NEURAL INFORMATION PROCESSING SYSTEMS 2025

EgoBlind: Towards Egocentric Visual Assistance for the Blind

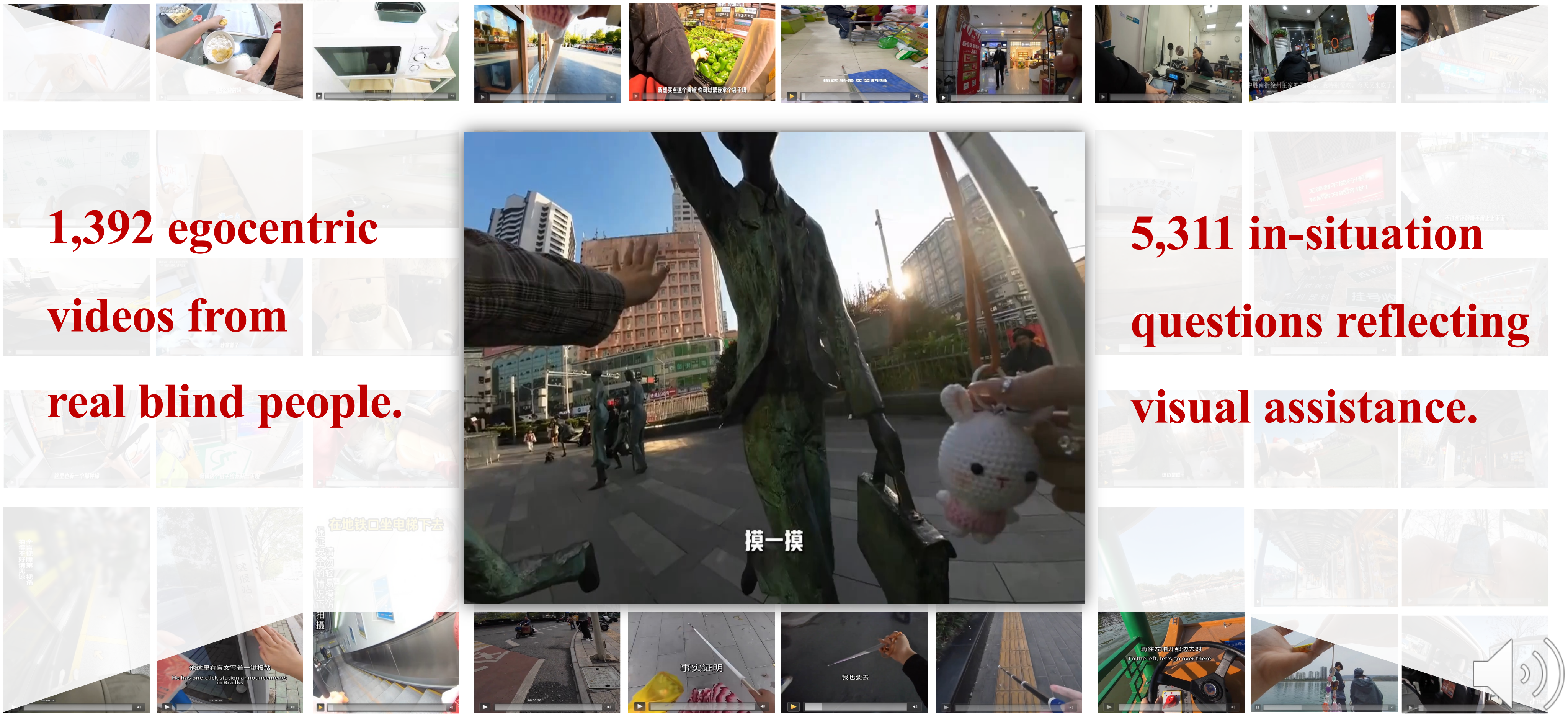
(Track on datasets and benchmarks)

Junbin Xiao*, Nanxin Huang*, Hao Qiu, Zhulin Zhang
Xun Yang, Richang Hong, Meng Wang, Angela Yao

Presenter: Junbin Xiao, NUS

EgoBlind
towards egocentric visual assistance for the blind

EgoBlind Dataset Overview



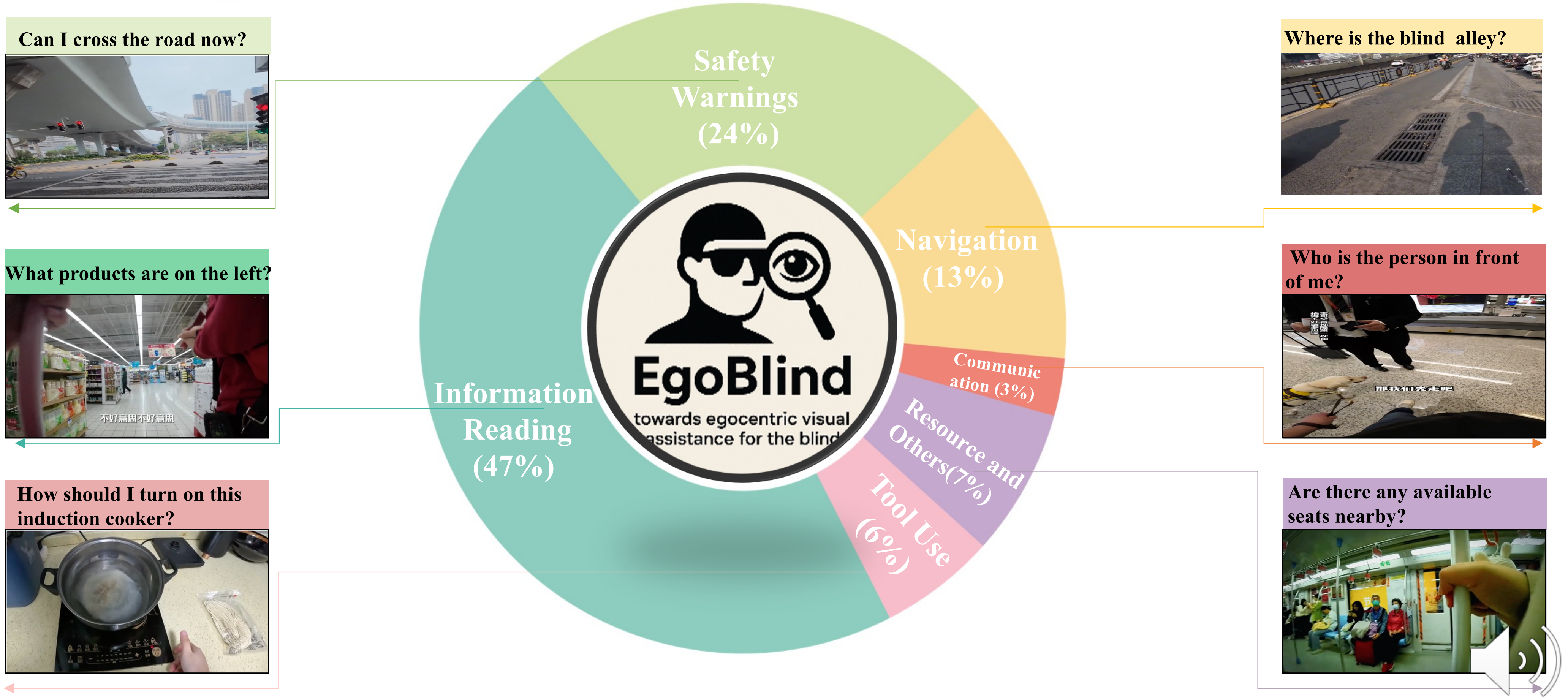
1,392 egocentric videos from real blind people.

5,311 in-situation questions reflecting visual assistance.

EgoBlind
towards egocentric visual assistance for the blind

EgoBlind Video Scenarios

EgoBlind Question Catgeories



Can I cross the road now?

What products are on the left?

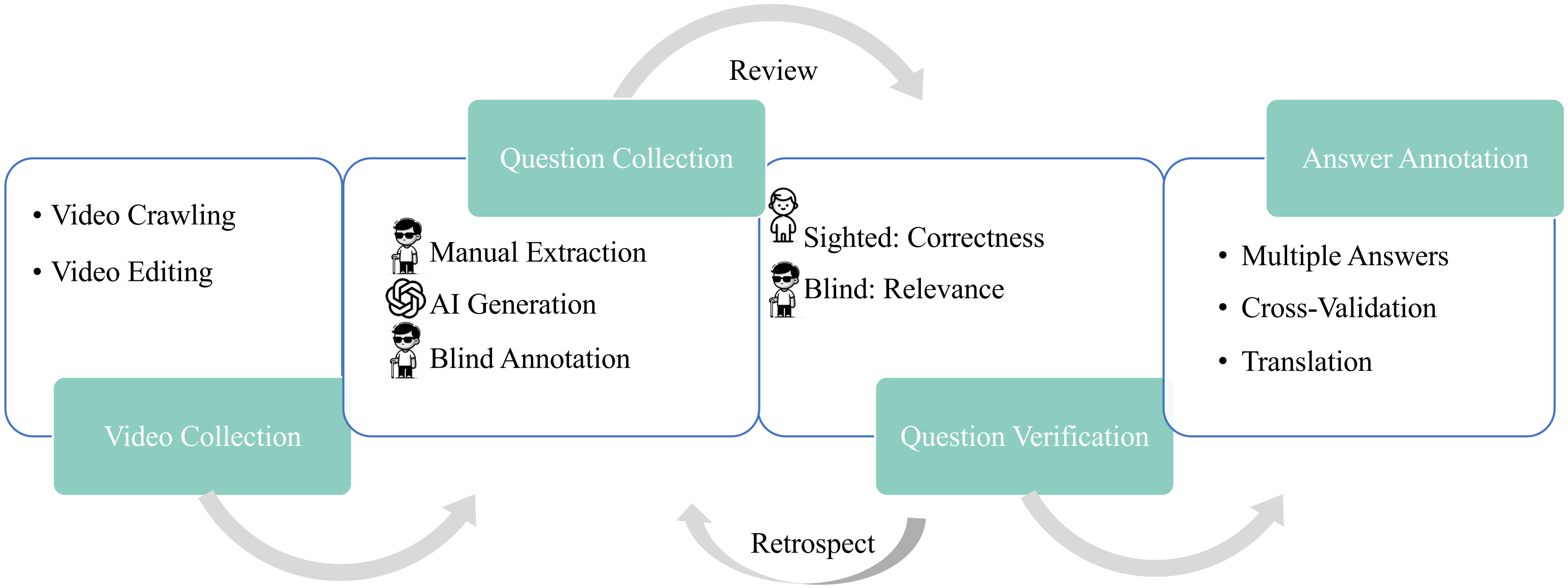
How should I turn on this induction cooker?

Where is the blind alley?

Who is the person in front of me?

Are there any available seats nearby?

EgoBlind Dataset Construction



Review

Question Collection

Answer Annotation

- Video Crawling
- Video Editing

Manual Extraction

AI Generation

Blind Annotation

Sighted: Correctness

Blind: Relevance

- Multiple Answers
- Cross-Validation
- Translation

Video Collection

Question Verification

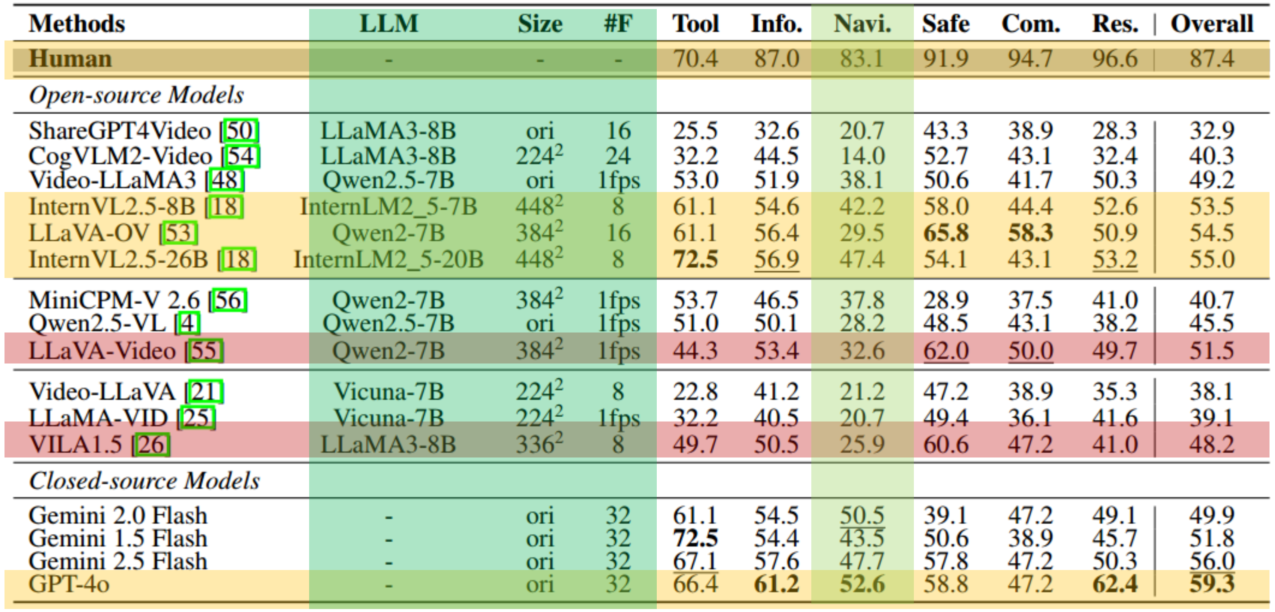
Retrospect

EgoBlind data construction pipeline.

Experiments – Overall Analysis

- None of the model achieves the desired level of performance on EgoBlind, all lagging behind human performance by a whopping 54%~28%.

- No single model wins across all question types. Answering “Navigation” questions is the most challenging task for almost all models.



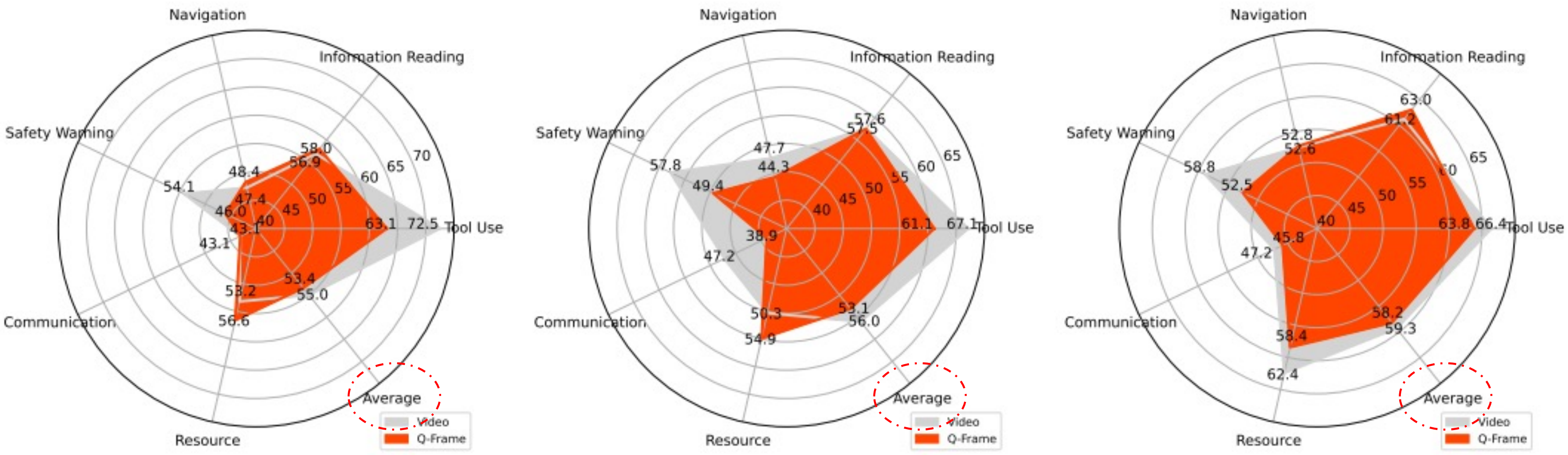
Methods	LLM	Size	#F	Tool	Info.	Navi.	Safe	Com.	Res.	Overall
Human	-	-	-	70.4	87.0	83.1	91.9	94.7	96.6	87.4
Open-source Models										
ShareGPT4Video [50]	LLaMA3-8B	ori	16	25.5	32.6	20.7	43.3	38.9	28.3	32.9
CogVLM2-Video [54]	LLaMA3-8B	224^2	24	32.2	44.5	14.0	52.7	43.1	32.4	40.3
Video-LLaMA3 [48]	Qwen2.5-7B	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**

- Stronger LLMs and larger visual resolution often bring better performance, while more frames do not always help

- 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.

EgoBlind
towards egocentric visual assistance for the blind

Experiments- Investigations



(a) InternVL2.5-26B.

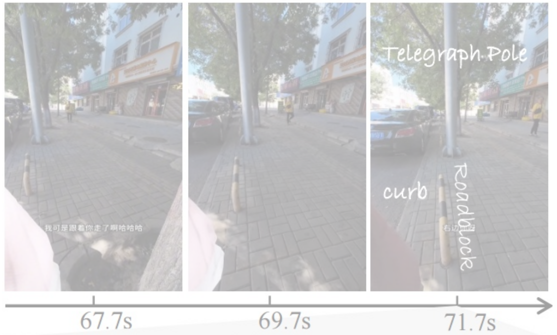
(b) Gemini 2.5 Flash.

(c) GPT-4o.

- Single frame input at the question moment hurts the overall performance, though it helps information reading.

EgoBlind
towards egocentric visual assistance for the blind

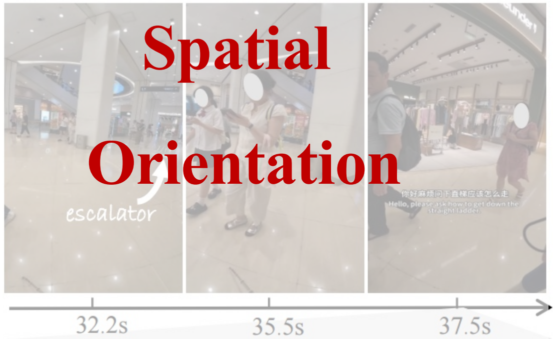
Experiments-Assist-related Challenges



[Safety] Is there a road ahead?
GT1: No, move to the right and then move forward.
GT2: No. **GT3**: There are many obstacles ahead, you should move to the right.

All models answer "Yes" and think there is a road ahead.

Spatial Orientation



[Navigation] How should I go to the escalator?
GT1: Behind you.
GT2: On your right rear.

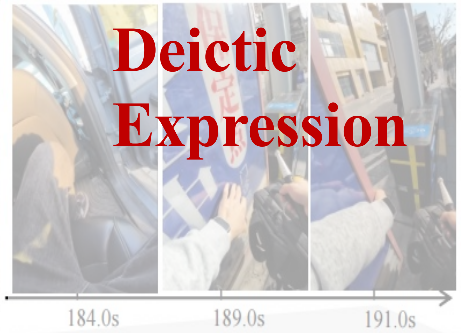
All models fail to answer the correct direction.

Scene-text

[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]

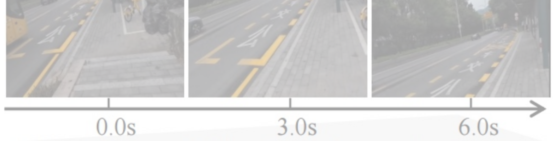
Deictic Expression



[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]

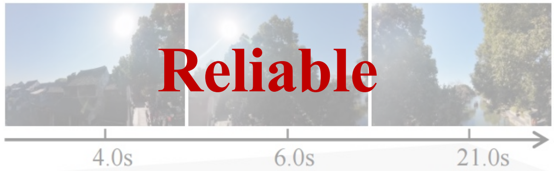
User Intention



[Other Resource] Where is the bus stop?
GT1: Directly in front of you.
GT2: Five to ten meters in front of you.
GT3: Directly in front.

All models answer that the bus stop is on the right side of the road or street.

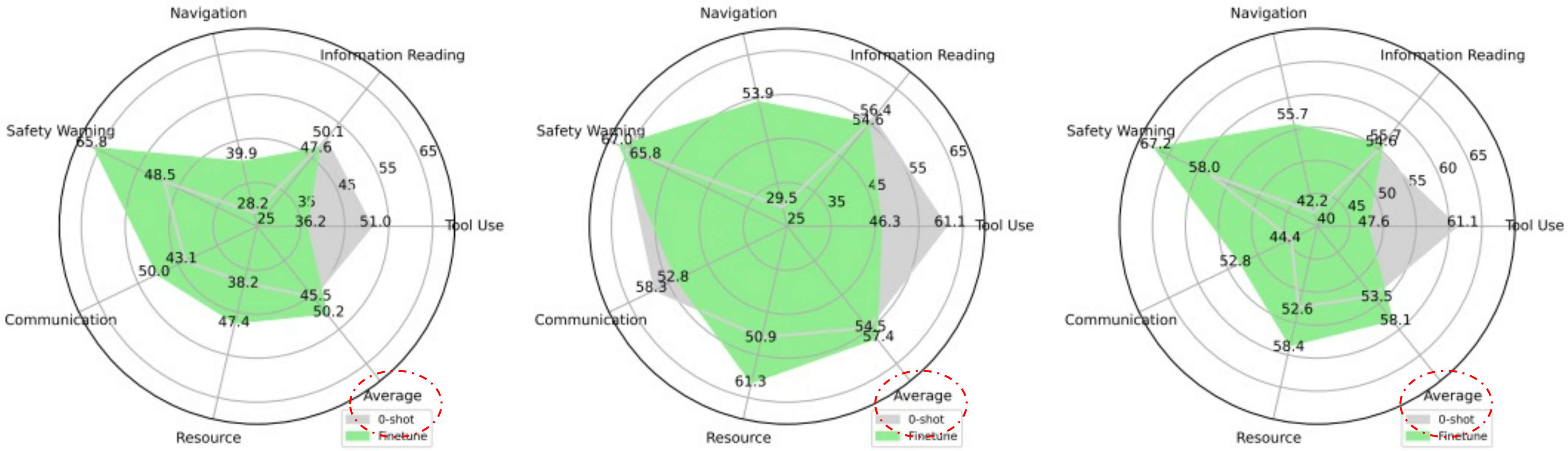
Reliable



[Navigation] How do I get to the nearest bridge?
GT1: You are on the bridge.
GT2: Standing on the bridge now.
GT3: You are on the bridge already..

All models do not know that the user is on the bridge, and give wrong and even malicious suggestions.

Experiments- Investigations



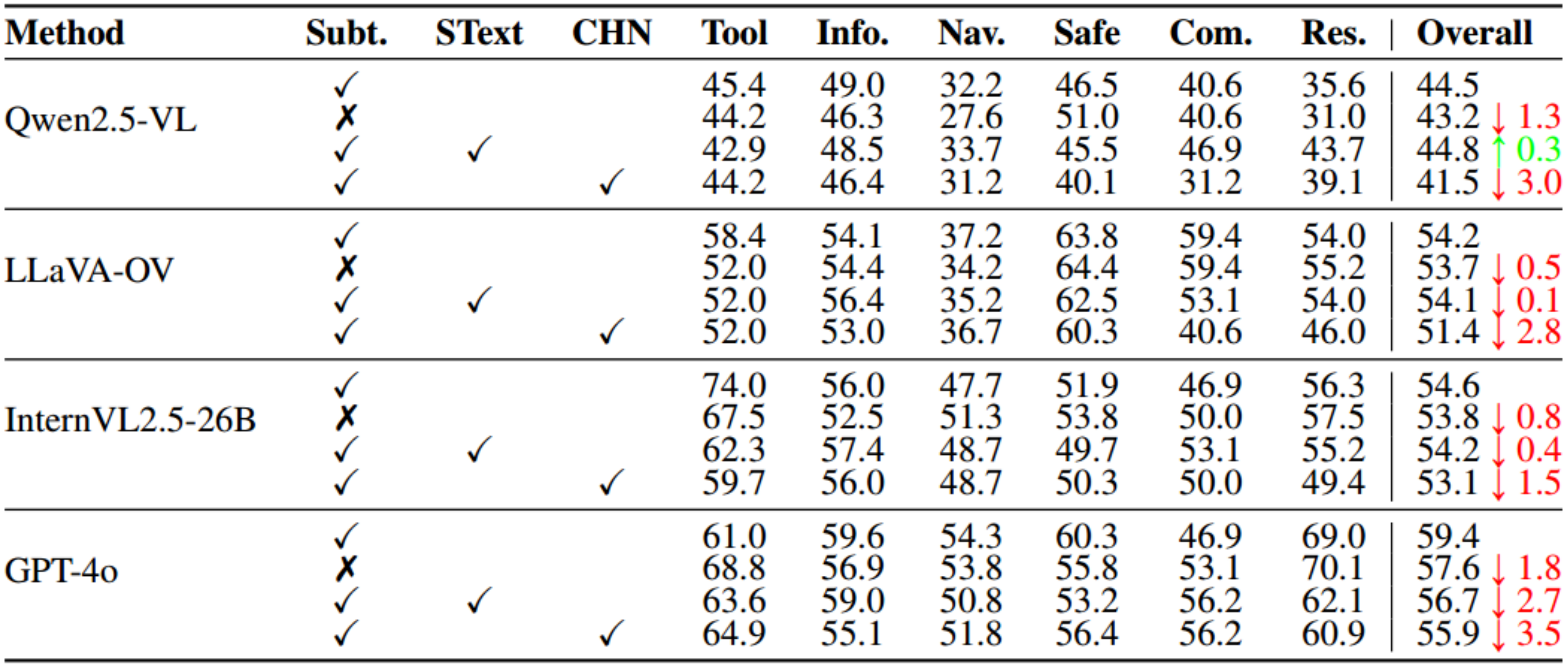
(a) Qwen2.5-VL.

(b) LLaVA-OV.

(c) InternVL-2.5-8B.

- Finetuning with EgoBlind training data significantly improves QA performance.

Experiments- Investigations



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

- Chinese-specific elements matter little the performance, though EgoBlind videos are collected from China.

Summary

- EgoBlind is the first egocentric VideoQA datasets collected from real-blind people.
- The videos and questions are diverse, reflecting blind users’ in-situation needs for visual assistance under various conditions.
- We provide an average 3 reference answers for each question for better evaluation.
- Existing models show significant performance gap to humans, indicting large room for improvements. EgoBlind training data are important.
- Limited location bias though the data are collected in China.

https://github.com/doc-doc/EgoBlind

Presenter:
junbin@nus.edu.sg