GuideDog: A Real-World Egocentric Multimodal Dataset for Blind and Low-Vision Accessibility-Aware Guidance

Junhyeok Kim[♣*] **Jaewoo Park**[♣*] **Junhee Park**[♣] **Sangeyl Lee**[♣]
Jiwan Chung[♣] **Jisung Kim**[♢†] **Ji Hoon Joung**[♡†] **Youngjae Yu**[♠]
[♣]Yonsei University [♢]LG AI Research [♡]Euler Robotics [♠]Seoul National University
junhyeok@yonsei.ac.kr
https://github.com/jun297/GuideDog

Abstract

For people affected by blindness and low vision (BLV), safe and independent navigation remains a major challenge, impacting over 2.2 billion individuals worldwide. Although multimodal large language models (MLLMs) offer new opportunities for assistive navigation, progress has been limited by the scarcity of accessibility-aware datasets, because creating them requires labor-intensive expert annotation.

To this end, we introduce GUIDEDOG, a novel dataset containing 22K image-description pairs (2K human-verified) capturing real-world pedestrian scenes across 46 countries. Our human-AI pipeline shifts annotation from generation to verification, grounded in established BLV guidance standards from experts and research, improving scalability while maintaining quality. We also present GUIDEDOGQA, an 818-sample benchmark evaluating object recognition and depth perception. Experiments reveal that depth perception and adherence to these standards remain challenging for current MLLMs.

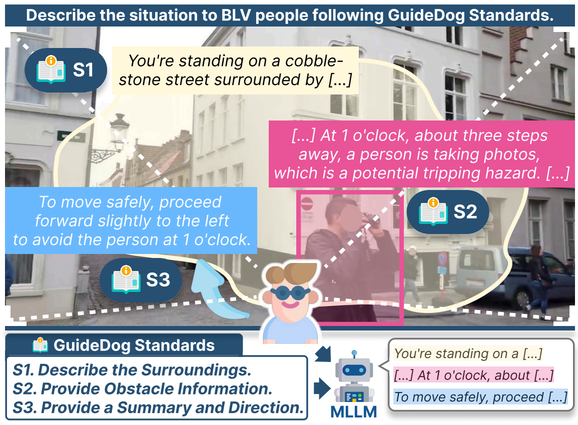


Figure 1: An overview of the accessibility-aware guidance generation task, wherein MLLMs describe the overall situation (*S1*), identify hazardous obstacles (*S2*), and summarize recommended actions for BLV. (*S3*)

1 Introduction

An estimated 2.2 billion individuals worldwide are affected by blindness and low vision (BLV), including approximately 36 million who are completely blind (Bourne et al., 2017). For this population, safe and independent mobility remains a significant daily challenge, as navigating unfamiliar or obstacle-filled environments poses substantial risks. Prior work has highlighted the severity of this issue, reporting that approximately 7% of visually impaired individuals experience falls at least once a month (Manduchi and Kurniawan, 2011). These challenges motivate the development of assistive technologies that support safe navigation and environmental understanding.

Early BLV assistive approaches, including electronic travel aids and computer vision–based systems, primarily focus on obstacle detection and avoidance (Kandalan and Namuduri, 2020; Lin et al., 2019). While effective for basic safety, these methods struggle to capture richer contextual information required in complex real-world environments (Liu et al., 2023b; Xie et al., 2024). Recent advances in multimodal large language models (MLLMs) (Liu et al., 2023a; Achiam et al., 2023) offer new opportunities for higher-level scene understanding, and have begun to be explored for BLV assistance (Xie et al., 2024; Merchant et al., 2024; Zhao et al., 2024; Hwang et al., 2023; Kim et al., 2025; Han et al., 2025).

However, effective guidance data for BLV assistance remains difficult to obtain. Sighted annotators often struggle to anticipate BLV-specific needs, making them unsuitable for BLV-aware annotation (Tigwell, 2021; Islam et al., 2024; Morris, 2020), and collecting diverse egocentric data requires costly manual exploration (Han et al., 2025). Consequently, annotation relies on a small pool of

[*] Equal Contribution.
[†]Work done while at SK Telecom.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5545–5574
July 2-7, 2026 ©2026 Association for Computational Linguistics

Dataset	# Samples	Modality	Source	Geo-Diverse	Annotation	BLV Involvement	Task
VizWiz (Gurari et al., 2018)	31K	Image	User photos	✗	Human	Captured by BLV	VQA
VIALM (Zhao et al., 2024)	200	Image	Web images	✗	Human	Expert manual	Guidance
Merchant et al. (2024)	48	Image	VizWiz	✗	Human	✗	Guidance
WalkVLM (Yuan et al., 2024)	12K	Video	Web + recorded	10 locations	Human	Survey	Guidance
EgoBlind (Xiao et al., 2025)	1.3K	Video	Web	✗	Human + AI	Annotated partial	Video QA
GUIDEDOG (Ours)	22K	Image	Web	183 locations	Human + AI	Standards-based	Guidance + QA

Table 1: Comparison of existing BLV guidance datasets. GUIDEDOG uniquely offers geographic diversity across 183 locations, standards-based BLV involvement, and comprehensive task coverage at scale.

Figure 2: Side-by-side comparison of examples from (a) GUIDEDOG and (b) VIALM (Zhao et al., 2024). GUIDEDOG consists of real-world scenes from a pedestrian viewpoint, while VIALM comprises home and supermarket images.

domain experts (Gurari et al., 2018; Zhao et al., 2024), resulting in BLV benchmarks that remain extremely small and limited in diversity, as summarized in Table 1 and illustrated in Figure 2 (Zhao et al., 2024; Merchant et al., 2024).

In this work, we introduce GUIDEDOG, a large-scale accessibility-aware dataset for blind and low vision (BLV) guidance, grounded in structured standards distilled from established BLV guidelines (see Figure 1). The dataset is constructed using a verification-centric human–AI pipeline that replaces free-form annotation with structured generation of *silver* labels followed by human verification into high-fidelity *gold* labels (Kim et al., 2024; Papadopoulos et al., 2016), enabling scalable data collection while maintaining annotation quality.

To capture realistic, in-the-wild scenarios reflective of BLV users' daily mobility experiences, GUIDEDOG samples frames from geographically and visually diverse 'walking videos'. These web-sourced videos are captured from an egocentric viewpoint across diverse real-world landscapes, mirroring the perspective of individuals with BLV. GUIDEDOG contains a 22K-image dataset (including 2K human-verified data) covering a broad range of real-world settings, each annotated under GUIDEDOG standards. Additionally, we construct GUIDEDOGQA, an evaluation subset featuring multiple-choice question-answer pairs. As GUIDEDOG contains real-world scenes, GUIDEDOGQA enables fine-grained evaluation of visual perception, such as verifying which objects are truly present (from correct and incorrect options) and determining relative depths among detected objects, an essential capability for both BLV assistance systems and broader egocentric applications like robotics.

Using GUIDEDOG, we evaluate several open-source and proprietary models. Our results show that proprietary models, such as GPT-4o, demonstrate superior zero-shot capabilities for BLV guidance. On GUIDEDOGQA, open-source models lead in object recognition but lag significantly in depth perception compared to proprietary counterparts. Notably, fine-tuning with silver labels further boosts performance. These findings underscore that accurate spatial understanding is critical for effective BLV assistance. We hope these benchmarks foster further research into MLLMs for BLV assistive technologies and real-world visual perception.

Our contributions are:

1. GUIDEDOG, an accessibility-aware dataset with 22K real-world scene image–description pairs designed for blind and low vision (BLV) users, along with GUIDEDOGQA, an 818-sample QA benchmark for evaluating fine-grained visual perception in real-world scenes.

2. A scalable dataset construction pipeline that shifts annotation from generation to human verification, using GUIDEDOG standards distilled from established BLV guidance.

3. Experimental results demonstrating that spatial understanding and adherence to BLV-specific standards remain key challenges for current MLLMs.

2 GUIDEDOG Dataset

We present GUIDEDOG, an egocentric multimodal dataset designed to evaluate MLLMs on BLV guidance in diverse real-world environments. To ensure scalability without compromising quality, we employ a two-stage pipeline that prioritizes verification over generation (Section 2.2). First, an automated pipeline generates high-quality *silver* labels guided by established GUIDEDOG standards; then, trained annotators verify and refine these labels to produce authoritative *gold* labels for evaluation following guidance from expert organizations and prior research.

2.1 GUIDEDOG Standards for BLV Guidance

Prior research has characterized BLV mobility challenges through extensive user studies (Merchant et al., 2024; Kuriakose et al., 2023; Song and Yang, 2010; Duh et al., 2020; Manduchi and Kurniawan, 2011; Madake et al., 2023), while organizations have codified practical protocols for sighted guides (Vision Australia, 2022; Emma Turner, Sense, 2023; Washington State University, 2025; Wisconsin Department of Health Services, 2023a,b; Vision Loss Resources, n.d.; Stevens, 2003; Be My Eyes, 2024). To structure these varied recommendations for MLLM evaluation, we consolidated them into three standards: *S1*, *S2*, and *S3*.

S1. Describe the Surroundings.
S2. Provide Obstacle Information.
S3. Provide a Summary and Direction.

S1. Describe the Surroundings. To orient BLV individuals in unfamiliar environments, assistive systems should provide contextual descriptions that establish spatial awareness (Merchant et al., 2024; Emma Turner, Sense, 2023; Wisconsin Department of Health Services, 2023b,a). This standard clearly identifies the user's current location and key environmental elements (Kuriakose et al., 2023; Song and Yang, 2010). For example, "You are on a busy pedestrian street with shops on both sides and a clear path for walking surrounded by people."

S2. Provide Obstacle Information. Obstacle awareness is a critical component of safe mobility for BLV individuals (Islam et al., 2024). This standard focuses on delivering comprehensive information about obstacles, including their type, location, and proximity, enabling BLV individuals to make informed decisions about their path (Kuriakose et al., 2023; Fernandes et al., 2019; Vision Australia, 2022; Emma Turner, Sense, 2023; Be My Eyes, 2024; Wisconsin Department of Health Services, 2023a,b). For instance, "Directly ahead at 12 o'clock, approximately 4 steps away, there is a sign post that could pose a risk of collision."

S3. Provide a Summary and Direction. Cognitive load management is essential for BLV navigation. This standard promotes offering concise summaries with intuitive measurements (e.g., "3 steps", "1 o'clock direction") instead of precise measurements (e.g., "3ft","5m") (Merchant et al., 2024; Duh et al., 2020; Wisconsin Department of Health Services, 2023b; Be My Eyes, 2024), while avoiding overly detailed explanations that may be overwhelming (Merchant et al., 2024; Be My Eyes, 2024). For example, "To navigate safely, proceed straight ahead with slight adjustments to avoid people at 12 and 2 o'clock."

These standards serve as the foundation for our annotation pipeline, ensuring high-quality guidance faithful to BLV best practices while enabling scalable dataset creation. Per-standard references are provided in Appendix A.

2.2 Dataset Construction

We present the data collection pipeline for GUIDEDOG, which consists of four key stages: (a) Scene Image Collection, (b) Scene Information Extraction, (c) Silver Label Generation, and (d) Gold Label Generation. The latter three stages (b, c, d) are illustrated in Figure 3.

Scene Image Collection. Diverse real-world settings are crucial for representing the daily experiences of BLV individuals. To address this, we source egocentric 'walking videos' from readily available YouTube channels that capture pedestrian perspectives. These channels typically provide extensive footage spanning multiple hours.

Formally, let $\mathcal{V}$ be the set of all videos. We apply a keyword-based filter to remove scenes irrelevant to BLV needs (e.g., those containing "drive", "car", or "bike") and eliminate non-relevant content such as shopping or vlogs, resulting in the filtered set of videos $\mathcal{V}_{\text{filtered}}$.

We next discard videos shorter than six minutes. For the remaining videos, we remove the first five minutes and the final minute, which often contains intro and outro, and exclude any live or restricted videos. From the filtered video set, $\mathcal{V}_{\text{filtered}}$, we then use GPT-4o (Achiam et al., 2023) to extract city and region information for each video, allowing us to maintain a geographically diverse distribution. Specifically, we sample at most 5 videos per city

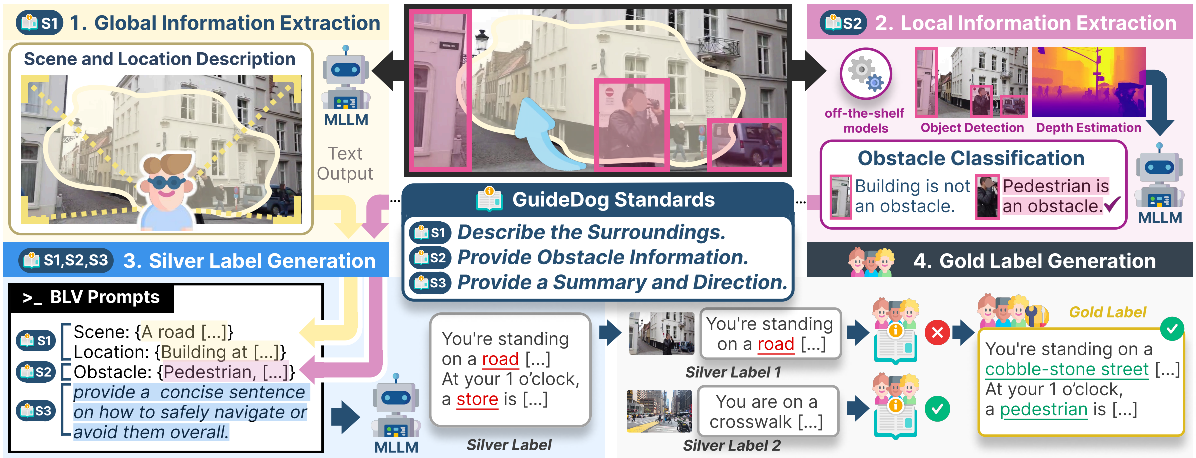


Figure 3: An overview of the GUIDEDOG generation pipeline, ensuring all stages adhere to GUIDEDOG standards (*S1*, *S2*, *S3*). For the collected scene image displayed at the center top, (1) in accordance with *S1*, the MLLM first extracts a comprehensive scene description; (2) following *S2*, both off-the-shelf models and the MLLM identify obstacles; (3) next, the extracted information is incorporated into a BLV-specific instruction designed to adhere to *S3*, generating a silver label that satisfies *S1*, *S2*, and *S3*; and (4) finally, human annotators assess and refine the silver labels to produce gold labels.

and at most 10 videos per region, producing a final video set $\mathcal{V}_{\text{sampled}}$.

From $\mathcal{V}_{\text{sampled}}$, we first allocate a maximum of 50,000 total frames across videos based on their respective city-region distribution, ensuring balanced geographic coverage. For shorter videos, we sample at a minimum rate of one frame every 10 seconds to avoid overly redundant coverage. For longer videos, we adaptively increase the sampling interval so that the total sampled frames remain within the 50,000 limit. We then remove near-identical frames using DINO (Caron et al., 2021) to reduce duplication further, resulting in a diverse collection of scene images for subsequent processing.

Scene Information Extraction. To generate *silver* labels that comply with GUIDEDOG standards, we extract two types of information: (1) global and (2) local. Each component provides information that enables MLLMs to generate *silver* labels in accordance with *S1* and *S2*.

For global information, we employ GPT-4o to filter irrelevant or noisy scenes (e.g., those where the camera points at the sky or ground or where the view is completely blocked). Then, GPT-4o extracts both a scene description $\mathcal{T}^{\text{s}}$ and a location description $\mathcal{T}^{\text{l}}$, capturing overall contextual details that satisfy *S1*. Formally, for the i-th scene, we obtain the global information $\mathcal{X}_i^{\text{global}} = (t_i^s, t_i^l)$.

For local information, we extract key objects and their associated spatial properties (location, distance, direction) to enable MLLMs to generate *silver* labels satisfying *S2*. MLLMs often struggle with fine-grained visual discrimination (Liang et al., 2024; Jiao et al., 2024) and exhibit inconsistent object naming (synonyms). To address this limitation, we predefine a curated set of 80 crucial objects based on previous research (Islam et al., 2024), enabling more consistent and reliable detection of obstacles and hazards relevant to BLV navigation.

We employ an open-world object detector (Cheng et al., 2024) to locate these objects in each i-th scene, yielding a set of k detected object classes $\mathcal{O}_i$ and corresponding bounding boxes $\mathcal{B}_i$. Next, a depth estimation model (Bochkovskii et al., 2024) generates a depth map m_i for the i-th scene. For each detected j-th object, its distance d_{ij} is computed as the median of depth values within the bounding box b_{ij}. Note that these distances serve as auxiliary spatial signals; human verification relies on relative depth ordering between objects rather than absolute metric values. To facilitate intuitive distance comprehension for BLV users, we convert these distances into step units (0.7 m per step) (Sekiya et al., 1997). To determine each object's direction l_{ij}, we assign a clock-face direction (10, 11, 12, 1, or 2 o'clock) to each bounding box center based on its horizontal position in the image.

From this process, we obtain the object set

Figure 4: Visualization of the worldwide (a) city and (b) region distribution of samples in the GUIDEDOG dataset.

$\mathcal{A}_i = \{(o_{ij}, b_{ij}, d_{ij}, l_{ij}) \mid j = 1, \ldots, n\}$ for the i-th scene. Finally, we feed the scene i and its associated information $\mathcal{A}_i$ into an MLLM to identify potentially crucial objects. Only those objects $\mathcal{X}_i^{\text{local}}$ (a subset of $\mathcal{A}_i$) are used to generate the final *silver* labels.

Silver Label Generation. In this stage, we combine the global information $\mathcal{X}_i^{\text{global}}$ and filtered local objects $\mathcal{X}_i^{\text{local}}$ with an accessibility-aware instruction for the MLLM to produce labels adhering to *S1*, *S2*, and *S3*. The resulting *silver* labels thus conform to key GUIDEDOG standards. We also apply EgoBlur (Raina et al., 2023) to safeguard privacy by blurring individual faces and license plates in the original images.

Collecting Human Annotations. To obtain high-quality *gold* labels for BLV guidance generation, we employ a human verification process for *silver* labels. Three sighted annotators verified and refined the *silver* labels by filtering out unsuitable images and low-fidelity GUIDEDOG standard annotations, while correcting any inaccuracies.

The verification applies two filters: image-level suitability (c_1) and standard-level adherence (c_2). Of the reviewed samples, 866 (26.5%) were rejected for image-level issues (e.g., unreadable frames, extreme angles, severe occlusion), and 265 (8.1%) for failing BLV guidance standards, confirming that the pipeline produces standards-compliant outputs when given suitable images.

Additionally, a randomly sampled subset of 150 images is annotated to construct GUIDEDOGQA. The annotators validate the detected object classes and their calculated distances. From these verified objects, we create multiple-choice question-answer pairs for two distinct tasks: (1) object recognition and (2) relative depth comparison. For object recognition, each verified object was presented alongside three distractor objects randomly selected from the 80 predefined classes not present in the image. For relative depth comparison, we randomly selected pairs of verified objects from each image and created questions about which object was closer to or farther from the camera. The final benchmark consists of 435 questions across 150 images for object recognition, and 383 questions across 135 images for relative depth comparison.

# Source Videos	269
# Total Frames in Source Videos	59.8M
Total Source Videos Duration	291 hours
# GUIDEDOG Samples	22,084
# GUIDEDOG Gold Label	2,106
# GUIDEDOGQA Samples	818
# Cities in GUIDEDOG	183
# Countries in GUIDEDOG	46

Table 2: GUIDEDOG data statistics. "Source videos" indicates that images in GUIDEDOG are sampled from video data.

2.3 Dataset Analysis

As presented in Table 2, the GUIDEDOG dataset comprises 22,084 samples in total, with 19,978 silver-labeled and 2,106 gold-labeled samples. GUIDEDOGQA further provides 818 QA samples for evaluating fine-grained visual perception. These samples are curated from 269 source videos with a combined duration of 291 hours, representing approximately 59.8 million candidate frames. GUIDEDOG can be easily scaled up by collecting more videos or sampling more frames. The dataset is not only substantial in size but also di-

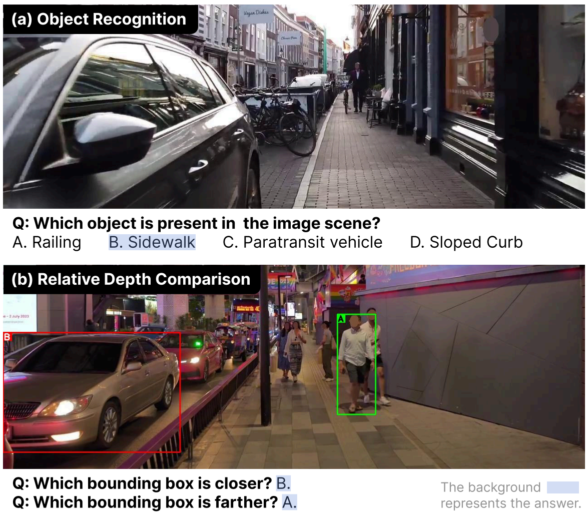


Figure 5: Examples of object recognition and relative depth comparison questions from GUIDEDOGQA.

verse in geographic coverage, spanning 183 cities across 46 countries. Our pipeline implements a balanced region-city sampling strategy to prevent regional biases. As illustrated in Figure 4, this approach prevents regional skews and produces a well-balanced global distribution. This geographic diversity ensures that GUIDEDOG captures a wide range of environments, essential for developing robust BLV assistance systems that can function effectively across diverse real-world settings.

3 Task Definition

We propose two complementary tasks to evaluate models' capabilities in assisting BLV individuals: (1) an accessibility-aware guidance generation task that provides comprehensive navigational assistance and (2) a real-world scene understanding task that assesses fundamental visual perception abilities through multiple-choice questions. The first task evaluates holistic BLV guidance according to established standards, while the second targets specific perceptual skills essential for effective navigation. We evaluate these tasks using our GUIDEDOG and GUIDEDOGQA datasets, respectively.

Guidance Generation. This task requires models to generate comprehensive guidance for BLV users from a given image, following the GUIDEDOG standards. To elicit standardized responses, we ask models to: (1) summarize the surroundings (*S1*), (2) concisely describe hazards using clock-face directions and step counts (*S2*), and (3) provide a final direction (*S3*).

Visual Perception QA. To diagnose failure modes in guidance generation and benchmark real-world scene perception, we assess two perceptual skills via multiple-choice questions. For object recognition, models select which of four candidate objects actually appears in the image. For depth comparison, models determine which of two bounding-boxed objects is closer or farther. Examples are shown in Figure 5.

4 Experiments

We evaluate models on two tasks: accessibility-aware guidance generation on GUIDEDOG and visual perception QA on GUIDEDOGQA. We first describe the evaluated models and then present results for each task.

Models We evaluate a diverse set of vision-language systems, including open-source, proprietary, and Socratic models (Zeng et al., 2022). For *open-source* models, we consider LLaVA 1.6 (Liu et al., 2024a), LLaVA-OneVision (Li et al., 2024), Qwen-2.5-VL (Bai et al., 2025), Cambrian-1 (Tong et al., 2025), and Molmo (Deitke et al., 2024). LLaVA 1.6, LLaVA-OneVision, and Qwen-2.5-VL represent high-performing open-source approaches, Cambrian-1 features a distinctive multi-encoder architecture, and Molmo is notable for its transparency. Additionally, we fine-tune Qwen-2.5-VL using LoRA (Hu et al., 2022) on our silver labels to investigate the impact of fine-tuning on performance. Among *proprietary models*, we include GPT-4o (Achiam et al., 2023) and Gemini 2.0 Flash (Google DeepMind, 2024), both recognized for their exceptional performance on general vision-language tasks.

Finally, we implement a Socratic Model (SM) (Zeng et al., 2022) using a two-stage framework: first, converting visual inputs into intermediate textual descriptions, and then performing text-only reasoning based on those descriptions and the associated object information $\mathcal{A}_i$ generated by off-the-shelf models (as described in Section 2.2). For image caption generation, we use LLaVA-1.6, while reasoning is performed using Gemini 2.0 Flash or GPT-4o. This approach enables us to assess the role of direct visual perception.

4.1 Accessibility-aware Guidance Generation

We evaluate models on the guidance generation task using GUIDEDOG, focusing on how well the

	Model	BLEU-2		BLEU-4		ROUGE-L		METEOR		GPT-Eval		Gemini-Eval
		0-shot	3-shot	0-shot	3-shot	0-shot	3-shot	0-shot	3-shot	0-shot	3-shot	0-shot
MLLM	Cambrian-1	0.218	0.329	0.087	0.175	0.221	0.323	0.267	0.375	0.219	0.307	0.277
	Molmo	0.139	0.360	0.042	0.205	0.205	0.350	0.277	0.423	0.268	0.334	0.323
	LLaVA-1.6	0.128	0.320	0.048	0.173	0.226	0.313	0.311	0.352	0.149	0.276	0.322
	LLaVA-OneVision	0.183	0.343	0.058	0.180	0.225	0.337	0.300	0.404	0.264	0.334	0.350
	Qwen2.5-VL	0.179	0.347	0.069	0.190	0.228	0.341	0.294	0.412	0.230	0.319	0.330
	↳ w/ finetuning	**0.408**	<u>0.418</u>	**0.235**	<u>0.246</u>	**0.399**	**0.400**	**0.471**	0.456	**0.541**	**0.529**	0.589
	Gemini 2.0 Flash	0.275	0.359	0.141	0.208	0.311	0.375	0.400	<u>0.463</u>	0.462	0.481	**0.664**
	GPT-4o	<u>0.339</u>	**0.425**	<u>0.188</u>	**0.252**	<u>0.359</u>	<u>0.399</u>	<u>0.445</u>	**0.474**	<u>0.490</u>	<u>0.505</u>	<u>0.651</u>
SM	Gemini 2.0 Flash	0.324	0.324	0.172	0.173	0.341	0.342	0.435	0.435	0.281	0.289	0.588
	GPT-4o	0.310	0.309	0.160	0.159	0.327	0.326	0.419	0.417	0.384	0.391	0.563

Table 3: Experimental results of the guidance generation task on GUIDEDOG. The best scores are **boldfaced**, and the second-best scores are <u>underlined</u>.

generated outputs align with the GUIDEDOG standards.

Metric Detailed performance metrics across all models and settings are shown in Table 3. To assess the alignment between model-generated outputs and ground-truth text, we employ standard natural language generation metrics, including BLEU-2, BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), and METEOR (Banerjee and Lavie, 2005).

To capture high-level semantic alignment with the GUIDEDOG standards beyond surface-level text similarity, we use LLM-based evaluation (Liu et al., 2023c) with GPT-4o as an automated evaluator to assess guidance quality based on BLV-specific criteria. Additionally, we conduct reference-based evaluation using GPT-4o to compare generated guidance against gold labels. GPT-4o measures the degree to which generated outputs align with the corresponding gold labels, and we report accuracy based on its assessments. To mitigate evaluator bias, we further validate with Gemini 2.0 Flash as an alternative evaluator, yielding consistent model rankings.

Visual Perception Matters for Guidance. GPT-4o exhibits the strongest zero-shot capabilities, while open-source models struggle significantly, with LLaVA-1.6 yielding the lowest scores across most metrics. SMs achieve competitive performance despite the absence of direct visual input processing, but remain inferior to MLLMs. This suggests that both direct visual perception and a robust language model foundation are essential for generating effective BLV guidance.

In-context Learning Helps to Adapt to Guidance Generation. In-context learning through few-shot examples substantially improves performance across nearly all models. In contrast, SM approaches show minimal performance changes between zero-shot and 3-shot settings. This suggests that in-context learning helps models understand the guidance generation task, while SM models already possess this understanding.

Model	Depth Comparison	Object Recognition
Random Chance	25.0	25.0
Cambrian-1	24.3	82.3
Molmo	28.5	34.0
LLaVA-1.6	30.0	80.9
LLaVA-OneVision	32.4	**87.4**
Qwen2.5-VL	22.2	<u>85.7</u>
↳ w/ finetuning	41.5	83.9
Gemini 2.0 Flash	<u>53.0</u>	65.7
GPT-4o	**67.1**	74.7

Table 4: Experimental results of the object recognition and relative depth comparison tasks on GUIDEDOGQA.

Model	S1	S2	S3	Avg
Cambrian-1	3.06	1.92	2.15	2.38
Molmo	3.02	2.41	2.36	2.60
LLaVA-OneVision	3.26	2.47	2.49	2.74
Qwen2.5-VL	2.93	2.27	2.23	2.48
↳ w/ finetuning	3.64	<u>3.45</u>	<u>3.67</u>	3.59
Gemini 2.0 Flash	<u>3.95</u>	<u>3.45</u>	3.40	<u>3.60</u>
GPT-4o	**4.15**	**3.8**	**3.76**	**3.90**
(SM) GPT-4o	2.88	2.02	2.7	2.53
Filtered Silver	4.76	4.62	4.52	4.63

Table 5: Results of the user study evaluated using a Likert scale ranging from 1 to 5, where 1 indicates poor adherence to the GUIDEDOG standard and 5 indicates high adherence. Best is **boldfaced**; second-best is <u>underlined</u>.

Fine-tuning Is Beneficial. Fine-tuning on BLV-specific data yields dramatic performance improvements, outperforming all other models. Moreover, fine-tuned Qwen2.5-VL demonstrates remarkable stability across zero-shot and few-shot settings, with only minimal additional gains from in-context examples. This suggests that fine-tuning on domain-specific data effectively embeds the BLV guidance principles that would otherwise need to be communicated through examples.

4.2 Question Answering

We evaluate models on the visual perception QA task using GUIDEDOGQA, which measures object recognition and relative depth comparison.

Metric We report accuracy on the multiple-choice QA task, which evaluates models' ability to identify objects present in the image and compare their relative depth. For depth comparison, we consider the answer correct only if the model correctly identifies both the closer and the farther objects to reduce the language prior. Detailed performance metrics for both object recognition and depth comparison tasks are presented in Table 4.

Open-source Models Recognize Objects Well. For object recognition, open-source models demonstrate surprisingly strong performance. Interestingly, proprietary models like GPT-4o and Gemini 2.0 Flash underperform in this task despite their superior guidance generation capabilities. The fine-tuned Qwen2.5-VL maintains strong object recognition performance, showing only a slight decrease (-1.8%) from its base model while substantially improving depth comparison ($+19.3\%$, see Table 4). This asymmetry suggests that occasional object-level noise (primarily hallucinated or misdetected objects in silver labels) is the main quality bottleneck of the silver-label pipeline, rather than structural or standard-adherence issues.

Depth Comparison Remains a Challenge. Depth perception presents a significant challenge across all models, especially for open-source models. Cambrian-1 and Qwen2.5-VL could not surpass random chance. The fine-tuned Qwen2.5-VL shows notable improvement in depth perception, indicating that BLV-specific training enhances spatial reasoning capabilities. This finding aligns with the needs of BLV users, who require not just information about what objects are present, but also their spatial relationships and proximity.

4.3 User Study

To complement our automated metrics, we conduct a subjective user study to assess how well different models adhere to the GUIDEDOG standards. As shown in Table 5, we recruited 14 participants to evaluate model-generated guidance across all three GUIDEDOG standards (*S1*: Describe the Surroundings, *S2*: Provide Obstacle Information, and *S3*: Provide Summary and Direction). The participants rated the outputs of each model on a 5-point Likert scale (Likert, 1932), where a score of 1 indicates poor adherence to the standard and a score of 5 indicates excellent adherence. We randomly sampled 86 images from the gold labels and collected guidance outputs from each model under evaluation.

Results. Our human evaluation results align with the automated metrics while providing additional insights into model performance across specific standards. GPT-4o consistently received the highest ratings across all three standards. Gemini 2.0 Flash ranks second overall, comparable to fine-tuned Qwen2.5-VL. For reference, silver labels evaluated by the same protocol achieved 4.76 (*S1*), 4.62 (*S2*), and 4.52 (*S3*). The relatively low scores for *S2* across most models highlight the challenge of accurately describing obstacles with appropriate spatial references that align with our depth perception results in Table 4. Socratic GPT-4o performs worse than GPT-4o, confirming our earlier finding that visual information is crucial for effective BLV guidance. Additionally, the Socratic approach performed relatively better on *S3*, suggesting that summary and guidance may rely more on high-level reasoning than on fine-grained visual details.

5 Related Works

MLLMs for BLV Assistance. Recent multimodal large language models (MLLMs) (Achiam et al., 2023; Liu et al., 2023a; Google DeepMind, 2024; Bai et al., 2025) have advanced vision-language capabilities beyond object recognition toward holistic and contextual scene understanding. Building on these advances, several works have explored MLLM-based assistance for blind and low vision (BLV) individuals, including navigation guidance (Zhao et al., 2024; Merchant et al., 2024; Xie et al., 2024), object localization (Liu et al., 2024b), robotic guide dog systems (Hwang et al., 2023; Kim et al., 2025; Han et al., 2025), and broader evaluations of MLLMs as visual assistants for BLV users (Karamolegkou et al., 2025). However, most existing approaches rely on small-scale datasets collected from limited domains, constraining their generalization to real-world BLV mobility scenarios.

Egocentric and BLV Datasets. Large-scale egocentric datasets such as Ego4D (Grauman et al., 2022) advance first-person scene understanding but are not tailored to BLV needs. BLV-specific efforts include ORBIT (Massiceti et al., 2021), an object recognition dataset directly captured by BLV users, and ViewQA (Song et al., 2024), which offers VQA with a 360-degree egocentric camera. EgoBlind (Xiao et al., 2025) further collects BLV-recorded egocentric videos from the web for visual assistance. VizWiz (Gurari et al., 2018, 2019; Huh et al., 2024) pioneered accessibility-aware VQA, while VIALM (Zhao et al., 2024) and Merchant et al. (2024) explored guidance generation, though both remain limited in scale due to costly expert annotation. An et al. (2025) studied BLV user preferences for navigation. WalkVLM (Yuan et al., 2024) contributes a video-based BLV walking dataset validated through surveys with visually impaired participants. GUIDEDOG instead grounds its standards in BLV guidelines and prior studies, and scales annotation through a verification-centric human-AI pipeline to 22K samples from 183 cities across 46 countries.

6 Conclusion

We introduce GUIDEDOG, a scalable accessibility-aware benchmark for evaluating MLLMs on BLV guidance against established mobility standards. Our analysis reveals that current models struggle to adhere to BLV-specific standards and lack the spatial understanding critical for safe navigation. We hope this work catalyzes future research into better spatial and temporal understanding, real-time navigation systems, and personalized assistance to better serve the global BLV community.

7 Limitations

Static Images We use static images to isolate guidance generation quality. Video inputs introduce confounds such as query timing and temporal coherence. Some hazards, including escalators and revolving doors, are motion-defined and may appear safe in a static frame; capturing such implicit motion blindness (Zhang, 2025) requires temporally grounded evaluation. Because GUIDEDOG is sourced from continuous walking tours, it can be extended to video-based navigation in future work.

Cross-Cultural Spatial Language Although GUIDEDOG spans 46 countries and reflects regional differences in object distributions and infrastructure (Section C.2), our annotations rely on universal spatial references and do not model cross-cultural variation in spatial language or object semantics (Karamolegkou et al., 2024). Modeling such variation remains an important direction for future work.

8 Ethical Considerations

License & Privacy We restricted data collection to videos published under the Creative Commons (CC-BY-SA) license. To safeguard privacy in public spaces, we applied EgoBlur (Raina et al., 2023) to automatically detect and blur faces and license plates, followed by a human verification stage to ensure complete anonymization.

Societal Impact GUIDEDOG aims to advance assistive technologies for the BLV community. We recognize that current MLLMs exhibit hallucinations and explicitly advise against deploying models trained on this data as standalone navigational aids without additional safety guardrails and extensive real-world validation.

Acknowledgments

This work was supported by the AI Seoul Tech Research Support Program of the Seoul Future Foundation. This work was partly supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean Government (MSIT) (No. RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)), the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00354218), and the Technology Innovation Program (RS-2025-25456760, Development of a humanoid robot specialized in chemical processes based on AI foundation model) funded by the Ministry of Trade, Industry and Resources (MOTIR, Korea). We thank the KAIT GPU project for computational support. The ICT at Seoul National University provides research facilities for this study.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Na Min An, Eunki Kim, Wan Ju Kang, Sangryul Kim, Hyunjung Shim, and James Thorne. 2025. Can lvlms and automatic metrics capture underlying preferences of blind and low-vision individuals for navigational aid? *arXiv preprint arXiv:2502.14883*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Be My Eyes. 2024. Tips for visual interpretation. https://support.bemyeyes.com/hc/en-us/articles/360005536638-Tips-for-Visual-Interpretation.

Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R Richter, and Vladlen Koltun. 2024. Depth pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073*.

Rupert RA Bourne, Seth R Flaxman, Tasanee Braithwaite, Maria V Cicinelli, Aditi Das, Jost B Jonas, Jill Keeffe, John H Kempen, Janet Leasher, Hans Limburg, and 1 others. 2017. Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: a systematic review and meta-analysis. *The Lancet Global Health*, 5(9):e888–e897.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660.

Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. 2024. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16901–16911.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, and 1 others. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*.

Ping-Jung Duh, Yu-Cheng Sung, Liang-Yu Fan Chiang, Yung-Ju Chang, and Kuan-Wen Chen. 2020. V-eye: A vision-based navigation system for the visually impaired. *IEEE Transactions on Multimedia*, 23:1567–1580.

Emma Turner, Sense. 2023. How to guide someone who is blind or partially sighted. https://www.sense.org.uk/blog/how-to-guide-someone-who-is-blind-or-partially-sighted.

Hugo Fernandes, Paulo Costa, Vitor Filipe, Hugo Paredes, and João Barroso. 2019. A review of assistive spatial orientation and navigation technologies for the visually impaired. *Universal Access in the Information Society*, 18:155–168.

Google DeepMind. 2024. Introducing gemini 2.0: our new ai model for the agentic era. https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, and 1 others. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18995–19012.

Danna Gurari, Qing Li, Chi Lin, Yinan Zhao, Anhong Guo, Abigale Stangl, and Jeffrey P Bigham. 2019. Vizwiz-priv: A dataset for recognizing the presence and purpose of private visual information in images taken by blind people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 939–948.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.

ByungOk Han, Woo-han Yun, Beom-Su Seo, and Jaehong Kim. 2025. Space-aware instruction tuning: Dataset and benchmark for guide dog robots assisting the visually impaired. *arXiv preprint arXiv:2502.07183*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Mina Huh, Fangyuan Xu, Yi-Hao Peng, Chongyan Chen, Hansika Murugu, Danna Gurari, Eunsol Choi, and Amy Pavel. 2024. Long-form answers to visual questions from blind and low vision people. *arXiv preprint arXiv:2408.06303*.

Hochul Hwang, Tim Xia, Ibrahima Keita, Ken Suzuki, Joydeep Biswas, Sunghoon I Lee, and Donghyun Kim. 2023. System configuration and navigation of a guide dog robot: Toward animal guide dog-level guiding work. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9778–9784. IEEE.

Md Touhidul Islam, Imran Kabir, Elena Ariel Pearce, Md Alimoor Reza, and Syed Masum Billah. 2024. Identifying crucial objects in blind and low-vision individuals' navigation. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–8.

Qirui Jiao, Daoyuan Chen, Yilun Huang, Yaliang Li, and Ying Shen. 2024. Enhancing multimodal large language models with vision detection models: An empirical study. *arXiv preprint arXiv:2401.17981*.

Roya Norouzi Kandalan and Kamesh Namuduri. 2020. Techniques for constructing indoor navigation systems for the visually impaired: A review. *IEEE Transactions on Human-Machine Systems*, 50(6):492–506.

Antonia Karamolegkou, Malvina Nikandrou, Georgios Pantazopoulos, Danae Sanchez Villegas, Phillip Rust, Ruchira Dhar, Daniel Hershcovich, and Anders Søgaard. 2025. Evaluating multimodal language models as visual assistants for visually impaired users. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25949–25982, Vienna, Austria. Association for Computational Linguistics.

Antonia Karamolegkou, Phillip Rust, Ruixiang Cui, Yong Cao, Anders Søgaard, and Daniel Hershcovich. 2024. Vision-language models under cultural and inclusive considerations. In *Proceedings of the 1st Human-Centered Large Language Modeling Workshop*, pages 53–66. ACL.

Hannah Kim, Kushan Mitra, Rafael Li Chen, Sajjadur Rahman, and Dan Zhang. 2024. Meganno+: A human-llm collaborative annotation system. *arXiv preprint arXiv:2402.18050*.

J Taery Kim, Morgan Byrd, Jack L Crandell, Bruce N Walker, Greg Turk, and Sehoon Ha. 2025. Understanding expectations for a robotic guide dog for visually impaired people. *arXiv preprint arXiv:2501.04594*.

Bineeth Kuriakose, Raju Shrestha, and Frode Eika Sandnes. 2023. Deepnavi: A deep learning based smartphone navigation assistant for people with visual impairments. *Expert Systems with Applications*, 212:118720.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Zijing Liang, Yanjie Xu, Yifan Hong, Penghui Shang, Qi Wang, Qiang Fu, and Ke Liu. 2024. A survey of multimodel large language models. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, pages 405–409.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yimin Lin, Kai Wang, Wanxin Yi, and Shiguo Lian. 2019. Deep learning based wearable assistive system for visually impaired people. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Ruiping Liu, Jiaming Zhang, Kunyu Peng, Junwei Zheng, Ke Cao, Yufan Chen, Kailun Yang, and Rainer Stiefelhagen. 2023b. Open scene understanding: Grounded situation recognition meets segment anything for helping people with visual impairments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1857–1867.

Ruiping Liu, Jiaming Zhang, Angela Schön, Karin Müller, Junwei Zheng, Kailun Yang, Kathrin Gerling, and Rainer Stiefelhagen. 2024b. Objectfinder: Open-vocabulary assistive system for interactive object search by blind people. *arXiv preprint arXiv:2412.03118*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023c. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Jyoti Madake, Shripad Bhatlawande, Anjali Solanke, and Swati Shilaskar. 2023. A qualitative and quantitative analysis of research in mobility technologies for visually impaired people. *IEEE Access*, 11:82496–82520.

Roberto Manduchi and Sri Kurniawan. 2011. Mobility-related accidents experienced by people with visual impairment. *AER Journal: Research and Practice in Visual Impairment and Blindness*, 4(2):44–54.

Daniela Massiceti, Luisa Zintgraf, John Bronskill, Lida Theodorou, Matthew Tobias Harris, Edward

Cutrell, Cecily Morrison, Katja Hofmann, and Simone Stumpf. 2021. Orbit: A real-world few-shot dataset for teachable object recognition. In *proceedings of the IEEE/CVF international conference on computer vision*, pages 10818–10828.

Zain Merchant, Abrar Anwar, Emily Wang, Souti Chattopadhyay, and Jesse Thomason. 2024. Generating contextually-relevant navigation instructions for blind and low vision people. *arXiv preprint arXiv:2407.08219*.

Meredith Ringel Morris. 2020. Ai and accessibility. *Communications of the ACM*, 63(6):35–37.

Dim P Papadopoulos, Jasper RR Uijlings, Frank Keller, and Vittorio Ferrari. 2016. We don't need no bounding-boxes: Training object class detectors using only human verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 854–863.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Nikhil Raina, Guruprasad Somasundaram, Kang Zheng, Sagar Miglani, Steve Saarinen, Jeff Meissner, Mark Schwesinger, Luis Pesqueira, Ishita Prasad, Edward Miller, Prince Gupta, Mingfei Yan, Richard Newcombe, Carl Ren, and Omkar M Parkhi. 2023. Egoblur: Responsible innovation in aria. *Preprint*, arXiv:2308.13093.

Noboru Sekiya, Hiroshi Nagasaki, Hajime Ito, and Taketo Furuna. 1997. Optimal walking in terms of variability in step length. *Journal of Orthopaedic & Sports Physical Therapy*, 26(5):266–272.

Inpyo Song, Minjun Joo, Joonhyung Kwon, and Jangwon Lee. 2024. Video question answering for people with visual impairments using an egocentric 360-degree camera. *arXiv preprint arXiv:2405.19794*.

Ji-Won Song and Sung-Ho Yang. 2010. Touch your way: haptic sight for visually impaired people to walk with independence. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '10, page 3343–3348, New York, NY, USA. Association for Computing Machinery.

Sue Stevens. 2003. Assisting the blind and visually impaired: guidelines for eye health workers and other helpers. *Community Eye Health*, 16(45):7.

Garreth W Tigwell. 2021. Nuanced perspectives toward disability simulations from digital designers, blind, low vision, and color blind people. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–15.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2025. Label Studio: Data labeling software. Open source software available from https://github.com/HumanSignal/label-studio.

Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, and 1 others. 2025. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356.

Vision Australia. 2022. Helpful resources for business consulting. https://www.visionaustralia.org/business-consulting/helpful-resources.

Vision Loss Resources. n.d. How to be a sighted guide. https://visionlossresources.org/resources/how-to-be-a-guide/.

Washington State University. 2025. Etiquette for interacting with people who are blind or have low vision. https://studentaffairs.vancouver.wsu.edu/access-center/etiquette-interacting-people-who-are-blind-who-have-low-vision.

Wisconsin Department of Health Services. 2023a. Dos and don'ts for people with vision loss. https://www.dhs.wisconsin.gov/obvi/adjustment/dos-donts.htm.

Wisconsin Department of Health Services. 2023b. Sighted guide techniques. https://www.dhs.wisconsin.gov/obvi/adjustment/sightedguidetech.htm.

Junbin Xiao, Nanxin Huang, Hao Qiu, Zhulin Tao, Xun Yang, Richang Hong, Meng Wang, and Angela Yao. 2025. EgoBlind: Towards egocentric visual assistance for the blind. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jingyi Xie, Rui Yu, He Zhang, Sooyeon Lee, Syed Masum Billah, and John M Carroll. 2024. Emerging practices for large multimodal model (lmm) assistance for people with visual impairments: Implications for design. *arXiv preprint arXiv:2407.08882*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Zhiqiang Yuan, Ting Zhang, Ying Deng, Jiapei Zhang, Yeshuang Zhu, Zexi Jia, Jie Zhou, and Jinchao Zhang. 2024. Walkvlm: Aid visually impaired people walking by vision language model. *arXiv preprint arXiv:2412.20903*.

Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico

Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, and 1 others. 2022. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*.

Xiantao Zhang. 2025. The escalator problem: Identifying implicit motion blindness in AI for accessibility. In *ICCV*.

Yi Zhao, Yilin Zhang, Rong Xiang, Jing Li, and Hillming Li. 2024. Vialm: A survey and benchmark of visually impaired assistance with large models. *arXiv preprint arXiv:2402.01735*.

A References Supporting the GuideDog Standards for BLV Guidance

This section lists the key references that inform the development of the GUIDEDOG Standards for BLV Guidance in Section 2.1. The supporting literature for each standard (S1–S3) is organized below.

S1. Describe the Surroundings (Merchant et al., 2024; Emma Turner, Sense, 2023; Wisconsin Department of Health Services, 2023b,a; Kuriakose et al., 2023; Song and Yang, 2010)

S2. Provide Obstacle Information (Merchant et al., 2024; Kuriakose et al., 2023; Fernandes et al., 2019; Vision Australia, 2022; Emma Turner, Sense, 2023; Washington State University, 2025; Wisconsin Department of Health Services, 2023a,b)

S3. Provide a Summary and Direction (Kuriakose et al., 2023; Fernandes et al., 2019; Vision Australia, 2022; Emma Turner, Sense, 2023; Washington State University, 2025; Wisconsin Department of Health Services, 2023a,b; Merchant et al., 2024; Be My Eyes, 2024)

B Automatic Pipeline Details

B.1 Gathering Videos

To construct a high-quality dataset of outdoor walking scenes, we implemented a systematic approach to video selection. We manually identified YouTube channels specializing in walking tours across urban and natural environments. All selected channels (listed in Table 6) satisfy three criteria: (1) content exclusively featuring walking tours, (2) availability under Creative Commons (CC) licensing, and (3) diverse geographical coverage. Following channel identification, we developed an automated pipeline to extract relevant video URLs. This pipeline is initialized by parsing channel metadata from a JSON configuration file and establishing a structured directory hierarchy for organizing filtered results. Our filtering mechanism applies multiple criteria to ensure content quality and relevance: (1) Exclusion of titles containing non-walking activities through regex-based keyword filtering (e.g., “drive”, “car”, “drone”, “shopping”, “market”, “VR”); (2) Removal of videos shorter than six minutes to ensure sufficient content depth; (3) Automatic skipping of live broadcasts and restricted content. After preprocessing, the resulting data distributions are visualized in Figure 4 using OpenStreetMap[1] Carto style[2] and Plotly[3].

B.2 Scene Information Extraction

To generate silver labels, we extract both global and local scene information using GPT-4o. For global extraction, the model first validates whether an input image represents a suitable street scene by providing a binary "Yes" or "No" assessment. Images are disqualified if (1) large objects obstruct the majority of the frame, (2) the focal point emphasizes elements peripheral to the street environment (e.g., store displays, signage), or (3) the camera perspective is oriented significantly upward or downward rather than at pedestrian level. After validation, the model produces a comprehensive scene description encompassing pedestrians, buildings, vehicles, and other salient elements. The spatial arrangement is documented using a clock-position reference system limited to the 10, 11, 12, 1, and 2 o’clock positions (with 10 o’clock corresponding to the leftmost portion of the image, 12 o’clock to the center, and 2 o’clock to the rightmost area). The prompt is shown in Figure 7.

For local information extraction, we employ an off-the-shelf object detection model combined with distance estimation. Detected objects are categorized into two zones based on a 5-meter proximity threshold. In the Complete Danger Zone, we evaluate whether objects directly intersect the user’s projected walking path, potentially resulting in a collision. In the Ordinary Zone, we focus primarily on dynamic elements such as approaching vehicles (motorcycles, cars), bicycles, and pedestrians. To refine this classification, we prompt the MLLM to provide explicit reasoning for each object’s danger designation. The prompt is shown in Figure 8.

B.3 Silver Label Generation

To generate silver labels that adhere to GUIDEDOG standards, we integrate the extracted global context information and filtered local object data using a specialized prompt designed to satisfy all three standards (*S1*, *S2*, and *S3*), as shown in Figure 9. During this stage, we also implement EgoBlur (Raina et al., 2023) technology to ensure privacy protection by automatically blurring faces and license plates in the original images.

[1] https://www.openstreetmap.org/
[2] https://carto.com/platform
[3] https://plotly.com/python/maps/

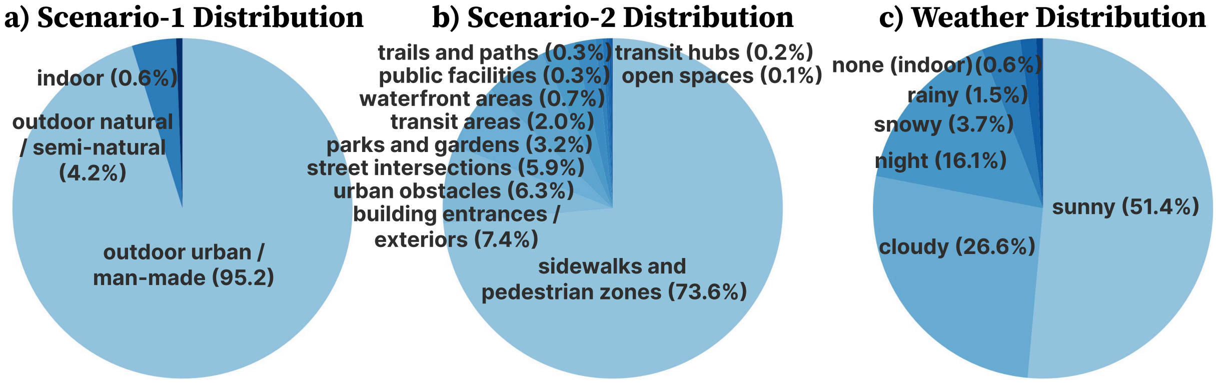


Figure 6: Scene distribution in the GUIDEDOG dataset based on a Places365-inspired taxonomy (Level-1 and Level-2 categories), along with overall weather category distribution.

B.4 Collecting Human Annotations

To establish gold-standard labels, we implemented a rigorous human verification protocol. Three sighted annotators reviewed the silver labels using Label Studio tools (Tkachenko et al., 2020-2025) (UI shown in Figure 18), filtering out unsuitable images and substandard annotations while correcting any inaccuracies. To ensure fair compensation, we paid the dedicated annotators approximately 1.5 times the local minimum hourly wage. Separately, the subjective user study was conducted with student volunteers. For GUIDEDOGQA construction, we randomly sampled 150 images from the verified dataset. Annotators validated both object classifications and distance measurements using the interface depicted in Figure 19. From these validated objects, we developed two categories of multiple-choice questions: (1) object recognition tasks, where each verified object is presented alongside three randomly selected distractors from the 80 predefined classes not present in the image; and (2) relative depth comparison tasks, where randomly selected object pairs prompt questions about relative proximity to the camera. The final benchmark comprises 435 object recognition questions across 150 images and 383 depth comparison questions spanning 135 images. Example gold labels for both GUIDEDOG and GUIDEDOGQA are illustrated in Figures 15, 16, and 17.

C Dataset Details

C.1 Fine-grained Analyses

We expand Section 2.3 by providing fine-grained dataset statistics. These include scene complexity (average of 9.3 objects per image) and description analysis (average length of 126 words, range: 52–212). As shown in Figure 6, we analyzed scene and scenario distributions using a Places365-inspired taxonomy. These were categorized and validated by GPT-4.1. One author manually verified 30 samples, confirming high agreement between GPT and human annotations (e.g., Level-1 scenario: 96% vs. 98%; Level-2 scenario: 86% vs. 86%; weather: 92% vs. 94%).

Regarding object distribution, beyond the most frequent category, person (29.5%), we observe a significant presence of vehicles (14.8%, including car, bus, bicycle) and other traffic- or street-related objects such as foldout signs (7.2%), lamp posts (5.5%), barrier posts (3.6%), and benches (2.0%). This distribution varies across regions and scenes, for example, vehicles are more common at street intersections (28.1%). Notably, in Vietnam, motorcycles (24.8%) significantly outnumber cars (13.2%).

C.2 Regional Object Distribution

To investigate whether GUIDEDOG reflects regional variation rather than averaging it out, we analyze object distributions across two regional groupings: East/Southeast Asia and four European subregions.

Southeast Asia exhibits a vehicle-dominant environment, with motorcycles dominant in Vietnam (77%), Cambodia (74%), and Thailand (50%). In contrast, East Asia shows higher prevalence of foldout signs (58% vs. 40%), indicating signage-dense pedestrian environments.

Western Europe shows a strong bicycle culture (32%; Netherlands 48%), while Southern Europe exhibits higher sidewalk (38%) and raised entryway (16%) frequencies, suggesting elevation-related walking risks. These differences indicate that GUIDEDOG structurally reflects meaningful

YouTube Channels	Channel URLs
Roma Walking Tour	https://www.youtube.com/@RomaWalkingTour
DuckTravel (World Tourist Channel)	https://www.youtube.com/@DuckTravel
Realistic Tourist	https://www.youtube.com/@RealisticTourist
POPtravel	https://www.youtube.com/@poptravelorg

Table 6: The list of YouTube channels from which videos were downloaded.

Object	Southeast Asia ($n = 186$)	East Asia ($n = 106$)
Motorcycle	42%	8%
Car	59%	25%
Foldout Sign	40%	58%
Road	33%	20%
Street Vendor	6%	3%

Table 7: Object frequency in Southeast Asia vs. East Asia samples.

Object	W. Europe ($n = 611$)	N. Europe ($n = 279$)	E. Europe ($n = 503$)	S. Europe ($n = 253$)
Foldout Sign	45%	28%	34%	30%
Bicycle	32%	11%	9%	8%
Trash bins	33%	16%	22%	21%
Building	57%	72%	51%	57%
Lamp Post	34%	41%	45%	37%
Sidewalk	33%	25%	29%	38%
Raised Entryway	15%	13%	9%	16%

Table 8: Object frequency across European subregions.

regional variations in mobility culture and urban infrastructure.

C.3 Viewpoint Comparison with BLV-Recorded Video

A natural concern with using sighted walking-tour footage is the gap between this viewpoint and that of a BLV user with a head-mounted assistive camera. To quantify this gap, we computed optical-flow-based motion statistics on GUIDEDOG source videos and on EgoBlind (Xiao et al., 2025), a BLV-recorded video dataset. We sampled one matched 20-second window per video ($N = 332$ per domain) and extracted frames at 2 FPS for dense optical-flow analysis.

EgoBlind exhibits a higher stationary ratio (median 16.6% vs. 8.3%), indicating more frequent static or slow-movement content. GUIDEDOG shows substantially stronger bottom-region motion (Cliff's $d = -0.48$, large), consistent with sustained forward walking and ground-plane dynamics. Jitter differences are negligible, suggesting comparable camera stability during active movement.

To explain these quantitative differences, we annotated 50 random clips per domain:

EgoBlind's higher stationary ratio reflects frequent standing (42%) and slow walking (52%), with visible cane motion absent in GUIDEDOG. GUIDEDOG predominantly shows slow walking (84%), typical of walking tour recordings. EgoBlind clips also contain video edits (30%) and subtitle overlays (60%), both largely absent in GUIDEDOG. Both datasets use body-mounted cameras (98%), indicating that the motion differences stem from navigation behavior rather than camera configuration.

These results confirm a measurable viewpoint gap, which we acknowledge as a limitation. The two datasets are complementary: EgoBlind captures authentic BLV motion patterns but is constrained by edits, limited scale, and limited geographic diversity, while GUIDEDOG offers clean, geo-diverse scenes aligned with smart-glasses-style assistive scenarios.

D Experiment Details

D.1 Training details

To perform a more comprehensive analysis of GUIDEDOG, we fine-tune Qwen-2.5-VL using LoRA on our silver labels. Specifically, we utilize Hugging Face AutoTrain[4] with a learning rate of 2e-5, a batch size of 2, and train for 1 epoch, taking approximately 1 hour on four NVIDIA A6000 GPUs.

D.2 Evaluation Details

In this study, we conduct experiments using the `lmms-eval` repository[5]. All evaluations are performed on a machine equipped with NVIDIA A100 GPUs in a single run.

[4] https://huggingface.co/docs/autotrain/en/index

[5] https://github.com/EvolvingLMMs-Lab/lmms-eval

Metric	KS stat	p-value	Cliff's d	Effect size
Flow Magnitude (median)	0.273	<0.001	-0.329	Small
Global Motion (mean)	0.253	<0.001	-0.281	Small
Global Motion (std / jitter)	0.134	0.005	0.118	Negligible
Stationary Ratio	0.278	<0.001	0.333	Medium
Bottom-1/3 Motion (mean)	0.391	<0.001	-0.480	Large

Table 9: Optical-flow motion statistics comparing GUIDEDOG source videos and EgoBlind ($N = 332$ per domain).

Property	EgoBlind ($n = 50$)	GUIDEDOG ($n = 50$)
Standing	21 (42%)	0 (0%)
Slow walk	26 (52%)	42 (84%)
Normal walk	1 (2%)	7 (14%)
Cane visible	21 (42%)	0 (0%)
Edit/cut visible	15 (30%)	0 (0%)
Subtitle occlusion	30 (60%)	2 (4%)

Table 10: Qualitative properties of 50 randomly sampled clips per domain.

D.3 Evaluation Prompts

We evaluate two main model types: the general MLLM and the SM model. For each model, we configured both 0-shot and 3-shot prompts. Figure 11, Figure 12, Figure 13, and Figure 14 show these four configurations, respectively. Additionally, we adopted GPT-Eval to semantically evaluate the outputs of these models. We designed the prompt based on Yu et al. (2023) to assign scores ranging from 0 to 1, as shown in Figure 10.

D.4 User Study

We conduct a subjective user study to assess how well different models adhere to the GUIDEDOG standards. We employ 14 human annotators to score the outputs of seven models per image using the Label Studio tool, as shown in Figure 20. We construct a gold label set for GUIDEDOG and GUIDEDOGQA, and sample annotations are provided in Figure 18 and Figure 19.

E AI Usage

We used Gemini to fix typos and polish the written sentences.

You are an expert tasked with determining whether an image depicts a **street scene**. Please follow the instructions below:
1. Review the image and decide if it represents a street scene ("Yes") or not ("No").
Filter out the following non-street cases from the label `{is_street}`:

- Certain objects block most of the screen, making it difficult to recognize the street.
- The viewer is looking at something other than the street (e.g., looking at the display glass of a store on the street, looking at a store's sign, etc.).
- The viewer is looking upwards or at the floor, not a straight sight of pedestrians.

And include the following cases:

- If the image is of a pedestrian's sight, even if it does not depict a designated street or a clear path, because this viewer is blind or has low vision, they could unknowingly enter a roadway.

2. If `"Yes"`:

- scene_description: Provide an overview of the street, including pedestrians, buildings, vehicles, or any key elements that make it a street.
- scene_location: Describe the location and surroundings using only 10, 11, 12, 1, and 2 o'clock positions. The leftmost part of the image is 10 o'clock, the center is 12 o'clock, and the rightmost is 2 o'clock.

3. If `"No"`:

- scene_description: Briefly explain why this is not a street scene.
- scene_location: Must be `"None"`.

4. Use only the JSON format below. 5. Do not include any text outside of this JSON format.

Output JSON example:

```
{
    "is_street": "Yes",
    "scene_description": "A detailed description of the street, including key
    elements such as pedestrians, shops, and vehicles.",
    "scene_location": "A positional overview using 10, 11, 12, 1, and 2 o'clock
    references."
}

Or:

{
    "is_street": "No",
    "scene_description": "Reason why this is not a street scene (e.g., highway,
    indoor, or empty road).",
    "scene_location": "None"
}
```

You must include exactly these three fields:

- `"is_street"`
- `"scene_description"`
- `"scene_location"`

No additional text beyond this JSON structure is allowed.

Figure 7: The global information extraction prompt filters out inappropriate images and extracts scene information and location details in accordance with *S1*. The inputs to the prompts are **boldfaced**.

You are an AI vision safety expert specializing in assisting visually impaired and low-vision users. Your task is to analyze a street image and determine which objects pose a danger to the user.

The object information is divided into two sections based on a distance threshold of 5 meters:

Complete Danger Zone (within 5 meters):
{in_object_info}

Ordinary Zone (beyond 5 meters):
{out_object_info}

Instructions:

1. Complete Danger Zone (within 5 meters):
   - Evaluate whether each object is directly in the user's walking path and could lead to a collision.
   - Mark an object as dangerous ("Yes") only if it poses a collision risk (e.g., curbs, potholes, poles, stairs, construction barriers, parked vehicles, etc.).
   - Otherwise, mark it as not dangerous ("No") and briefly explain why.

2. Ordinary Zone (beyond 5 meters)
   - Focus on moving objects in this zone (e.g., approaching motorcycles, cars, bicycles, pedestrians).
   - Mark an object as dangerous ("Yes") if it is moving toward the user and could pose a threat.
   - Otherwise, mark it as not dangerous ("No") and provide a brief explanation.

For each object, provide a JSON entry with the following three fields:

- `"object"`: the object's identifier or name.
- `"is_dangerous"`: "Yes" if dangerous, "No" if not.
- `"why_dangerous"`: a clear explanation for your assessment.

Output your entire response strictly in JSON format following this example:

```
[
  {
    "object": "object_name",
    "is_dangerous": "Yes",
    "why_dangerous": "This object is in the user's path and may cause a
    collision."
  },
  {
    "object": "object_name",
    "is_dangerous": "No",
    "why_dangerous": "This object is stationary and not in the user's immediate
    path."
  }
]
```

Do not include any text outside of the JSON structure.

Figure 8: The local information extraction prompt, which evaluates detected objects and classifies each as either dangerous or not, in order to satisfy *S2*. The inputs to the prompts are **boldfaced**.

You are an expert guide for visually impaired individuals. Your task is to provide a concise explanation based on the following guidelines, delivering the content as if speaking naturally without section breaks.

Guidelines: 1) Surroundings and Position: Summarize where the person is, the general environment, their current position, and any nearby landmarks in 1-2 sentences.
2) Hazards:

- For each direction (10, 11, 12, 1, and 2 o'clock), combine all hazards in that direction into exactly one sentence, mentioning approximate distance(s) and reason(s) they are dangerous.
- Follow the order of 10, 11, 12, 1, and 2 o'clock.

3) Navigation: After describing all hazards, provide a single, concise sentence on how to safely navigate or avoid them overall.
Potential Hazards:
{object_info}

Scene Information:
{scene_info}

Remember to provide a single, flowing explanation without labeled sections, as if talking directly to the visually impaired individual.

Figure 9: The silver label generation prompt leverages classified dangerous objects and extracted scene information to generate silver labels. The input of the prompts are **boldfaced**.

Compare the ground truth and prediction from AI models, to give a correctness score for the prediction. <AND> in the ground truth means it is totally right only when all elements in the ground truth are present in the prediction, and <OR> means it is totally right when any one element in the ground truth is present in the prediction. The correctness score is 0.0 (totally wrong), 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, or 1.0 (totally right). Just complete the last space of the correctness score.

gpt_query_prompt |Ground truth |Prediction |Correctness
What is x in the equation? |-1 <AND>-5 |x = 3 |0.0
What is x in the equation? |-1 <AND>-5 |x = -1 | 0.5
What is x in the equation? |-1 <AND>-5 |x = -5 |0.5
What is x in the equation? |-1 <AND>-5 |x = -5 or 5 |0.5
What is x in the equation? |-1 <AND>-5 |x = -1 or x = -5 |1.0

Can you explain this meme? |This meme is poking fun at the fact that the names of the countries Iceland and Greenland are misleading. Despite its name, Iceland is known for its beautiful green landscapes, while Greenland is mostly covered in ice and snow. The meme is saying that the person has trust issues because the names of these countries do not accurately represent their landscapes. |The meme talks about Iceland and Greenland. It's pointing out that despite their names, Iceland is not very icy and Greenland isn't very green. |0.4

Can you explain this meme? |This meme is poking fun at the fact that the names of the countries Iceland and Greenland are misleading. Despite its name, Iceland is known for its beautiful green landscapes, while Greenland is mostly covered in ice and snow. The meme is saying that the person has trust issues because the names of these countries do not accurately represent their landscapes. |The meme is using humor to point out the misleading nature of Iceland's and Greenland's names. Iceland, despite its name, has lush green landscapes while Greenland is mostly covered in ice and snow. The text 'This is why I have trust issues' is a playful way to suggest that these contradictions can lead to distrust or confusion. The humor in this meme is derived from the unexpected contrast between the names of the countries and their actual physical characteristics. |1.0

Write a single, natural-sounding passage with no section breaks. Begin by providing a 1–2 sentence summary of the person's surroundings and position, including any nearby landmarks. Then, for each direction from 10 to 2 o'clock, combine all hazards in exactly one sentence, specifying approximate distances and why they are dangerous in the order of 10, 11, 12, 1, and 2 o'clock. Finally, offer a single concise sentence explaining how to safely navigate or avoid all hazards. |**{gold_label}** |**{model_generated_output}**

Figure 10: The GPT-Eval prompt, which evaluates how well it adheres to GUIDEDOG standards. The inputs for the prompt are **boldfaced**.

You are an expert guide for visually impaired individuals. Your task is to provide a concise explanation based on the following guidelines, delivering the content as if speaking naturally without section breaks.

Guidelines: 1) Surroundings and Position: Summarize where the person is, the general environment, their current position, and any nearby landmarks in 1-2 sentences.
2) Hazards:

- For each direction (10, 11, 12, 1, and 2 o'clock), combine all hazards in that direction into exactly one sentence, mentioning approximate distance(s) and reason(s) they are dangerous.
- Follow the order of 10, 11, 12, 1, and 2 o'clock.

3) Navigation: After describing all hazards, provide a single, concise sentence on how to safely navigate or avoid them overall.

Remember to provide a single, flowing explanation without labeled sections, as if talking directly to the visually impaired individual.

Figure 11: The zero-shot prompt designed for MLLM to generate accessibility-aware guidance generation in GUIDEDOG. The inputs of the prompts are **boldfaced**.

You are an expert guide for visually impaired individuals. Your task is to provide a concise explanation based on the following guidelines, delivering the content as if speaking naturally without section breaks.

Guidelines: 1) Surroundings and Position: Summarize where the person is, the general environment, their current position, and any nearby landmarks in 1-2 sentences.
2) Hazards:

- For each direction (10, 11, 12, 1, and 2 o'clock), combine all hazards in that direction into exactly one sentence, mentioning approximate distance(s) and reason(s) they are dangerous.

- Follow the order of 10, 11, 12, 1, and 2 o'clock.

3) Navigation: After describing all hazards, provide a single, concise sentence on how to safely navigate or avoid them overall.

Examples:

- You're on a bustling city street with buildings on your left, and the sidewalk and storefronts on your right. At 10 o'clock, about five steps away, there's a moving car which is potentially dangerous if you stray off the sidewalk. To navigate safely, stay on the sidewalk, maintaining a safe distance from the road.

- You're standing in a lively marketplace with stalls under umbrellas at 12 o'clock and buildings in the background; there are parked vehicles to your sides. At 10 o'clock, approximately 5 steps away, there's a parked car that could obstruct any movement in that direction. At 11 o'clock, there are no immediate hazards. Directly ahead, at 12 o'clock, the market stalls might pose a minor obstacle if you walk too closely. At 1 o'clock, there is a car about 4 steps away, posing a potential obstacle. At 2 o'clock, no significant hazards are present. To navigate safely, proceed slowly towards 12 o'clock while veering slightly to your right to avoid the car at 1 o'clock.

- You are in a public plaza with buildings directly ahead at 12 o'clock and an art structure at 2 o'clock, with pedestrians around. At 10 o'clock, there is a barrier post and a pole about 3 steps away which could obstruct your path. At 11 o'clock, trash bins are 4 steps away, which might be a tripping hazard. Directly ahead at 12 o'clock, a foldout sign is 3 steps away, posing a risk of collision. At 2 o'clock, a barrier post is 4 steps away, which could also cause a trip. To safely navigate the area, move slightly to your left and proceed forward, avoiding the central obstacles.

Remember to provide a single, flowing explanation without labeled sections, as if talking directly to the visually impaired individual.

Figure 12: The 3-shot prompt designed for MLLM to generate accessibility-aware guidance generation in GUIDEDOG. The inputs of the prompts are **boldfaced**.

You are an expert guide for visually impaired individuals. Your task is to provide a concise explanation based on the following guidelines, delivering the content as if speaking naturally without section breaks.

Guidelines: 1) Surroundings and Position: Summarize where the person is, the general environment, their current position, and any nearby landmarks in 1-2 sentences.
2) Hazards:

- For each direction (10, 11, 12, 1, and 2 o'clock), combine all hazards in that direction into exactly one sentence, mentioning approximate distance(s) and reason(s) they are dangerous.
- Follow the order of 10, 11, 12, 1, and 2 o'clock.

3) Navigation: After describing all hazards, provide a single, concise sentence on how to safely navigate or avoid them overall.

Remember to provide a single, flowing explanation without labeled sections, as if talking directly to the visually impaired individual.

Scene Description: **{llava_output**}
Object Info:
{object_info}
Guidance:

Figure 13: The zero-shot prompt designed for SM to generate accessibility-aware guidance generation in GUIDEDOG. The inputs of the prompts are **boldfaced**.

You are an expert guide for visually impaired individuals. Your task is to provide a concise explanation based on the following guidelines, delivering the content as if speaking naturally without section breaks.

Guidelines:
1) Surroundings and Position: Summarize where the person is, the general environment, their current position, and any nearby landmarks in 1-2 sentences.
2) Hazards:

- For each direction (10, 11, 12, 1, and 2 o'clock), combine all hazards in that direction into exactly one sentence, mentioning approximate distance(s) and reason(s) they are dangerous.
- Follow the order of 10, 11, 12, 1, and 2 o'clock.

3) Navigation: After describing all hazards, provide a single, concise sentence on how to safely navigate or avoid them overall.

Remember to provide a single, flowing explanation without labeled sections, as if talking directly to the visually impaired individual.

Scene Description:
Object Info:
Guidance: You're on a bustling city street with buildings on your left, and the sidewalk and storefronts on your right. At 10 o'clock, about five steps away, there's a moving car which is potentially dangerous if you stray off the sidewalk. To navigate safely, stay on the sidewalk, maintaining a safe distance from the road.

Scene Description:
Object Info:
Guidance: You're standing in a lively marketplace with stalls under umbrellas at 12 o'clock and buildings in the background; there are parked vehicles to your sides. At 10 o'clock, approximately 5 steps away, there's a parked car that could obstruct any movement in that direction. At 11 o'clock, there are no immediate hazards. Directly ahead, at 12 o'clock, the market stalls might pose a minor obstacle if you walk too closely. At 1 o'clock, there is a car about 4 steps away, posing a potential obstacle. At 2 o'clock, no significant hazards are present. To navigate safely, proceed slowly towards 12 o'clock while veering slightly to your right to avoid the car at 1 o'clock.

Scene Description:
Object Info:
Guidance: You are in a public plaza with buildings directly ahead at 12 o'clock and an art structure at 2 o'clock, with pedestrians around. At 10 o'clock, there is a barrier post and a pole about 3 steps away which could obstruct your path. At 11 o'clock, trash bins are 4 steps away, which might be a tripping hazard. Directly ahead at 12 o'clock, a foldout sign is 3 steps away, posing a risk of collision. At 2 o'clock, a barrier post is 4 steps away, which could also cause a trip. To safely navigate the area, move slightly to your left and proceed forward, avoiding the central obstacles.

Scene Description: **{vlm_output}**
Object Info:
{object_info}
Guidance:

Figure 14: The 3-shot prompt designed for SM to generate accessibility-aware guidance generation in GUIDEDOG. The inputs of the prompts are **boldfaced**.

Gold Label: You're on a cobblestone path in a lively street surrounded by historical buildings, with shops to your left and people exploring the area. At your 10 o'clock, about three steps away, are shop displays that might protrude slightly, making the path narrower. From 11 to 1 o'clock, pedestrians are walking closely at about four to five steps ahead, presenting a risk of bumping due to the crowd. To navigate safely, move slightly to your right where there's more open space and maintain a steady pace, staying aware of nearby people and obstacles.

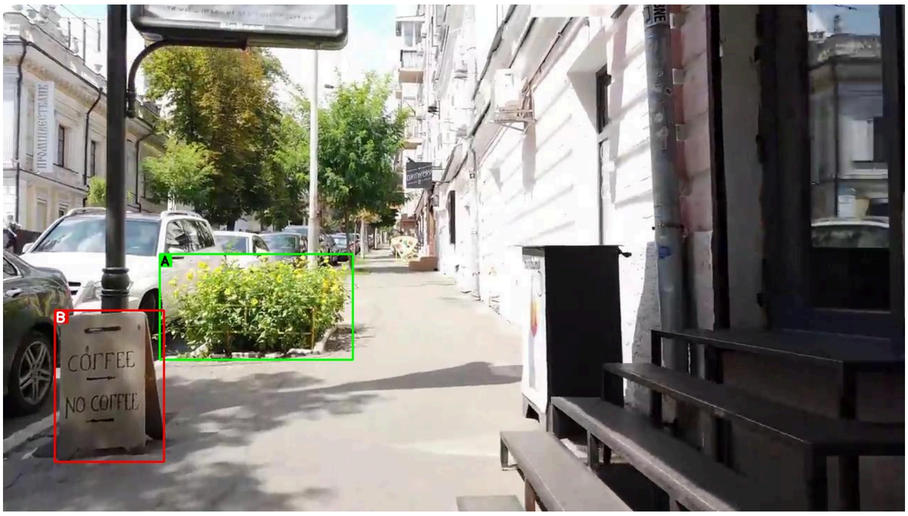
Figure 15: An example of a gold label in GUIDEDOG.

Which object is **present** in the image scene?

A. Lamp Post
B. Wheelchair
C. Train Tracks
D. Car

Answer: **D**

Figure 16: An example of an object recognition task on GUIDEDOGQA.



A bounding box is an annotated rectangle surrounding an object. The edges of bounding boxes should touch the outermost pixels of the object that is being labeled. Given the two bounding boxes on the image, labeled by A and B, which bounding box is **closer/farther** to the camera? Select from the following choices." Answer with the option's letter from the given choices directly.

A. Vegetation
B. Foldout Sign

In closer QA, answer : **B**
In farther QA, answer: **A**

Figure 17: An example of a depth recognition task on GUIDEDOGQA.

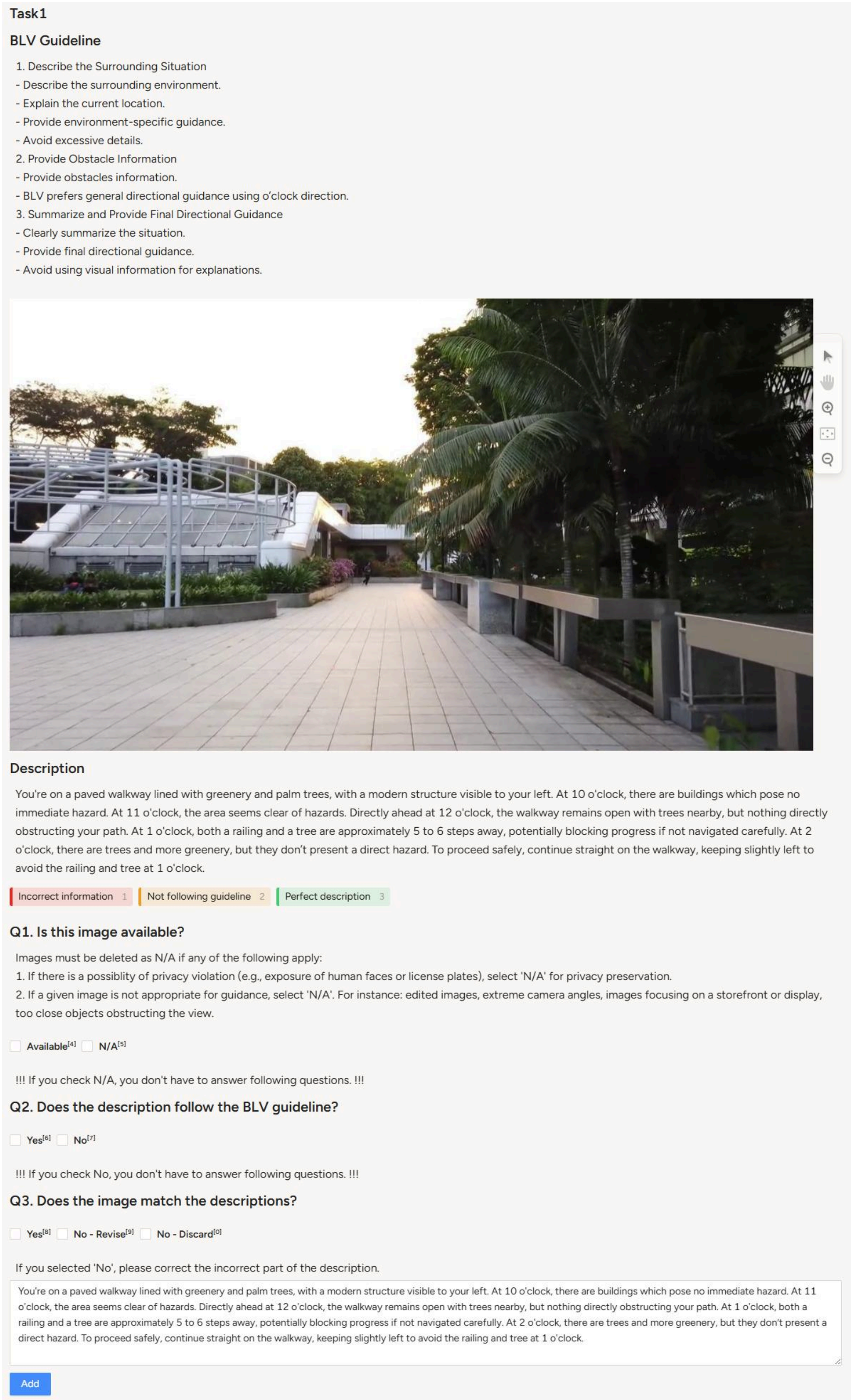


Figure 18: An interface for evaluating and refining silver labels into gold labels.

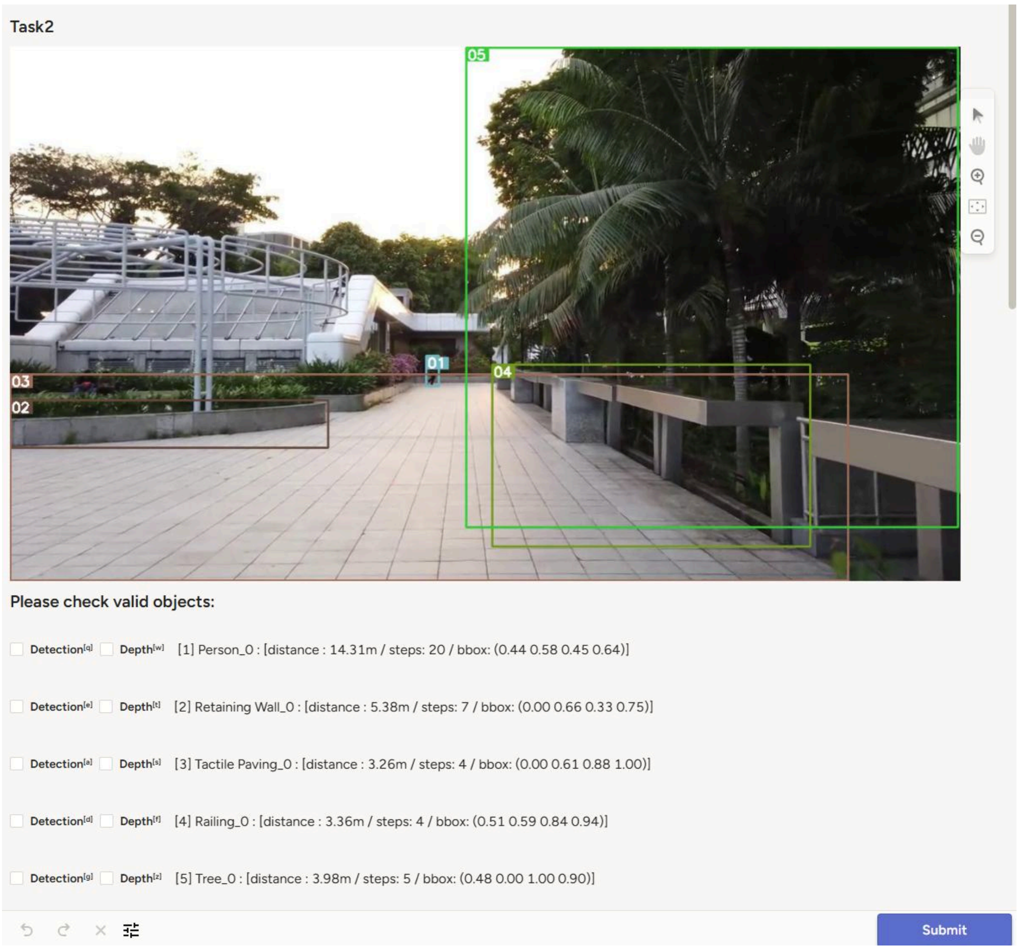
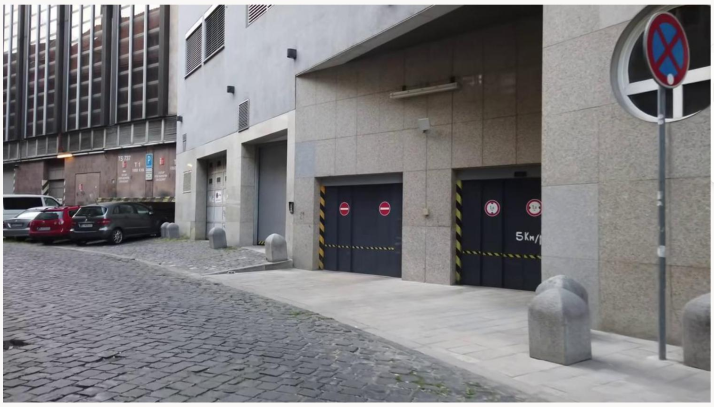


Figure 19: An interface for validating object detection and depth estimation.



Ground Truth

You are standing on a cobblestone street with parked cars on your left at 10 o'clock and a building with garage entrances on your right at 1 o'clock. There are no significant hazards to note at 10, 11, 12, and 2 o'clock; however, at 1 o'clock, about 6 steps ahead, there's a barrier stump that could cause tripping. To safely navigate, keep to your left towards the parked cars and away from the barriers on your right.

Model generated description

The person is standing near a building with large garage doors, surrounded by parked cars and a street sign. The building has a modern design with large windows and a concrete facade. To the left, there's a bus stop with a red sign indicating parking restrictions. The person is facing the garage doors, which are closed. The sidewalk is made of cobblestones, and there are concrete bollards along the edge. The person should be cautious of the parked cars and the garage doors when navigating around this area.

Q1. Is the generated text describing the surrounding situation?

- Describe the surrounding environment.
- Explain the current location.
- Provide environment-specific guidance.
- Avoid excessive details.

1. Poor[1] 2[2] 3[3] 4[4] 5. Perfect[5]

Q2. Is the generated text providing obstacle information?

- Provide obstacles information.
- BLV prefers general directional guidance using o'clock direction.

1. Poor[6] 2[7] 3[8] 4[9] 5. Perfect[0]

Q3. Is the generated text providing final directional guidance?

- Clearly summarize the situation.
- Provide final directional guidance.
- Avoid using visual information for explanations.

1. Poor[q] 2[w] 3[e] 4[t] 5. Perfect[a]

Figure 20: An interface for human evaluation of model-generated descriptions.