EgoBlind: Towards Egocentric Visual Assistance for the Blind

Junbin Xiao[1]*, **Nanxin Huang**[2*], **Hao Qiu**[2], **Zhulin Tao**[2†], **Xun Yang**[3], **Richang Hong**[4], **Meng Wang**[4], **Angela Yao**[1†]
[1] National University of Singapore, [2] Communication University of China
[3] University of Science and Technology of China, [4] Hefei University of Technoloy

Abstract

We present EgoBlind, the first egocentric VideoQA dataset collected from blind individuals to evaluate the assistive capabilities of contemporary multimodal large language models (MLLMs). EgoBlind comprises 1,392 first-person videos from the daily lives of blind and visually impaired individuals. It also features 5,311 questions directly posed or verified by the blind to reflect their in-situation needs for visual assistance. Each question has an average of 3 manually annotated reference answers to reduce subjectiveness. Using EgoBlind, we comprehensively evaluate 16 advanced MLLMs and find that all models struggle. The best performers achieve an accuracy near 60%, which is far behind human performance of 87.4%. To guide future advancements, we identify and summarize major limitations of existing MLLMs in egocentric visual assistance for the blind and explore heuristic solutions for improvement. With these efforts, we hope that EgoBlind will serve as a foundation for developing effective AI assistants to enhance the independence of the blind and visually impaired. Data and code are available at `https://github.com/doc-doc/EgoBlind`.

1 Introduction

The rapid advancement of multimodal large language models (MLLMs) [1–6] has significantly improved the performance of visual question answering (VQA). However, existing VQA datasets primarily focus on the third-person perspective [7, 8] or general-purpose image and video understanding [9–14]. Applications such as visual assistance for the visually-impaired [15], who constitute an overwhelming 2.2 billion people around the world [16] have been received less attention. Research in assisting the blind from a first-person perspective is especially scarce [17].

In light of this, we construct EgoBlind – the first egocentric VideoQA dataset designed to benchmark and advance MLLMs towards egocentric visual assistance for the blind. EgoBlind comprises 1,392 egocentric videos that capture the visual experiences of blind individuals and 5,311 questions that are directly posed or automatically generated and verified by blind users to reflect their assistive needs in exploring their surroundings. To categorize these needs for better analysis, the questions are grouped into six key types: "Information Reading", "Safety Warning", "Navigation","Social Communication", "Tool Use", and "Other Resources". We set the QA task as online (timestamp restricted) and open-form answer generation to better align with its live assistance nature. Multiple reference answers with well-aligned evaluation prompts are also provided for effective assessment.

Our focus on blind users' needs for egocentric visual assistance allows us to explore several key challenges in video-language learning, including egocentric dynamic scene understanding with

*Equal Contribution.
†Corresponding Authors.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

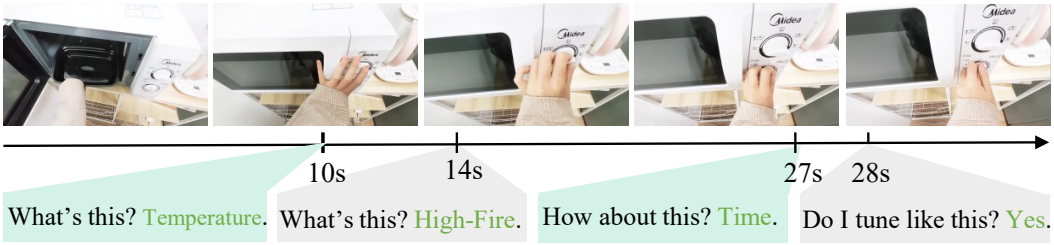


Figure 1: Example from EgoBlind about a blind user demonstrating egocentric visual assistance. As she places her hands on various microwave dials, she asks a series of questions about what the dial controls, its position and settings and how to adjust it.

poor visual quality (*e.g.*, unstable motion, object blur, and occlusions), in-situation user intention reasoning, assistive and reliable answer generation. Take Figure 1 as an example; engaging in the user's activity from a first-person perspective is crucial to understanding and answering the questions. This specification requires capturing the gaze area, hand motions and reasoning about related visual content spatio-temporally. Finally, this should be achieved with poor quality visual inputs, as the regions of interest may be out of focus or occluded (*e.g.*, the control panel in Figure 1).

With EgoBlind, we benchmark 16 recent MLLMs, covering the advanced open-source models (*e.g.*, InternVL2.5 [18], Qwen2.5-VL [4]) and the closed-source ones (*e.g.*, GPT-4o [5] and Gemini [6, 19]). For open-source models, we consider models that 1) achieve the state-of-the-art (SOTA) on common video QA benchmarks [20, 8, 21, 22], and egocentric understanding benchmarks [11, 10, 23]. Additionally, we include 3 models (Video-LLaVA [24] and LLaMA-VID [25], VILA1.5 [26]) that achieve SOTA on the blind image QA dataset VizWiz [15].

Our experimental results show that all models struggle on EgoBlind. The best performer (GPT-4o) achieves an accuracy of 59.3% and falls behind human performance of 87.4% by a whopping $\sim$28%. Interestingly, close-sourced models (*e.g.*, Gemini 1.5 and 2.0), which often surpass open-source models in general-purpose VQA [8, 11, 22] perform worse than top-performing open-source models (*e.g.*, InternVL2.5). Also, models superior on egocentric VQA are not necessarily the best-performing. Our further analyses and investigations lead to the following primary observations:

- No single model wins across all assistance types, and most models are poor in egocentric navigation, safety warning, and communication, indicating research deficiency in these areas.
- Models may correctly answer questions about the visual scene, though the answers may fail to meet users' needs for assistance, indicating a weakness in reasoning about user intentions.
- Models struggle in reasoning the change of spatial orientation relative to the users from the sequence of egocentric visual inputs.
- Models, especially the open-source ones, are sycophantic [27]; they hallucinate wrong and potentially malicious answers when the users' questions deviate the visual facts due to blindness.
- Finetuning with EgoBlind training data effectively benefits model performances, though a large gap remains compared to human capabilities.

Our work tackles significant challenges to construct the first VideoQA dataset (EgoBlind) to benchmark and promote research toward egocentric visual assistance for the blind. We have comprehensively analyzed the behaviors of the leading MLLMs, revealing limitations and areas for improvements. With these efforts, we hope that EgoBlind will serve as a foundation for developing effective AI assistants to enhance the independence of the blind and visually impaired.

2 Related Work

VQA for the Blind. Visual question answering (VQA) is the task of answering user questions about images and videos [28–30]. One promising application area is to support the visually impaired. Yet, the majority of advancements have been made in general-purpose settings [7, 31, 8, 20] with images and videos captured from third-person perspectives. The closest in aim is VizWiz [15], which collects

data from real human-powered VQA systems for helping blind users (*e.g.*, BeMyEye[3]). VizWiz opens up the potential for visual assistance of the blind, but is limited in handling static images and object-centric questions; it fails to cater to the real-time and broader assistive needs of blind users in exploring the dynamic surroundings. Recent work VIEW-QA [17] captures the daily challenges faced by visually impaired individuals via using 360-degree egocentric wearable cameras. However, its videos and questions are collected from seeing actors who simulate the experiences of the blind, and thus may not reflect the daily lives and true needs of the blind for visual assistance, especially for those who have been blind for a long time.

Egocentric VQA. Egocentric VQA has gained interest for its application value towards embodied assistance [32, 29]. Early effort EgoVQA [33] is small scale and limits its questions in challenging action recognition from first-person perspective with the absence of the camera wearer in the footage. Subsequent advancement EgoTaskQA [9] uses machine-generated questions to evaluate models' task understanding capabilities. AssistQ [34] is close to our aim for embodied assistance, but it limits to instructional videos for demonstration of tool use. Ever since the release of Ego4D [35], lots of real-world egocentric QA tasks are explored [36–38, 11, 39, 40, 10, 23]. However, they mostly aim at general-purpose egocentric visual understanding, or focus on a specific aspect of assistance, *e.g.*, QAEgo4D [38] for episodic memory, EgoTextVQA [41] for scene text understanding and EgoLifeQA [42] for long context life assistants. To our best knowledge, there is so far no existing egocentric VQA dataset specially collected from real blind individuals.

MLLMs for VQA. By integrating visual information into powerful LLMs through lightweight connection modules (*e.g.*, MLP [43]), MLLMs extend the capabilities of LLMs to converse with images and videos naturally like humans. MLLMs have significantly improved the landscape of VQA. This brings breakthrough over traditional VQA which answers questions by either multi-choice selection [44, 8] or close-vocabulary classification [31, 20]. While most MLLMs are developed for VQA from third-person views [29, 21, 45, 24, 46, 25, 47, 4, 48], recent advancements [36, 37, 49, 50, 18, 23] specifically add egocentric videos for understanding. Nonetheless, all of they target a general-purpose egocentric visual understanding. In this paper, we will comprehensively examine their capabilities towards egocentric visual assistance for the blind users.

3 EgoBlind Dataset

3.1 Video Collection

We collect videos from video sharing platforms such as Bilibili and TikTok and design an annotation pipeline as illustrated in Figure 2. Specifically, we download 478 long-form egocentric videos (in batches) of blind and visually impaired content creators. These videos, captured using GoPro or mobile phones, document their daily lives while traveling, cooking, navigating through town, shopping, and in social gatherings. It provides near-real-world data on how visually impaired individuals perform daily tasks and solve problems in dynamic environments. Since a single long video often contains an array of visual content, *e.g.*, a summarized record of a visually impaired creator's week, we manually partition the videos into distinct segments based on the source edits. We exclude video moments with sharp scene transitions or strong staging and production indications (*e.g.*, large subtitles, visual special effects, etc.) to focus primarily on life-logging. After partioning and filtering, we obtain 1,329 video clips with an average duration of 40 seconds.

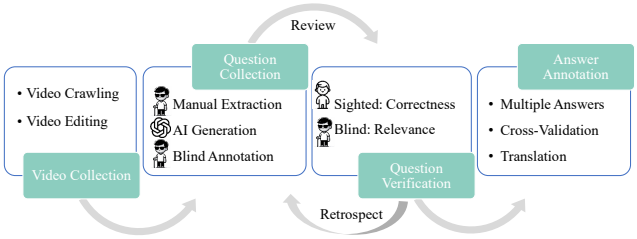


Figure 2: Data annotation pipeline.

3.2 Question Collection

We obtain the questions via three different approaches: 1) **Manual Extraction**: We extract visually-assistive questions posed by the blind content creators in their videos. 2) **AI-Generation**: We prompt GPT-4o [5] to act as a blind user and generate questions by engaging in their egocentric perspective.

[3]http://www.bemyeyes.org/

Table 1: QA examples categorized by different assistance needs.

Category	#Q / #V	QA Examples (Multiple reference answers)
Information Reading	2,464 / 1,040	Q: What floor is the elevator currently on? A: 1st floor / At this time on the first floor.
Safety Warnings	1,260 / 686	Q: Is it safe to cross the street now? A: No. / No, there are cars on both sides. / A car is about to pass through the street, it is recommended to wait a little.
Navigation	751 / 438	Q: Where is the entrance to the building? A: Just a few steps ahead of you. / Directly in front.
Social Communication	153 / 131	Q: Who is the person talking to me? A: He is a delivery guy. / Delivery man.
Tool Use	288 / 197	Q: How do I turn on the stove? A: Turn the knob on the front of the stove clockwise until you feel a click. / By clock in the switch button on the stove. / Rotate the button below. / Turn the switch to the left.
Other Resources	395 / 280	Q: Is there anyone nearby that I can ask for directions? A: Yes. / Yes, there is a person sitting in the front. / Yes, there is a security guard ahead. / Yes, there is front-desk security.

The generated questions are further verified and edited by both seeing and blind people, ensuring their correctness and alignment with blind users' true needs. 3) **Blind User Annotation**: For some videos, we describe the contents to blind annotators and ask them pose assistive questions as if they were in such a setting. The actual annotation is done together with the verification stage for AI generated questions. Our collection details are as follows:

Manual Extraction We first manually watch the video clips and extract the vision-assistive questions from the camera wearers, the timestamps of the questions are also recorded. It is worth noting that these in-video questions are answered by the blind users' sighted partners or others in the videos. Thus, to prevent answer leakage and ensure real-time QA, models can only access visual content up to the question timestamps to answer a specific question. We obtain a total of 541 questions during this process since qualified questions are sparse in the videos.

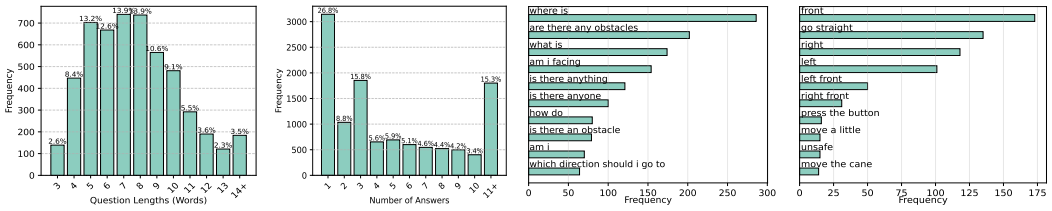
AI-Generation. To enrich the annotations, we consider generating questions by prompting GPT-4o [5] to act as blind questioners, followed by rigorous verification with real blind individuals. We categorize the assistance scenarios into six groups based on our observation of video contents and the extracted questions: *Information Reading*, *Safety Warning*, *Navigation*, *Social Communication*, *Tool Use*, and *Other Resources*. Typical questions from each group is listed in Table 1. To ensure a diverse set of examples for each assistance type, the generation is conducted categorically with tailored prompts for each question type. In addition, we also prompt GPT to generate a reference answer for each question.

In our prompts, the generated QA pairs are ensured to be context-aware and aligned with the needs of the blind: 1) **Ego-centric**: All questions should be framed from the perspective of the visually impaired individual (camera wearers). 2) **Practical**: The questions must reflect practical needs of the blind in the given situation. 3) **Video-Level Questions**: Questions should reflect temporal events in the video, emphasizing dynamic characteristics and requiring multiple frames to answer. 4) **Online Context**: Questions must pertain to the online context of the video and exclude any external information beyond the visible content up to the question time stamp. Other details are presented in the Appendix A.1.

3.3 Manual Verification and Answer Annotation

Our **verification** is done in three stages involving both seeing and visually impaired people. The first three authors are engaged in all stages for quality control. Concretely, **in the first stage**, the authors check the QA quality after generating a limited number of questions for each assistance type. Related issues are recorded and reflected in the updated prompts for subsequent generations to improve the generation quality. In such an alternate way, we generate 16,560 questions in batches.

In the second stage, we invite 63 volunteers seeing individuals to thoroughly review the questions. In this process, we remove questions that are 1) redundant in meaning; 2) vague and difficult to answer



(a) Question length. (b) Answer length. (c) Top-10 questions. (d) Top-10 answers*.

(e) Word cloud of questions. (f) Word cloud of answers*. (g) Distribution of video durations.

Figure 3: Statistic analysis of EgoBlind. *: We omit 1,123 (22.6%) "yes/no" and 541 (11.0%) "do not know" answers in (d) and (f) for better presentation. (Please zoom in for better view.)

directly (*e.g.*, "How do I use my phone?"); and 3) meaningless (*e.g.*, "How do I use my white cane?"). We also remove a moderate number of questions that are 1) not related to visual inputs (*e.g.*, "How heavy is this box?") and 2) not engaged in first-person perspective (*e.g.*, "what is the man (camera wearer) doing?"). Note that all ambiguous questions are further cross-validated by the three authors to either remove or edit them. After this stage, 5,080 questions remained; more than 80% of them have been human-edited by this point. Additionally, the 541 extracted questions also undergo a check for language translation errors, resulting in a total of 5,621 valid questions after this stage.

In the third stage, we sample 30% (1.5K) of the questions and invite 111 blind volunteers to review them for relevance to blind users' needs for visual assistance. The sampling are visual situation based to cover all visual situations in EgoBlind (see details in Appendix A.2). Specifically, the blind volunteers are requested to score (from the lowest 0 to the highest 5) the confidence that a question would be asked under a specific visual scenario. Each question is scored by 10 to 33 volunteers. Additionally, the bind volunteers are encouraged to pose additional new questions. Afterwards, we remove the questions with an average confidence score lower than 1, and add approximately 300 new questions directly from the blind volunteers (**Blind User Annotation**). The result observations (what question to be deleted or added) are then reflected throughout the whole dataset by the authors. Finally, we obtain 5,311 valid questions. Their distributions are listed in Table 1.

Multiple Answer Annotation. For each question, we annotate multiple ground-truth answers by inviting 21 university students to watch the related video. To enhance the persuasiveness of the standard answers and reduce subjectivity, each question is assigned to three or four annotators for answering. Only one may see the generated answer (for generated questions only) and must verify and edit it as needed. The others are required to directly answer the question for assistance purpose. If there is no generated answer, *e.g.*, for the manually extracted or annotated question, the annotators need to answer the question themselves. We resolve contradictory answers through consensus between the authors and the majority of annotators. Redundant answers are avoided by merging similar responses. Ultimately, each question has 1 to 4 valid answers (3 on average). Over 67% of the generated answers are edited by human annotators, and most unedited answers are deterministic, such as yes/no and numbers. Examples of answers are shown in Table 1.

3.4 Statistic Analysis

Data Split. The data collection and annotation process takes about 10 months. We finalize 5,311 questions across 1,329 videos. For effective and efficient evaluation, we split half of the videos (656) along with their QA pairs (2,565) as a test set, while leaving the remaining for instruction tuning towards model development. The split ensures that video clips from the same source video do not

Table 2: Statistics of the EgoBlind datasets

Videos			Questions		
Train	Test	All	Train	Test	All
673	656	1,329	2,746	2,565	5,311

Table 3: Ego-VQA dataset comparison. BV: Video captured by the blind. BQ: Question posed or verified by the blind. RT: Answer based on the video content up to the question timestamp. MA: Multiple answers for each question. QC: Question Category. OE/MC: Open-Ended/Multi-Choice.

	Datasets	# V	# Q	VLen(s)	BV	BQ	RT	MA	QC	Task
General Purpose	EgoTaskQA [9]	2K	40K	25.0	✗	✗	✗	✗	✓	OE
	VidEgoThink [10]	217	600	23.4	✗	✗	✗	✗	✓	OE
	EgoSchema [11]	5K	5K	180.0	✗	✗	✗	✗	✗	MC
	EgoMemoria [23]	629	7K	858.5	✗	✗	✗	✗	✗	MC
Assist	QAEgo4D [38]	1.3K	14.5K	495.1	✗	✗	✗	✗	✗	OE
	AssistQ [34]	100	531	115.0	✗	✗	✗	✗	✓	MC
	EgoTextVQA [41]	1.5K	7.0K	101.7	✗	✗	✓	✗	✓	OE
Blind	VizWiz [15]	-	-	-	✓	✓	✗	✓	✗	OE
	VIEW-QA [17]	1.0K	4.1K	34.4	✗	✗	✗	✗	✓	OE
	EgoBlind (Ours)	1.3K	5.3K	40.0	✓	✓	✓	✓	✓	OE

appear in both sets (we also study a user-specific data-split strategy in Appendix B.3). Additionally, we use Gemini 2.0 to probe if there is a significant discrepancy of model behaviors between the test set and the whole dataset. We find that the prediction accuracy differences are quite small, ranging from 0.2% to 0.5% across different question types. This suggests that the test set is sufficient for reliable evaluation. The detailed numbers are presented in Table 2.

Questions. Figure 3a shows that more than 80% of the questions are shorter than 10 words, with an average length of 7.9 words. The relatively short questions could be attributed to the use of more spoken language in practice. The word cloud in Figure 3e shows that the questions are from blind users' first-person perspective ("I", "me", "my") and feature referential concepts, such as "this", "there", "that", "it" and "now", which requires context-specific interpretation. Figure 3c shows that blind people are often interested in locating something ("where is") and checking for potential safety issues ("are there any obstacles").

Answers. Figure 3b shows that the answers have an average length of 5.7 words. Near half of the answers have more than 3 words, with 19.4% over 10 words. A detailed study shows that the answers for tool use are significantly longer and more complex than other assistive questions; the average answer length is 12.1 words for tool use and 4 to 6 words for others. The most frequent answers are "yes" (13.0%), followed by "I do not know" (10.7%) and "no" (9.6%). The answer of "I do not know" is due to poor visual quality or the answer not framed by the blind users. Additionally, it may be attributed to our online QA task setting, *i.e.*, the answer is not visible up to the question moment. We keep these unanswerable questions to evaluate if models can reject to answer rather than hallucinate potentially malicious answers. For better analysis, we omit these answers and analyze the remaining frequent answers and answer word clouds in Figure 3d and 3f, respectively. We find that the answers are mostly about directions, navigation and locations. It is worth mentioning that the locations and directions are often relative to the camera wearers (questioners).

Videos. Figure 3g shows that the vast majority of the videos (82.3%) are within 1 minute, with 17.7% exceeding 1 minute and 7.1% running longer than 2 minutes. The mean video duration is 40.0 seconds, reflecting the average duration each time a visually-impaired person recording their daily moment. Additionally, we conduct a comparison between the egocentric video content captured by sighted people (*e.g.*, Ego4D [35] videos) and those are blind (examples are presented in Appendix A.3). Differences arise from 1) **Composition and Focus:** Blind individuals often produce footage where subjects are off-center or out of focus (refer to Figure 1 and Appendix A.3). 2) **Camera Orientation and Stability:** Blind individuals struggle to maintain consistent camera orientation, leading to potentially tilted or unstable footage. 3) **Environmental Awareness:** Blind individuals inadvertently capture obstructed views or poorly lit scenes. Nevertheless, the video resolution is higher, possibly because the videos are crawled from content creators on video sharing platforms.

3.5 Dataset Comparison

Unlike most Ego-VQA datasets (the first block of Table 3) which focus on general-purpose visual understanding, EgoBlind emphasizes assisting blind individuals. Compared to other assist-oriented datasets (the middle block), EgoBlind covers the wide aspects of visual assistance for the blind in daily life, versus episodic memory in QAEgo4D [38], tool use in AssistQ [34], manipulation in HoloAssist [51] (proactive assistance), or scene-text reading in EgoTextVQA [41].

Table 4: QA accuracy of different models on EgoBlind.

Methods	LLM	Size	#F	Tool	Info.	Navi.	Safe	Com.	Res.	Overall
Human	-	-	-	70.4	87.0	83.1	91.9	94.7	96.6	87.4
Open-source Models										
ShareGPT4Video [50]	LLaMA3-8B	ori	16	25.5	32.6	20.7	43.3	38.9	28.3	32.9
CogVLM2-Video [54]	LLaMA3-8B	224^2	24	32.2	44.5	14.0	52.7	43.1	32.4	40.3
Video-LLaMA3 [48]	Qwen2.5-7B	ori	1fps	53.0	51.9	38.1	50.6	41.7	50.3	49.2
InternVL2.5-8B [18]	InternLM2_5-7B	448^2	8	61.1	54.6	42.2	58.0	44.4	52.6	53.5
LLaVA-OV [53]	Qwen2-7B	384^2	16	61.1	56.4	29.5	**65.8**	**58.3**	50.9	54.5
InternVL2.5-26B [18]	InternLM2_5-20B	448^2	8	**72.5**	56.9	47.4	54.1	43.1	53.2	55.0
MiniCPM-V 2.6 [56]	Qwen2-7B	384^2	1fps	53.7	46.5	37.8	28.9	37.5	41.0	40.7
Qwen2.5-VL [4]	Qwen2.5-7B	ori	1fps	51.0	50.1	28.2	48.5	43.1	38.2	45.5
LLaVA-Video [55]	Qwen2-7B	384^2	1fps	44.3	53.4	32.6	62.0	50.0	49.7	51.5
Video-LLaVA [21]	Vicuna-7B	224^2	8	22.8	41.2	21.2	47.2	38.9	35.3	38.1
LLaMA-VID [25]	Vicuna-7B	224^2	1fps	32.2	40.5	20.7	49.4	36.1	41.6	39.1
VILA1.5 [26]	LLaMA3-8B	336^2	8	49.7	50.5	25.9	60.6	47.2	41.0	48.2
Closed-source Models										
Gemini 2.0 Flash	-	ori	32	61.1	54.5	50.5	39.1	47.2	49.1	49.9
Gemini 1.5 Flash	-	ori	32	**72.5**	54.4	43.5	50.6	38.9	45.7	51.8
Gemini 2.5 Flash	-	ori	32	67.1	57.6	47.7	57.8	47.2	50.3	56.0
GPT-4o	-	ori	32	66.4	**61.2**	**52.6**	58.8	47.2	**62.4**	**59.3**

Compared to the blind-related QA datasets (the bottom block), EgoBlind advances VizWiz [15] by enabling egocentric live QA with more visual information about real-world dynamics and episodic memory support. Although the videos in the VIEW-QA [17] dataset offer 360-degree visual information, the videos and questions are simulated by seeing actors. In contrast, the videos in EgoBlind are entirely filmed by blind or visually impaired individuals themselves, authentically reflecting their real-life scenarios. Moreover, the questions in EgoBlind are posed or verified by blind individuals of different blind ages, further ensuring the dataset's practicality and relevance. Finally, EgoBlind supports an online QA setting with timestamp annotations for each question; answers are based on the video contents prior to the question timestamp.

4 Experiment

Evaluation. Following popular practices for LLMs [21, 52], we use the GPT score as a metric for evaluating generated answers. Specifically, we prompt GPT-4o mini to assess the semantic similarity between a models' prediction and the ground truth (reference answers), and answer with 'yes' if they are judged as the same. We then obtain the accuracy (0-100%) as the percentage of 'yes' answers evaluated. Noteworthy, for each question, a correct prediction is identified if it achieves *yes*-response with any one of the reference answers. Moreover, the evaluation prompts are manually finetuned to reach a maximal agreement (0.88) between human and AI reviewers (details in Appendix B.1).

Model Setup. To benchmark the challenges carried by EgoBlind, we comprehensively analyze 16 contemporary MLLMs, including 12 open-source models and 4 closed-source ones (via APIs). The choice of the open-source models are based on the following criteria: 1) Achieve SOTA results on common video QA benchmarks such as NExT-QA [8], VideoChatGPT-Bench [21] and Video-MME [22]. Corresponding models are LLaVA-OV [53], CogVLM2-Video [54], Video-LLaMA3 [48], InternVL2.5 [18]; 2) Achieve SOTA results on general-purpose egocentric VideoQA benchmarks such as EgoSchema [11], VidEgoThink [10] and EgoMemoria [23]. Corresponding models are LLaVA-Video [55], MiniCPM-V 2.6 [56] and Qwen2.5-VL[4]; 3) Achieve SOTA on the image blind QA dataset VizWiz [15], such as Video-LLaVA [21], LLaMA-VID[25] and VILA 1.5 [26]. Notably, for each question, we uniformly sample the video content up to the question timestamp for answer prediction, thus to match the live QA task setting. Moreover, we design customized prompts tailored for answering blind users' questions for better performance. Additionally, we obtain human performance by inviting 3 university students who did not participant in annotation.

4.1 Benchmarking Analysis

Table 4 presents the performances of different models. We summarize the following observations: **(1)** None of the models achieves the desired level of performance on EgoBlind, all lagging behind human performance by a whopping 54%$\sim$28%, suggesting significant room for improvements.

Figure 4: Common failure cases of tested MLLMs. The models fail to (a) reason user intention, (b) understand real-time spatial orientation, provide (c) assistive and (d) reliable answers.

(2) The models that are superior at general-purpose egocentric VQA (*e.g.*, LLaVA-Video) and image blind-VQA (*e.g.*, VILA1.5) are not the best-performing, depicting unique challenges of EgoBlind. **(3)** No single model wins across all question types. Answering "Navigation" questions is the most challenging task for almost all models, indicating a significant limitation of MLLMs in this field. **(4)** While most other models struggle in answering questions about tool use, Gemini 1.5 even surpasses human performance, demonstrating its rich knowledge outweighing human individuals. **(5)** Stronger LLMs and larger visual resolution often bring better performance, while more frames do not always help (*e.g.*, 8 frames are enough for InternVL to surpass other open-source models.).

4.2 Assistance-Related Challenges

We further reveal some specific challenges pertaining to egocentric visual assistance for the blind by analyzing the common failure cases.

User Intention. Reasoning about user intentions behind the questions is key for effective assistance. Yet, all models fall short in such a capability. For example, in Figure 4(a) and (c), all models fail to generate helpful answers for the blind, though the answers are objectively correct to the visual contents. It is worth mentioning that this is despite explicitly prompting the models to answer questions of blind individuals to provide visual assistance.

Spatial Orientation Change. MLLMs exhibit significant shortcomings in weaving together the temporal frames to reason the spatial orientations relative to the users' real-time location. A typical example is shown in Figure 4(b), where indoor navigation is requested by the blind user. The escalator was framed prior to the question moment, and the models have to retrieve the related moment and reason the users' orientation change after that moment, which shows extreme challenge to all MLLMs.

Table 5: Single frame inputs.

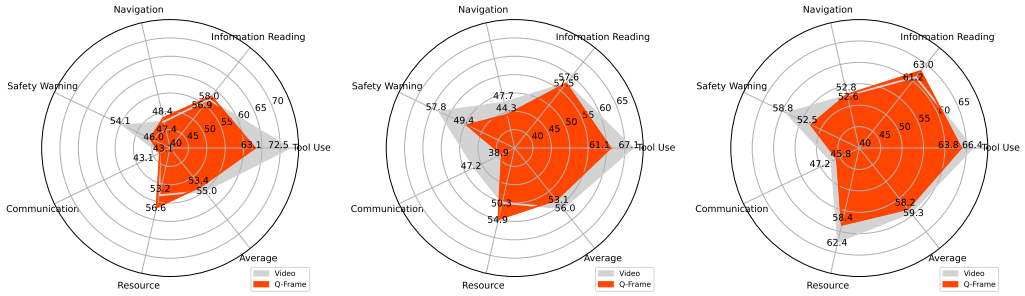
Models	Acc. (bef)	Acc. (aft)
InternVL2.5-26B	55.0	53.4 ↓ 1.6
VILA1.5	48.2	47.2 ↓ 1.0
GPT-4o	59.3	58.2 ↓ 1.1

Table 6: Normal QA prompts.

Models	Acc. (bef)	Acc. (aft)
InternVL2.5-26B	55.0	54.0 ↓ 1.0
Gemini 2.5 Flash	56.0	55.2 ↓ 0.8
GPT-4o	59.3	58.1 ↓ 1.2

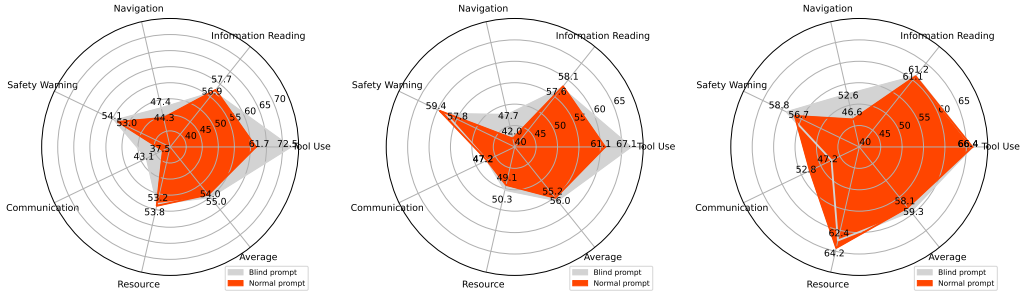
Table 7: Instruction Tuning.

Models	Acc. (bef)	Acc. (aft)
Qwen2.5-VL	45.5	50.2 ↑ 4.7
LLaVA-OV	54.5	57.4 ↑ 2.9
InternVL2.5-8B	53.5	58.1 ↑ 4.6



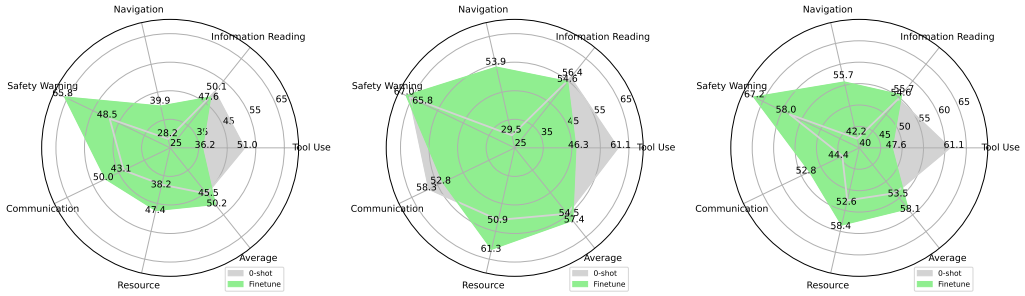
(a) InternVL2.5-26B. (b) Gemini 2.5 Flash. (c) GPT-4o.

Figure 5: Uniform sampling *vs.* single frame (Q-Frame) input at the question timestamp. The overall QA accuracy declines slightly when replacing video with a Q-Frame.



(a) InternVL2.5-26B. (b) Gemini 2.5 Flash. (c) GPT-4o.

Figure 6: Blind aware prompt (Appendix Table 12) *vs.* Normal VQA prompt (Appendix Table 11). The overall accuracy declines without blind-specific prompting.



(a) Qwen2.5-VL. (b) LLaVA-OV. (c) InternVL-2.5-8B.

Figure 7: Simple finetuning with EgoBlind training data can notably boost QA performances, but it is ill-suited for answering "tool-use" questions.

Reliable Answers. Providing reliable answers is of crucial importance in egocentric visual assistance. However, all MLLMs tend to be sycophantic when the users ask something that deviates the visual facts due to blindness. Take Figure 4(d) as an example, the blind user is already on the bridge but ask for navigating it. All models fail to remind the user that he is on the bridge, but either answer "I do not know", or give wrong or even malicious suggestions, such as "take the boat", "walk straight".

We also analyze other challenges related to streaming VQA, scene text recognition, and referential words in spoken languages in Appendix B.2.

4.3 Other Investigations

Our investigations aim to answer three questions: 1) Is a single frame at the question timestamp sufficient to answer the question? 3) Does explicitly prompting models to assist blind users benefit performance? 2) Can instruction tuning with our training data improve performance?

Table 5 shows that replacing multi-frame inputs with a single frame at the question timestamp degenerates model performances, though results remain competitive (which is reasonable as people often ask about their current visual environment.). Figure 5 further suggests that video-level modeling is key for answering questions of "Safety Warning", "Tool Use", and "Communication", while a single frame seems to be sufficient for information reading. Table 6 shows that substituting the blind-specific prompt with a normal VQA prompt jeopardizes model performance, highlighting the unique challenges of blind-oriented QA, especially in navigation and tool using as shown in Figure 6. Table 7 and Figure 7 demonstrate that using our training data for instruction tuning (LoRA [57] finetune) can remarkably improve performances, yet the gap compared with human remains significant (Implementation details are presented in Appendix Sec. B.3). Additionally, Figure 7 shows that existing finetuning will hurt the performances on "tool use" questions, likely because of overfitting on the limited training data for this category.

5 Conclusion

We present the construction of a VideoQA dataset EgoBlind to reflect on the challenges towards egocentric visual assistance for the blind. EgoBlind is the first of its kind in that the videos are taken by the visually-impaired individuals from a first-person perspective, with questions posed and verified by blind users to ensure close alignment with their true assistant needs. We further comprehensively benchmark the challenge with multiple prominent multimodal LLMs and unveil their significant limitations in various aspects. We conduct thorough analysis and share many key insights for future advancements in this direction. By formulating and bringing the challenge to the vision-language community, our primary goal is to push the MLLM research towards live egocentric visual assistance for the overwhelming number of visually impaired people around the world.

Importantly, we have ensured that the dataset collection and user studies adhere to IRB standards. Informed consent was obtained from all the content creators, who agreed to their content being used for non-commercial research purposes. Access to the dataset will be granted with the condition that video sources will be cited for distribution.

References

[1] 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," *Advances in neural information processing systems*, vol. 35, pp. 23716–23736, 2022.

[2] 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 *International conference on machine learning*, pp. 19730–19742, PMLR, 2023.

[3] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, 2024.

[4] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, *et al.*, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *arXiv preprint arXiv:2409.12191*, 2024.

[5] OpenAI, A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, *et al.*, "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024.

[6] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[7] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the v in vqa matter: Elevating the role of image understanding in visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

[8] J. Xiao, X. Shang, A. Yao, and T.-S. Chua, "Next-qa: Next phase of question-answering to explaining temporal actions," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9777–9786, 2021.

[9] B. Jia, T. Lei, S.-C. Zhu, and S. Huang, "Egotaskqa: Understanding human tasks in egocentric videos," *Advances in Neural Information Processing Systems*, vol. 35, pp. 3343–3360, 2022.

[10] S. Cheng, K. Fang, Y. Yu, S. Zhou, B. Li, Y. Tian, T. Li, L. Han, and Y. Liu, "Videgothink: Assessing egocentric video understanding capabilities for embodied ai," *arXiv preprint arXiv:2410.11623*, 2024.

[11] K. Mangalam, R. Akshulakov, and J. Malik, "Egoschema: A diagnostic benchmark for very long-form video language understanding," in *NeurIPS*, pp. 46212–46244, 2023.

[12] X. Yang, F. Feng, W. Ji, M. Wang, and T.-S. Chua, "Deconfounded video moment retrieval with causal intervention," in *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pp. 1–10, 2021.

[13] X. Yang, S. Wang, J. Dong, J. Dong, M. Wang, and T.-S. Chua, "Video moment retrieval with cross-modal neural architecture search," *IEEE Transactions on Image Processing*, vol. 31, pp. 1204–1216, 2022.

[14] X. Shang, D. Di, J. Xiao, Y. Cao, X. Yang, and T.-S. Chua, "Annotating objects and relations in user-generated videos," in *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pp. 279–287, 2019.

[15] D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J. P. Bigham, "Vizwiz grand challenge: Answering visual questions from blind people," in *CVPR*, pp. 3608–3617, 2018.

[16] G. W. H. Organization, "World report on vision." `https://www.who.int/publications/i/item/world-report-on-vision`, 2019.

[17] I. Song, M. Joo, J. Kwon, and J. Lee, "Video question answering for people with visual impairments using an egocentric 360-degree camera," *arXiv preprint arXiv:2405.19794*, 2024.

[18] Z. Chen, W. Wang, H. Tian, S. Ye, Z. Gao, E. Cui, W. Tong, K. Hu, J. Luo, Z. Ma, *et al.*, "How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites," *Science China Information Sciences*, vol. 67, no. 12, p. 220101, 2024.

[19] G. Comanici, E. Bieber, M. Schaekermann, I. Pasupat, N. Sachdeva, I. Dhillon, M. Blistein, O. Ram, D. Zhang, E. Rosen, *et al.*, "Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities," *arXiv preprint arXiv:2507.06261*, 2025.

[20] Z. Yu, D. Xu, J. Yu, T. Yu, Z. Zhao, Y. Zhuang, and D. Tao, "Activitynet-qa: A dataset for understanding complex web videos via question answering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 9127–9134, 2019.

[21] M. Maaz, H. Rasheed, S. Khan, and F. Khan, "Video-chatgpt: Towards detailed video understanding via large vision and language models," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12585–12602, 2024.

[22] C. Fu, Y. Dai, Y. Luo, L. Li, S. Ren, R. Zhang, Z. Wang, C. Zhou, Y. Shen, M. Zhang, *et al.*, "Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis," *arXiv preprint arXiv:2405.21075*, 2024.

[23] H. Ye, H. Zhang, E. Daxberger, L. Chen, Z. Lin, Y. Li, B. Zhang, H. You, D. Xu, Z. Gan, *et al.*, "Mm-ego: Towards building egocentric multimodal llms," *ICLR*, 2025.

[24] B. Lin, Y. Ye, B. Zhu, J. Cui, M. Ning, P. Jin, and L. Yuan, "Video-llava: Learning united visual representation by alignment before projection," *arXiv preprint arXiv:2311.10122*, 2023.

[25] Y. Li, C. Wang, and J. Jia, "Llama-vid: An image is worth 2 tokens in large language models," in *European Conference on Computer Vision*, pp. 323–340, Springer, 2024.

[26] J. Lin, H. Yin, W. Ping, P. Molchanov, M. Shoeybi, and S. Han, "Vila: On pre-training for visual language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26689–26699, 2024.

[27] Y. Zhao, R. Zhang, J. Xiao, C. Ke, R. Hou, Y. Hao, Q. Guo, and Y. Chen, "Towards analyzing and mitigating sycophancy in large vision-language models," *arXiv preprint arXiv:2408.11261*, 2024.

[28] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

[29] J. Xiao, N. Huang, H. Qin, D. Li, Y. Li, F. Zhu, Z. Tao, J. Yu, L. Lin, T.-S. Chua, and A. Yao, "Videoqa in the era of llms: An empirical study," *IJCV*, 2025.

[30] X. Yang, J. Zeng, D. Guo, S. Wang, J. Dong, and M. Wang, "Robust video question answering via contrastive cross-modality representation learning," *Science China Information Sciences*, vol. 67, no. 10, p. 202104, 2024.

[31] D. Xu, Z. Zhao, J. Xiao, F. Wu, H. Zhang, X. He, and Y. Zhuang, "Video question answering via gradually refined attention over appearance and motion," in *Proceedings of the 25th ACM international conference on Multimedia*, pp. 1645–1653, 2017.

[32] C. Plizzari, G. Goletto, A. Furnari, S. Bansal, F. Ragusa, G. M. Farinella, D. Damen, and T. Tommasi, "An outlook into the future of egocentric vision," *International Journal of Computer Vision*, pp. 1–57, 2024.

[33] C. Fan, "Egovqa-an egocentric video question answering benchmark dataset," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 0–0, 2019.

[34] B. Wong, J. Chen, Y. Wu, S. W. Lei, D. Mao, D. Gao, and M. Z. Shou, "Assistq: Affordance-centric question-driven task completion for egocentric assistant," in *ECCV*, pp. 485–501, Springer, 2022.

[35] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, *et al.*, "Ego4d: Around the world in 3,000 hours of egocentric video," in *CVPR*, pp. 18995–19012, 2022.

[36] K. Q. Lin, J. Wang, M. Soldan, M. Wray, R. Yan, E. Z. Xu, D. Gao, R.-C. Tu, W. Zhao, W. Kong, *et al.*, "Egocentric video-language pretraining," *Advances in Neural Information Processing Systems*, vol. 35, pp. 7575–7586, 2022.

[37] S. Pramanick, Y. Song, S. Nag, K. Q. Lin, H. Shah, M. Z. Shou, R. Chellappa, and P. Zhang, "Egovlpv2: Egocentric video-language pre-training with fusion in the backbone," in *ICCV*, pp. 5285–5297, 2023.

[38] L. Bärmann and A. Waibel, "Where did i leave my keys? - episodic-memory-based question answering on egocentric videos," in *CVPR Workshops*, pp. 1560–1568, 2022.

[39] S. Di and W. Xie, "Grounded question-answering in long egocentric videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12934–12943, 2024.

[40] S. Cheng, Z. Guo, J. Wu, K. Fang, P. Li, H. Liu, and Y. Liu, "Egothink: Evaluating first-person perspective thinking capability of vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14291–14302, June 2024.

[41] S. Zhou, J. Xiao, Q. Li, Y. Li, X. Yang, D. Guo, M. Wang, T.-S. Chua, and A. Yao, "Egotextvqa: Towards egocentric scene-text aware video question answering," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3363–3373, 2025.

[42] J. Yang, S. Liu, H. Guo, Y. Dong, X. Zhang, S. Zhang, P. Wang, Z. Zhou, B. Xie, Z. Wang, *et al.*, "Egolife: Towards egocentric life assistant," *arXiv preprint arXiv:2503.03803*, 2025.

[43] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024.

[44] Y. Jang, Y. Song, Y. Yu, Y. Kim, and G. Kim, "Tgif-qa: Toward spatio-temporal reasoning in visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2758–2766, 2017.

[45] H. Zhang, X. Li, and L. Bing, "Video-llama: An instruction-tuned audio-visual language model for video understanding," *arXiv preprint arXiv:2306.02858*, 2023.

[46] J. Xiao, A. Yao, Y. Li, and T.-S. Chua, "Can i trust your answer? visually grounded video question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13204–13214, 2024.

[47] Z. Cheng, S. Leng, H. Zhang, Y. Xin, X. Li, G. Chen, Y. Zhu, W. Zhang, Z. Luo, D. Zhao, *et al.*, "Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms," *arXiv preprint arXiv:2406.07476*, 2024.

[48] B. Zhang, K. Li, Z. Cheng, Z. Hu, Y. Yuan, G. Chen, S. Leng, Y. Jiang, H. Zhang, X. Li, *et al.*, "Videollama 3: Frontier multimodal foundation models for image and video understanding," *arXiv preprint arXiv:2501.13106*, 2025.

[49] Y. Zhao, I. Misra, P. Krähenbühl, and R. Girdhar, "Learning video representations from large language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6586–6597, 2023.

[50] L. Chen, X. Wei, J. Li, X. Dong, P. Zhang, Y. Zang, Z. Chen, H. Duan, B. Lin, Z. Tang, *et al.*, "Sharegpt4video: Improving video understanding and generation with better captions," *arXiv preprint arXiv:2406.04325*, 2024.

[51] X. Wang, T. Kwon, M. Rad, B. Pan, I. Chakraborty, S. Andrist, D. Bohus, A. Feniello, B. Tekin, F. V. Frujeri, *et al.*, "Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20270–20281, 2023.

[52] E. Song, W. Chai, G. Wang, Y. Zhang, H. Zhou, F. Wu, H. Chi, X. Guo, T. Ye, Y. Zhang, *et al.*, "Moviechat: From dense token to sparse memory for long video understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18221–18232, 2024.

[53] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu, *et al.*, "Llava-onevision: Easy visual task transfer," *arXiv preprint arXiv:2408.03326*, 2024.

[54] W. Hong, W. Wang, M. Ding, W. Yu, Q. Lv, Y. Wang, Y. Cheng, S. Huang, J. Ji, Z. Xue, *et al.*, "Cogvlm2: Visual language models for image and video understanding," *arXiv preprint arXiv:2408.16500*, 2024.

[55] Y. Zhang, J. Wu, W. Li, B. Li, Z. Ma, Z. Liu, and C. Li, "Video instruction tuning with synthetic data," *arXiv preprint arXiv:2410.02713*, 2024.

[56] Y. Yao, T. Yu, A. Zhang, C. Wang, J. Cui, H. Zhu, T. Cai, H. Li, W. Zhao, Z. He, *et al.*, "Minicpm-v: A gpt-4v level mllm on your phone," *arXiv preprint arXiv:2408.01800*, 2024.

[57] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, *et al.*, "Lora: Low-rank adaptation of large language models.," *ICLR*, vol. 1, no. 2, p. 3, 2022.

[58] M. M. L., "Interrater reliability: the kappa statistic," *Biochemia medica*, pp. 276—-282, 2012.

Appendix

A EgoBlind Dataset Creation

A.1 QA Generation with GPT-4o

To enrich the QA diversity and reduce the annotation burden, we prompt GPT-4o to act as blind users to generate part of questions and invite both blind and sighted people to check and edit them. For generation, we first decode the video into frames at 3 fps. We then adopt an adaptive sampling strategy over the decoded frames to obtain the final frames to be fed to GPT-4o for QA generation. We sample at a ratio of every 18 frames for long videos (*e.g.*, 180s) and 6 frames for short ones (*e.g.*, 30s). To simulate real-time QA generation, each frame is accompanied by its timestamp and a refined prompt (shown in Table 15). The generation process is divided into batches by question categories. This means that each batch contains only one type of question, which can help in organizing and enriching the dataset, ensuring a wide coverage of different topics and improving overall quality. The specific prompts for each category are attached in Table 15.

A.2 Blind User Study

We categorize manually corrected QA pairs (translated into Chinese) based on video scenes and invite 111 blind individuals to evaluate the **blind-relevance** of the questions across different scenarios. The evaluation scale ranges from 0 to 5, with higher scores indicating preferable questions and lower scores reflecting unwanted questions (as illustrated in Figure 8, a real example is shown in Figure 18). Additionally, blind individuals are encouraged to provide further questions and insights to better meet the needs of individuals with visual impairments.

The cohort study (Figure 9a) exhibits substantial demographic heterogeneity, encompassing participants ranging from teenagers to older adults (Figure 9b). Male participants account for a dominant proportion of 73.3%, with young adults aged 18-30 constituting the largest age cohort. Notably, for most respondents, the duration of visual impairment corresponded with their biological age, suggesting congenital origin - a finding consistent with the observation that over 80% presented Grade 1 Blindness (total visual acuity $\leqslant$0.02 or visual field $< 5°$). The remaining participants acquire visual impairment postnatally, though all reported living with the condition for long time.

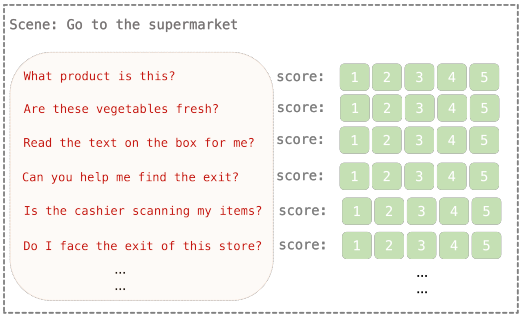
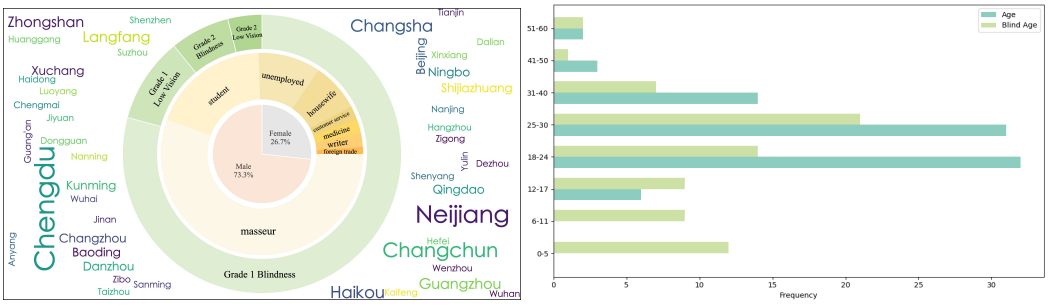


Figure 8: The scoring example



(a) Location distribution.

(b) Age distribution.

Figure 9: Distribution of blind participants

The occupational distribution reveals a pronounced concentration in specialized massage services (55.6%), reflecting a predominant vocational pattern among the visually impaired population of China. Students form the secondary occupational group (18.9%), followed by unemployed individuals (10.0%) and other professions (15.5%). Geographically, the sample represents 46 urban centers across China's hierarchical city-tier system, ranging from first-tier megacities (Beijing, Guangzhou,

Shenzhen) to emerging technology hubs (Hangzhou, Chengdu, Changsha) and smaller urban cities (Neijiang, Zhongshan, Langfang).

This demographic and geographic diversity ensures comprehensive coverage of life scenes across varying urban contexts, from high-density metropolitan networks to community-level systems.

We found that when blind people visit supermarkets, stores, and shopping malls (as shown in Figure 10); their primary concerns are navigation and positioning issues related to location. Specifically, they want to know where they are and how to get to their desired destination("How to get to XXX from here?") . Their second main concern is obtaining product information, such as price ("Where can I find the price of this product?"), style, shelf life, etc. Additionally, some blind individuals are also interested in understanding the range of products sold at specific stores and the layout of the mall, including questions like "What does this store sell?" and "Which products are displayed on the counter?" At the same time, we found that blind individuals tend to be less concerned about the presence and location of people around them. Questions related to the number of people nearby ("How many people are in front of me?") and their specific whereabouts ("Is the cashier directly in front of me?") received a significantly higher proportion of low scores compared to high scores. Additionally, they showed minimal interest in information about road obstacles. The proportion of low scores for questions such as "What are the obstacles on the road?" and "How many floors are there?" was considerably higher than the proportion of high scores.

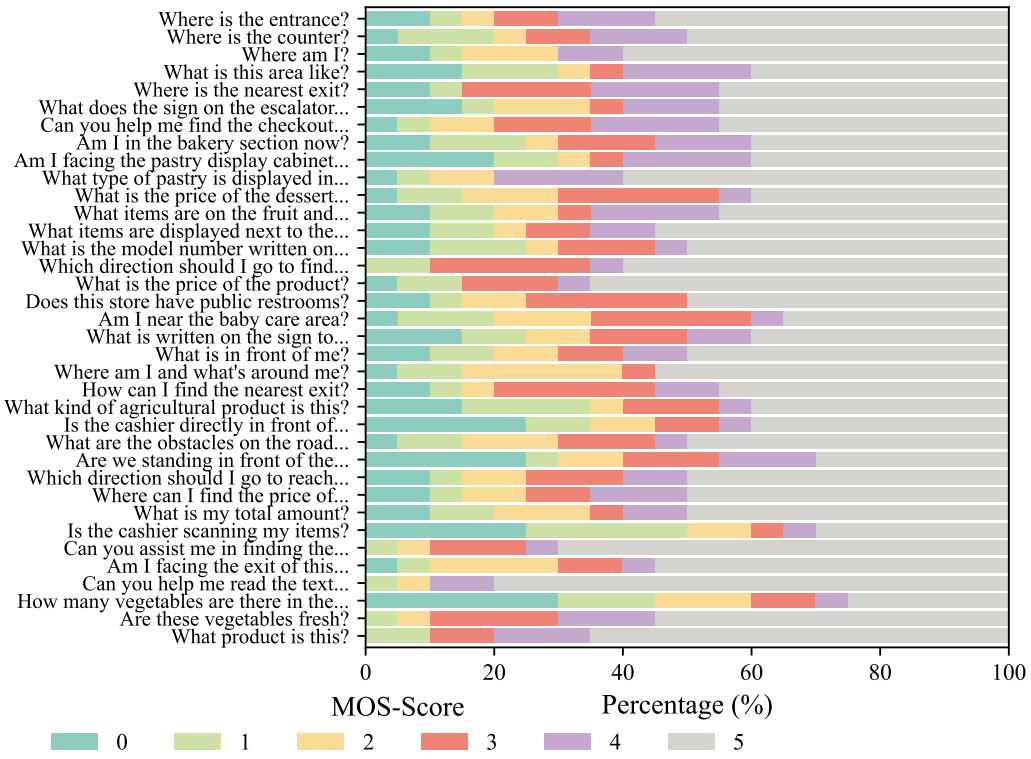


Figure 10: MOS score distribution of blind people in shopping scenarios. The stacked parts of each color represent the distribution proportions of different scores (same for below).

When blind people use transportation such as cars, buses, and other vehicles (as shown in Figure 11), over 60% of the respondents assigned full scores to questions like "Is this my taxi?" and "Which bus is coming?" while waiting for a taxi. This indicates that their primary concern is the current location of the vehicle. Meanwhile, respondents uniformly gave low scores to the question "Is the car being moved?" suggesting that blind individuals are capable of perceiving vehicle movement. Furthermore, approximately 80% of respondents gave low scores to questions such as "What does the bus stop advertisement say?" and "What brand is the car?" indicating that blind people often show little interest in irrelevant information surrounding their waiting position.

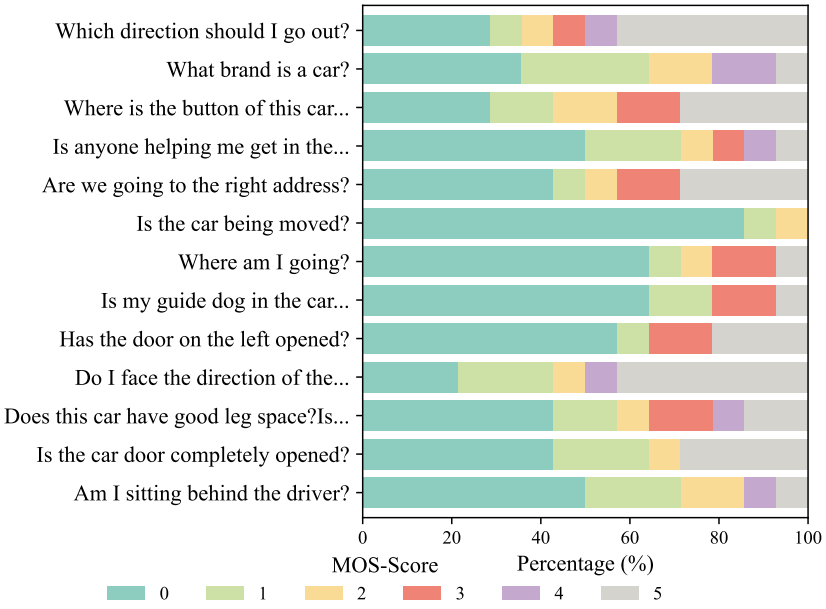


Figure 11: MOS score distribution of blind people using transportation.

In outdoor navigation scenarios, such as finding destinations, crossing roads, or walking, blind individuals prioritize safety above all else (as shown in Figure 12). For instance, the question “Is it safe to cross the road?" receives the highest score of 5 from all respondents. Similarly, over 80% of respondents rate questions related to safety during movement, such as “How can I safely get through this entrance?" and “Is it safe to go straight?", with scores of 4 or higher. Blind individuals also place significant emphasis on directional and positional information. For questions about accessing public facilities, such as “How to reach the nearest sidewalk, zebra crossing, or blind road?", over 70% of respondents gave scores of 4 or higher. Direction-related questions like “Which direction should I take to reach the intersection?" and “Is the supermarket on my left?" received scores above 3 from more than 60% of respondents.

However, respondents show less concern for questions about the characteristics of the surrounding environment, such as “Please describe the environment around me" and “Is traffic busy in this area?". Similarly, questions about weather conditions (“Is it sunny now?") or information about nearby pedestrians (“Who is the person standing in front?", “Are there many people around us gathered?", “Is there a pedestrian coming towards me?") received scores below 3 from more than 50% of respondents, indicating a general lack of interest. Interestingly, questions about obstacles did not receive as high a score as might be intuitively expected. For example, fewer than 50% of respondents positively rated questions like “Can you describe the obstacles ahead?" and “To avoid obstacles, which direction should I go?", while a large proportion gave neutral scores (3). This may be due to differences in the proficiency of using white canes among blind individuals, which influences their reliance on such information. Similarly, questions such as “Am I walking on a blind road” also exhibited a polarized response, with respondents either assigning very low scores (0 or 1) or very high scores (4 or 5). This polarization suggests that reliance on certain types of information may vary significantly depending on individual preferences and skills.

We also observed characteristics of blind individuals’ activities in familiar environments that contradict common assumptions. As shown in Figure 11, when blind individuals are at home, we found that the proportion of questions receiving low scores (0, 1, or 2) exceeded 60% across almost all queries. Respondents explained that although they cannot see, their brains form a mental “ma” of the environment based on prior familiarity. This mental representation allows them to navigate and plan their movements seamlessly within the environment. Unless there are significant changes in the environment, they typically do not encounter any issues.

To validate the rationality of question design in hands-on activity scenarios for visually impaired individuals, we conducted a study where participants rated questions in scenarios such as cooking

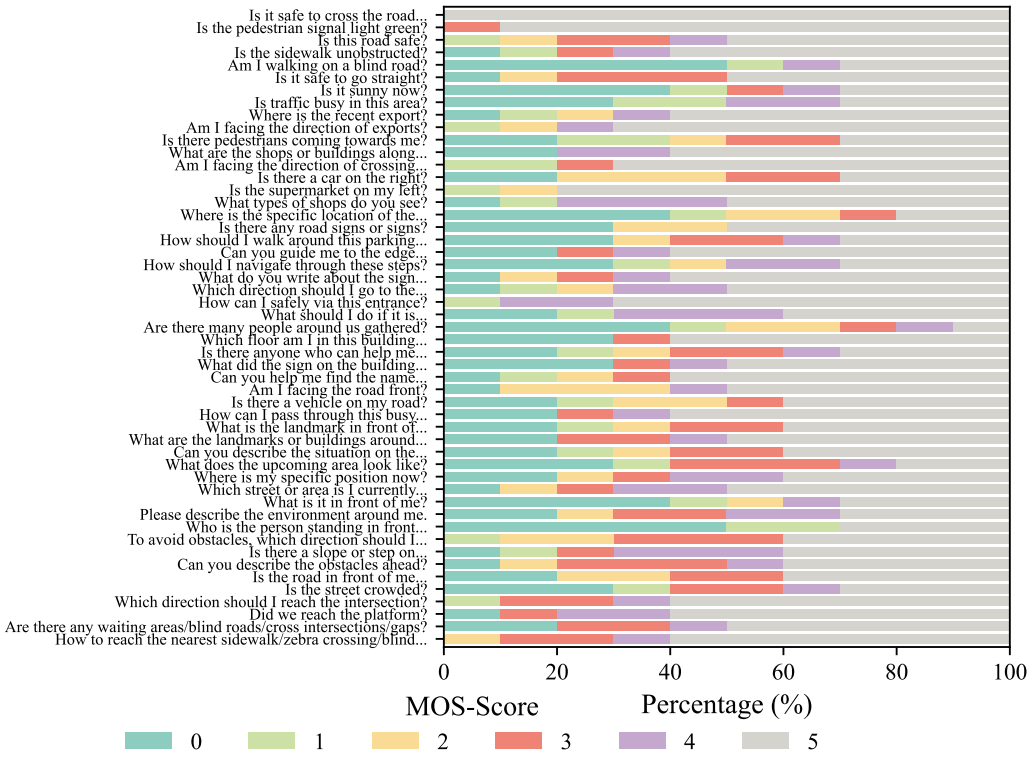


Figure 12: MOS score distribution of blind people navigating outdoor environments.

and crafting (As shown in Figure 14). Taking these two scenarios as examples, we found that, despite the differences in context, respondents consistently gave high scores (4 and 5) to core questions that directly influence task completion. For instance, "How much time is left?" (Cooking) and "Please describe the items on the shelves" (Doing crafts). At the same time, due to the differences in activities, visually impaired individuals tended to give a higher proportion of low scores (0 or 1) to cooking-related questions compared to crafting. This can be attributed to the fact that cooking typically occurs in familiar environments where blind individuals rely on their spatial memory and experience. In contrast, in the crafting scenario, respondents showed a neutral or low level of interest (0–3) in questions about basic item information such as appearance or shape, exemplified by questions like "Is it a cup?" and "What does this cup look like?".

A.3 Analysis of Videos

We analyze more details about the videos in EgoBlind. First, unlike the famous Ego4D [35] videos which are framed in a top-down view of sighted people via head-mounted camera devices (*e.g.*, smart glasses), the videos in EgoBlind are captured by GoPros mounted in front of the blind users' chests or mobile phones held in their hands. Specific challenges resulting from EgoBlind videos are: 1) The objects of interest are often off-center and out-of-focus due to blindness, as shown in Figure 16(a), 2) the objects are shown in challenging viewpoint (blind users tend to walk alongside a street curb or a building for safety.), and 3) The scene could be largely occluded by their canes, mobile phones, guide dogs or other people.

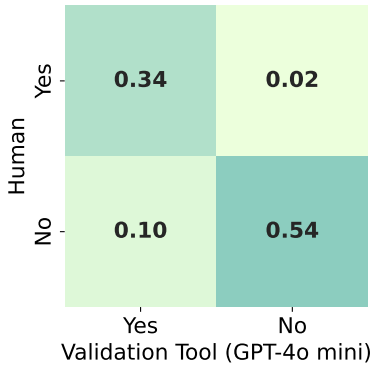


Figure 15: Consistency between hu-

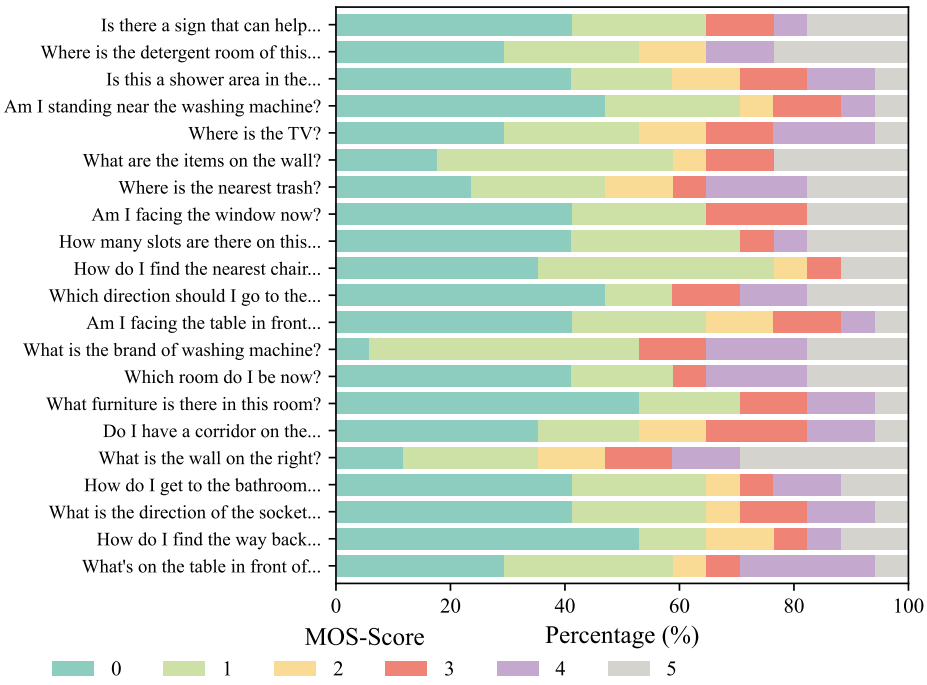


Figure 13: MOS score distribution of blind people in shopping scenarios.

B Experiment

B.1 Evaluation Metric

To enhance the reliability of the AI evaluation tool (*i.e.*, GPT-4o mini), we sample 30% of the test data to obtain inference results of Gemini 2.0. We invite human volunteers to score the models' inference results on this subset. Based on the results of human scoring, we refine the prompts to guide the machine's evaluation results closer to the human scoring outcomes. Ultimately, we calculate the Cohen Kappa coefficient [58] (for inter-rater reliability) between human and GPT scoring results, which is found to be 0.73, indicating a high degree of consistency between the two. Through heatmap analysis (shown in Figure 15), we observe that samples where human and machine ratings are consistent accounted for 88% of the total test data. For samples where ratings are inconsistent, we find that most of they are due to an information bias – machine cannot see the videos for answer scoring. Our final evaluation prompts are attached in Table 14.

B.2 Benchmark Analysis

Other Common Challenges. Figure 17 shows other common failure cases of the current Video-LLMs towards egocentric visual assistance for the blind. The example in Figure 17(a) demonstrates that the models are hard to perform effective live QA with streaming visual inputs. They often rely on the wrong frames for responses, even though we have explicitly prompted that the question is posed at the last frame (refer to our prompts at Table 12). The example in Figure 17(b) indicates that the models are poor at scene text recognition, as all model fail to answer the 3rd floor which was pressed by the user. The example also shows that the models are weak in performing temporal context reasoning. Because human can easily know the answer as the user moves her hand upward from button "1" to the 3rd button, even though the button "3" is partly blocked by the hand. Finally, the example in Figure 17(c) suggests that the models are struggling in linking the referential words (*e.g.*,"this") with the visual contents. This also necessitates spatio-temporal context understanding, *e.g.*, following the hand motion and the object it interacted with.

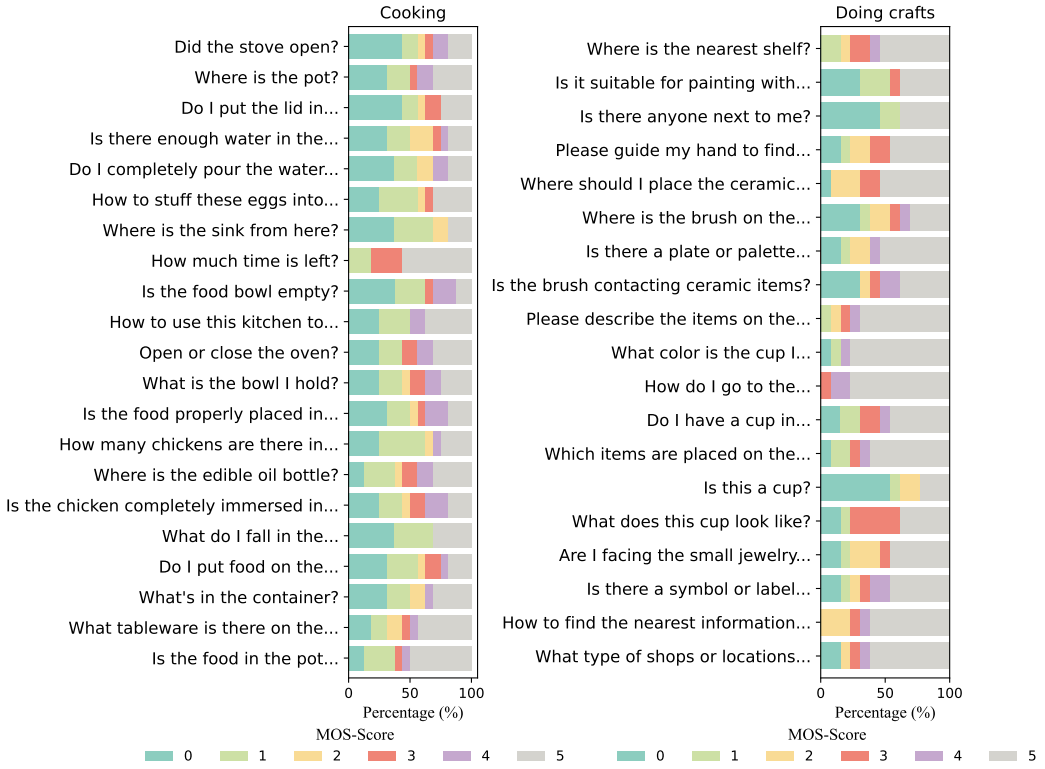


Figure 14: MOS score distribution of blind people in cooking and doing craft scenarios.

Table 8: Overall QA accuracy (%) of user-agonistic data split *vs.* user-aware data split.

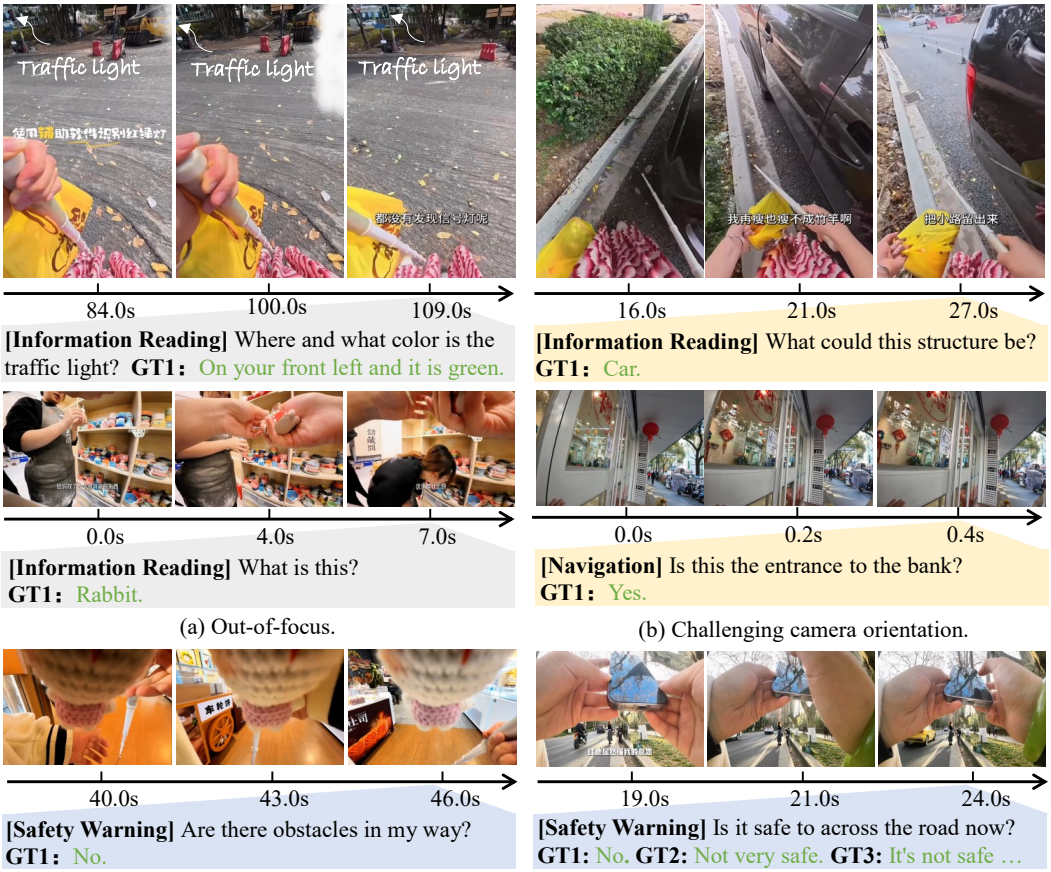
Models	User-Agnostic		User-Aware	
	Zero-shot	Finetune	Zero-shot	Finetune
Qwen2.5-VL	45.5	50.2 ↑ 4.7	47.7	52.9 ↑ 5.2
LLaVA-OV	54.5	57.4 ↑ 2.9	56.4	62.0 ↑ 5.6
InternVL2.5-8B	53.5	58.1 ↑ 4.6	56.2	62.3 ↑ 6.1

Table 9: Results of category-specific prompting for each question group (half of test data).

Method	**category prompt**	**Tool**	**Info.**	**Nav.**	**Safe**	**Com.**	**Res.**	**Overall**
LLaVA-OV	✗	58.4	54.1	37.2	63.8	59.4	54.0	54.2
	✓	50.6	53.2	30.2	64.1	56.2	52.9	52.2 ↓ 2.0
InternVL2.5-26B	✗	74.0	56.0	47.7	51.9	46.9	56.3	54.6
	✓	66.2	55.5	52.8	43.0	53.1	63.2	53.1 ↓ 1.5
GPT-4o	✗	61.0	59.6	54.3	60.3	46.9	69.0	59.4
	✓	79.2	63.3	58.8	58.6	50.0	64.4	62.2 ↑ 2.8

B.3 Other Investigations

Implementation Details for Finetuning. The fine-tuning procedures are conducted on two NVIDIA A800 GPUs. Most hyper-parameters are kept at their default configurations as specified in the fine-tuning scripts. For InternVL2.5-8B, optimal performance is achieved by increasing the training epochs to 2 while finetuning only the MLP layer. Regarding LLaVA-OV, we employ LoRA with reduced rank dimensionality (r=16) to optimize computational efficiency. Due to constraints of compute resource, we truncate the maximum sampled frames to 16, and cape the model's maximum sequence length to 8192, and freeze the vision encoder. Other refinements involve extending the number of epochs to 2, increasing gradient accumulation steps to 4, and adjusting weight decay to 0.05. Finally, for Qwen2.5-VL, we find that the best results are achieved by simply training the model 1 epoch while preserving other parameters unchanged.



(a) Out-of-focus.

(b) Challenging camera orientation.

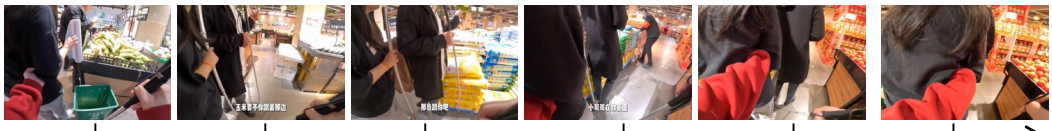
(c) Obstruction.

Figure 16: Visualization of EgoBlind examples.

Table 10: Investigation of Chinese-specific elements (half of test data). Subtitles and scene texts (SText) play little role in EgoBlind, while all model performance decline significantly when translating the questions into Chinese (CHN).

Method	Subt.	SText	CHN	Tool	Info.	Nav.	Safe	Com.	Res.	Overall
Qwen2.5-VL	✓			45.4	49.0	32.2	46.5	40.6	35.6	44.5
	✗			44.2	46.3	27.6	51.0	40.6	31.0	43.2 ↓ 1.3
	✓	✓		42.9	48.5	33.7	45.5	46.9	43.7	44.8 ↑ 0.3
	✓		✓	44.2	46.4	31.2	40.1	31.2	39.1	41.5 ↓ 3.0
LLaVA-OV	✓			58.4	54.1	37.2	63.8	59.4	54.0	54.2
	✗			52.0	54.4	34.2	64.4	59.4	55.2	53.7 ↓ 0.5
	✓	✓		52.0	56.4	35.2	62.5	53.1	54.0	54.1 ↓ 0.1
	✓		✓	52.0	53.0	36.7	60.3	40.6	46.0	51.4 ↓ 2.8
InternVL2.5-26B	✓			74.0	56.0	47.7	51.9	46.9	56.3	54.6
	✗			67.5	52.5	51.3	53.8	50.0	57.5	53.8 ↓ 0.8
	✓	✓		62.3	57.4	48.7	49.7	53.1	55.2	54.2 ↓ 0.4
	✓		✓	59.7	56.0	48.7	50.3	50.0	49.4	53.1 ↓ 1.5
GPT-4o	✓			61.0	59.6	54.3	60.3	46.9	69.0	59.4
	✗			68.8	56.9	53.8	55.8	53.1	70.1	57.6 ↓ 1.8
	✓	✓		63.6	59.0	50.8	53.2	56.2	62.1	56.7 ↓ 2.7
	✓		✓	64.9	55.1	51.8	56.4	56.2	60.9	55.9 ↓ 3.5

Category-aware Prompting. We investigate whether designing category-specific prompt for each question group can enhance QA performance. To reduce API costs for closed-source models, the experiments are conducted on a randomly selected half of the test set. Results in Table 9 show that category-specific prompting (see Table 13) effectively improves GPT-4o's performance, but not



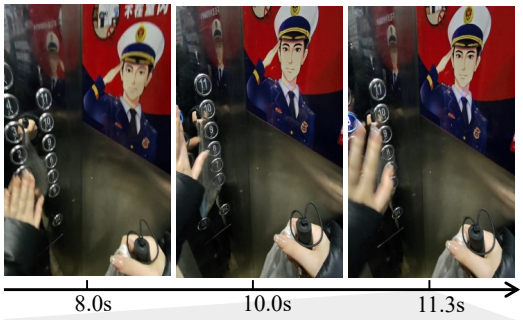
[Information Reading] Which section of the store are we in right now?
GT1: Rice section. **GT2:** Rice area. **GT3:** Area where rice is sold.

GPT-4o: produce near corn. [C: No, S: 0]
Gemini 1.5: produce. [C: No, S: 0]
InternVL2.5: produce. [C: No, S: 0]
MiniCPM V2.6: produce. [C: No, S: 0]
Qwen2.5 VL: produce. [C: No, S: 0]
VILA 1.5: grocery. [C: No, S: 0]
VideoLLaMA3: produce. [C: No, S: 0]
CogVLM2: produce. [C: No, S: 0]
LLaVA-OV: produce. [C: No, S: 0]
LLaVA-Video: supermarket. [C: No, S: 0]
Video-LLaVA: produce. [C: No, S: 0]
LLaMA-VID: vegetable. [C: No, S: 0]

[Information Reading] What products are on my right side?
GT1: Fruits. **GT2:** Tomatoes.

GPT-4o: cooking oils and soy sauce. [C: No, S: 1]
Gemini 1.5: rice or noodles. [C: No, S: 0]
InternVL2.5: rice and goods.. [C: No, S: 0]
MiniCPM V2.6: I don't know. [C: No, S: 0]
Qwen2.5 VL: vegetables and goods.. [C: No, S: 2]
VILA 1.5: apples. [C: No, S: 1]
VideoLLaMA3: I don't know. [C: No, S: 0]
CogVLM2: not clearly visible. [C: No, S: 0]
LLaVA-OV: bags of rice. [C: No, S: 0]
LLaVA-Video: rice. [C: No, S: 0]
Video-LLaVA: corn and eggs. [C: No, S: 1]
LLaMA-VID: vegetables. [C: No, S: 1]

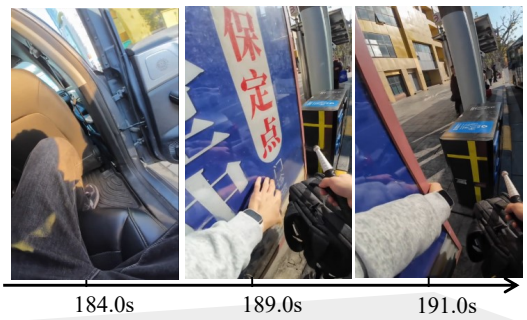
(a) Streaming VQA.



[Information Reading] Which floor button did I press? **GT1**: 3rd floor.

GPT-4o: the first floor. [C: No, S: 0]
Gemini 1.5: the 1st floor. [C: No, S: 0]
InternVL2.5: 1st floor. [C: No, S: 0]
MiniCPM V2.6: I don't know. [C: No, S: 0]
Qwen2.5 VL: 10 floor. [C: No, S: 0]
VILA 1.5: 10. [C: No, S: 0]
VideoLLaMA3: I don't know. [C: No, S: 0]
CogVLM2: the 1st floor. [C: No, S: 0]
LLaVA-OV: … floor 1. [C: No, S: 0]
LLaVA-Video: I don't know. [C: No, S: 0]
Video-LLaVA: I don't know. [C: No, S: 0]
LLaMA-VID: 10th floor. [C: No, S: 0]

(b) Scene texts .



[Information Reading] What is this? **GT1**: Billboard.
GT2: Advertising road sign. **GT3**: bus stop billboard

GPT-4o: mobile phone. [C: No, S: 0]
Gemini 1.5: taxi seat headrest. [C: No, S: 0]
InternVL2.5: car key fob. [C: No, S: 0]
MiniCPM V2.6: I don't know. [C: No, S: 0]
Qwen2.5 VL: car interior. [C: No, S: 0]
VILA 1.5: car. [C: No, S: 0]
VideoLLaMA3: I don't know. [C: No, S: 0]
CogVLM2: car. [C: No, S: 0]
LLaVA-OV: This video shows …. [C: No, S: 0]
LLaVA-Video: The video shows …. [C: No, S: 0]
Video-LLaVA: The video shows. [C: No, S: 0]
LLaMA-VID: car door handle. [C: No, S: 0]

(c) Referential words.

Figure 17: Visualization of failure cases. C: Correctness, S: Score. We only show the key answer words for brevity (same for below).

for other open-source models. We speculate that the open-source models are not strong enough to understand longer and complex prompts.

User-aware Dataset Split. Additionally, we re-split our EgoBlind dataset into train and test sets by ensuring videos token by the same user not appearing in both, *i.e.*, half users in train and half users in test. We obtain both the zero-shot and finetuned results of three representative open-source

models on the new test set in Table 8. The results show that finetuning can remarkably improve the model performance under different data splits. To our surprise, all models perform better with the user-aware split, regardless of zero-shot or finetuning. We speculate that some user videos and questions are not that challenging to answer.

Chinese Contents. While EgoBlind aims for "in-the-wild" visual assistance for the blind, the videos are majorly sourced from Chinese video sharing platforms. Hence, we study if the model performances are sensitive to Chinese-specific elements in the videos. Specifically, we first remove all subtitles from videos by using off-the-shelf subtitle remover [4]. Results in Table 10 show that all model performances drop slightly by 0.5% to 1.8%. A manual check of the new failure cases reveals that the vast majority (95.5% to 98.4% for different models) are irrelevant to subtitles and cloud be due to visual input variation resulted from subtitle removal (*e.g.*, visual blur), not because of no subtitles. Alternatively, we extract scene texts from videos by using advanced OCR tool[5]. The scene texts are used as additional prompting contents for MLLMs to make decisions. The results in Table 10 show that the additional scene texts do not help model predictions but slightly hurt the performance. The results and analyses suggest that Chinese subtitles and scene texts matter little in answering EgoBlind questions. Finally, we translate all questions into Chinese and find that all model performance decline significantly, reflecting that all models are weaker in understanding Chinese compared to English.

C Limitations

While we take significant challenge to collect the videos and QAs, the data are limited to the mainland of China so far. Also, the training set is not that big compared with other video QA datasets from sighted people, since there is not much egocentric video data from the blind online. However, we have shown the effectiveness and importance of the training data for better performances. Also, we are continuously collecting related data and aim to extend the dataset by cooperating with the blind association, possibly enclosing blind VQA data from all over the world. In addition, the MLLM techniques are evolving rapidly and we may have missed some most recent models for testing. We will setup an evaluation server and maintain a leader-board to trace the technique advancements in egocentric visual assistance for the blind.

D Societal Impacts

Positive Impacts: EgoBlind is the first VideoQA dataset that benchmarks vision-language research towards egocentric visual assistance for the blind. The research and outcomes can support developing assistive technology for blind and visually impaired individuals. EgoBlind can advance intelligent systems capable of understanding and describing real-world environments from a first-person perspective, thereby enhancing autonomy, safety, and quality of life for blind users. By enabling more natural and interactive communication between the blind users and the assistive system, VideoQA technologies can help bridge accessibility gaps in daily activities such as navigation, information reading, tool use, social interaction and other resource acquisition, thus holding immerse value towards enlightening the blind individual's life.

Negative Impacts: Despite its significance. Privacy concerns may arise as egocentric recordings inherently capture sensitive and personal information about both the blind user and bystanders. Additionally, biases in training data or system limitations may lead to inaccurate or misleading responses, potentially compromising user trust or safety. Addressing these challenges is essential to ensure the ethical and inclusive deployment of MLLMs in practical visual assistance for the blind.

[4] https://github.com/YaoFANGUK/video-subtitle-remover

[5] https://github.com/PaddlePaddle/PaddleOCR

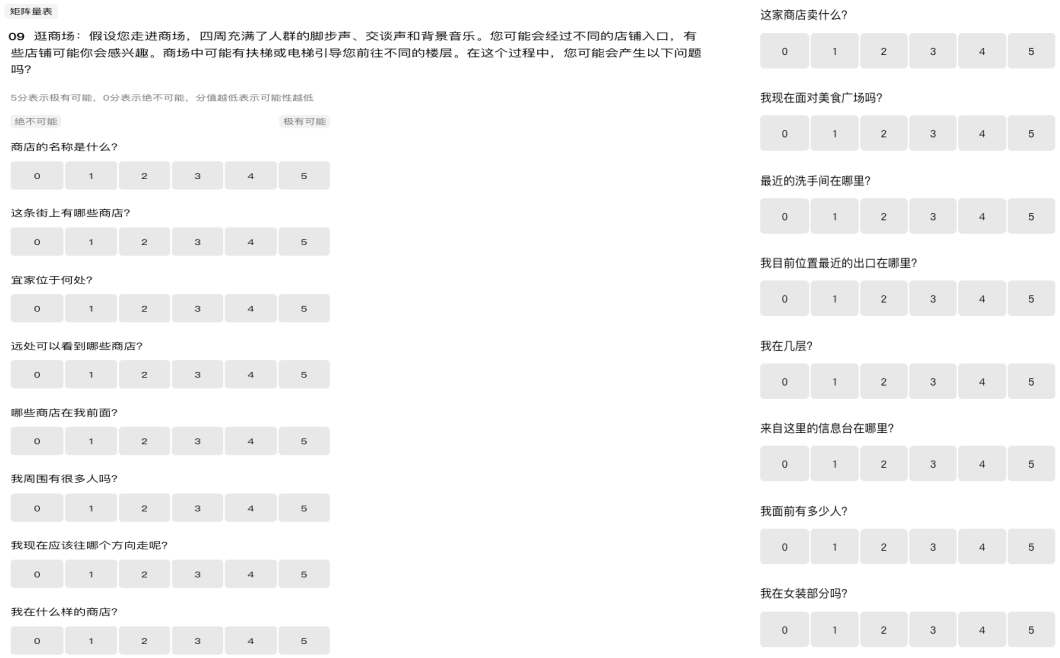


Figure 18: Exampe of questionnaire for blind study. Questions are translated into Chinese.

Table 11: Normal prompts for models to perform question answering on EgoBlind.

Model	General Prompts
InternVL2.5-26B	I will provide you with a video each time and one question; Your task is to answer the question which was posed by the people in the last frame of the video. The answer needs to be based on the content of the picture and the objective characteristics. If the question cannot be answered, you can say I don't know. Do not include Chinese characters in your response. The question is: row['question']
GPT-4o	I will provide you with several pictures each time and one question; Your task is to answer the question. The answer needs to be based on the content of the picture and the objective characteristics. If the question cannot be answered, you can say I don't know. Please generate the response with keys 'question', 'question_id', 'prediction', and 'type'. The question is: {question}, Question ID: {question_id}, Type: {question_type}. The content of the key 'question' is the question you received. The content of the key 'prediction' is the answer to the corresponding question you generated. The key 'question_type' represents which category the question type you received. Your response should look like this example: {'question':'Is there a wet floor caution sign near the door?', 'question_id': 'v_87_1_2', 'prediction':'yes.', 'type': 'information reading'}
Gemini 1.5 Flash	I will provide you with a video each time and one question; Your task is to answer the question which was posed by the people in the last frame of the video. The answer needs to be based on the content of the picture and the objective characteristics. If the question cannot be answered, you can say I don't know. Do not include Chinese characters in your response. Please generate the response with keys 'question', 'question_id', 'prediction', and 'type'. The question is: {question},Question ID: {question_id},Type: {question_type}. The content of the key 'question' is the question you received. The content of the key 'prediction' is the answer to the corresponding question you generated. The key 'question_type' represents which category the question type you received. Your response should look like this example: {{"question":"Is there a wet floor caution sign near the door?", "question_id": "v_87_1_2", "prediction":"yes.", "type": "information reading"}}

Table 12: Blind-aware prompts for models to perform question answering on EgoBlind.

Model	General Prompts
Open Source MLLMs	I will provide you with a video each time and one question; These questions are all questions raised by the blind person in the video from his own first-person perspective in the current scene. Your task is to answer the blind person's question which was posed in the last frame of the video. The answer needs to be based on the content of the picture and the objective characteristics that the blind person cannot see. If the question cannot be answered, you can say I don't know. Do not include Chinese characters in your response. The question is: {question}
GPT-4o	I want you to act as a visual assistant for the blind. I will provide you with several pictures each time and one question; These questions are all raised by the blind person in the video from his own first-person perspective in the current scene. Your task is to answer the blind person's question. The answer needs to be based on the content of the picture and the objective characteristics that the blind person cannot see. If the question cannot be answered, you can say I don't know. Please generate the response with keys 'question', 'question_id', 'prediction', and 'type'. The question is: {question},Question ID: {question_id},Type: {question_type}. The content of the key 'question' is the question you received. The content of the key 'prediction' is the answer to the corresponding question you generated. The key 'question_type' represents which category the question type you received. Your response should look like this example: {'question':'Is there a wet floor caution sign near the door?', 'question_id': 'v_87_1_2', 'prediction':'yes.', 'type': 'information reading'}
Gemini	I want you to act as a visual assistant for the blind. I will provide you with a video each time and one question; These questions are all questions raised by the blind person in the video from his own first-person perspective in the current scene. Your task is to answer the blind person's question which was posed in the last frame of the video. The answer needs to be based on the content of the picture and the objective characteristics that the blind person cannot see. If the question cannot be answered, you can say I don't know. Do not include Chinese characters in your response. Please generate the response with keys 'question', 'question_id', 'prediction', and 'type'. The question is: {question},Question ID: {question_id},Type: {question_type}. The content of the key 'question' is the question you received. The content of the key 'prediction' is the answer to the corresponding question you generated. The key 'question_type' represents which category the question type you received. Your response should look like this example: {"question":"Is there a wet floor caution sign near the door?", "question_id": "v_87_1_2", "prediction":"yes.", "type": "information reading"}

Table 13: Category-specific prompts. They are properly inserted into the QA prompts in Table 12.

Question Category	Prompts
Tool Use	The questions focus on tool use, helping the blind understand how to operate tools around them.
Information Reading	The questions focus on information reading, helping the blind understand visible facts such as what their surroundings look like.
Navigation	The questions focus on navigation, guiding the blind to move safely or find directions in their environment.
Safety Warnings	The questions focus on safety warnings, reminding the blind of possible dangers of their surroundings.
Social Communication	The questions focus on social communication, recognizing or describing people interacting with the blind.
Resource	The questions focus on other resources, identifying people or things nearby that may offer help.

Table 14: Prompts for GPT-4o mini to evaluate MLLMs on EgoBlind.

Evaluation Prompts

You are an intelligent chatbot designed for evaluating the correctness of generative outputs for question-answer pairs. Your task is to compare the predicted answer with the correct answer and determine if they match meaningfully. Here's how you can accomplish the task:

##INSTRUCTIONS:

- **Focus on meaningful matches:** Assess whether the predicted answer and the correct answer have a meaningful match, not just literal word-for-word matches.
- **Criteria for Correctness:** The predicted answer is considered correct if it reasonably matches any of the four standard answers, recognizing that synonyms or varied expressions that convey the same meaning are acceptable.
- **Allow for Paraphrasing:** Understand that different wording that conveys the same fundamental idea is valid. Evaluate if the essence of the predicted answer captures the core information of the correct answer.
- **Flexibility in Evaluation:** Use judgment to decide if variations in the predicted answer still correctly address the question, even if they do not directly replicate the correct answer's phrasing.

For example, when the correct answer is 'Left front', Predicted Answer: 'About ten meters to your left front', these two answers match.

Please evaluate the following video-based question-answer pair:

```
Question: {question}
Correct Answer0: {answer0}
Correct Answer1: {answer1}
Correct Answer2: {answer2}
Correct Answer3: {answer3}
Predicted Answer: {pred}
```

Provide your evaluation only as a yes/no and score where the score is an integer value between 0 and 5, with 5 indicating the highest meaningful match. Please generate the response in the form of a Python dictionary string with keys `'pred'` and `'score'`, where the value of `'pred'` is a string of 'yes' or 'no' and the value of `'score'` is in INTEGER, not STRING.

DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only provide the Python dictionary string. For example, your response should look like this: `{'pred': 'yes', 'score': 4.8}`.

Table 15: Example prompts for GPT-4o to generate QAs.

Social Communication

I want you to act as a blind person. I will provide you with several pictures each time and the pictures have two characteristics: 1. pictures have temporal correlation between them and the order of the pictures represents the order of time. 2.The pictures are all taken from your first point of view, which is equivalent to the picture recorded by the camera hanging on your chest.

Your task is to imagine what questions you would ask to meet your needs if you were in the situation of the picture provided. You need to focus on asking questions about 'communication and interaction'. Below I will give a detailed explanation of 'communication and interaction'.

Communication and Interaction: Describe the interaction between blind people and surrounding people/companions, and observe other people's status information for multi-person collaborative activities. For example: Did my companion help me carry my luggage onto the conveyor belt? /Has my guide dog entered the elevator safely? /Who is talking to me?

Here's how you can accomplish the task: 1. use your imagination and only ask about 'communication and interaction' questions; 2.Please ask questions about each picture; 3. The questions you ask need to be real-time and do not contain any off-site information; 4.The questions you ask need to be answered through picture content; 5.If you cannot answer based on the picture, you can answer 'I don't know'; 6. Ask as many open-ended questions as possible, not just 'yes-no' questions; 7. Some of the questions generated need to reflect the characteristics of the video, that is, you need to watch several consecutive frames to get the answer. For example:'What is my companion doing? 8. 'communication and interaction' questions should be based on the real life scenarios of blind people, and the questions must be practical (that is, questions that blind people would really ask in the current situation). 9. Please don't ask bad examples like 'What is my companion saying to me ?'/'Is there an announcement about the next station or any other important information?'

Please generate the response with 'timestamp', 'question', 'answer' and 'questoin_type'. The content of the key 'timestamp' the time value corresponding to the current picture, for example: the first image corresponds to the first element in timestamp list, and so on. The timestamp value can be found here: 'f' Timestamp list: time.