

# Is it safe to cross? Interpretable Risk Assessment with GPT-4V for Safety-Aware Street Crossing

Hochul Hwang<sup>1</sup>, Sunjae Kwon<sup>1</sup>, Yekyung Kim<sup>1</sup>, and Donghyun Kim<sup>1</sup>

**Abstract**—Safely navigating street intersections is a complex challenge for blind and low-vision individuals, as it requires a nuanced understanding of the surrounding context – a task heavily reliant on visual cues. Traditional methods for assisting in this decision-making process often fall short, lacking the ability to provide a comprehensive scene analysis and safety level. This paper introduces an innovative approach that leverages vision-language models (VLMs) to interpret complex street crossing scenes, offering a potential advancement over conventional traffic signal recognition techniques. By generating a safety score and scene description in natural language, our method supports safe decision-making for blind and low-vision individuals. We collected crosswalk intersection data that contains multiview egocentric images captured by a quadruped robot and annotated the images with corresponding safety scores based on our predefined safety score categorization. Grounded on the visual knowledge, extracted from images and text prompts, we evaluate a VLM for safety score prediction and scene description. Our findings highlight the reasoning and safety score prediction capabilities of the VLM, activated by various prompts, as a pathway to developing a trustworthy system, crucial for applications requiring reliable decision-making support.

## I. INTRODUCTION

Safety-aware scene understanding and risk assessment are critical for effective robot navigation, particularly in safety-critical contexts such as guiding blind or low-vision (BLV) individuals to cross a street. While strides in autonomous vehicle technology have led to the release of publicly available large-scale datasets for training deep learning models [1], [2], [3], their direct application poses challenges. This stems from the inherent differences between road and sidewalk environments [4]. Moreover, such vehicle-centric datasets often miss out on elements vital for blind navigation, such as tactile pavings.

Our research introduces a novel framework for assisting BLV individuals in safely navigating crosswalk intersections. Unlike traditional approaches, primarily focused on traffic light signal recognition, we leverage a vision-language model (VLM) to enhance contextual scene understanding. Recently, there has been research enabling zero-shot transfer for various computer vision tasks, including classification [5], segmentation [6], and multimodal reasoning [7], utilizing transformers [8], [9] and large language models (LLMs) such as ChatGPT and GPT-4 [10]. Here, we utilize a vision-language model for a more comprehensive and nuanced

understanding of safe street crossing. This advancement potentially enables the provision of safety-critical information in natural language, significantly enhancing the decision-making process for BLV individuals.

The essence of our approach lies in the collection and usage of a custom outdoor street crossing dataset, uniquely gathered via the Unitree Go1 quadruped robot. This dataset, annotated with predefined safety scores, serves as a cornerstone in the enhancement of prompts for VLMs. By integrating *visual knowledge* – encompassing object detection bounding boxes, segmentation masks, and optical flow – with textual information, we explore the capability of the VLM, with a particular focus on GPT-4V<sup>1</sup>. This investigation aims to evaluate the model's proficiency in accurately interpreting the intricacies of crosswalk environments, thereby informing decisions for safe street crossing.

Our findings highlight the effectiveness of the VLM and its potential for making safety-aware decisions, advancing the domain of autonomous robot navigation in pedestrian environments. This contribution is particularly geared towards developing a robotic mobility aid for BLV individuals, offering implications for enhancing their autonomy and mobility. We summarize our contributions as follows:

- Data collection of multiview egocentric images in a crosswalk environment for street crossing.
- Safety score categorization on egocentric images.
- Evaluation of GPT-4V for safety score prediction and scene description using visual and text prompts.

## II. RELATED WORK

### A. Assistive Technology for Blind Street Crossing

BLV people primarily rely on their residual senses, such as hearing traffic sounds to decide when to cross streets. However, these methods do not guarantee accurate alignment with pedestrian crossings [11]. Crossing streets safely poses a significant challenge for blind individuals, who rely primarily on auditory cues [12]. Studies reveal that blind pedestrians face a higher risk and typically encounter significant delays compared to sighted pedestrians, particularly in complex traffic settings such as roundabouts [13].

Recognizing safety is paramount for street crossing, yet it may be equally crucial to provide the robot's user with a detailed scene description. Unlike traditional travel aids such as white canes and guide dogs, which do not provide feedback to the user, recent studies explore the potential of scene description through advancements in computer

<sup>1</sup> Manning College of Information and Computer Sciences, University of Massachusetts Amherst, 140 Governors Dr, Amherst, MA 01002, U.S. {hochulhwang, sunjaekwon, yekyungkim, donghyunkim}@cs.umass.edu

<sup>1</sup><https://platform.openai.com/docs/guides/vision>

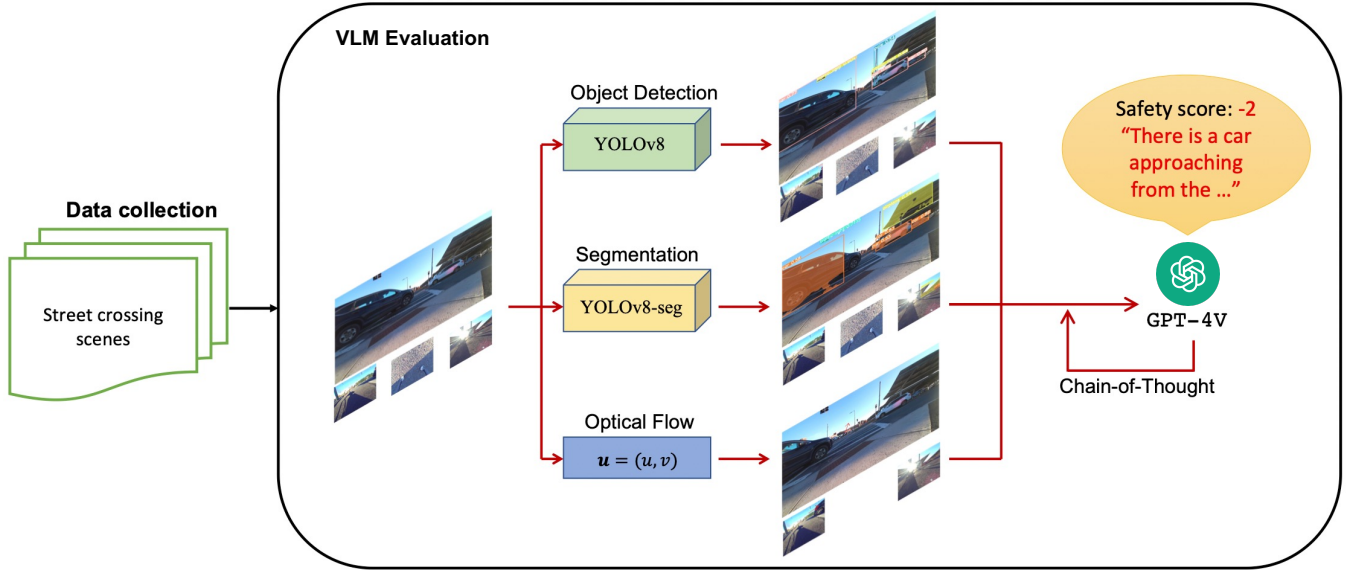


Fig. 1. **Illustrative concept of VLM evaluation pipeline for safety-aware street crossing.** Visual knowledge of object detection bounding boxes, segmentation masks, and optical flow is extracted from the robot’s multiview egocentric images. This is then provided to the VLM along with the text prompts. The VLM outputs the safety score and scene description.

vision [14], [15], [16]. Hoogsteen et al. [17] investigate the factors that influence the importance of describing specific objects in the environment and explore effective methods for communicating these descriptions. Their study also compares the preferences of early and late blind participants regarding object descriptions. More recent vision assistive mobile applications such as OKO [18] and Be My Eyes [19] show their practical applications in conveying useful information to the user. Such research aligns with the overarching theme of our paper, which focuses on integrating vision language models in robots to function as mobility aids, moving beyond the capabilities of traditional travel aids.

### B. Vision Language Models

With the advent of ChatGPT [20], generative large language models (LLMs) have been increasingly applied to complex inferential tasks beyond the reach of natural language processing models. These tasks include solving mathematical problems [21], definition modeling [22], and navigating domains that require expert knowledge (e.g., medical and financial sectors [23], [24]). In particular, models such as GPT-4 have demonstrated superior annotation quality compared to untrained human annotators (e.g., Amazon Mechanical Turk [25]). This opens up the possibility of using LLMs as an evaluator, offering a cost-effective alternative to human evaluations with enhanced instruction-following capabilities [26], [27], [28], [29].

Current research on LLMs is expanding their application scope beyond textual analysis to include other modalities such as vision and speech [30], [31], [7], [32]. We extend this exploration by incorporating the visual data recognition capabilities of VLMs into the annotation process. Demonstrating exceptional capability, our study employs GPT-4V to eval-

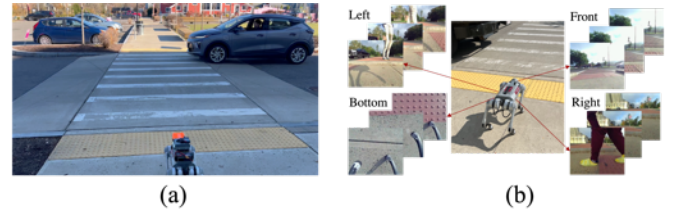


Fig. 2. **Data collection using a quadruped robot.** (a) The remote controlled robot collected RGB and depth data in crosswalk settings. One of the researchers is driving the car to simulate an unsafe scenario for street crossing. (b) Various onboard sensors enable egocentric multiview data.

uate the safety of crossing streets. This approach uniquely integrates various visual knowledge with text prompts, laying the groundwork for informed and safety-aware decision-making in street crossing scenarios.

## III. METHOD

In this section, we introduce our evaluation pipeline for risk assessment in a street crossing scenario, as illustrated in Fig. 1. Our framework consists of two major parts: visual knowledge extraction and evaluation with VLM using visual and text prompts.

### A. Visual Knowledge as Prompts

1) *Crosswalk data collection:* To evaluate the capabilities of the vision-language model, we collected multiview egocentric images of real-world crosswalk scenarios including situations that are not safe for pedestrians to cross the street as illustrated in Fig. 2. All researchers participated in data collection while one researcher remotely controlled the robot and another drove a vehicle to create realistic scenarios. We utilized three onboard stereo cameras to capture the right,

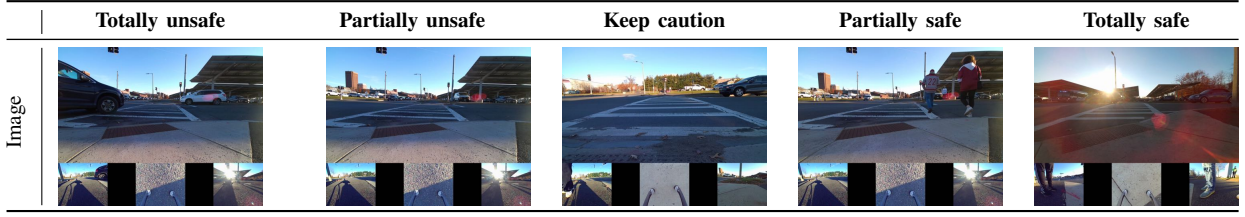


TABLE I  
VISUALIZATION OF EGOCENTRIC IMAGES FROM EACH SAFETY LEVEL.

Safety Score	Safe to cross	Moving car	Parallel traffic light	Pedestrian signal	Crossing pedestrian
Totally unsafe (-2)	✗	yes	all	all	all
Partially unsafe (-1)	✗	no, not visible	all	stop	all
Keep caution (0)	-	no, not visible	red, yellow, not visible	go	no, not visible
Partially safe (1)	✓	no, not visible	red, yellow, not visible	go, not visible	yes
		no, not visible	green	go	no, not visible
Totally safe (2)	✓	no, not visible	green	go, not visible	yes

TABLE II  
SAFETY SCORE CATEGORIZATION.

left, and bottom viewpoints of the robot, and one Azure Kinect camera to capture the front viewpoint. While depth information was also captured, our study focused exclusively on the RGB data. Data collection occurred at various times throughout the day to ensure a diverse representation of lighting and environmental conditions.

2) *Safety score categorization*: Safety score is categorized into five levels (see Table I) to be used as robot feedback. We employed a detailed approach to label image-safety score pairs, as outlined in Table II. This labeling was based on predefined rules, taking into account various critical factors necessary for ensuring safe street crossing. Specifically, the safety level was assigned based on (1) presence of moving vehicles toward the crosswalk: {yes, no, not visible}; (2) status of the parallel traffic light: {red, yellow, green, not visible}; (3) state of the pedestrian signal: {go, stop, not visible}; (4) and presence of pedestrians in the crosswalk: {yes, no, not visible}.

Contrary to previous studies that categorized street crossing as a binary decision [33], our method introduces a nuanced safety score,  $s \in \{-2, -1, 0, 1, 2\}$ , with the highest value indicating the safest situation for street crossing. To ensure accuracy and consistency in the labeling, our team meticulously reviewed each piece of data.

### 3) Visual knowledge extraction:

a) *Optical flow*: As a component of visual knowledge, optical flow was integrated to provide motion data, which is crucial for detecting moving objects, such as cars. We utilized the Lucas-Kanade [34] method for precise optical flow estimation, which enhances feature point tracking across consecutive frames. This method operates under the assumption that flow remains consistent in a pixel's local vicinity, solving optical flow equations for each pixel in this area. Our qualitative analysis revealed that employing the average of

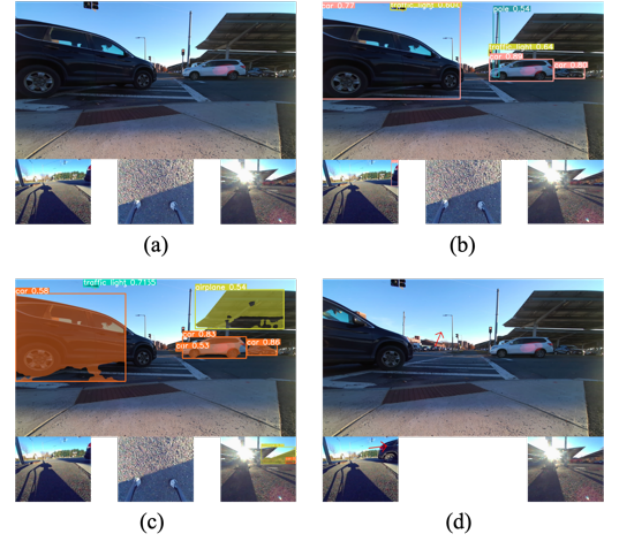


Fig. 3. **Visualization of visual knowledge.** (a) Raw multiview images contain the egocentric front (top patch), left (bottom left patch), bottom (bottom center patch), and right (bottom right patch) viewpoints of the robot. (b) Bounding boxes are added using an object detection algorithm. (c) Segmentation masks are added using an instance segmentation algorithm. (d) The average optical flow for each viewpoint (except for the bottom) is represented by a red arrow.

optical flows yields more effective visual cues for the VLM. This visually implies a shift from multiple arrows depicting individual flows to a single averaged arrow, supposedly offering a more concise and informative representation of an object's movement as shown in Fig. 3 (d).

b) *Segmentation mask*: We employed YOLOv8-Seg, a variant of YOLOv8, to produce segmentation masks for input images. This model incorporates CSPDarknet53 [35] for feature extraction, followed by a C2f module and dual



#### Instruction Prompt

You are given an image of a crosswalk, sidewalk and road. Image consists of different viewpoints of a walker. The top of the image shows the front angle of the walker, in the bottom left shows the left angle, the image in the bottom center shows the bottom angle, and the image in the bottom right displays the right angle. Your task is to assess the safeness of crossing street intersections based on image. Please make sure you read and understand these instructions carefully. You should provide a score of safety and reason why.

Fig. 4. Prompt for instructing of safety measurement

#### Auto Chain-of-Thought for Risk Assessment

1. Check if the cars on the road are approaching the crosswalk or just stopped. Check the sign of parallel traffic light of crosswalk
2. Check if the pedestrian signal shows a crossing signal or not.
3. Check if there are any crossing pedestrians around the crosswalk.
4. Repeat this to the whole angles (front, left, bottom, right) of the image.
5. Assign a score for safety on a scale of 1 to 5, where higher score means it is safer to cross the road.

Fig. 5. Prompt for Auto Chain-of-Thought

segmentation heads. YOLOv8-Seg was selected for its high-speed performance and accuracy in instance segmentation tasks. A value of 0.7 was used for the `iou` hyperparameter. For confidence scores, 0.5 was used for the front viewpoints while 0.25 was used for the other viewpoints.

c) *Bounding box*: We utilized YOLOv8 [36] to generate bounding boxes with two pixel coordinates identifying objects within a scene. The class labels and confidence score are depicted in the output to provide additional visual information to the VLM as shown in Fig. 3 (b). However, initial qualitative assessments revealed that YOLOv8, pretrained on the COCO [37] dataset encompassing 80 classes, was inadequate for recognizing objects from a robot's egocentric perspective in sidewalk environments. To address this, we finetuned YOLOv8 on the SideGuide [38] dataset which significantly improved the model's ability to discern features unique to sidewalk environments, leading to enhanced performance of object detection for the VLM. Hyperparameters of 0.7 and 0.25 were used for the values of intersection over union (`iou`) and minimum confidence threshold (`conf`)<sup>2</sup>.

<sup>2</sup><https://docs.ultralytics.com/usage/cfg>

#### Evaluation Criteria

Safety level (-2 to 2) - An assessment of whether the current situation is safe for pedestrians to cross the crosswalk. Each score means the following status:

Score -2 - Totally dangerous: Any car coming toward the crosswalk.

Score -1 - Partially dangerous: There is no car approaching the crosswalk, but the pedestrian light is at a stop sign.

Score 0 - Keep caution: Traffic light is red, yellow or not visible and pedestrian light is a crossing sign.

Score 1 - Partially safe: you can cross the road because there are no cars, traffic light is red, yellow or not visible, the pedestrian light is a crossing sign or not visible, and pedestrians are crossing. Or, there are no cars, the traffic light is green, the pedestrian light is a crossing sign or not visible.

Score 2 - Totally safe: you can cross the road because there are no cars, the traffic light is green, the pedestrian light is crossing sign or not visible and other pedestrian traffic.

Fig. 6. Prompt for evaluating safety score criteria

## B. Vision-Language Model for Risk Assessment

Motivated by prior successful endeavors in generating automatic metrics using VLMs [28], [29], we employ GPT-4V to automatically measure safety scores with inputs of both visual knowledge and text. Note that our approach is based on zero-shot prompting and does not require labeled data to measure scores. Our approach is a prompt-based evaluator with three key components: (1) a prompt including the definition of the task and input details, (2) a series of intermediate instructions designed to solve problems by outlining detailed evaluation steps, and (3) a prompt for predicting a safety score based on the given criteria.

1) *Instruction prompt for safety evaluation*: A prompt is a natural language instruction that outlines a task, providing details on inputs, desired outputs, evaluation criteria, and reasoning verification. In our prompt, we initially present task specifications and instructions to obtain both the safety score and the corresponding reasoning. Moreover, to leverage multiview information, we incorporate explanations of distinct information based on the image patch location as it is important to augment instructions when using additional feature inputs (see Fig. 4).

2) *Auto Chain-of-Thought for safety measurement*: Chain-of-Thought (CoT) is a sequential set of intermediate steps designed to solve a problem systematically [39]. To enhance reasoning capabilities, we incorporate the Auto CoT [29], which furnishes intermediate instructions to mea-

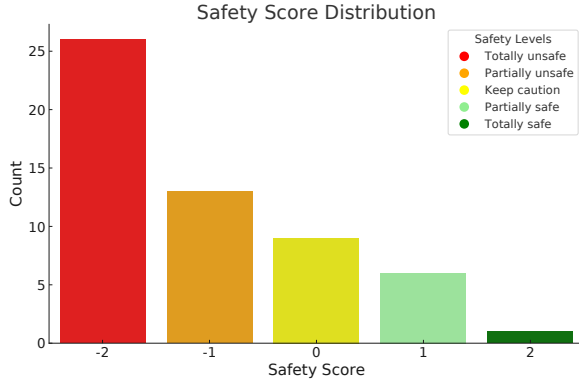


Fig. 7. **Safety score distribution of annotated data.** A large portion of the data contains unsafe scenarios.

Method	Accuracy	Spearman's $\rho$
Baseline	0.2545	0.3602
+ CoT <sup>1</sup>	0.1636	0.2532
+ bbx	0.3272 <sup>2</sup>	0.1214
+ mask	<b>0.4545</b>	0.2757
+ flow	<b>0.4545</b>	<b>0.4348</b>

TABLE III  
MODEL EVALUATION RESULTS.

<sup>1</sup> Chain-of-Thought.

<sup>2</sup> The **best** and second best performances.

sure scores. In the context of vision-language models and their applications, it demonstrates superior performance, particularly in explaining the evaluation process. Fig. 5 illustrates one of our Auto CoT prompts.

3) *Evaluating safety with criteria and scoring function:* To obtain a desired output, the prompt should contain detailed criteria for measurement. Based on the predefined criteria (Table II), we provide the safety score and description for the safety levels in natural language as shown in Fig. 6.

We provide a multiview image  $I$  and its extracted visual knowledge of bounding box  $I_{bbx}$ , optical flow  $I_{flow}$ , and segmentation mask  $I_{mask}$  to the VLM to get a safety score. We develop a reason-verification prompt  $E$  that instructs the VLM to measure the safety score  $s$  with specified criteria and to verify the reason. Concretely, we provide the image  $I$  with its visual knowledge and feed the prompt  $E$  to measure the safety score,  $s = VLM(E, I, I_{bbx|flow|mask|none})$ . As a baseline, we use the  $I_{none}$  and  $E$  prompts, which represent raw images without visual knowledge and text prompts without the Auto CoT prompt, respectively.

#### IV. EXPERIMENTAL RESULTS

In this section, we evaluate the effectiveness of GPT-4V on real-world street crossing data. Our approach marks a pioneering effort in autonomously calculating safety scores by integrating vision and language in a zero-shot scheme.

##### A. Experiment Setup

a) *Dataset:* Due to the lack of publicly available datasets on safety metrics, we generated a unique dataset

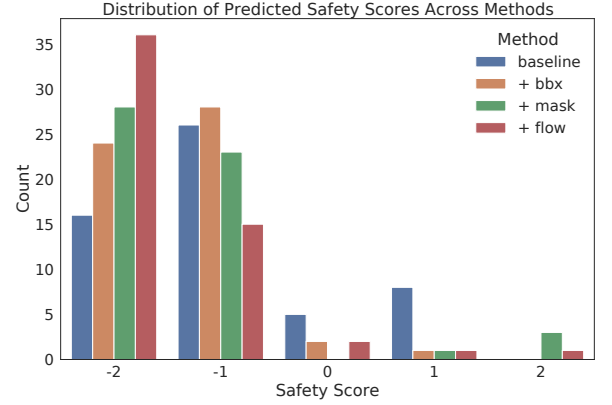


Fig. 8. **Safety score distribution of predictions.** The distribution of safety score predictions from different methods is illustrated with each represented by a unique color.

comprising a subset of 55 images collected via a quadruped robot. Although our framework does not require labeled data, we labeled the safety score based on the aforementioned criteria to verify performance. Three researchers independently annotated the data and selected the majority vote for consensus. In cases where no majority was reached, the median annotation was chosen. The Fleiss'  $\kappa$  was calculated to be 0.61, indicating substantial agreement among raters. Fig. 7 illustrates the annotated safety score distribution.

b) *Implementation details:* We use GPT-4V to measure the safety score, with the temperature set to 1 for answer generation. We conducted a series of experiments by varying the composition of prompts and types of visual information to assess their individual effects on safety score prediction.

c) *Evaluation metric:* We use accuracy to measure the performance of safety score prediction as mentioned in Sec. III-B. However, as the distribution of labels is skewed as shown in Fig. 7, we additionally verify with the Spearman's  $\rho$  correlation of the output and score labels.

##### B. Evaluation Results

Table III comprehensively outlines the results of various experiments conducted under diverse settings. Initially, the baseline setting is introduced, where the model's performance metrics are quantified in terms of accuracy and correlation degree, with values of 0.2545 and 0.3602 respectively. Upon comparison, the Chain of Thought (CoT) approach demonstrated inferior results, lagging behind the baseline in both accuracy and correlation degree.

Further analysis reveals interesting insights regarding the bounding box information. While it offers a marginal improvement in accuracy, it concurrently leads to a significant reduction in correlation. This discrepancy poses a challenge in drawing definitive conclusions regarding the performance increase.

In a contrasting scenario, the segmentation mask feature significantly elevates the accuracy from the baseline's 0.2545 to an impressive 0.4545. However, this improvement comes with a trade-off, as the correlation degree experiences a substantial decrease, dropping from 0.3602 to 0.2757.

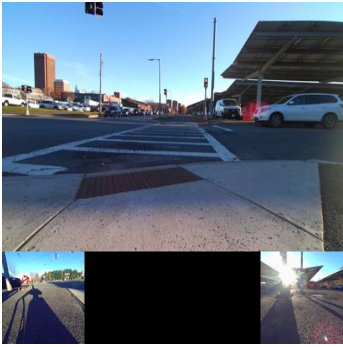
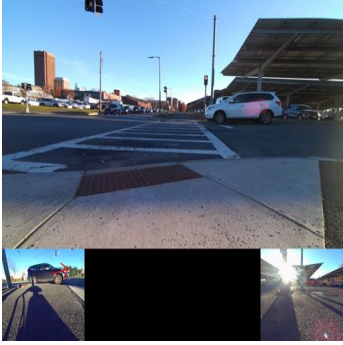
Images	Baseline	+ Optical Flow
	There is no car approaching the crosswalk directly, but the pedestrian light is showing a red hand, indicating a stop signal for pedestrians.	There is a vehicle visibly approaching the crosswalk as indicated by the red arrow showing average optical flow towards the crosswalk area.
	There is no car approaching the crosswalk directly, but the pedestrian signal is red, indicating a stop sign for pedestrians, which makes it partially dangerous to cross.	There is a vehicle shown in the bottom left image with an arrow indicating its movement toward the crosswalk, which suggests it is unsafe for pedestrians to cross at this moment

TABLE IV  
QUALITATIVE RESULTS ON SCENE DESCRIPTION

The optical flow information stands out in the experimental results. It not only achieves the highest accuracy among all tested features but also maintains a relatively high correlation degree. This may underscore the significant role of temporal information, as captured by the optical flow feature, in enhancing the suitability and effectiveness for the targeted task. The analysis of these diverse experimental settings provides a nuanced understanding of how different features and methodologies impact the overall performance in task-specific scenarios.

## V. DISCUSSION

### A. The Effect of Visual Knowledge

*a) Bounding Box:* The analysis reveals that the bounding box feature exhibits relatively low correlation degree and acts as noise when compared to the baseline. This is because the information that can be extracted from object detection is already sufficiently captured by the VLM. Moreover, the bounding boxes often obscure important information, such as traffic signals, which complicates the model's ability to make accurate judgments.

*b) Segmentation Mask:* The inclusion of segmentation mask feature has enhanced accuracy while diminishing the correlation degree. This phenomenon correlates with the data distribution observed in Fig. 7. Our analysis determined that the model becomes more responsive to risk factors upon receiving a segmentation mask as input, leading to a higher classification of scenes into unsafe categories, as depicted in

Fig. 8. Thus, accuracy improves because the data is skewed to unsafe cases, resulting in a reduced correlation degree.

*c) Optical Flow:* Among visual features, optical flow has shown the highest performance improvement in terms of accuracy. This is a reasonable result considering the impact of the direction and speed of pedestrians and vehicles on safety due to the road environment. In particular, as can be seen in Table IV, using the optical flow feature, the risk can be accurately analyzed by recognizing vehicles turning right toward a crosswalk. This result supports our assumption that temporal information in the scene, especially vehicle direction information, is an important factor in assessing risk. Furthermore, the correlation degree is observed to be higher than the baseline unlike other visual features.

### B. Does Chain-of-Thought Really Help?

In contrast to the findings of Liu et al. [29], the incorporation of CoT into the prompt did not enhance performance in our task; rather, it resulted in degradation. We hypothesize two potential reasons for this outcome after analyzing the reasoning outputs.

First, the reasoning produced through our method is notably more verbose than that of the baseline, displaying excessive sensitivity to keywords embedded within the CoT instructions. For example, while the baseline typically generates concise explanations comprising two to three sentences, the responses from CoT-informed prompts extend beyond three sentences. Moreover, in scenarios involving a car positioned a considerable distance from the crosswalk, the

focus tends to be on the mere existence of the car, leading to an interpretation of an unsafe situation.

Second, our CoT framework might not be optimally configured for the task at hand. The effectiveness of the generated content relies on the specifics of the prompt, suggesting that our current CoT approach may not be ideally tailored for assessing safety. Furthermore, despite the proven success of CoT methodologies in NLP tasks, their applicability and performance in interpreting multimodal data have yet to be conclusively established.

## VI. CONCLUSIONS

We evaluate a vision-language model for safety-aware scene understanding in robot navigation, particularly to support BLV individuals to safely cross streets. By leveraging GPT-4V, we evaluate the capabilities of a comprehensive and nuanced understanding of street crossing environments, aiming for the development of a reliable mobility aid. We share experimental results that provide valuable insights into the effects of various visual features. Optical flow emerged as a significant feature, underscoring the importance of temporal information in enhancing performance.

While our findings underscore the effectiveness of the VLM in facilitating safety-aware decisions for street crossing, this is not without its limitations. Notably, the application of the Chain-of-Thought methodology did not yield the anticipated performance improvements. This discrepancy highlights the need for further refinement in prompt engineering and adaptation strategies for visual data. Furthermore, our dataset's sensitivity to specific risk factors and its tendency to overestimate risk in certain scenarios highlight the need for greater diversity and quantity in data collection.

Future research will focus on several key areas: investigating the impact of different viewpoints on model performance, exploring the importance of temporal understanding in navigation decisions, comparing the performance of our model with vision-only models, increasing inference speed, and fine-tuning the model based on user-specific preferences. Addressing these areas is crucial for advancing our goal of developing a reliable mobility aid that ultimately supports the mobility of blind and low-vision (BLV) individuals.

## REFERENCES

- [1] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3354–3361.
- [2] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.
- [3] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2636–2645.
- [4] A. Zhang, C. Eranki, C. Zhang, J.-H. Park, R. Hong, P. Kalyani, L. Kalyanaraman, A. Gamare, A. Bagad, M. Esteva, *et al.*, "Towards robust robot 3d perception in urban environments: The ut campus object dataset," *arXiv preprint arXiv:2309.13549*, 2023.
- [5] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.
- [6] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.
- [7] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *arXiv preprint arXiv:2304.08485*, 2023.
- [8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
- [10] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.
- [11] D. A. Guth, E. W. Hill, and J. J. Rieser, "Tests of blind pedestrians' use of traffic sounds for street-crossing alignment," *Journal of Visual Impairment & Blindness*, vol. 83, no. 9, pp. 461–468, 1989.
- [12] S. E. Hassan, "Are normally sighted, visually impaired, and blind pedestrians accurate and reliable at making street crossing decisions?" *Investigative ophthalmology & visual science*, vol. 53, no. 6, pp. 2593–2600, 2012.
- [13] D. H. Ashmead, D. Guth, R. S. Wall, R. G. Long, and P. E. Ponchillia, "Street crossing by sighted and blind pedestrians at a modern roundabout," *Journal of Transportation Engineering*, vol. 131, no. 11, pp. 812–821, 2005.
- [14] R. Ichikawa, B. Zhang, and H.-o. Lim, "Voice expression system of visual environment for a guide dog robot," in *2022 8th International Symposium on System Security, Safety, and Reliability (ISSSR)*. IEEE, 2022, pp. 191–192.
- [15] S. Liu, A. Hasan, K. Hong, R. Wang, P. Chang, Z. Mizrahi, J. Lin, D. L. McPherson, W. A. Rogers, and K. Driggs-Campbell, "Dragon: A dialogue-based robot for assistive navigation with visual language grounding," *arXiv preprint arXiv:2307.06924*, 2023.
- [16] J. Guerreiro, D. Sato, S. Asakawa, H. Dong, K. M. Kitani, and C. Asakawa, "Cabot: Designing and evaluating an autonomous navigation robot for blind people," in *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, 2019, pp. 68–82.
- [17] K. M. Hoogsteen, S. Szpiro, G. Kreiman, and E. Peli, "Beyond the cane: describing urban scenes to blind people for mobility tasks," *ACM Transactions on Accessible Computing (TACCESS)*, vol. 15, no. 3, pp. 1–29, 2022.
- [18] AYES. (2023) Meet oko — the ai-powered navigation app for all pedestrians. [Online]. Available: <https://www.ayes.ai/>
- [19] B. M. Eyes. (2023) Be my eyes - see the world together. [Online]. Available: <https://www.bemyeyes.com/>
- [20] P. P. Ray, "Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope," *Internet of Things and Cyber-Physical Systems*, 2023.
- [21] I. Poola and V. Božić, "Guiding ai with human intuition for solving mathematical problems in chat gpt," *Journal Homepage: http://www.ijmra.us*, vol. 11, no. 07, 2023.
- [22] S. Kwon, R. Garodia, M. Lee, Z. Yang, and H. Yu, "Vision meets definitions: Unsupervised visual word sense disambiguation incorporating gloss information," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 1583–1598. [Online]. Available: <https://aclanthology.org/2023.acl-long-88>
- [23] H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz, "Capabilities of gpt-4 on medical challenge problems," *arXiv preprint arXiv:2303.13375*, 2023.
- [24] Y. Liang, J. Zhang, H. Li, X. Liu, Y. Hu, Y. Wu, J. Zhang, Y. Liu, and Y. Wu, "Breaking the bank with chatgpt: Few-shot text classification for finance," in *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, 2023, pp. 74–80.



- [25] F. Gilardi, M. Alizadeh, and M. Kubli, "Chatgpt outperforms crowdworkers for text-annotation tasks," *arXiv preprint arXiv:2303.15056*, 2023.
- [26] Y. Dubois, X. Li, R. Taori, T. Zhang, I. Gulrajani, J. Ba, C. Guestrin, P. Liang, and T. Hashimoto, "Alpacafarm: A simulation framework for methods that learn from human feedback," *ArXiv*, vol. abs/2305.14387, 2023. [Online]. Available: <https://api.semanticscholar.org/CorpusID:258865545>
- [27] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. Gonzalez, and I. Stoica, "Judging llm-as-a-judge with mt-bench and chatbot arena," *ArXiv*, vol. abs/2306.05685, 2023. [Online]. Available: <https://api.semanticscholar.org/CorpusID:259129398>
- [28] S. Min, K. Krishna, X. Lyu, M. Lewis, W. tau Yih, P. W. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi, "Factscore: Fine-grained atomic evaluation of factual precision in long form text generation," *ArXiv*, vol. abs/2305.14251, 2023. [Online]. Available: <https://api.semanticscholar.org/CorpusID:258841470>
- [29] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "G-eval: Nlg evaluation using gpt-4 with better human alignment," *ArXiv*, vol. abs/2303.16634, 2023. [Online]. Available: <https://api.semanticscholar.org/CorpusID:257804696>
- [30] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," *arXiv preprint arXiv:2304.10592*, 2023.
- [31] S. Wu, H. Fei, L. Qu, W. Ji, and T.-S. Chua, "Next-gpt: Any-to-any multimodal llm," *arXiv preprint arXiv:2309.05519*, 2023.
- [32] J. Chen, D. Z. X. S. X. Li, Z. L. P. Zhang, R. K. V. C. Y. Xiong, and M. Elhoseiny, "Minigpt-v2: Large language model as a unified interface for vision-language multi-task learning," *arXiv preprint arXiv:2310.09478*, 2023.
- [33] N. Radwan, W. Burgard, and A. Valada, "Multimodal interaction-aware motion prediction for autonomous street crossing," *The International Journal of Robotics Research*, vol. 39, no. 13, pp. 1567–1598, 2020.
- [34] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *IJCAI'81: 7th international joint conference on Artificial intelligence*, vol. 2, 1981, pp. 674–679.
- [35] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.
- [36] G. Jocher, A. Chaurasia, and J. Qiu, "YOLO by Ultralytics," Jan. 2023. [Online]. Available: <https://github.com/ultralytics/ultralytics>
- [37] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.
- [38] K. Park, Y. Oh, S. Ham, K. Joo, H. Kim, H. Kum, and I. S. Kweon, "Sideguide: a large-scale sidewalk dataset for guiding impaired people," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 10 022–10 029.
- [39] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. H. hsin Chi, F. Xia, Q. Le, and D. Zhou, "Chain of thought prompting elicits reasoning in large language models," *ArXiv*, vol. abs/2201.11903, 2022. [Online]. Available: <https://api.semanticscholar.org/CorpusID:246411621>