Exploring LLM-based Assistants with Smart Glasses for the Visually Impaired

Zhe-Yu Lee

AIINS@NTHU
Ambient Intelligence for
Immersive Networked Systems Lab

國立清華大學
NATIONAL TSING HUA UNIVERSITY

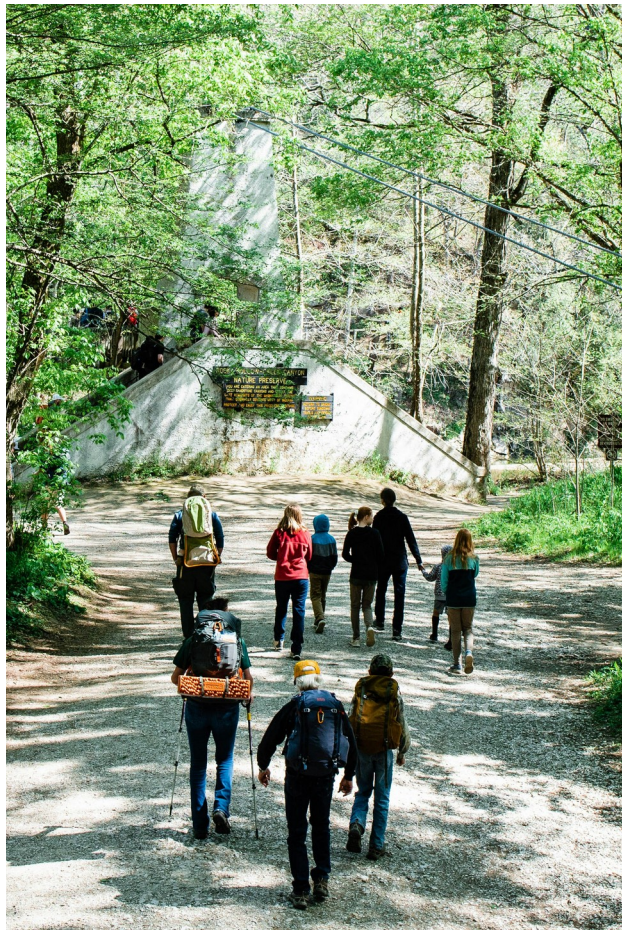
Outline

- Motivation
- Related Work
- System Architecture
- Research Problem
- Experiments
- Conclusion & Future Work

The Challenges of Visually Impaired

Safety

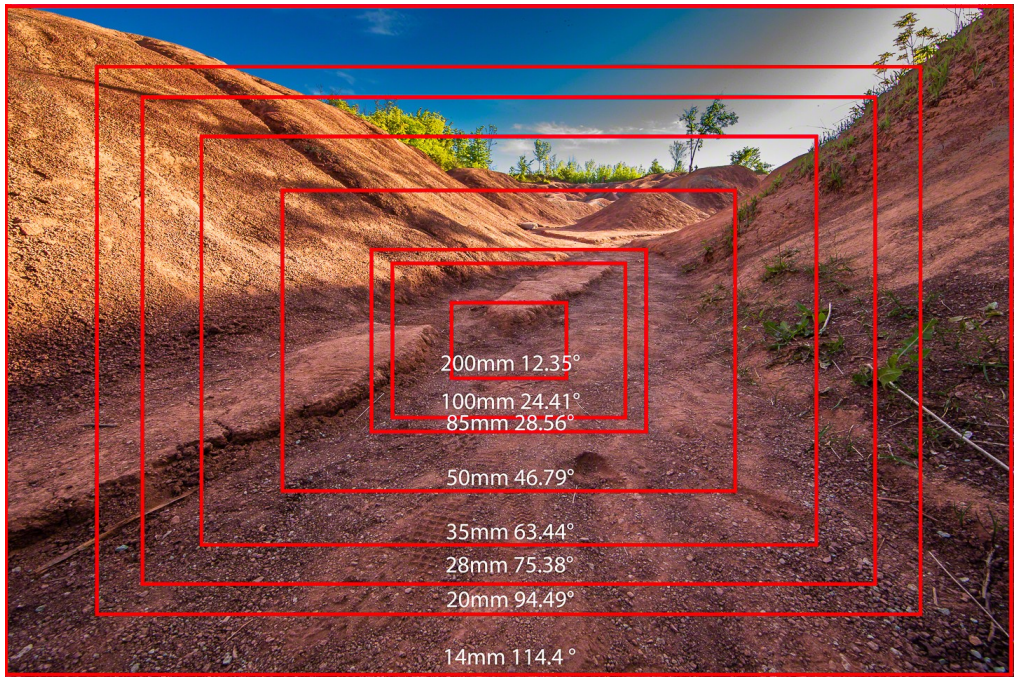
Navigation



Scene Understanding

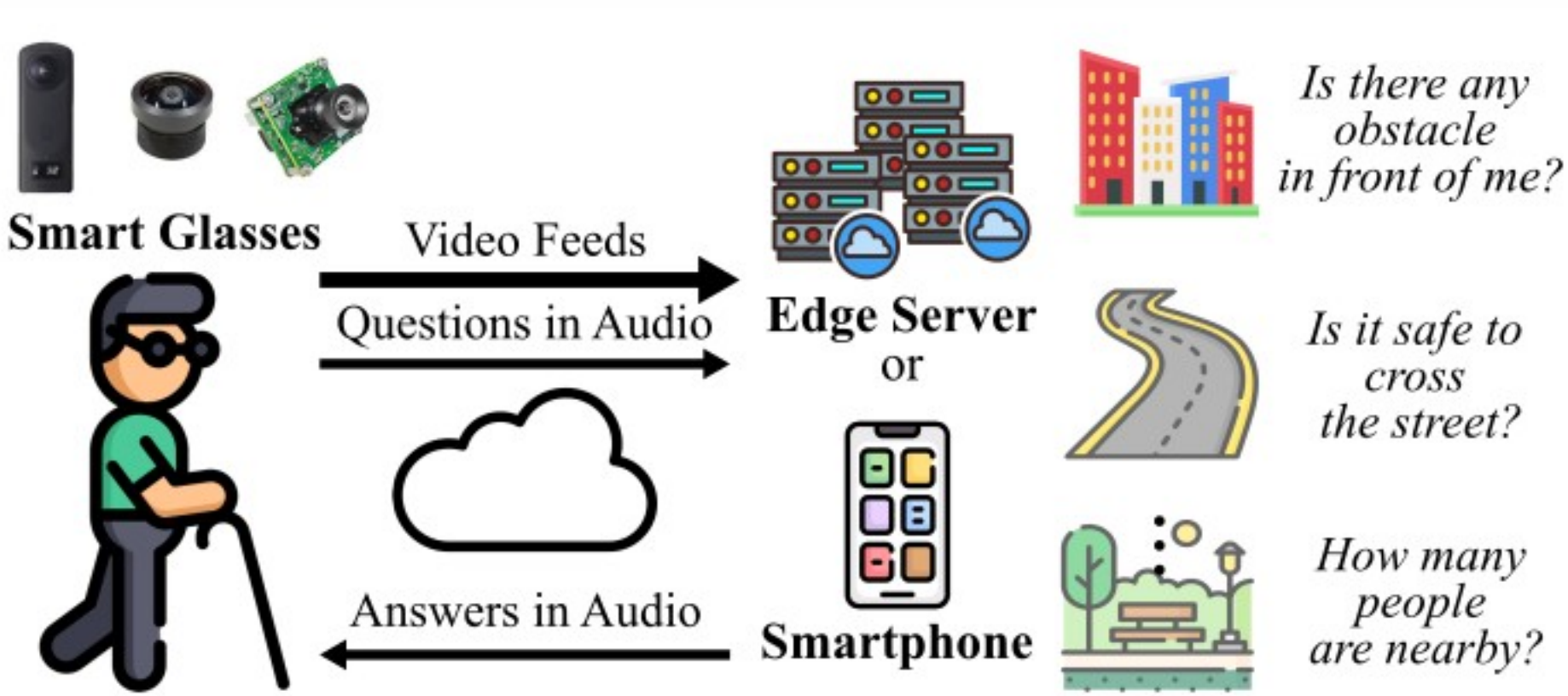
AI Wearable Assistants

- Popular smart glasses that show the potential of AI assistance



- Limitations:
  - Camera Angle of View (AoV)
  - Computing power

Real-World Assistance



To enable such an assistive system in real-world environments, several key challenges must be addressed.

Three Core Challenges

To realize our vision, we must overcome three key challenges:

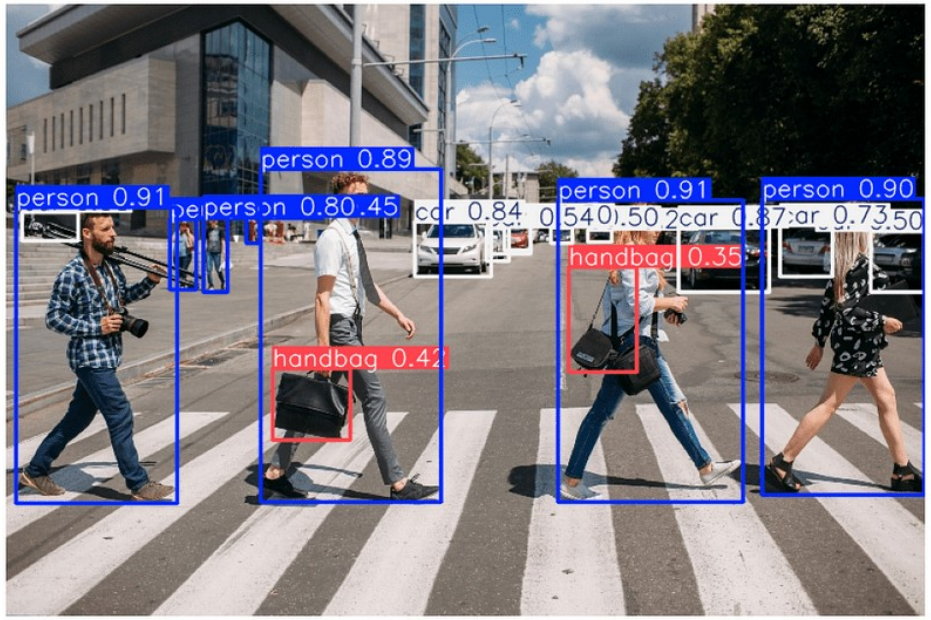
- A narrow AoV restricts scene understanding
  - Critical context is often missing or outside the frame
- A limited wearable computation constrains LLM processing
  - Offloading introduces latency and network dependency
- A lack of suitable datasets limits AoV analysis
  - No benchmark covers varying AoVs in assistive scenarios

Outline

- Motivation
- Related Work
- System Architecture
- Research Problem
- Experiments
- Conclusion & Future Work

Related Work

- Smart Cane [1]: Lacked semantic understanding
- Object Detection [2, 3]: Bound by task-specific models
- LLM-Based Assistance [4]: Assumed smaller AoVs

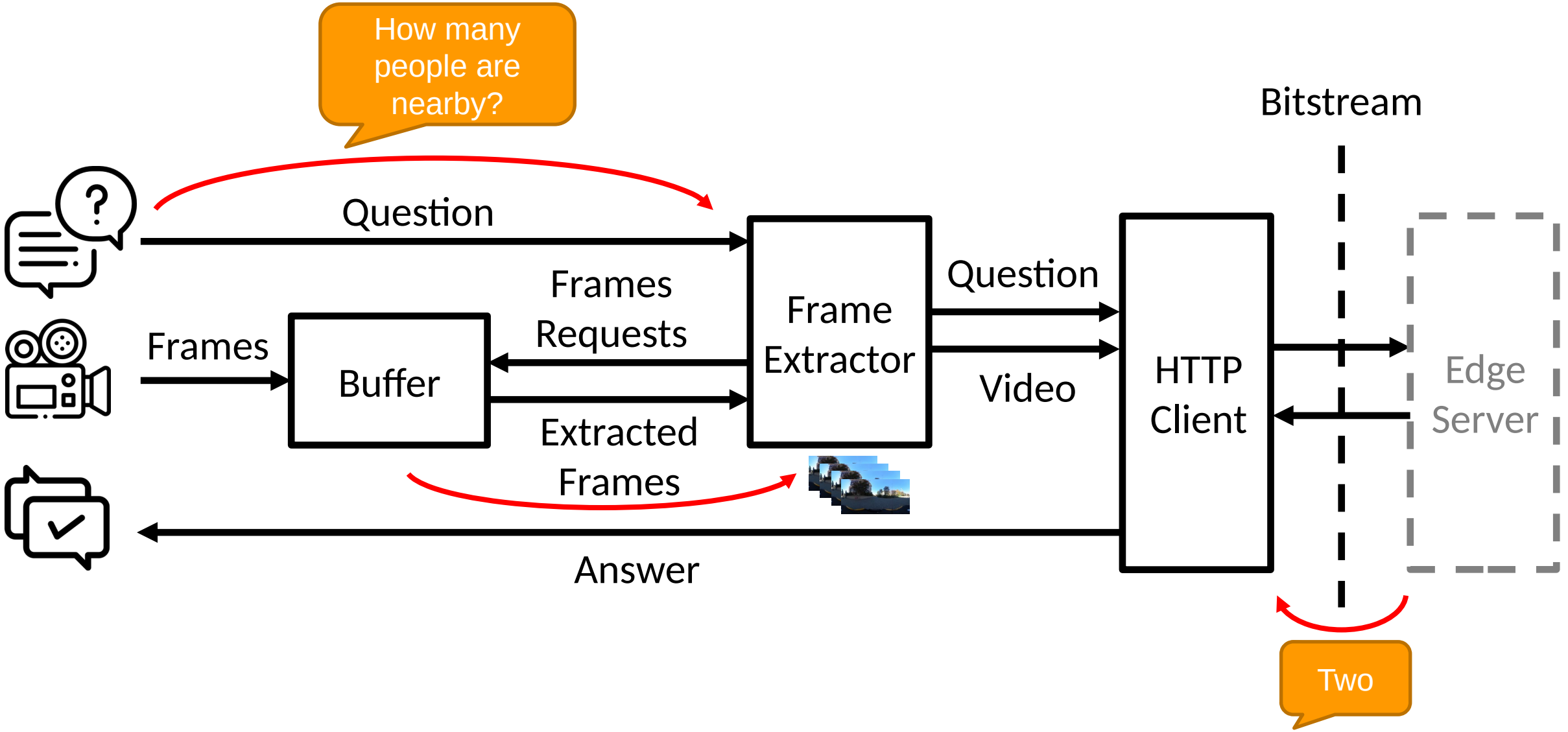


1. Wahab, Mohd Helmy Abd, et al. "Smart Cane: Assistive Cane for Visually-Impaired People." *arXiv preprint arXiv:1110.5156 (2011).*
2. Campos, et al. "Orb-slam3: An Accurate Open-source Library for Visual, Visual–Inertial, and Multimap Slam“ *In Proc. of the IEEE transactions on robotics (T-RO)*
3. Yang, Wenyan, et al. “Object Detection in Equirectangular anorama.” *2018 In Proc. of the International Conference on Pattern Recognition (ICPR).*
4. Xiao, Junbin, et al. "EgoBlind: Towards Egocentric Visual Assistance for the Blind People." *arXiv preprint arXiv:2503.08221 (2025).*

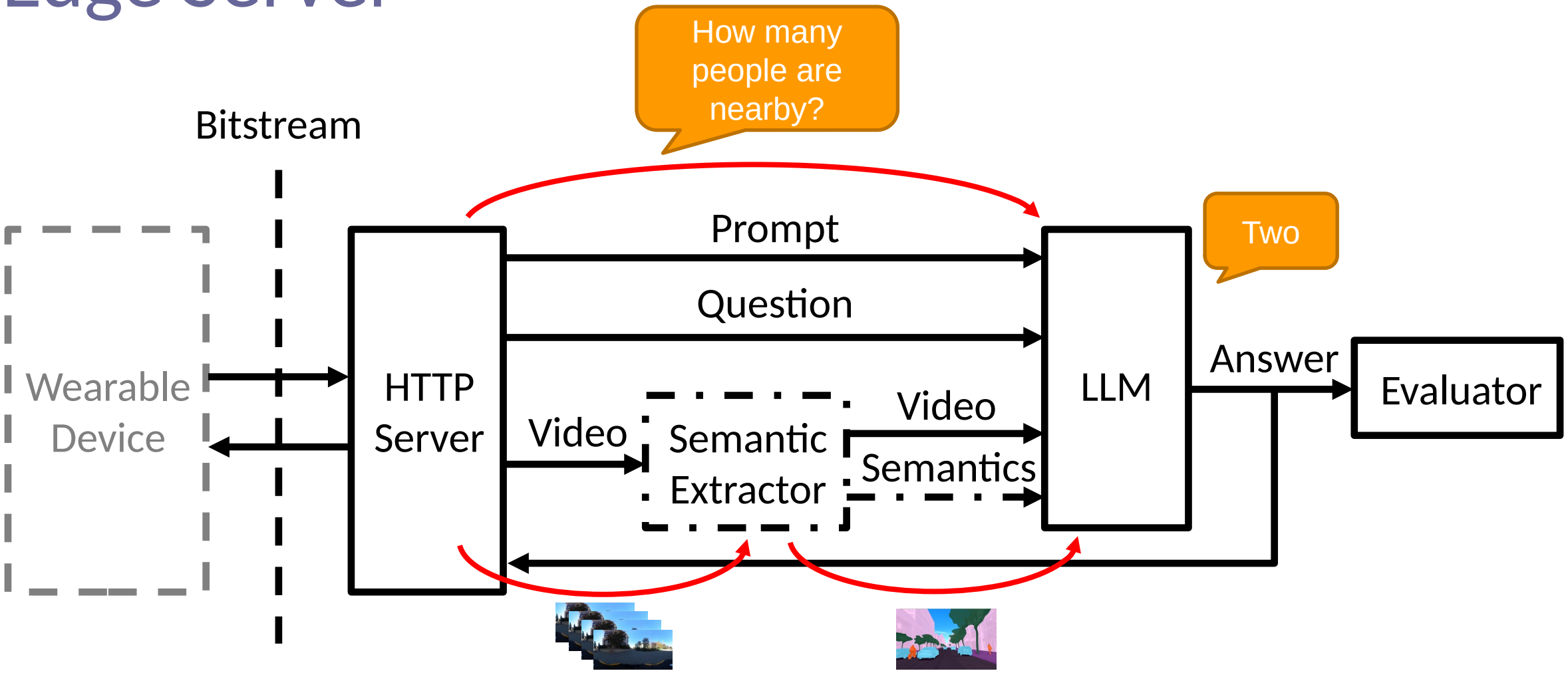
Outline

- Motivation
- Related Work
- System Architecture
- Research Problem
- Experiments
- Conclusion & Future Work

Wearable Device



Edge Server

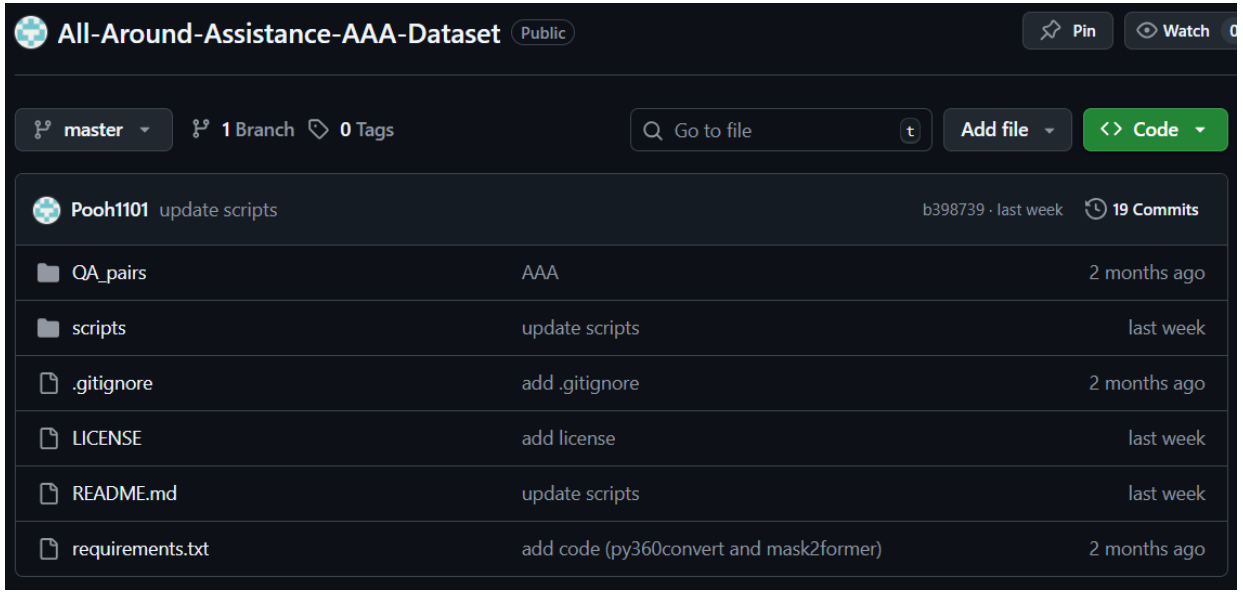

How many people are nearby?
Bitstream
Prompt
Question
Two
Wearable Device
HTTP Server
Video
Semantic Extractor
Video
Semantics
LLM
Answer
Evaluator

Outline

- Motivation
- Related Work
- System Architecture
- Research Problem
- Experiments
- Conclusion & Future Work

Challenge 1: Lack of Dataset

- We want to evaluate how camera **AoV** impacts the assistant
  - Lack of relevant datasets
- Solution: Construct a new dataset
  - Our approach is to **add Q&A pairs** to an existing 360° video dataset



Relevant Datasets

Dataset	Temporal Dynamic	360° AoV	Visually Impaired	Publicly Available	No. Seq.	Total Durations(m)
VizWiz [14]	✗	✗	✓	✓	–	–
EgoK360 [4]	✓	✓	✗	✓	127	1397
EgoBlind [37]	✓	✗	✓	✓	1329	886
VIEW-QA [33]	✓	✓	✓	✗	1030	600
AAA (Ours)	✓	✓	✓	✓	184	1153

- VizWiz [1]: Dataset based on static images
- EgoK360 [2]: Not designed for the visually impaired
- EgoBlind [3]: Cannot be used to compare different AoVs
- VIEW-QA [4]: Not publicly available

VizWiz

EgoK360

EgoBlind

1. Danna Gurari, et al. 2018. Vizwiz GrandChallenge: Answering Visual Questions from Blind People. *In Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). 3608–3617.*
2. Keshav Bhandari, et al. 2020. Egok360: A 360 Egocentric Kinetic Human ActivityVideo Dataset. *In Proc. of the IEEE International Conference on ImageProcessing (ICIP). 266–270.*
3. Jun Xiao, et al.2023. EgoBlind: Egocentric Video Question Answering for Blind As-sistance. *In Proc. of the IEEE/CVF Conference on Computer Vision andPattern Recognition (CVPR). 14000–14009.*
4. Inpyo Song, et al. 2024.Video Question Answering for People with Visual Impairments Usingan Egocentric 360-degree Camera. *arXiv preprint arXiv:2405.19794(May 2024)*

GND Dataset [1]

- We needed a dataset with **360° video sequences**
  - 184 video sequences
  - Captured from 9 universities
- We downsampled 360° videos to
  - 70°, 110°, and 180° ← to quantify the implications of AoV

1. Liang, et al. GND: Global Navigation Dataset with Multi-Modal Perception and Multi-Category Traversability in Outdoor Cam-pus Environments. arXiv:2409.14262

Samples from the GND dataset [1]

American University

Catholic University of America

George Mason University

Georgetown University

George Washington University

Marymount University

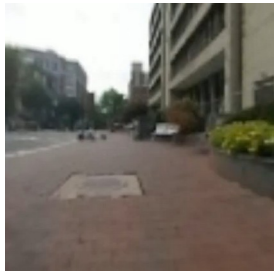
Northern Virginia Community College

University of the District of Columbia

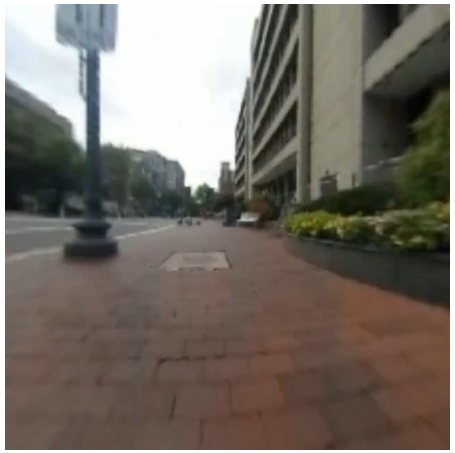
University of Maryland

1. Liang, et al. GND: Global Navigation Dataset with Multi-Modal Perception and Multi-Category Traversability in Outdoor Cam-pus Environments. arXiv:2409.14262

Sample Videos with Diverse AoVs



70°



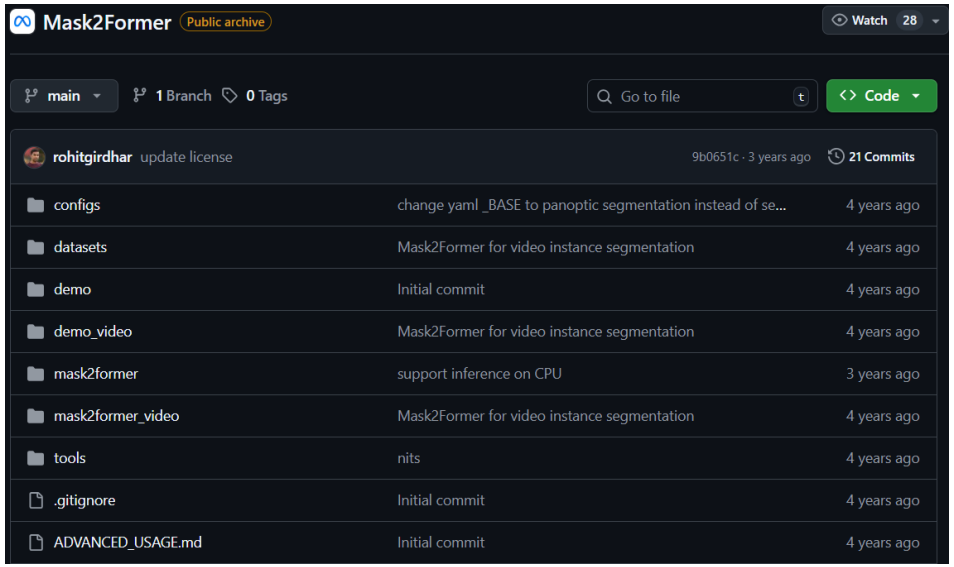
110°

180°

360°

Mask2Former [1]

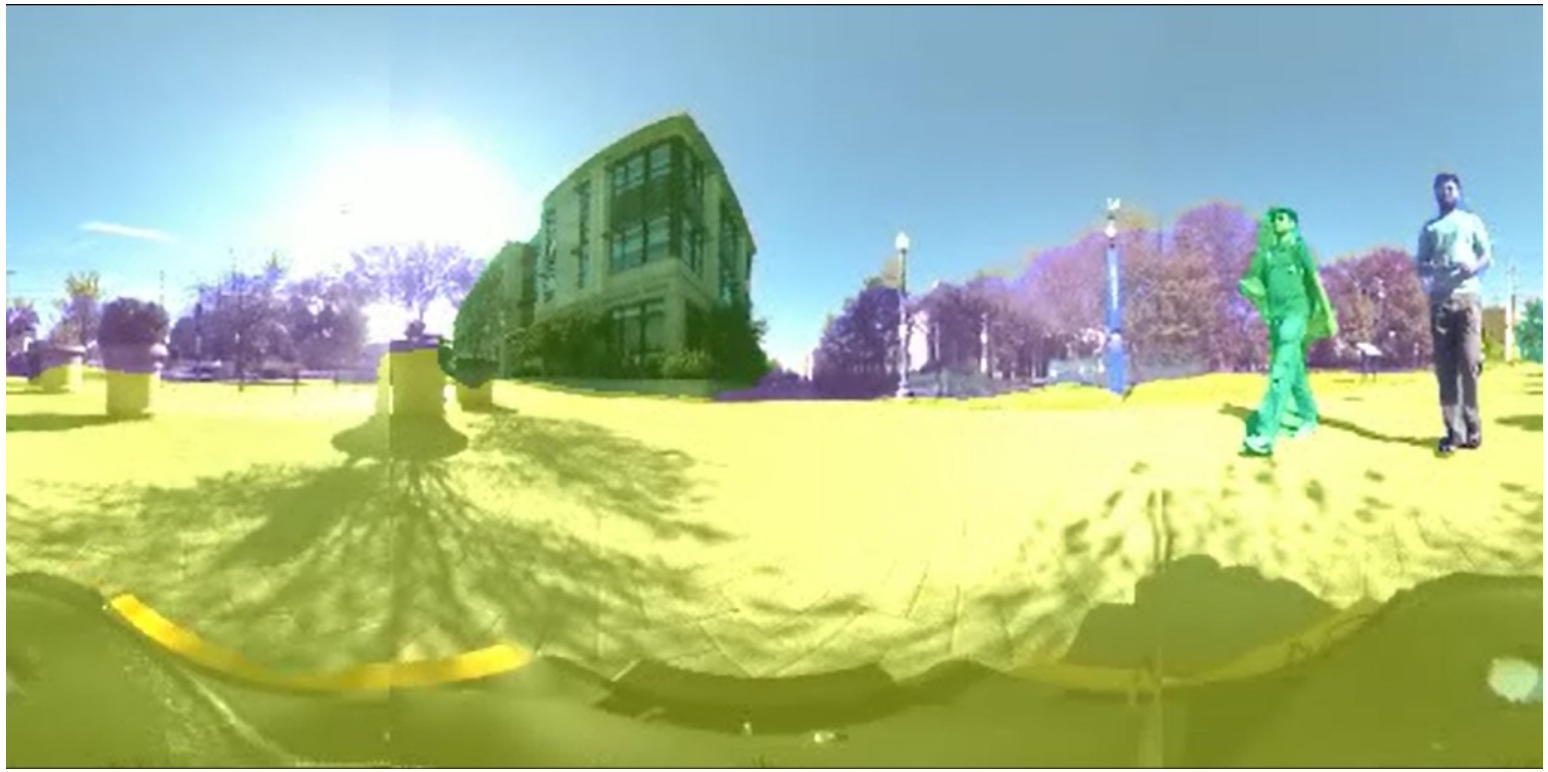
- We use **Mask2Former** for our **semantic analysis**
- Assigns every pixel in a frame to either
  - Objects (e.g. car, person)
  - Regions (e.g. road, pavement)



1. Bowen Cheng, et al. Masked-Attention Mask Transformer forUniversal Image Segmentation. *In Proc. of the IEEE/CVF Conference onComputer Vision and Pattern Recognition (CVPR). 1290–1299.*

Sample of Mask2Former

We preserved detailed annotations for each detected object.



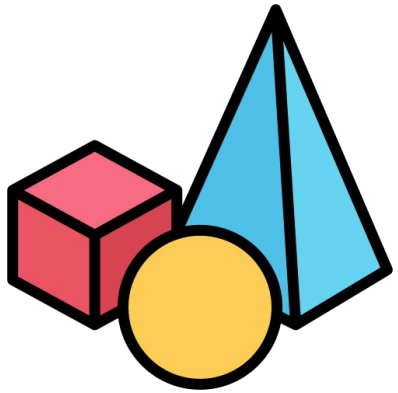
Direction

- We divide 360° frame into four main directional regions
  - front
  - left
  - right
  - back

- When an object is detected, we determine which region the object's mask has the maximum overlapping area with

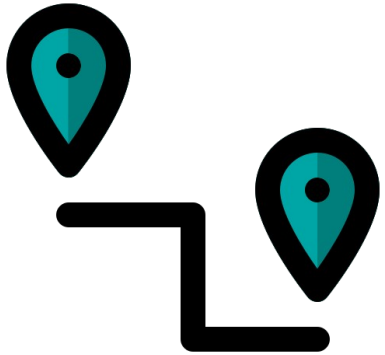
Q&A Categories

Inspired by earlier datasets for visually impaired individuals, we divide our Q&A into **four categories:**

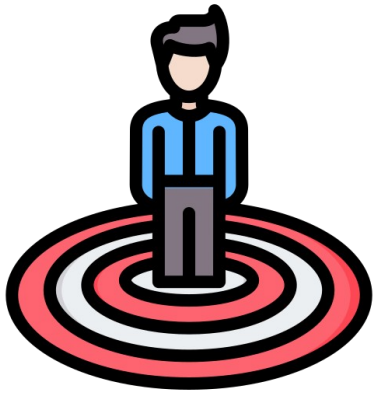


Object

Safety



Navigation



Surrounding

Templates & Triggers

We designed two Q&A triggering mechanisms

- Object-driven
  - employed whenever a new object appear
- Time-triggered
  - created once every fixed number of frames

All questions are Multiple-Choice Questions

Question	Type
Object	
How many o are there?	Object
Is the o on my p?	Object
Safety	
Should I avoid the o ahead?	Object
Is an o blocking the r?	Object
Are there obstacles on the r?	Object
Is it safe to walk on the r ahead?	Object
What is blocking the r ahead?	Time
Navigation	
What is the relative position of the o?	Object
Am I still walking on the r?	Object
Which region should I follow?	Object
Is the o visible from here?	Object
Which of the following is closest in front of me?	Time
Which of the following is the farthest in front of me?	Time
Surrounding	
Is there an o in front of me?	Object
Which object is nearby?	Time
Which object is on my p?	Time

Q&A Pairs

Our Q&A pairs record

- Question
- Answer
- Options
- Template's ID
- Category
- Trigger type

ID: 16, Time Trigger, Surrounding

Q: Which object is on my right?

A: person B: couch

C: traffic light D: dog

ID: 8, Object Trigger, Navigation

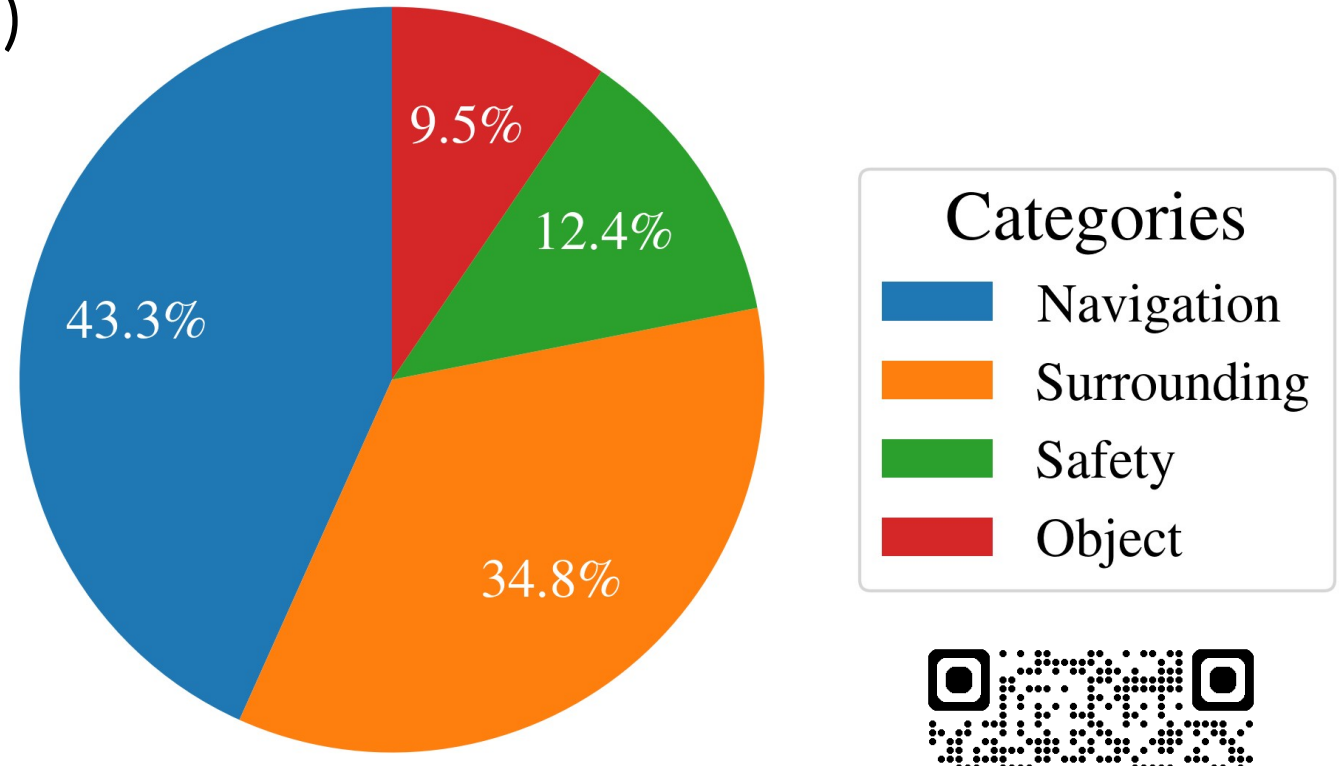
Q: What is the relative position of the person?

A: left B: back

C: front D: right

All Around Assistance (AAA) Dataset

- **184** video sequences
  - 1,153 minutes (~19 hours)
- **4,438** total Q&A pairs
  - Navigation: 1,940
  - Surrounding: 1,560
  - Safety: 556
  - Object: 426



- Quality: **512×256** resolution at **15 fps**

Challenge 2: Limited Angle of View (AoV)

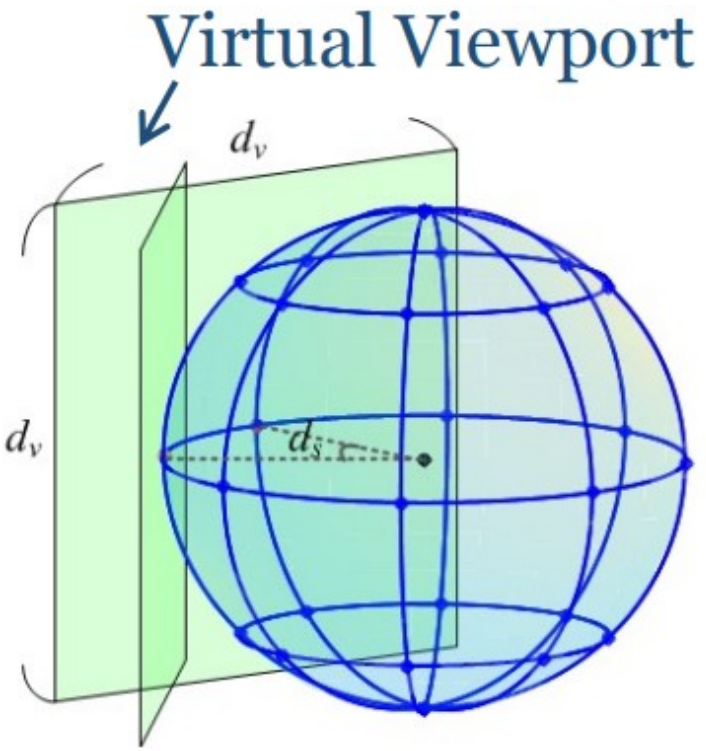
- Popular smart glasses (e.g., Meta Ray-Ban) are constrained by a **limited camera AoV**
- Solution: using **360° cameras** to capture the full surrounding scene
  - Introduces **image distortion** and **discontinuous objects**

360° Video Distortion

- Equirectangular projection causes shape distortion near the poles of a 360° video

Multi-tile Approach

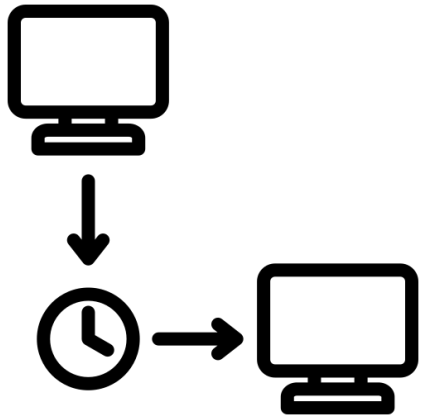
- To avoid 360° distortion, we adopt the multi-tile approach from Fan et al. [1] to get accurate semantic annotations
  - Basically, project into multiple virtual viewports
  - Next, perform semantic segmentation
  - Then, merge all identified annotations
- Key parameters:
  - d_v: The **AoV** of each individual tile
  - d_s: The **overlapping angles** between two adjacent tiles



1. Wen-Chih Lo, Ching-Ling Fan, et al. 2017. 360 Video Viewing Dataset in Head-Mounted Virtual Reality. In Proc. of the ACM International Conferenceon Multimedia Systems (MMSys). 211–216.

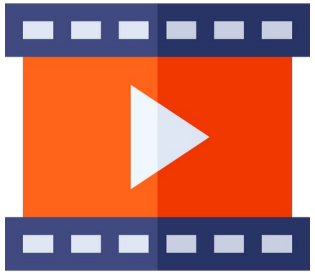
Challenge 3: Computation & Latency

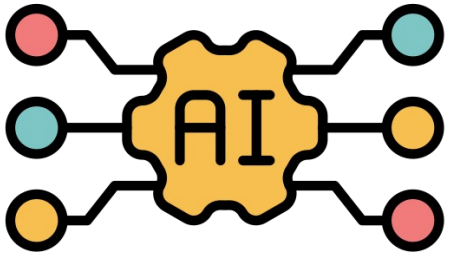
- Smart glasses have limited computing power
- Solution: offload heavy computations (LLMs) to edge servers or smartphones
- Research problems:
  - Analyze the Accuracy vs. Latency trade-off
  - Compare different LLMs
  - Test different access networks



Offload LLM and Semantic Extractor

- Wearable Device
  - Focuses on video capture and **H.264 compression** to minimize bandwidth usage
- Edge Server
  - Handles heavy computations: **Large Scale LLMs** and **Semantic Segmentation**





Outline

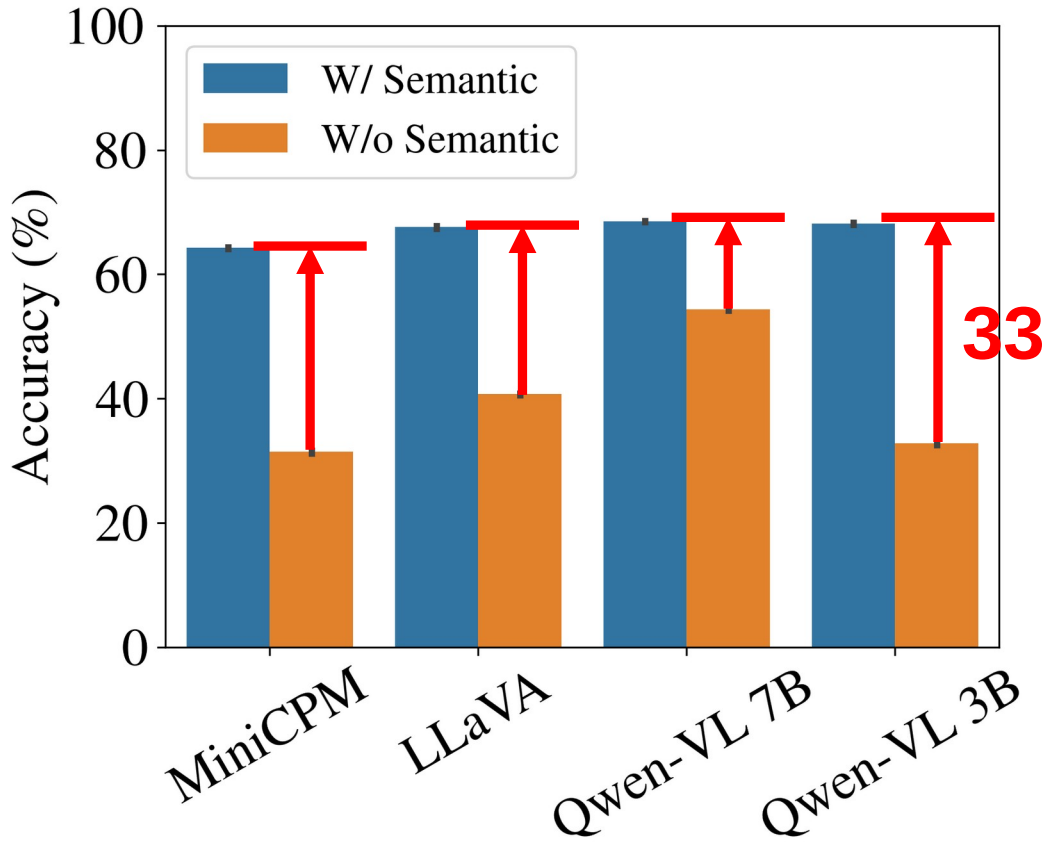
- Motivation
- Related Work
- System Architecture
- Research Problem
- Experiments
- Conclusion & Future Work

Experimental Setup

- Dataset: AAA Dataset
- Hardware Setup:
  - Wearable Device: A laptop
  - Edge Server: An NVIDIA RTX-3090 GPU workstation
- Considered (vision) LLMs:
  - MiniCPM-O 2.6 8B
  - LLAVA-OneVision 7B
  - Qwen2.5-VL 3B/7B
- Metrics:
  - **Accuracy:** The fraction of correctly answered questions
  - **Delay:** Processing, Transmission, Semantic, and Inference delays

Semantic Extractor Significantly Improves Accuracy

- Semantic extractor improves accuracy by **up to 33.77%**, with an average gain of **26.76%** across all LLMs

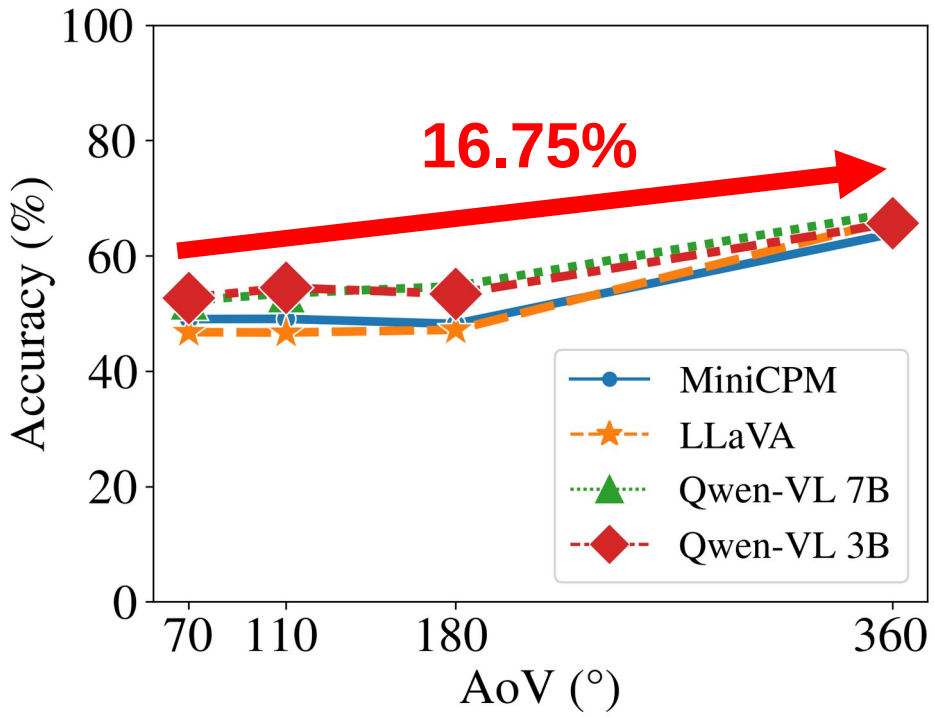


Reliance on LLMs alone for visual processing is insufficient

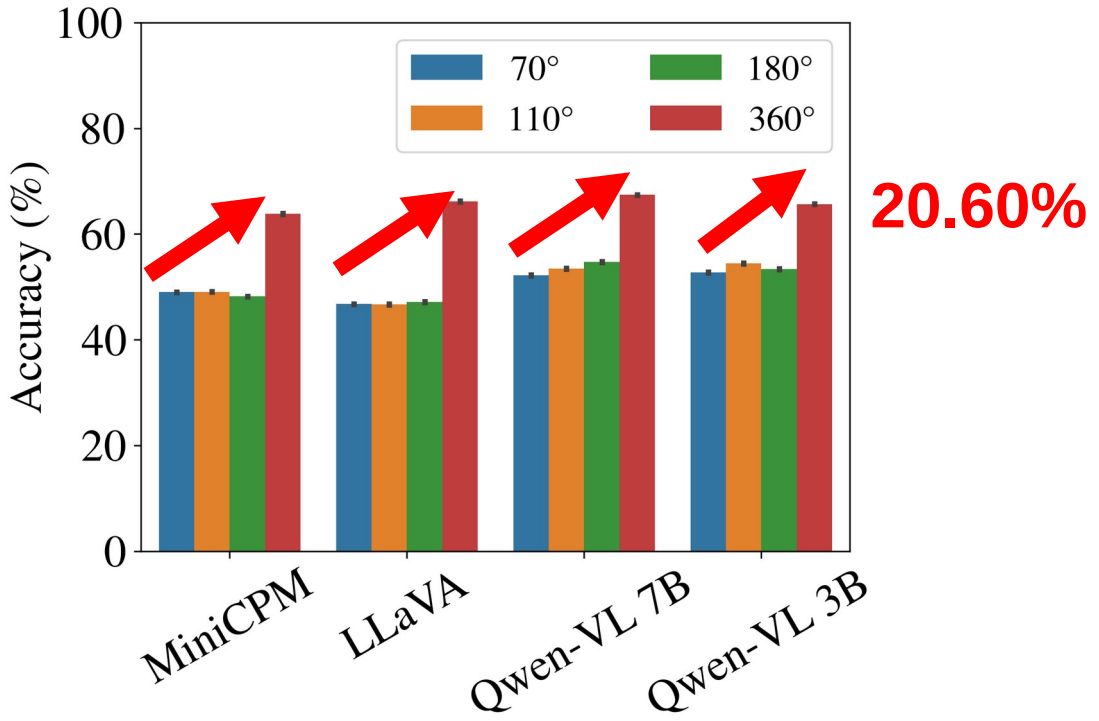
AoV is Critical

- 360° achieve an average accuracy of **66.17%**
- Improvement of **20.60%** over the 70° cameras

AoV creates a physical bottleneck, imposing a hard limit on accuracy regardless of the model's capability



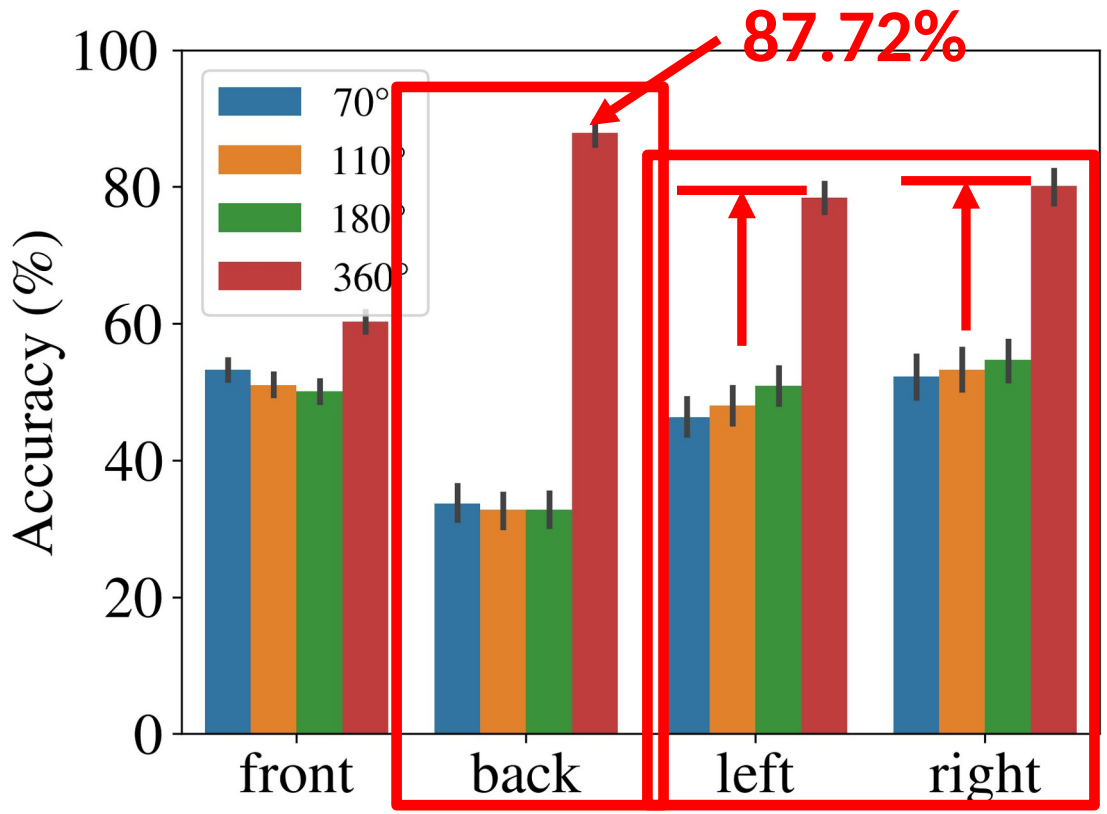
American University



All univ.

Why 360° Cameras are Needed

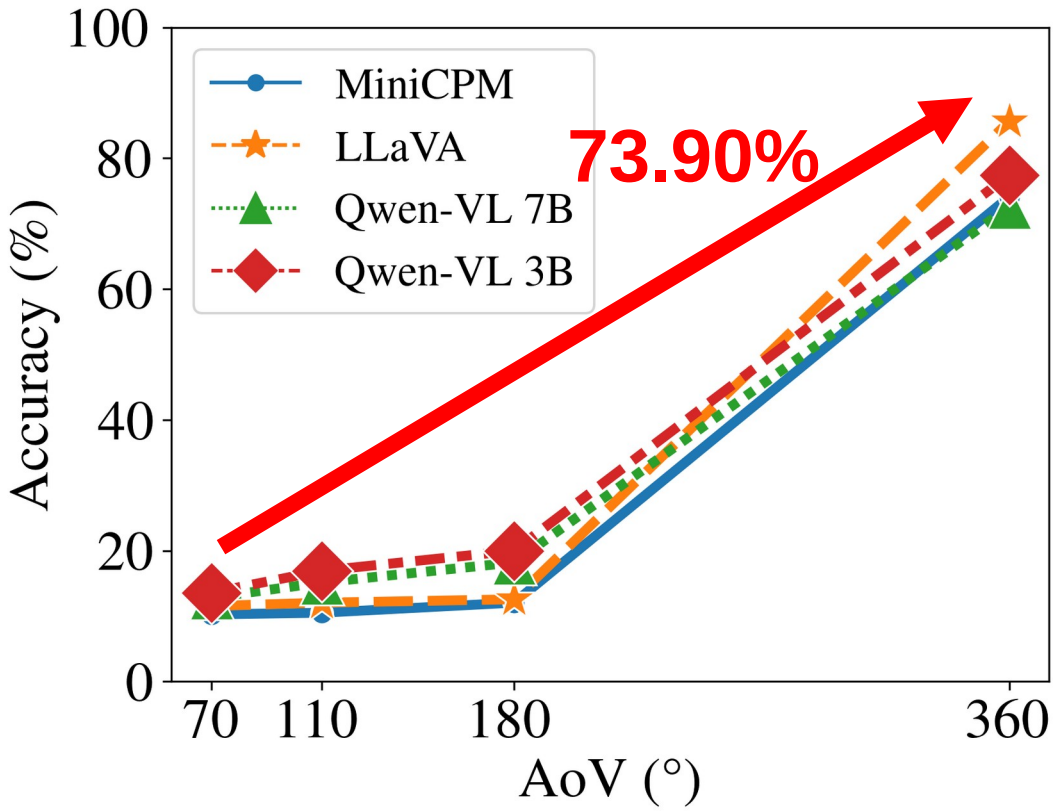
- 360° achieves the highest accuracy
  - Accuracy in “back” direction reached **87.72%**
- Demonstrates that 360° cameras are essential



360° cameras provide consistent reliability across all directions

Questions Related to Different Directions

- 360° cameras achieve an accuracy as high as **85.64%**
- Compared to 70° cameras, accuracy improvement of up to **73.90%**



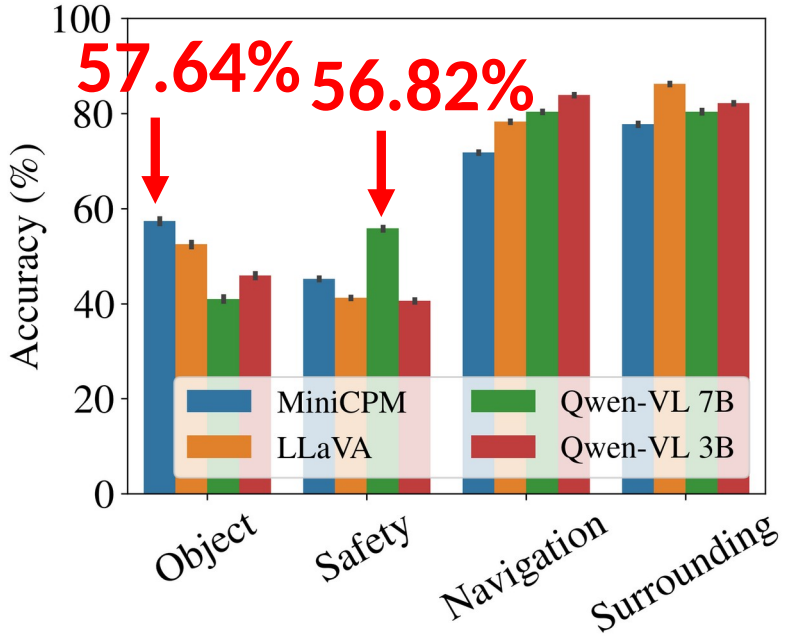
How many [objects] are

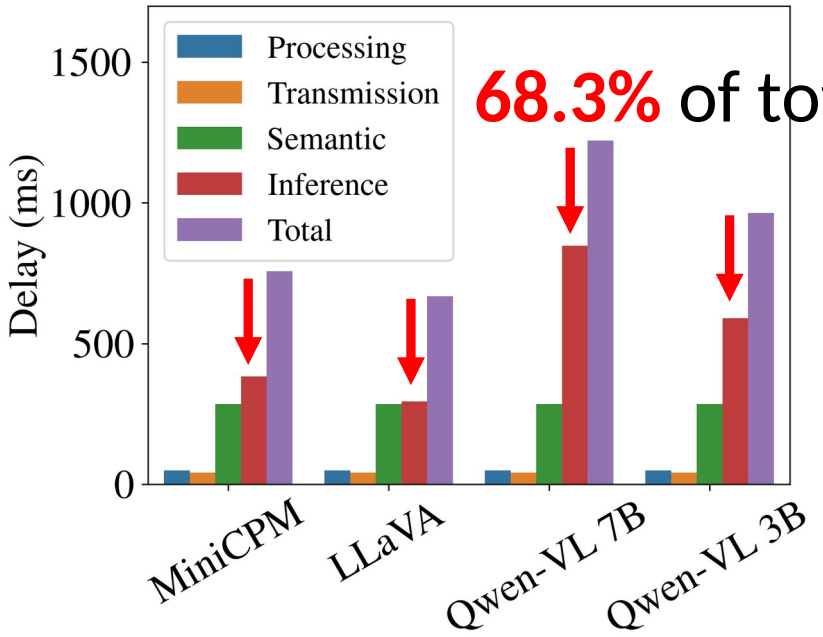
Fragmented vision directly leads to a breakdown in spatial understanding

Different LLMs Performance

Model selection involves trade-offs, as no single model dominates all tasks

- Different LLMs excel at different question types
  - Object: MiniCPM achieved the highest accuracy of **57.64%**
  - Safety: Qwen achieved the highest accuracy of **56.82%**
- Inference delay dominates the total delay
  - Qwen-VL 7B inference takes **68.3%** of total delay





68.3% of total

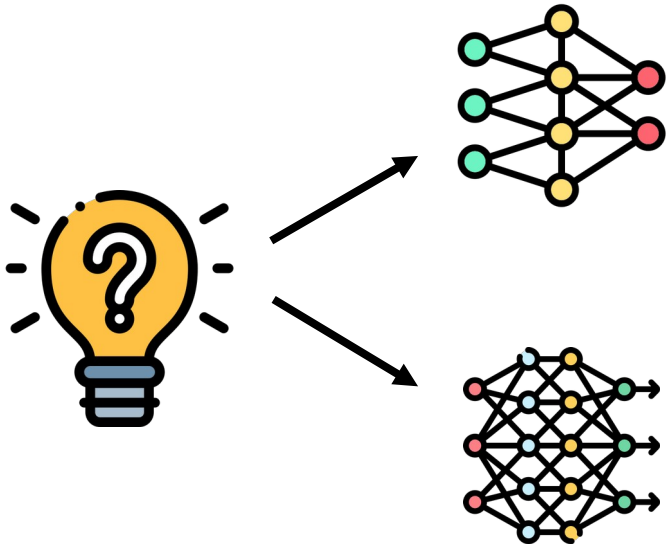
Outline

- Motivation
- Related Work
- System Architecture
- Research Problem
- Experiments
- Conclusion & Future Work

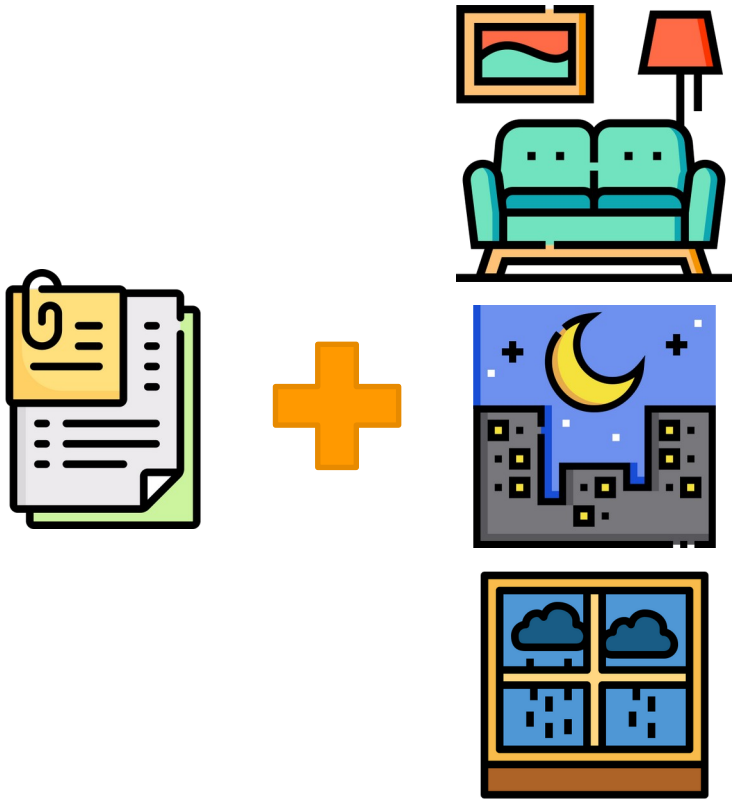
Conclusion

- We studied the integration of diverse-AoV smart glasses with LLMs for visually impaired assistance
- **We constructed AAA**, the first public dataset to benchmark the impact of different AoVs in LLM-based assistants
- Key Findings:
  - Semantic Extractor Helps: improves accuracy by **up to 33.77%**
  - AoV is Critical: 360° cameras improve accuracy by **20.60% on average**
  - Model Selection Matters: different LLMs excel at different question types

Future Work



Dynamic Model Selection



Dataset Enhancement

Real-world User Study

Thank you for listening!

Thanks for the help of Prof. Hsu, Yuan-Chun Sun, Yee-Nam Wong

Publications:

1. Z. Lee, Y. Sun, Y. Wong, and C. Hsu. Exploring LLM-based assistants with smart glasses for the visually impaired. In Proc. of International Workshop on Immersive Computing at the Edge (ImmerEdge), December 2025.