ASEC
2025
Conference

The 6th International Electronic Conference on Applied Sciences

09–11 December 2025 | Online

MDPI

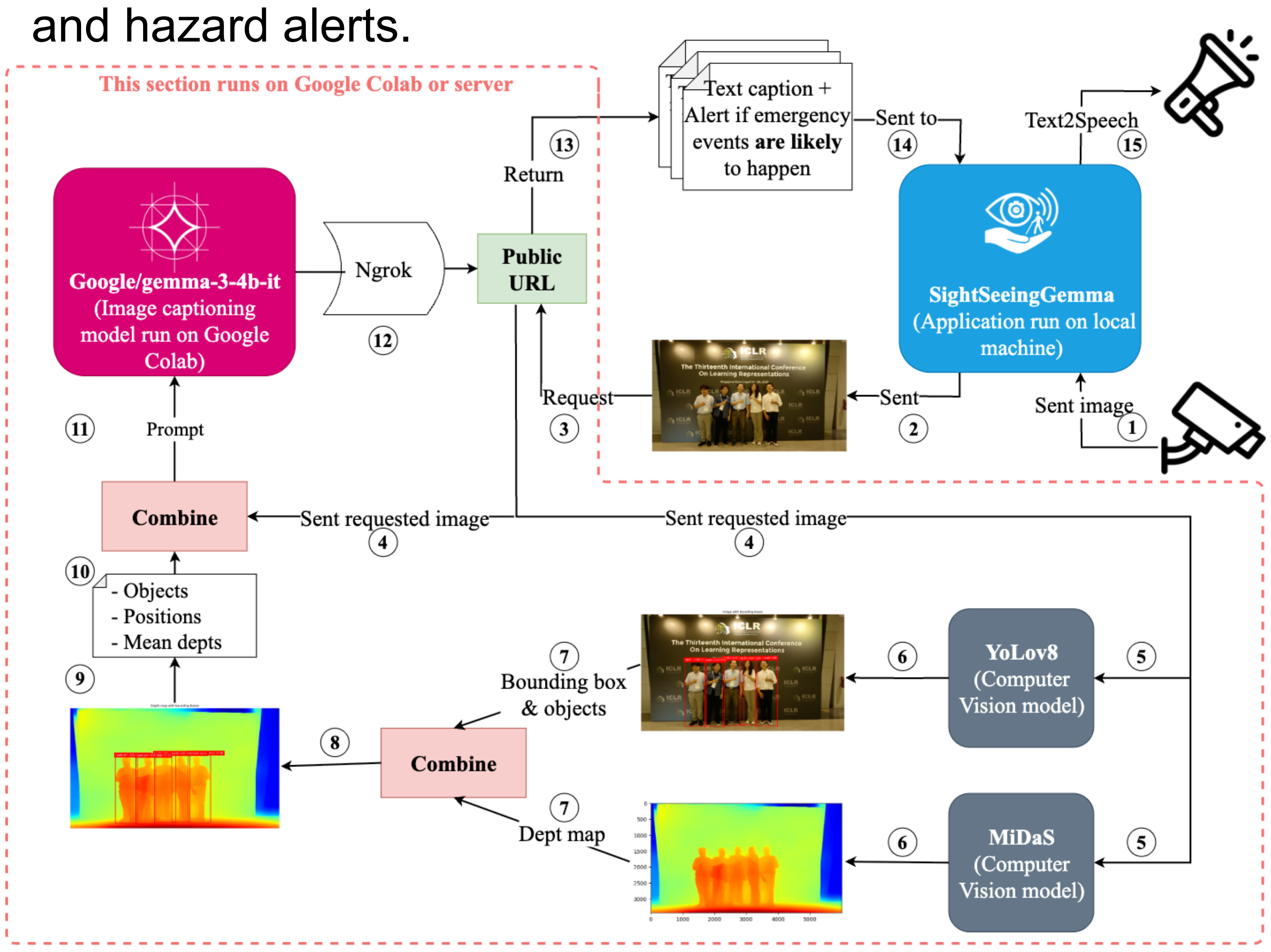
SightSeeingGemma: Enhancing Assistive AI for the Visually Impaired via Object Detection and Monocular Depth Estimation with Language-Based Scene Understanding

ANH DINH TRAC DUC, TAI TA TIEN, MINH TUE HUA, TRI TRINH HUU, THO QUAN

Unlimited Research Group of AI (URA), Ho Chi Minh City University of Technology (HCMUT), Vietnam National University – Ho Chi Minh City (VNU-HCM)

INTRODUCTION & AIM

In Vietnam, assistive tools often lack cultural and language adaptation. SightSeeingGemma fills this gap by combining object detection, depth estimation, and a vision-language model to provide real-time Vietnamese scene descriptions and hazard alerts.


This section runs on Google Colab or server
Google/gemma-3-4b-it (Image captioning model run on Google Colab)
Ngrok
(12)
Public URL
Return
(13)
Text caption + Alert if emergency events are likely to happen
Sent to
(14)
Text2Speech
(15)
SightSeeingGemma (Application run on local machine)
Sent image
(1)
Sent
(2)
Request
(3)
(11)
Prompt
Combine
Sent requested image
(4)
Sent requested image
(4)
(10)
- Objects
- Positions
- Mean depts
(9)
(8)
Combine
(7)
Bounding box & objects
(7)
Dept map
(6)
YoLov8 (Computer Vision model)
(5)
(6)
MiDaS (Computer Vision model)
(5)

METHOD

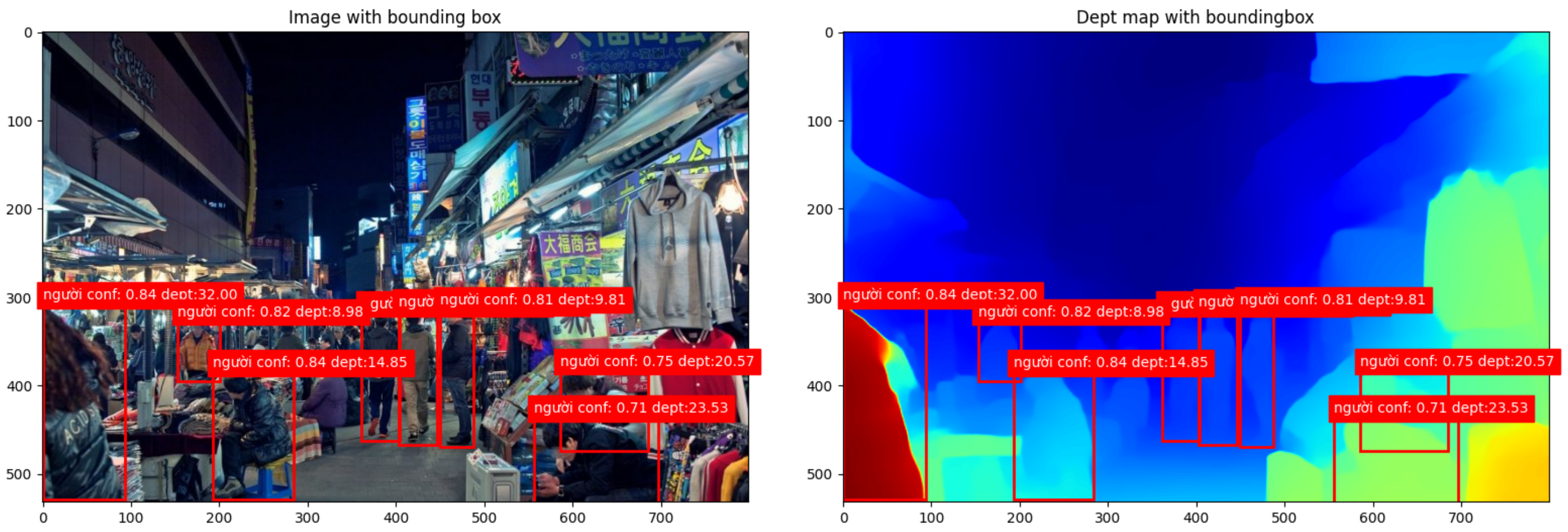


Image with bounding box
Dept map with boundingbox
0
100
200
300
400
500
0
100
200
300
400
500
600
700
người conf: 0.84 dept:32.00
người conf: 0.82 dept:8.98
gư người người conf: 0.81 dept:9.81
người conf: 0.84 dept:14.85
người conf: 0.75 dept:20.57
người conf: 0.71 dept:23.53

Each detected object is defined as:

$$o = (c, s, B, d)$$

Where:

- c: class label (e.g cat, car…)
- s: onfidence score from the detector.
- B: bounding box coordinates.
- d: mean depth value within the bounding box.

The mean depth d is computed from the depth map $D(x, y)$ as:

$$d = \frac{1}{|B|} \sum_{(x,y)\in B} D(x, y)$$

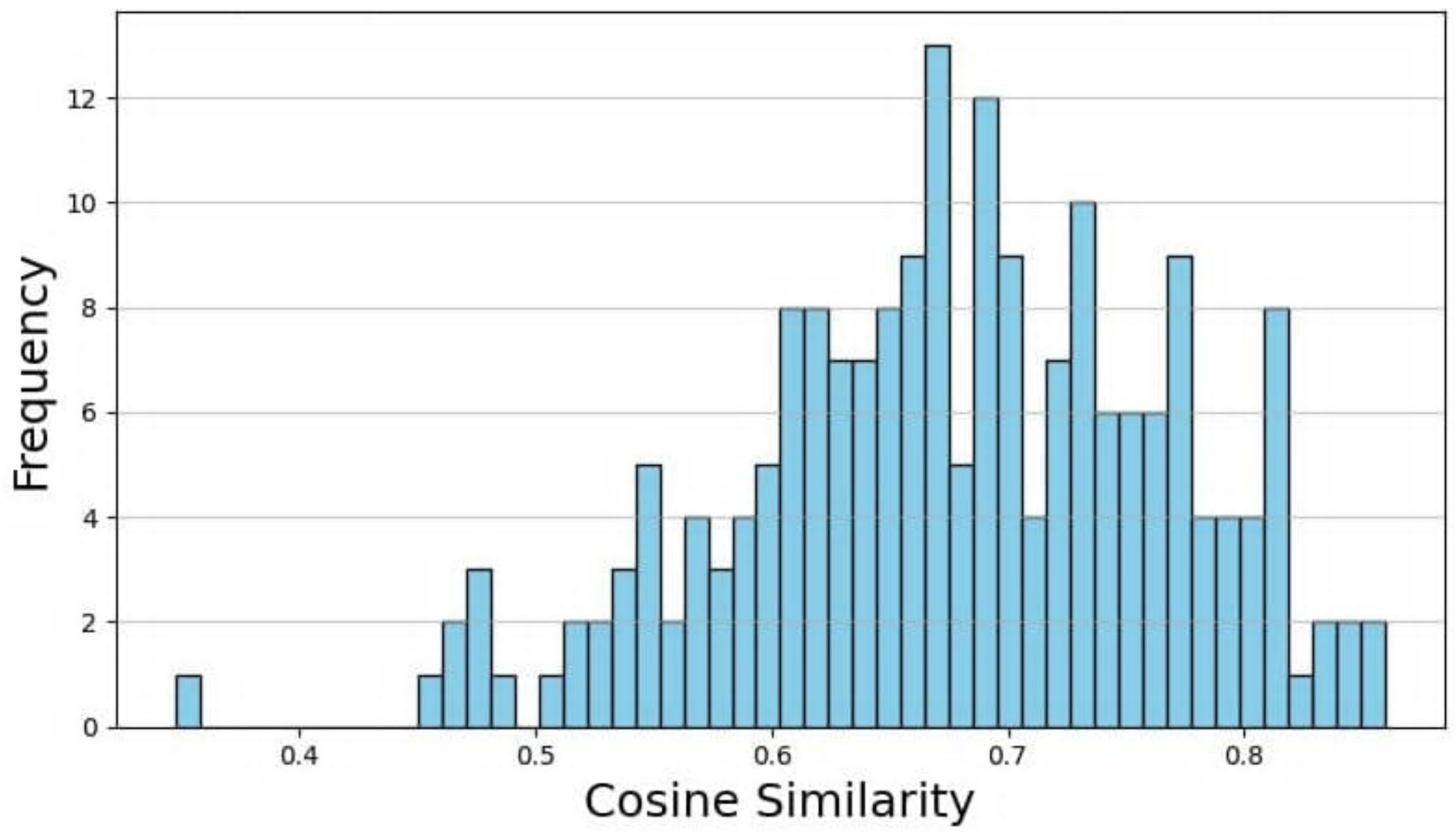
These are fused into a prompt:

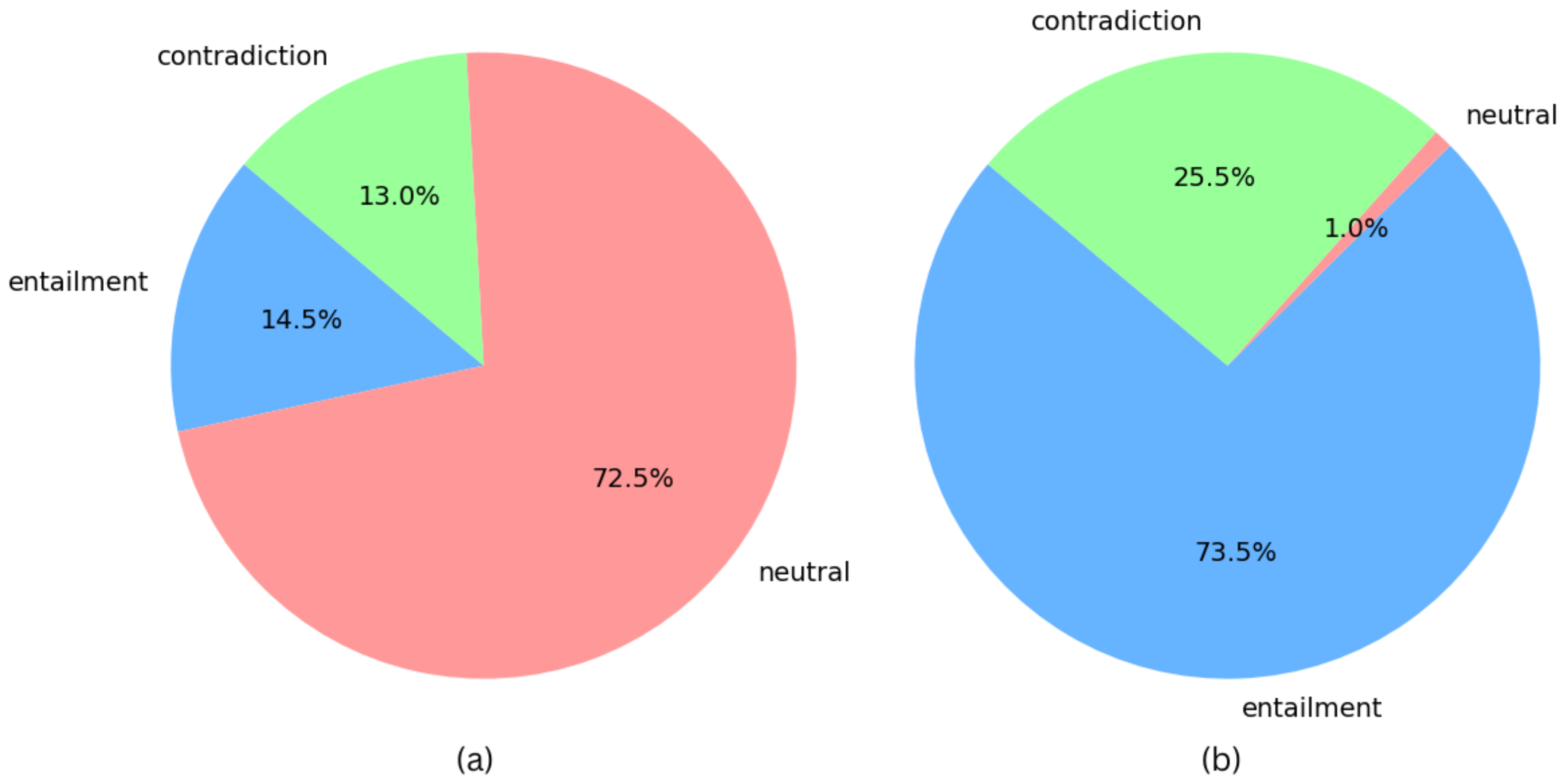
$$P \quad = \text{"There is a "} + \sum_{i=1}^{n} (c_i + \text{" at depth "} + d_i)$$

This prompt is translated into Vietnamese and fed to Gemma-3-4b-it, which generates spoken descriptions and hazard warnings via TTS.

RESULTS & DISCUSSION

SightSeeingGemma was evaluated on 200 image-description pairs, including 42 with hazards and 63 from Vietnam, with an average response time of 15–17 seconds. To assess semantic similarity between model- and human-written descriptions, Vietnamese SBERT embeddings and cosine similarity were used. The results show over 95% of scores above 0.5. NLI analysis confirmed 73.5% entailment in hazard warnings, highlighting the system's reliability in safety-critical scenarios.





(a) (b)

Chart (a) shows that most general descriptions are classified as neutral, suggesting the model captures the overall meaning but may miss minor details. In contrast, chart (b) shows that 73.5% of hazard warnings are entailments, confirming the system's high reliability in safety-critical situations.

CONCLUSION

In conclusion, SightSeeingGemma combines vision and language to support visually impaired users, showing strong semantic accuracy and reliable hazard alerts. Despite current latency, it shows promise for real-world use. Future work aims to reduce response time and enable on-device processing.

https://sciforum.net/event/ASEC2025