined templates. These templates are derived from a survey with professional guide dog trainers on diverse outdoor navigation cases.

Following these advances, related studies have explored visual assistance for visually-impaired individuals from a variety of perspectives. EgoBlind [10] provides egocentric video with question–answer pairs that capture everyday assistive needs—such as reading labels or recognizing household objects—well beyond walking guidance. The SideGuide dataset [11] offers RGB–depth images and detailed annotations of obstacles in real walking environments, supporting perception models for navigation. DRAGON [9] proposes a conversational guiding robot that grounds user commands to the environment, enabling visually impaired users to navigate to landmarks and receive spoken descriptions of their surroundings in an everyday indoor environment. Closer to direct walking guidance, the safety-aware street crossing method [12] addresses decision-making for visually impaired pedestrians at complex intersections. VIALM (Visually Impaired Assistance with Large Models) [13] is a benchmark that evaluates large multi-modal models by generating detailed step-by-step instructions in indoor environments.

However, while these studies provide valuable datasets and assistive methods, they rarely examine how well essential

[1] ETRI (Electronics and Telecommunications Research Institute), Republic of Korea
[2] UST (University of Science and Technology), Republic of Korea
ByungOk Han* is a corresponding author.{byungok.han}@etri.re.kr
This work was partly supported by the Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (RS-2023-00215760, Guide Dog: Development of Navigation AI Technology of a Guidance Robot for the Visually Impaired Person, 70%) and by the National Research Council of Science & Technology(NST) grant by the Korea government(MSIT) (No. GTL25041-000, 30%). This research (paper) used datasets from 'The Open AI Dataset Project (AI-Hub, S. Korea)'. All data information can be accessed through 'AI-Hub (www.aihub.or.kr)'.

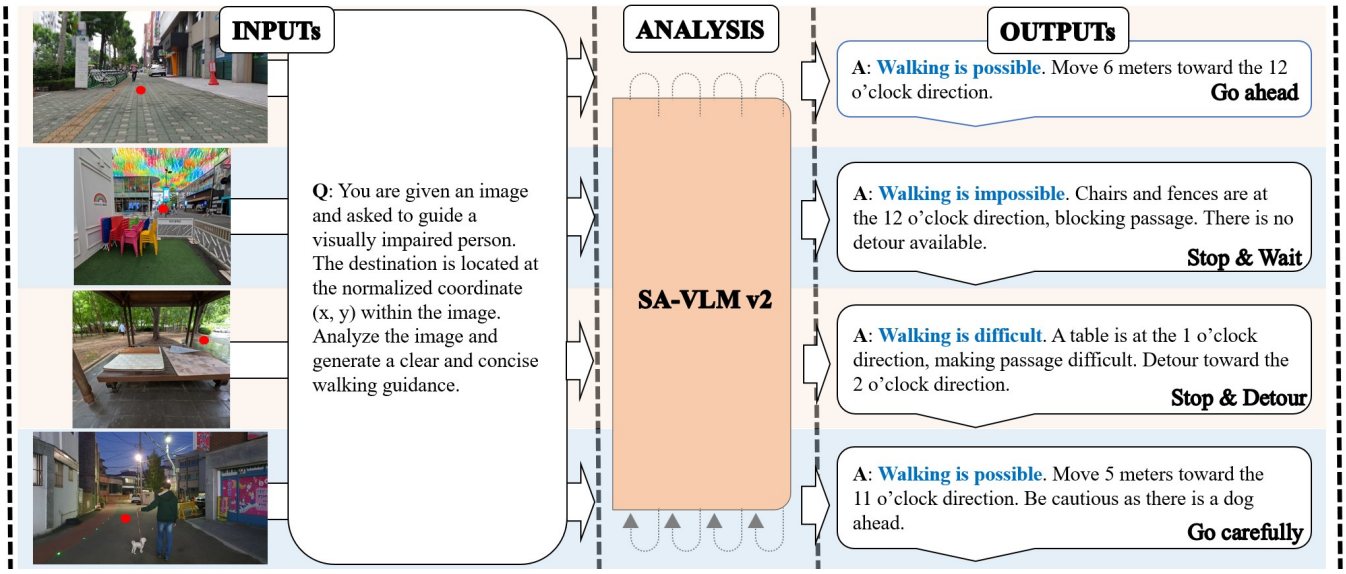


Fig. 2. **Overview of SA-VLMv2.** For each walking situation, the model takes an ego-centric scene and a specified goal position as input, identifies the situation, and generates concise walking guidance in one of four preferred formats, which were derived through surveys with professional guide dog trainers. These formats correspond to distinct walking situations: "Go ahead", "Go carefully", "Stop and Detour", and "Stop and Wait".

information is delivered to visually impaired users so that guidance is truly useful, sufficiently comprehensive, and appropriately concise. Their generated explanations are often unstructured and inconsistent in detail: some are unnecessarily long, while others omit key facts, making it difficult for users to obtain exactly what they need. Because large-scale, real-world evaluations with visually impaired participants are limited, it is still unclear how to consistently provide navigation guidance that is useful in practice, comprehensive enough to ensure safety, and concise enough to avoid overload, a challenge central to the focus of this study.

To address this challenge, SA-VLM (Space-Aware VLM) [14], our previous work, defined a structured walking-guidance task for visually impaired individuals. We introduced an automatic data-generation pipeline that extracts depth maps and captures rich spatial context, created the Space-Aware Instruction Tuning (SAIT) dataset, and proposed an evaluation benchmark to assess how well VLMs can provide precise and concise guidance descriptions. These contributions demonstrated that incorporating spatial awareness into VLMs enables more reliable guidance.

Building on the previous study, which introduced a structured walking-guidance task and an automated data-generation pipeline, this work extends that foundation to improve the quality and delivery of guidance for visually impaired individuals as illustrated in Fig. 1. We conducted a systematic user study with professional guide-dog trainers who reviewed a diverse set of real navigation scenarios and provided detailed evaluations of alternative explanations, highlighting what information is most useful, comprehensive, and concise in practical situations. Drawing on insights from our user study, we distilled the essential informational elements-walking possibility, obstacle details, and path information-and organized them into a set of canonical guidance templates designed to match the way visually impaired users understand and follow navigation guidance as shown in Fig. 2. Building upon these templates, we collected, manually annotated, and curated a large structured dataset and fine-tuned a VLM to generate clear, consistent guidance descriptions aligned with expert preferences. Extensive evaluations with both holistic and factor-wise assessments demonstrate that the resulting model delivers significantly clearer and more reliable guidance than strong multi-modal baselines, highlighting the effectiveness of our expert-informed task refinement and subsequent model training.

This work makes the following contributions:

- We derived four structured templates for walking guidance through systematic evaluation with professional guide dog trainers.
- We collected, manually annotated, and curated a dataset of 19,945 structured guidance samples and proposed SA-VLMv2, a fine-tuned Qwen2.5VL model trained to generate concise, template-based walking instructions.
- We conducted extensive evaluations using both holistic metrics (METEOR, ROUGE, BERTScore, LLM-as-a-Judge) and factor-wise assessments (possibility, obstacle information, format, distance, direction), demonstrating SA-VLMv2's superiority over state-of-the-art baselines.

II. Method

A. Task Definition

When walking outdoors with a living guide dog, people with visual impairments may encounter moments when the dog suddenly stops or changes pace for reasons the handler cannot immediately know. A guide dog cannot explain these cues, but a guide-dog robot equipped with SA-VLM [14] can convey real-time information about the situation, helping the user understand why movement has paused and how to proceed. Following the previous work, we consider an ego-centric RGB image I captured by a camera mounted

on the guide-dog robot. The user's intended destination is specified as a goal position (x, y) within this image. A VLM f generates the walking instruction t by processing the image I together with a goal-oriented query prompt q_{goal} that includes the goal position $\mathbf{p}_{\text{goal}}$. Formally,

$$t = f(I, q_{\text{goal}}) \tag{1}$$

In the previous study [14], the walking guidance t consisted of information about the destination, the left and right surroundings, the path, and a recommended action along with the reasoning.

B. Challenges in Systematic Evaluation

Systematic investigations into how visually impaired individuals prefer guidance explanations in different scenarios, and which explanation styles are most effective across various walking contexts, are crucial. However, such studies have rarely been conducted across diverse environmental settings, including situations that require moving forward or stopping, as well as walking on safe paths, caution-required paths, sidewalks, or shared pedestrian–vehicle roads. This gap exists largely because it is practically difficult to carry out experiments with visually impaired participants in a wide range of walking environments. While presenting scenario-based images can be an effective method to evaluate guidance under different conditions, conducting large-scale studies directly with visually impaired participants remains challenging due to time, environmental, and ethical constraints.

C. User Study Setup

To address these limitations, we conducted a user study with professional guide dog trainers, who possess deep expertise in the mobility needs of visually impaired individuals and can visually inspect images. A total of 11 trainers participated, and the study employed 150 scenario images covering diverse walking environments. For each image, two different guidance explanations were presented side by side with equal probability, and participants rated each explanation on three criteria—usefulness, comprehensibility, and conciseness—using a five-point Likert scale. In addition to quantitative ratings, participants were invited to provide detailed written feedback for each explanation. The survey was conducted over five days, with 30 items per day, and on the final day participants were asked for overall satisfaction and suggestions for improvement. Each participant received an honorarium of 400,000 KRW, and the entire study was approved by the Institutional Review Board (IRB)[1]. The interface used for the user study is illustrated in Figure 3.

1) 150 Scenario Images: From the 1,000 outdoor walking images in [14], we selected 150 representative scenarios for the evaluation. Specifically, we applied k-means clustering separately to 500 "walkable" and 500 "non-walkable" images, and then sampled 75 representative images from each set. As a result, we obtained a balanced set of 150 images reflecting diverse walking environments, including crosswalks, sidewalks, and shared pedestrian–vehicle roads. Example images are shown in Fig. 4.

Fig. 3. User study interface. For each walking scenario image, two explanations were presented side by side with equal probability on the left and right. Participants rated each explanation on three criteria (usefulness, comprehensibility, conciseness) using a five-point scale and could provide additional written feedback. (The survey was conducted in Korean, and for readability in the paper, all content is translated into English.)

2) Two Explanation Versions: For each scenario image, participants were presented with two different explanations.

Version 1 (Ver1): Based on explanations collected in the preliminary study of SA-VLM [14]. These included destination information, left/right and path obstacle descriptions, and explicit statements of walkability with supporting reasons. Each explanation was manually written and then reconstructed into a single sentence using GPT-4o.

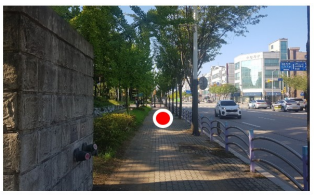
Version 2 (Ver2): Generated using GPT-4o[2], given the scenario image and target coordinates. Because GPT-4o often failed to correctly assess walkability, we explicitly provided the walkability label in the prompt. We also instructed the model to exclude unnecessary visual cues and generate concise explanations.

To prevent penalization by evaluators due to factual errors, both versions were manually corrected during the final review when obvious mistakes were identified. Examples of Ver1 and Ver2 explanations are shown in Fig. 4.

3) Evaluation Criteria: Participants evaluated explanations along three dimensions:

[1]IRB protocol number: 2025-HR-0001, approval number: N01-202502-01-003

[2]GPT-4o-2024-05-13



Ver1: The destination is the sidewalk. On the left there is a wall, and on the right a fence, but there are no obstacles along the path, so walking is possible.

Ver2: From the starting point, go straight. Walk along the sidewalk with a road on the right and trees and a wall on the left. Keep going straight until you reach the destination.

Ver1: The destination is a park path. There are trees on the left and right. A no-entry rope is across the path, making it difficult to walk.

Ver2: If you move forward from your current position, the path is blocked. This route is unsafe, so you should stop and find another way.

Fig. 4. Examples of images and explanations used in the user study. Explanations were originally presented in Korean but translated into English here for clarity. To improve the visual clarity of the red circle representing the destination, we redrew it for the paper.

Usefulness: The extent to which the explanation provides practical walking assistance.

Comprehensibility: Whether the explanation is intuitive and easy to understand.

Conciseness: Whether the explanation avoids unnecessary details and conveys only essential information.

All criteria were scored on a five-point Likert scale (1 = disagree, 5 = agree). Additionally, participants provided free-text comments suggesting improvements for each explanation.

D. Quantitative Results

Among the 11 trainers, four had participated in the preliminary SA-VLM study [14] that informed the design of Ver1. To ensure unbiased comparison, quantitative analyses were based on the ratings of the remaining seven trainers. However, the overall trends were consistent between the two groups, with the seven-trainer subset giving slightly higher ratings to Ver1.

Table I reports the average scores and standard deviations. Across all three criteria, Ver1 outperformed Ver2, with lower variance as well. The largest difference was observed in conciseness, where Ver1 scored on average 0.86 points higher. The score distributions in Fig. 5 further confirm this trend. Ver1 explanations were consistently rated in higher ranges across all criteria, with conciseness showing the clearest advantage. Statistical testing using the Wilcoxon signed-rank test confirmed that differences across all three criteria were significant ($p < 0.05$). These results provide strong evidence that SA-VLM style explanations offer more effective and consistent guidance than those generated by GPT-4o.

TABLE I

COMPARISON OF AVERAGE SCORES (MEAN ± STD) FOR VER1 (SA-VLM-BASED) AND VER2 (GPT-BASED) EXPLANATIONS.

-	Usefulness	Comprehensibility	Conciseness
Ver1	4.15 ± 0.85	4.13 ± 0.85	4.40 ± 0.65
Ver2	3.89 ± 1.03	3.96 ± 0.95	3.54 ± 1.11

E. Summary of Improvement Suggestions

Although Ver1 received higher ratings than Ver2, both versions showed room for improvement. To address this, we

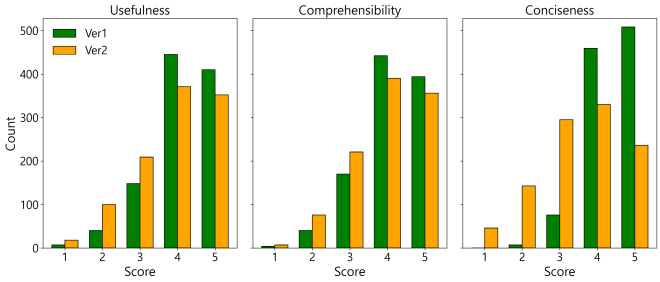


Fig. 5. Score distributions for Ver1 and Ver2. Ver1 consistently received higher ratings across all criteria, with the largest margin in conciseness.

organized the collected feedback and used it to refine the final guidance formats. For every scenario image, at least one improvement comment was provided, resulting in a total of 259 comments for Ver1 and 334 for Ver2. These comments were thematically categorized into five groups:

1) Order and focus of information delivery: Guidance should begin with walkability (possible or not) followed by supporting details. For walkable paths, concise movement instructions (e.g., "Go straight 10 meters") are preferred over unnecessary left/right descriptions. For non-walkable paths, more detailed situational explanations are necessary.

2) Clarity and specificity: When describing the environment, clear and specific terms (e.g., "crosswalk", "flowerbed") should be used. If such compact terms are not available, description with directions using left/right/forward or clock-based references should be given (e.g., "An outdoor exercise equipment is at 1 o'clock, a playground equipment is at 11 o'clock. Walking is impossible as the path is blocked by a no-parking sign and a vehicle in front"). Walking instructions should always include both distance (in meters) and direction (clock-based) (e.g., "Go straight for 10 meters toward the 11 o'clock.").

3) Focus on the path itself: Guidance should emphasize the current path rather than the destination. Important transitions (e.g., sidewalk-road changes, curbs, crosswalks, slopes) must be explicitly described.

4) Criteria for describing surrounding entities: Dynamic or reactive entities (e.g., pets) could be mentioned, while irrelevant objects or passersby should generally be omitted unless they significantly block the path.

5) Conciseness of expression: Redundant expression is unnecessary. Value judgments (e.g., "it is unsafe, so avoid it") are meaningless, and alternatives are preferred. References to objects that can only be identified visually are not helpful.

F. Derived Templates

Minority opinions that only applied to niche cases were excluded from the final design to avoid unnecessary complexity. Based on the aggregated feedback, we derived four structured templates for walking guidance that capture the essential elements of mobility. While these canonical templates provide standardized formats, evaluators also expressed diverse preferences for explanation styles-for example, some did not prefer explicit cautionary notes, while others did not require separate detour instructions. Accordingly, the four canonical templates should be regarded as baseline forms

that can be flexibly adapted or customized to accommodate the varying needs of individual users.

(1) Walking possible + normal path (no obstacles/-path change)

Format: Walking is possible. Move [distance] meters toward the [clockwise direction].
Examples:

- Walking is possible. Go straight 15 meters forward.
- Walking is possible. Move 10 meters toward the 11 o'clock direction.

(2) Walking possible + path/surface change or caution present

Format: Walking is possible. Move [distance] meters toward the [clockwise direction]. Be cautious as there is [caution element].
(Examples of caution elements: surface change, hazard, slope, path transition, crosswalk entry, etc.)
Examples:

- Walking is possible. Move 12 meters forward. At the end, enter the crosswalk.
- Walking is possible. Move 5 meters toward the 1 o'clock direction. There is a dog ahead, so be cautious.

(3) Walking difficult (detour available)

Format: Walking is difficult. [Obstacle] is at the [front/clockwise direction], making passage difficult. Detour toward the [clockwise direction].
Examples:

- Walking is difficult. A construction fence is ahead, making passage difficult. Detour toward the 9 o'clock direction.
- Walking is difficult. A truck is at the 12 o'clock direction, making movement difficult. Detour toward the 3 o'clock direction.

(4) Walking impossible (no detour)

Format: Walking is impossible. [Obstacle] is at the [front/clockwise direction], blocking passage. There is no detour available.
Examples:

- Walking is impossible. A truck is at the 12 o'clock direction, and both the 3 o'clock and 9 o'clock directions are blocked by walls. There is no detour available.
- Walking is impossible. A steel fence is ahead, with a road on the left and bushes on the right, blocking passage. There is no detour available.

G. Method

In this study, we considered two approaches for building a Vision-Language Model (VLM) capable of generating walking guidance explanations based on the derived templates. The first approach involves prompting, where the VLM is guided to produce responses in the desired format through carefully designed instructions. The second approach involves curating a dataset that reflects the structured templates and fine-tuning the VLM with this data. In the following sections, we focus on the second approach, while the first approach is discussed later in the experimental evaluation.

TABLE II
SUMMARY OF COLLECTED IMAGES ACROSS WALKABLE AND NON-WALKABLE SCENARIOS.

Scenario	Environment	Source	# images
Walkable	Outdoor	SideGuide[11]	8,999
	Indoor	Collected	1,033
Non-walkable	Outdoor	Collected	9,107
	Indoor	Collected	806

H. Dataset Collection and Curation

To the best of our knowledge, no existing public dataset directly incorporates the guidance templates defined in this study. Therefore, we adapted the SAIT dataset from the most relevant prior work [14]. This dataset contains outdoor egocentric images and instruction-tuning Q&A pairs for robotic guide dog navigation. For a given target coordinate (x, y), it provides structured guidance across five elements: the destination, left side, right side, and path, alongside the navigability status (i.e., recommended action such as "go" or "stop") and its associated reasoning. However, SAIT dataset is biased toward walkable scenarios and lacks diverse non-walkable cases. Additionally, its automated annotation pipeline introduces label noise and textual inaccuracies.

To address these limitations, we additionally collected images of non-walkable scenarios and manually annotated the information following the same five elements in [14]: destination, left side, right side, path, and navigability status with associated reasoning. Specifically, we adopted 8,999 outdoor walkable images from the SideGuide dataset [11]. To compensate for the lack of indoor walkable scenarios, we additionally collected 1,033 indoor images. For non-walkable cases, we collected a total of 9,913 images, including 9,107 outdoor cases and 806 indoor cases. The indoor images in both walkable and non-walkable cases were captured in various environments such as building lobbies, airports, shopping centers, and subway stations. Table II summarizes the collected images.

In particular, the information for destination, left, right, and path was annotated in two variants: simple and complex. The simple version only specifies the type of obstacle, while the complex version includes not only the obstacle type but also its distance and direction. In our study, we integrated the complex descriptions of destination, left, right, and path along with the navigability status with associated reasoning into a single unified sentence. During this process, we utilized GPT-4o to transform the original descriptions into the defined template format, ensuring that no information beyond what was provided in the dataset was artificially introduced. Through this procedure, we constructed a training dataset comprising a total of 19,945 samples.

I. Training

From the curated dataset, a total of 19,945 samples were converted into a conversation format and used for VLM train-

ing. Each query includes the goal position coordinates (x, y) and image tokens, while the response was required to follow one of the four predefined formats, incorporating walking possibility and the corresponding information required by each format (e.g., movement distance and direction, obstacle details, caution, or detour instructions). The trained VLM must be able to accurately interpret the given image and goal position, determine the appropriate walking situation, select the corresponding format accordingly, and generate a guidance sentence consistent with that format.

III. Experiments

A. Training Details

For the pretrained base VLM model, we used Qwen2.5VL [8], which supports various capabilities such as document parsing, object grounding, and visual question answering. The training framework was LLaMA-Factory [15]. All parts in 7B model including the vision encoder, vision-language adapter, and LLM were trained for a total of 5 epochs. We adopted LoRA-based supervised fine-tuning with a cosine scheduler, a learning rate of 1.0×10^{-4}.

We trained two versions of the model: (1) **SA-VLMv2**, which was trained only on the curated dataset of 19,945 samples; (2) **SA-VLMv2+**, which was trained on the same curated dataset combined with additional public datasets such as LLaVA-Instruct-150K [16], OpenOrca [17], and VQASynth [18] and OpenSpaces [19] to further enhance conversational ability and spatial reasoning. Training was conducted on 4 NVIDIA RTX 6000 Ada GPUs. The training time was approximately 8 hours for SA-VLMv2 and 35 hours for SA-VLMv2+.

B. Evaluation Metrics

The evaluation was conducted on the SA-Bench dataset [14], which provides a uniform format covering the destination, left, right, path, and navigability status. For this study, we manually restructured the dataset into our four predefined guidance templates. For ambiguous scenarios with multiple valid formats or navigation options (e.g., detouring left vs. right), we annotated the dataset with multiple ground-truth references. During evaluation, the model's prediction was compared against all available references, and the maximum score was recorded as the final score.

Performance comparison was divided into two categories: (1) holistic evaluation of the entire walking description, and (2) factor-wise evaluation of individual components within the description.

For holistic evaluation, we followed the same methodology as SA-VLM [14], employing METEOR [20], ROUGE [21], BERTScore [22] and LLM-as-a-Judge [23]. In addition to these quality measures, we also assessed the conciseness of responses by counting the number of words used.

The factor-wise evaluation was conducted according to the five criteria summarized in Table III. The evaluation procedure employed GPT-4o as the LLM-as-a-Judge [23] to compare candidate and reference descriptions. To compute fine-grained scores, the customized evaluation prompt was used, incorporating explicit scoring rules with predefined bonuses and penalties for each factor following the criteria in Table III. This approach enabled not only the calculation of an overall score but also the identification of specific components where performance differences emerged.

TABLE III
SUMMARY OF FACTOR-WISE EVALUATION CRITERIA.

Evaluation Factor	Criteria
Walking possibility accuracy (3 points)	Correct categorization among walking possibilities (possible / difficult / impossible)
Non-possibility information accuracy (3 points)	Completeness of additional information such as caution, obstacle, or detour (excluding walking possibility)
Format and structure (2 points)	Consistency with one of the four template structures and sentence format
Distance accuracy (2 points)	Numerical accuracy of distance (within $\pm 10\%$ or ± 1m)
Direction accuracy (2 points)	Accuracy of movement or obstacle/detour direction (tolerance: ± 1 hour on the clock face)

C. Prompting

In addition to training-based approaches, prompting was also employed to guide the VLM to select one of the four predefined templates and generate walking instructions in the correct structured form. Specifically, the four templates (possible, possible with caution, difficult, and impossible) were explicitly provided, along with the destination coordinates (x, y) and the image input. The prompt also included basic rules necessary for walking guidance for visually impaired individuals.

The prompt was organized into five key components:

(1) Role Definition: The assistant provides walking guidance for visually impaired individuals based on an image and target coordinates.

(2) Input Specification: An image and normalized coordinates (x, y), with (0.0, 0.0) at the top-left and (1.0, 1.0) at the bottom-right.

(3) Task Definition: Analyze the image and generate guidance to the target location using one of four predefined templates.

(4) Language Constraints: Instructions must be clear, concise, and free of ambiguity or vision-dependent terms. Use directions (e.g., "12 o'clock") and distances (e.g., "10 meters") when relevant.

(5) Output Constraints: Only the final sentence in a predefined template format, without explanation about reasoning or adding extra words.

Through this procedure, the VLM was encouraged to avoid unnecessary explanations or reasoning, and instead produce concise natural-language outputs that strictly follow one of the four predefined formats.

D. Holistic Evaluation

We conducted performance comparisons across three categories of models. First, proprietary generative models such as Claude 3.5 Sonnet, Gemini 2.5 Flash, and GPT-4o were evaluated using the same prompting strategy described in Section III-C, where each model was instructed to select one of the four predefined templates. Second, the base model

Qwen2.5VL was assessed under two prompt settings: (1) Qwen-Gen, where the prompting strategy in Section III-C was applied, and (2) Qwen2.5VL, where the same input prompt used for SA-VLMv2 queries was applied without explicitly providing the four templates. Notably, in Qwen2.5VL, the absence of explicit template guidance led to unconstrained free-form responses. Third, we evaluated SA-VLMv2, trained with our curated dataset, and SA-VLMv2+, trained with both the curated dataset and additional public datasets to enhance conversational and spatial reasoning abilities. For both SA-VLMv2 and SA-VLMv2+, we used a prompt which is *"You are given an image and asked to guide a visually impaired person. The destination is located at the normalized coordinate (x, y) within the image. Analyze the image and generate a clear and concise walking guidance."*.

The holistic evaluation results are summarized in Table IV. Overall, models with higher LLM Judge scores also achieved higher results on METEOR, BertScore, and ROUGE, demonstrating consistent evaluation trends across metrics.

SA-VLMv2 and SA-VLMv2+ demonstrated superior performance, reaching a METEOR score of up to 0.64 and an LLM Judge score of 7.45 and 7.34, respectively, while both achieving a BertScore of 0.82 and a ROUGE-L of 0.65. These results indicate that the performance improvements cannot be achieved by prompting alone, but rather are enabled by fine-tuning on a domain-specific dataset that enforces structured guidance generation.

In terms of response length, SA-VLMv2 and SA-VLMv2+ produced outputs averaging around 16 words, which is more concise than proprietary models (18–22 words). This conciseness, combined with higher accuracy, aligns with the needs of visually impaired users who prefer essential information without unnecessary verbosity. In contrast, Qwen2.5VL generated excessively long responses (84 words on average), reflecting the lack of structural constraints when explicit template guidance is absent. On the other hand, Qwen-Gen produced shorter outputs (9 words on average) but achieved lower accuracy across metrics, showing that brevity alone does not guarantee quality.

Finally, we observed little difference between SA-VLMv2 and SA-VLMv2+. This suggests that the curated dataset itself is sufficiently effective for training high-quality walking guidance models. Nevertheless, the inclusion of additional open datasets in SA-VLMv2+ may provide broader conversational and spatial reasoning capabilities, making it more adaptable to real-world deployment scenarios.

TABLE IV

COMPARISON OF MODEL PERFORMANCE.

Model	MET	Bert	ROUGE	Judge	#Words
Claude	0.46	0.72	0.45	4.43	21.51
Gemini	0.55	0.78	0.55	6.01	17.95
GPT-4o	0.56	0.78	0.56	5.48	20.11
Qwen-Gen	0.26	0.68	0.36	2.60	**9.08**
Qwen2.5VL	0.11	0.47	0.08	0.01	84.76
SA-VLMv2	**0.64**	**0.82**	**0.65**	**7.45**	15.99
SA-VLMv2+	0.63	**0.82**	**0.65**	7.34	15.59

MET = METEOR, Bert = BertScore, ROUGE = ROUGE-L, Judge = LLM_Judge, #Words = Num of Words.
Claude = Claude 3.5 Sonnet, Gemini = Gemini 2.5 Flash.

E. Factor-wise Evaluation

The factor-wise evaluation results are presented in Table V. This evaluation decomposed the quality of walking descriptions into five independent factors: (1) walking possibility, (2) non-possibility information (e.g., obstacle, caution, detour), (3) format and structural consistency, (4) distance accuracy, and (5) direction accuracy.

In terms of the overall score, SA-VLMv2 and SA-VLMv2+ achieved the highest performance with scores of 7.45 and 7.34, respectively, outperforming all proprietary models, including Gemini (6.01), GPT-4o (5.48), and Claude (4.43). Notably, Qwen-Gen reached only 2.60 – approximately half the performance of proprietary models and roughly one-third of the score achieved by SA-VLMv2. In the case of Qwen2.5VL, the absence of strict template constraints led to overly long and free-form responses, which hindered a proper evaluation across individual factors. The detailed factor-wise analysis is as follows:

- **Possibility (Poss.)**: SA-VLMv2 and SA-VLMv2+ scored 2.79 and 2.78, respectively, close to the maximum of 3.0, whereas proprietary models ranged between 1.86 and 2.33. This indicates that the SA-VLM models provide more reliable classification of walking possibility or impossibility.
- **Non-possibility info (NonPoss.)**: The SA-VLM models achieved 1.18–1.25, higher than the performance of proprietary models (0.60–1.03). This reflects a superior ability to identify and describe critical safety details, such as obstacles, cautions, and detour availability.
- **Format (Fmt.)**: SA-VLMv2 and SA-VLMv2+ achieved 1.77–1.79, demonstrating much stronger consistency with the predefined template structures compared to other models (1.22–1.45).
- **Distance (Dist.)**: All models performed poorly, with SA-VLMv2/v2+ reaching only 0.75–0.77. This highlights the inherent difficulty of accurately estimating distance from a single RGB image.
- **Direction (Dir.)**: Both SA-VLM models (1.46–1.47) and proprietary models (1.11–1.51) showed relatively stable performance. While Gemini achieved the highest score (1.51), the SA-VLM series maintained competitive accuracy in spatial orientation.

Overall, SA-VLMv2 and SA-VLMv2+ consistently demonstrated superior performance across nearly all factors, with particular strengths in walking possibility classification and structural consistency. The negligible performance gap between SA-VLMv2 and SA-VLMv2+ further indicates that the curated dataset alone is sufficient for high-quality walking guidance generation, while SA-VLMv2+ offers additional benefits in broader conversational and spatial reasoning tasks.

Distance estimation was the lowest-performing factor, largely due to its higher output resolution. While direction was categorized into seven discrete positions (covering the forward-facing range), distance was predicted in meters; this increased resolution, combined with range-dependent

difficulty, made precise estimation inherently challenging. Error patterns showed high accuracy within 5 meters, but increasing errors at longer distances, which accounts for the overall performance drop. This limitation highlights the necessity of integration with real-time depth sensors to enhance the system's reliability in long-range spatial perception.

TABLE V

MODEL COMPARISON RESULTS WITH FACTOR-WISE EVALUATION. PERFORMANCE WAS MEASURED BY DECOMPOSING GUIDANCE QUALITY INTO MULTIPLE FACTORS.

Model	Overall	Poss.	NonPoss.	Fmt.	Dist.	Dir.
Claude	4.43	1.86	0.60	1.22	0.04	1.11
Gemini	6.01	2.33	1.03	1.36	0.29	**1.51**
GPT-4o	5.48	2.14	0.82	1.45	0.05	1.35
Qwen-Gen	2.60	1.59	0.11	0.49	0.09	0.69
Qwen2.5VL	0.01	0.00	0.00	0.00	0.00	0.01
SA-VLMv2	**7.45**	**2.79**	**1.25**	**1.79**	0.75	1.47
SA-VLMv2+	7.34	2.78	1.18	1.77	**0.77**	1.46

Overall (range: 0–12), Poss. = Possibility (0–3), NonPoss. = Non-possibility info (0–3), Fmt. = Format (0–2), Dist. = Distance (0–2), Dir. = Direction. (0–2) Claude = Claude 3.5 Sonnet, Gemini = Gemini 2.5 Flash.

IV. CONCLUSIONS

In this paper, we introduced SA-VLMv2, a space-aware vision-language model tailored to generate structured walking guidance for visually impaired individuals. Through expert-driven template derivation, dataset construction, and fine-tuning of Qwen2.5VL, our model delivers concise, accurate, and user-aligned instructions. Quantitative experiments confirmed that SA-VLMv2 consistently outperforms both proprietary LLMs and base VLMs in holistic and factor-wise evaluations, while producing shorter, more efficient outputs.

Although our approach substantially improves guidance quality, limitations remain in precise distance estimation from single RGB images. In future work, we plan to develop a depth-aware integration framework that incorporates real depth-sensor data into the responses of SA-VLMv2, while exploring advanced multimodal reasoning to further enhance the system's reliability. Furthermore, we aim to conduct real-world user studies with visually impaired users to evaluate the practical utility of SA-VLMv2. Specifically, we will first identify high-priority navigation scenarios through in-depth discussions with these users and then deploy our system in those specific contexts. By analyzing user feedback and behavioral data, we intend to refine the model to better align with the diverse and nuanced needs of its end-users. We believe SA-VLMv2 represents a step toward deployable assistive AI systems, bridging the gap between powerful vision-language models and the practical navigation needs of visually impaired users.

ACKNOWLEDGMENT

The authors used ChatGPT and Gemini for manuscript polishing, English translation, and code debugging. These tools were also employed to verify the completeness and logical flow of Section II. The authors reviewed all AI outputs and maintain full responsibility for the final content. We also thank YoungBin Kim and SunAh Son for helping construct the survey system in Section II-C.

REFERENCES

[1] S. Saegusa, Y. Yasuda, Y. Uratani, E. Tanaka, T. Makino, and J.-Y. J. Chang, "Development of a guide-dog robot: Leading and recognizing a visually-handicapped person using a lrf," *J. Adv. Mech. Des. Syst. Manuf.*, vol. 4, no. 1, pp. 194–205, 2010.

[2] Y. Wei and M. Lee, "A guide-dog robot system research for the visually impaired," in *ICIT*, 2014, pp. 800–805.

[3] 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 *ICRA*. IEEE, 2021, pp. 11 470–11 476.

[4] B. Hong, Y. Guo, M. Chen, Y. Nie, C. Feng, and F. Li, "Collaborative route map and navigation of the guide dog robot based on optimum energy consumption," *AI & SOCIETY*, pp. 1–7, 2024.

[5] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, *et al.*, "Flamingo: a visual language model for few-shot learning," in *NeurIPS*, vol. 35, 2022, pp. 23 716–23 736.

[6] H. You, H. Zhang, Z. Gan, X. Du, B. Zhang, Z. Wang, L. Cao, S.-F. Chang, and Y. Yang, "Ferret: Refer and ground anything anywhere at any granularity," *arXiv preprint arXiv:2310.07704*, 2023.

[7] 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 *ICML*, 2023, pp. 19 730–19 742.

[8] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin, "Qwen2.5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.

[9] S. Liu, A. Hasan, K. Hong, R. Wang, P. Chang, Z. Mizrachi, J. Lin, D. L. McPherson, W. A. Rogers, and K. Driggs-Campbell, "Dragon: A dialogue-based robot for assistive navigation with visual language grounding," *IEEE Robot. Autom. Lett.*, p. 3712–3719, Apr. 2024.

[10] J. Xiao, N. Huang, H. Qiu, Z. Tao, X. Yang, R. Hong, M. Wang, and A. Yao, "Egoblind: Towards egocentric visual assistance for the blind," 2025. [Online]. Available: https://arxiv.org/abs/2503.08221

[11] K. Park, Y. Oh, S. Ham, K. Joo, H. Kim, H. Kum, and I. S. Kweon, "Sideguide:a large-scale sidewalk dataset for guiding impaired people," in *IROS*, 2020.

[12] H. Hwang, S. Kwon, Y. Kim, and D. Kim, "Is it safe to cross? interpretable risk assessment with gpt-4v for safety-aware street crossing," in *UR*, 2024.

[13] Y. Zhao, Y. Zhang, R. Xiang, J. Li, and H. Li, "Vialm: A survey and benchmark of visually impaired assistance with large models," 2024. [Online]. Available: https://arxiv.org/abs/2402.01735

[14] B. Han, W. han Yun, B.-S. Seo, and J. Kim, "Space-aware instruction tuning: Dataset and benchmark for guide dog robots assisting the visually impaired," in *ICRA*, 2025.

[15] Y. Zheng, R. Zhang, J. Zhang, Y. Ye, Z. Luo, Z. Feng, and Y. Ma, "Llamafactory: Unified efficient fine-tuning of 100+ language models," in *ACL*, 2024.

[16] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *NeurIPS*, 2023.

[17] W. Lian, B. Goodson, E. Pentland, A. Cook, C. Vong, and "Teknium", "Openorca: An open dataset of gpt augmented flan reasoning traces," https://huggingface.co/datasets/Open-Orca/OpenOrca, 2023.

[18] remyxai, "Vqasynth," 2024, gitHub repository. [Online]. Available: https://github.com/remyxai/VQASynth/tree/main

[19] B. Chen, Z. Xu, S. Kirmani, B. Ichter, D. Sadigh, L. Guibas, and F. Xia, "Spatialvlm: Endowing vision-language models with spatial reasoning capabilities," in *CVPR*, 2024, pp. 14 455–14 465.

[20] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proc. ACL Workshop Intrinsic Extrinsic Eval. Meas. MT Summarization*, 2005, pp. 65–72.

[21] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[22] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019.

[23] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, "Judging llm-as-a-judge with mt-bench and chatbot arena," in *NeurIPS Datasets and Benchmarks Track*, 2023.