2026 IEEE/SICE International Symposium on System Integration (SII)
January 11-14, 2026. Cancún, México

Texting-While-Walking Detection in Real-World Environments Using Vision-Language Models with Prompt Engineering

Seungpyo Choi[1], Jiaxu Wu[2], Qi An[1] and Atsushi Yamashita[1]

Abstract— **Smartphone-induced "texting while walking" poses growing safety risks not only in public shared spaces but also in robot navigation scenarios where humans and robots coexist. To mitigate these risks, recent studies have developed pedestrian behavior detection models that aim to recognize when people are distracted by their smartphones. However, these models still suffer from high false-positive rates and reduced detection accuracy when visually similar poses or occlusions occur. To address this issue, we propose a Vision-Language Model (VLM)-based behavior detector that exploits VLMs pretrained on large image-text datasets and capable of global-context inference. Specifically, we leverage LLaVA-7B and systematically evaluate three prompt-engineering schemes—chain-of-thought and self-consistency under zero-shot settings, and few-shot prompting under few-shot settings. We conducted the dataset generation experiment in a typical indoor hall with a centrally placed table that intermittently occluded the robot's view. During each session, four to six participants walked freely while performing nine everyday actions, resulting in 11,815 annotated pedestrian images captured from the robot's perspective. Experimental results show that our VLM-based pipeline significantly reduces false-positive detections and improves both precision and overall F1-score compared to a conventional pose-based LSTM baseline. These gains demonstrate that combining large-scale VLM reasoning with specially designed prompts can overcome long-standing misclassification issues in existing approaches. Our curated dataset and prompt-analysis results provide a foundation for extending VLM-based perception to a wide range of camera-based monitoring and navigation systems.**

I. INTRODUCTION

In recent years, the deployment of autonomous mobile robots has expanded across various sectors of society, from logistics and healthcare to public safety and customer service. These robots frequently share space with pedestrians, making robust collision avoidance not just desirable but essential. Pedestrians exhibit a spectrum of behaviors, from attentive walking to chatting with companions. Among these, walking while focused on a smartphone has recently drawn particular concern due to its elevated collision risk. Specifically, pedestrians engaged with their smartphones often fixate on their screens. Consequently, they may fail to perceive changes in their surroundings, thereby increasing the risk of collision [1]. Moreover, such pedestrians exhibit unpredictable and erratic movements, necessitating sudden evasive maneuvers by nearby pedestrians [2]. Thus, accurately identifying such distracted pedestrians is essential to ensure collision avoidance for mobile robots operating in dynamic, pedestrian-shared environments. Although considerable effort has been devoted to detecting such pedestrians, no efficient method for characterizing their behavior has yet been established, and significant limitations remain [3]. Therefore, developing an effective detection approach for texting pedestrians is a crucial foundation for achieving safe mobile robot navigation.

Specifically, existing approaches that rely on traditional deep learning models share a common limitation: they lack training data representative of real-world scenarios, which leads to markedly degraded performance when detecting texting pedestrians in practice. Although several datasets commonly used for autonomous driving research include pedestrian samples, they are primarily designed for limited intention-prediction tasks (e.g., forecasting pedestrian crossings) rather than detailed behavior recognition [4], [5]. Consequently, datasets suitable for training models on pedestrian behavior are extremely limited, rendering adequate model training challenging. As a result, these models experience significant drops in accuracy under real-world conditions characterized by varied occlusion patterns and diverse pedestrian motions, leading to sharply increased false-positive rates (FPR).

To address these limitations, we propose a pedestrian behavior detector based on a vision-language model (VLM) that has the ability to reason over the visual context of an image and flexibly extract task-relevant cues in response to diverse natural language queries. Furthermore, we propose a novel prompt engineering approach to enhance the model's sensitivity to visual cues and reduce the high FPR observed in prior methods. In particular, targeted textual prompts effectively guide the VLM's attention toward relevant pedestrian behaviors, substantially lowering FPR and improving the reliability of behavior recognition in complex visual environments.

II. RELATED WORK

In human-robot interaction (HRI), seamless operation in real-world environments demands not only accurate perception of pedestrians' positions and velocities but also reliable recognition of their behaviors, since achieving a positive perception of safety is essential for robots to be accepted as partners and co-workers [6]. Among these behaviors,

*This work was not supported by any organization.

*This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the University of Tokyo under Application No. 24-524, and performed in line with the University of Tokyo.

[1]. S. Choi, Q. An and A. Yamashita are with the Department of Human and Engineered Environmental Studies, Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8563, Japan {choi, qi, yamashita}@robot.t.u-tokyo.ac.jp

[2]. J. Wu is with the Institute of Engineering Innovation, Graduate School of Engineering, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8563, Japan jiaxuwu@robot.t.u-tokyo.ac.jp

detecting pedestrians who are texting while walking is especially important, because they pose a high risk of collisions and navigation failures in shared spaces [7]. Consequently, detecting texting-while-walking has garnered significant research attention. Prior research has primarily focused on using RGB (or RGB-D) cameras and LiDAR, with the goal of detecting pedestrian behaviors before robots approach them.

LiDAR-based methods first acquire point clouds from LiDAR sensors and then estimate pedestrians' poses based on the obtained data. Furthermore, LiDAR-based methods, which employ laser pulses to acquire object shape information, remain robust under challenging optical conditions, including backlighting and low-light scenarios. Wu et al. interpret LiDAR point clouds to analyze pedestrian behaviors, demonstrating relative invariance in detection accuracy under extreme optical variations [8]. However, shape information obtained from LiDAR point clouds is inherently sparser than that of RGB imagery, often resulting in lower detection precision.

RGB camera-based methods, on the other hand, typically use motion-related features such as keypoint trajectories or optical flow that are extracted from RGB images for pose estimation. For instance, Kumamoto et al. derive body keypoints from preprocessed RGB frames to recognize pedestrian actions, achieving superior accuracy compared to LiDAR-based approaches when image resolution is sufficient [9]. However, these keypoint-based techniques often struggle to differentiate visually similar behaviors—such as texting while walking versus other forward-leaning postures (e.g., holding an umbrella)—resulting in elevated FPR.

To address misclassification in RGB-based detection, Rangesh et al. incorporate gaze-tracking cues alongside body keypoints to enhance pose estimation, thereby reducing false positives among visually similar actions [10]. However, the misclassification rate increases noticeably when pedestrians interact with objects or perform hand gestures that were not represented in the training dataset.

On the other hand, a major limitation of one-shot keypoint-based approaches is their susceptibility to misclassification when body parts are occluded by obstacles. To address this issue, studies have leveraged temporal data to exploit contextual information across frames. Building on this idea, Terao et al. propose a pose detector that processes time-series keypoint data and extracts feature vectors via a Variational Autoencoder (VAE) [11]. They first generate clean 3D motion-capture sequences and synthetically occlude them by selectively removing segments, then use both original and occluded sequences as ground-truth inputs to train the model. During inference, sequential keypoint frames are fed into an LSTM encoder to capture temporal context so that occlusions in any given frame can be overcome using information from surrounding frames. Although this method performs well on motion-capture benchmarks, its accuracy degrades significantly under real-world conditions due to various occlusion patterns and insufficient training data.

To overcome the limitations of previous research—which often rely on domain-specific datasets and struggle to recognize unseen behaviors [10], [11]—our approach integrates a VLM, pretrained on large-scale image-text pairs and capable of reasoning over the entire visual context of each image, into the behavior detection pipeline. Specifically, motivated by related work showing that prompt engineering is a dominant factor in VLM performance [12], we design a diverse set of prompt-engineering strategies to steer the pretrained VLM's knowledge to accurately detect and classify complex human behaviors from visual data. Furthermore, by leveraging the rich semantic grounding of VLMs, we not only address the shortage of labeled examples but also achieve more robust and generalizable detection in unconstrained, real-world environments.

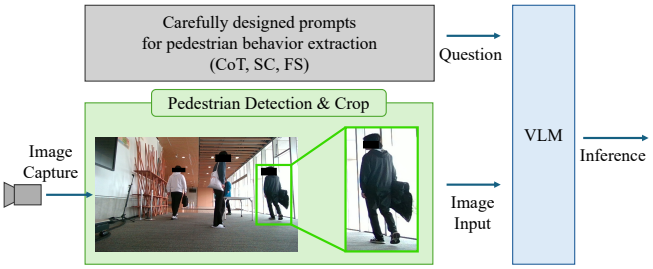


Fig. 1: Overview of the proposed two-stage pedestrian behavior detection pipeline.

III. PROPOSED METHOD

Prior efforts in pedestrian behavior detection suffer from a drastic drop in accuracy when deployed in real-world scenes, primarily due to the scarcity and narrow diversity of annotated training data. To overcome this limitation, we propose a novel two-stage pipeline (Fig. 1) that leverages a pretrained VLM as a zero-shot detector. To effectively leverage the VLM's broad visual-textual grounding, we design three complementary evidence streams: (i) object-level visual cues extracted via Chain-of-Thought (CoT) reasoning [13], (ii) consensus signals obtained through Self-Consistency (SC) querying [14], and (iii) class-specific prior knowledge provided by Few-Shot (FS) exemplars [15]. Each stream offers a distinct perspective on the scene—appearance features of hand-held objects (CoT), reliability estimated from answer agreement (SC), and exemplar similarity to curated references (FS). Throughout our experiments, each perspective individually produces robust inferences even under occlusion or when distinguishing visually similar poses such as texting versus umbrella-holding.

A. Pedestrian Detection and Frame Extraction

To ensure that the VLM focuses its inference specifically on pedestrians, we first employ a pedestrian detector to identify and crop the Region of Interest (ROI) for each individual. The isolated ROI is then provided as the sole visual input to the VLM during the behavior analysis stage, thereby directing the model's attention exclusively to pedestrian behavior.

B. Prompt Design for Behavior Inference with VLM

In the second stage, cropped patches of pedestrians are fed into LLaVA-7B, an open-source VLM that integrates a CLIP-based visual encoder and a Vicuna 7B language decoder [16]. Trained on multi-modal instruction-following data, LLaVA supports both zero-shot and few-shot reasoning, making it well-suited for fine-grained behavior inference tasks [17]. Specifically, we design three prompting strategies to fully leverage the VLM's visual-textual grounding, guiding the model to focus on pedestrian behavior:

1) Chain-of-Thought (CoT) Prompting [13]
2) Self-Consistency (SC) Prompting [14]
3) Few-shot (FS) Prompting [15]

Below, we describe each method in detail, along with the prompts used.

1) Chain-of-Thought (CoT) Prompting

In CoT prompting, we decompose the reasoning process into two steps: hand-object interaction analysis and behavior inference. We first craft targeted prompts that highlight distinct visual features of hand-object interactions. Specifically, we ask the VLM to identify an object held in the pedestrian's hand using the following prompt:

Prompt 1: Object Identification

Q: What is the person holding in their hand? Answer concisely.

A: "Mobile phone"

Based on this intermediate answer (e.g., "Mobile phone" in the upper case), we construct a follow-up question to infer the pedestrian's behavior as follows:

Prompt 2: Behavior Inference Under Object Detection Results

Q: The person is holding "Mobile phone" in their hand. Is the person using a mobile phone? Answer only 'Yes' or 'No'.

A: "Yes" (Expected)

Through this step-by-step reasoning pipeline, we use the extracted object-level information to generate follow-up questions that infer the pedestrian's behavior. This design explicitly grounds the model's final judgment in intermediate visual cues, thereby improving behavioral inference accuracy.

2) Self-Consistency (SC) Prompting

To enhance the reliability of behavioral inference under pose ambiguity, we adopt a self-consistency (SC) decoding strategy, as illustrated in Fig. 2. Specifically, we query the model ten times with the identical prompt, using its default temperature, a hyper-parameter that controls sampling randomness—lower values make the output more deterministic,

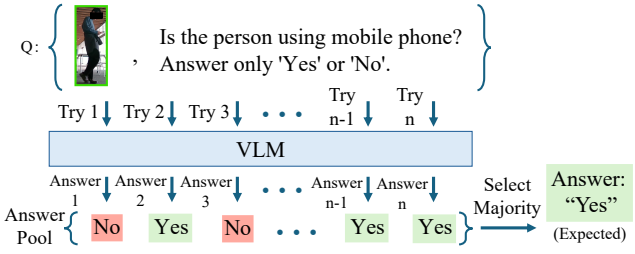


Fig. 2: Example of self-consistency prompting (n times repetition)

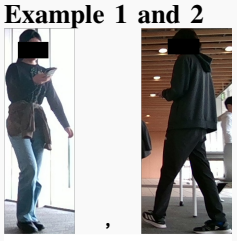
while higher values yield more diverse responses. In this study, we adopted a sufficiently large sampling count of ten queries per task to maximize detector performance and enhance prediction stability. The final prediction is determined by majority voting over the ten sampled outputs, reducing stochastic variation and promoting more robust and trustworthy predictions. In practice, the optimal number of queries should be selected by considering the trade-off between desired accuracy and allowable inference latency in the target application.

3) Few-shot (FS) Prompting

To further enhance behavioral inference, we present the VLM with a small set of labeled exemplar images for each class (texting and non-texting) before querying it about the target pedestrian. This FS prompting method aims to guide the model toward learning subtle visual distinctions between the two behaviors. Specifically, we supply the labeled exemplar images with the following prompts:

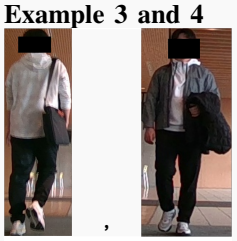
Few-shot Prompt (Examples 1-4)

Example 1 and 2



User: Is the person using a mobile phone? Answer only 'Yes' or 'No'.
Assistant: Yes

Example 3 and 4



User: Is the person using a mobile phone? Answer only 'Yes' or 'No'.
Assistant: No

After providing these examples, we query the target pedestrian image with the following prompt:

Target Prompt (for Inference)

Q: Is the person using a mobile phone? Answer only 'Yes' or 'No'.

A: "Yes" (Expected)

This prompting pipeline enables the model to learn and apply visual cues acquired from the provided examples, thereby improving behavioral inference accuracy even for visually ambiguous actions. To implement this strategy within the detector system's computational resources and latency requirements, we supply two exemplar images per scenario—one frontal and one rear view of the pedestrian—for both texting and non-texting cases, as described above. The number of exemplar images can, however, be dynamically adjusted to accommodate the constraints of the inference environment and the inference results.

IV. Experiment And Results

A. Experimental Setup

To evaluate the model's inference performance, we utilized a dataset previously collected by our research group [11]. The dataset was obtained by mounting a RealSense D455 RGB-D camera on a four-legged robot (Unitree Go1) and conducting navigation experiments in an indoor environment populated with pedestrians. The environment included furniture such as desks and chairs, which naturally introduced occlusions and visual clutter. During navigation, the robot continuously captured front-facing RGB images at 30 frames per second.

A total of 11 participants took part in the experiment, with four to six individuals participating per session. Each participant was instructed to perform behaviors commonly observed in real-world pedestrian scenarios.

B. Data Construction

To reflect the diversity of pedestrian behaviors encountered in the real world, participants performed various activities while walking, including the following activities (Fig. 3):

- Texting while walking
- Carrying a briefcase or shoulder bag
- Simply walking
- Holding clothes while walking
- Walking with an umbrella
- Holding a cup of coffee while walking
- Holding a bottle while walking
- Talking on the phone while walking
- Holding a phone without using it

In particular, the "holding phone while walking" class was defined as cases where the participant visibly carried a phone in hand without any interaction, while keeping their gaze forward. Likewise, participants performing "talking on the phone while walking" were also instructed to keep their gaze forward during the activity, simulating attentive walking behavior.

Fig. 3: Overview of pedestrian activities.

The "texting while walking" class, which served as the positive class in this study, was defined as cases where the participant continuously interacted with a smartphone while walking, with their gaze directed toward the screen.

From the recorded videos, all frames were extracted into individual images. Image patches containing pedestrians were then cropped and manually labeled based on the observed behaviors. As a result, the dataset consisted of 2,980 frames labeled as Texting, and 8,835 frames labeled as other. The dataset will not be made publicly available, as the experimental protocol was approved on the condition that participant data remain confidential and be used exclusively for internal research purposes.

C. Evaluation Metrics

We evaluated three prompt-engineering strategies proposed in this study using the constructed dataset. The evaluation metrics included Accuracy, Precision, Recall, and F1-score, which provided a comprehensive understanding of model performance under realistic conditions.

In addition to these metrics, we also computed the FPR for each method to assess how well the proposed strategies reduced erroneous detections. This allowed us to quantitatively compare the improvement over prior approaches, which typically exhibited high FPR in complex or ambiguous real-world scenarios.

D. Behavior Inference with VLM

To fully leverage these capabilities, we employ the three prompting strategies described in Section III-*B* to guide the

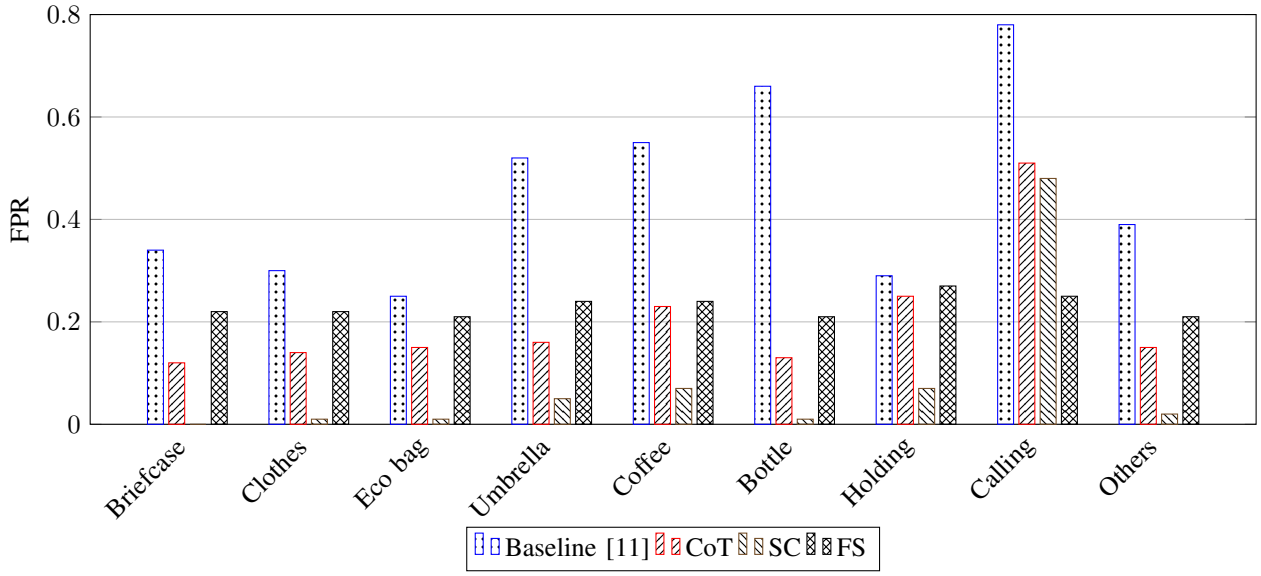


Fig. 4: Comparison of FPR across four prompting methods (Baseline, CoT, SC, FS) for nine object categories.

TABLE I: Evaluation results for each prompting method.

Prompt	Recall	Precision	Accuracy	Specificity	F1
CoT	**0.54**	0.51	0.75	0.82	0.53
SC	0.48	**0.78**	**0.83**	**0.95**	**0.59**
FS	0.27	0.29	0.65	0.77	0.28

model's attention toward pedestrian behavior. Each strategy—CoT, SC, and FS prompting—is applied independently, and their outputs are not combined. Finally, we evaluate the behavioral inference performance of each method, using the evaluation metrics defined in Section IV-*C*, calculated on the model's predicted labels. All metrics were derived from the aggregated confusion matrix across the entire test set, and thus represent a micro-averaged evaluation.

E. Results

Table I presents the evaluation results for each prompting method explained in section III-*B*. Among the three prompting strategies, CoT prompting achieved the highest recall, demonstrating that this approach effectively enhances the detector's sensitivity to actual smartphone-use behaviors. This suggests that guiding the model through intermediate reasoning steps enabled it to better capture critical visual cues from pedestrians' hands, allowing it to distinguish between texting and non-texting behaviors.

SC prompting exhibited the best performance in terms of accuracy and F1-score, which reflect the overall balance between precision and recall, thereby indicating that this strategy achieved both stable detection and reduced false alarms. These results imply that sampling multiple responses and aggregating them via majority voting effectively stabilized the model's output and improved overall classification reliability.

In contrast, FS prompting performed the worst across all evaluation metrics. One possible reason is that the limited number of provided examples may have biased the model's response space, constraining its flexibility in interpreting unseen behavioral cues. Providing more diverse or well-curated examples might help mitigate this issue, though further investigation would be required.

To better understand how each prompting strategy affects FPR, we conducted a FPR focused comparison against the baseline [11]. To address the class imbalance between the texting and non-texting classes, which contain 2,980 and 8,835 samples respectively, we evaluated FPR for each pedestrian behavior separately to ensure fair and balanced comparison across categories. We evaluated the FPR for each pedestrian behavior under three prompting methods, all of which effectively reduced it relative to the baseline. The results are summarized in Fig. 4.

Among the methods tested, SC Prompting achieved the lowest overall FPR across all pedestrian action frames. We attribute this performance to the generation of ten independent reasoning chains per inference, which increases the diversity of samples, reduces noise, and minimizes the randomness of individual responses.

Next, both SC and CoT exhibited significant FPR variability depending on the type of action frame. For example, FPR of SC almost approached zero when detecting "carrying briefcase" actions but rose to 0.07 for "carrying coffee" frames. In contrast, FS Prompting maintained a stable FPR between 0.21 and 0.27 regardless of action type. We believe this stability stems from FS supplying concrete "walking while texting" examples, enabling the VLM to internally learn the distinctions between confusable actions. Notably, when using the generic prompt "Is the person using a mobile phone?", SC and CoT showed a sharp FPR increase—up to 0.5—for "phone call" actions, likely misclassifying them as mobile phone use. In contrast, FS did not exhibit this increase, remaining at 0.25, suggesting that the VLM inter-

nally learned that "phone call" is not the same as "using a mobile phone". However, FS resulted in the highest absolute FPR among the three methods. This indicates that, by selecting example frames tailored to the robot's operating environment, FS Prompting can efficiently separate visually similar actions despite a higher baseline FPR.

Overall, the proposed detector demonstrated strong performance across all three prompting strategies, particularly in identifying non-texting behaviors, effectively reducing the FPR and achieving very high specificity. One plausible explanation for this robustness is that large-scale image-text datasets used in VLM pretraining (e.g., web-scraped captions) likely contain a disproportionately large number of non-texting examples, making the detector more conservative when judging ambiguous hand-held objects as "texting" and reducing FPR as a result. Thus, by fine-tuning the detector with additional "texting" samples to balance the dataset, the detector's sensitivity to actual texting behaviors can be improved.

Finally, despite these strengths, the proposed model achieves a throughput of 1-4 inferences per second on an RTX 4070 Ti GPU. The latency penalty is most pronounced for the SC strategy, which issues multiple stochastic decodings per frame to attain higher accuracy. These findings highlight the need for future work on model distillation to increase inference speed without compromising predictive performance.

V. CONCLUSION

In this paper, we proposed a VLM-based framework for pedestrian behavior detection, incorporating CoT prompting, SC voting and FS prompting to enhance robustness and interpretability. Our method leverages VLMs' strong multimodal reasoning capabilities, guided by carefully designed prompts to recognize visual cues especially from the pedestrian's hand-object interaction.

Experimental results demonstrated that CoT prompting achieved the highest recall by enabling the model to focus on intermediate reasoning steps, while SC prompting yielded the best accuracy and F1-score, highlighting the effectiveness of majority voting in stabilizing predictions. Notably, both methods significantly outperformed the baseline in terms of FPR, suggesting that VLMs' prior knowledge can contribute to reducing over-predictions. In the long term, a detector with a lower FPR would allow autonomous mobile robots to avoid unnecessary evasive maneuvers or abrupt stops. This, in turn, would reduce collision risk and enable them to reach their destinations quickly through more efficient path planning.

While our FS prompting strategy underperformed, it nevertheless highlights the importance of prompt design and sampling strategy when using VLMs for fine-grained behavior classification. In this study, the example images were randomly selected from the texting and non-texting classes without a specific criterion, even though the performance of few-shot prompting can vary significantly depending on which examples are provided [18]. Future research should therefore aim to establish a systematic method for selecting representative examples and investigate how such a method could further enhance few-shot performance in real-world behavior detection tasks.

Beyond the few-shot setting, we also plan to fine-tune the VLM on pedestrian-specific datasets to improve accuracy in ambiguous or occluded scenes. Revisiting the prompt formulation—such as explicitly targeting texting behavior—may further enhance reliability. We believe our approach can serve as a foundation for robust and generalizable behavior recognition in real-world environments.

REFERENCES

[1] F. Obayashi and K. Kozuka, "Sight Property at the Time 'Texting While Walking' by the Gaze Measurement, and Its Influence to Walking," IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, vol. J100-A, no. 9, pp. 338-345, 2017.

[2] H. Murakami, C. Feliciani, Y. Nishiyama and K. Nishinari, "Mutual Anticipation Can Contribute to Self-Organization in Human Crowds," Science Advances, vol. 7, no. 12, p. eabe7758, 2021.

[3] D. Ridel, E. Rehder, M. Lauer, C. Stiller and D. Wolf., "A Literature Review on the Prediction of Pedestrian Behavior in Urban Scenarios," Proceedings of the IEEE International Conference on Intelligent Transportation Systems, pp. 3105-3112, 2018.

[4] A. Rasouli, I. Kotseruba and J. K. Tsotsos, "Are They Going to Cross? A Benchmark Dataset and Baseline for Pedestrian Crosswalk Behavior," Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 206-213, 2017.

[5] A. Rasouli, I. Kotseruba, T. Kunic and J. K. Tsotsos, "PIE: A Large-Scale Dataset and Models for Pedestrian Intention Estimation and Trajectory Prediction," Proceedings of the International Conference on Computer Vision (ICCV), 2019.

[6] M. Rubagotti, I. Tusseyeva, S. Baltabayeva, D. Summers and A. Sandygulova, "Perceived Safety in Physical Human-Robot Interaction - A Survey," arXiv preprint arXiv:2105.14499, 2021.

[7] J. Nasar and D. Troyer, "Pedestrian Injuries due to Mobile Phone Use in Public Places," Accident Analysis & Prevention, vol. 57, pp. 91-95, 2013

[8] J. Wu, Y. Wang, H. woo, A. Moro and A. Yamashita, "Smartphone Zombie Detection from LiDAR Point Cloud for Mobile Robot Safety," IEEE Robotics and Automation Letters (IEEE RA-L), vol. 5, no. 2, pp. 2256-2263, 2020.

[9] K. Kumamoto and K. Yamada, "Detecting Pedestrian Interaction with Smartphones Based on Body Keypoints," Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, pp. 3261-3266, 2018.

[10] A. Rangesh and M. M. Trivedi, "Recognizing Phone-Based Activities of Pedestrians," IEEE Transactions on Intelligent Vehicles, vol. 3, no. 2, pp 218-227, 2018.

[11] H. Terao, J. Wu, Q. An and A. Yamashita, "Detection of Texting While Walking in Occluded Environment Using Variational Autoencoder for Safe Mobile Robot Navigation," IEEE Robotics and Automation Letters, vol. 10, no. 7, pp. 7675 - 7682, 2025.

[12] K. Zhou, J. Yang, C. Loy and Z. Liu, "Learning to Prompt for Vision-Language Models," Int J Comput Vis, vol 130, pp. 2337-2348, 2022.

[13] J. Wei et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," Advances in Neural Information Processing Systems, vol. 35, pp. 24824-24837, 2022.

[14] X. Wang et al., "Self-Consistency Improves Chain of Thought Reasoning in Language Models," arXiv:2203.11171, 2022.

[15] T. Brown et al., "Language Models are Few-Shot Learners," Advances in Neural Information Processing Systems, vol. 33, pp. 1877-1901, 2020.

[16] H. Liu, C. Li, Y. Li and Y. Lee, "Improved Baselines with Visual Instruction Tuning," arXiv:2310.03744, 2023.

[17] H. liu, C. Li, Q. Wu and Y. Lee, "Visual Instruction Tuning," arXiv:2304.08485, 2023.

[18] S. Kong, A. Madan, N. Peri, and D. Ramanan, "Revisiting Few-Shot Object Detection with Vision-Language Models," Proceedings of the 38th Conference on Neural Information Processing Systems, pp. 19547-19560, 2024.