Vision-language Fusion for Road Marking Detection in Autonomous Driving

Shaofan Sheng[a], Nicolette Formosa[b], Mohammed Quddus[a,*]

[a] < Department of Civil and Environmental Engineering, Imperial College London>, <London>, <United Kingdom>
shaofan.sheng21@imperial.ac.uk, m.quddus@imperial.ac.uk
[b] <National Highways>, <Birmingham>, <United Kingdom>
nicolette.formosa@nationalhighways.co.uk
* Corresponding author

Keywords: Road marking detection, Decision-making, Vision language models, Large language models

1 INTRODUCTION

Accurate detecting of road markings is indispensable for autonomous vehicles (AVs) to navigate safely and efficiently. However, current autonomous driving systems still face challenges in perceiving static and dynamic road information leading to inaccurate trajectory and path planning processes, especially in complex driving scenarios. To address these challenges, recent years have seen a predominant reliance on deep learning and attention mechanisms for road marking detection (Jayasinghe et al., 2022). However, persistent challenges such as high computational demands, extensive dataset requirements for training, and manual data annotation issues necessitate exploring alternative approaches. Recent advancements in Large Language Models (LLMs) and Vision Language Models (VLMs) present the possibility of addressing these challenges (Sha et al., 2023). While LLMs have demonstrated remarkable capabilities in reasoning tasks, VLMs offer a distinct advantage by integrating visual and textual information, thereby providing a more comprehensive understanding of complex scenarios (Zhang et al., 2024). Consequently, VLMs are instrumental in interpreting complex driving scenarios and enhance autonomous vehicles' decision-making processes in real-time (Sha et al., 2023).

As a result, this paper proposes a new road marking detection model based on VLM to address the aforementioned challenges. This model leverages both image and label information of road markings, achieving optimal detection results even with relatively small datasets. Furthermore, in conventional autonomous driving, perception and decision-making are two relatively independent processes, and to fully utilise the results of perception to assist the decision-making process, a comprehensive end-to-end pipeline is established that uses GPT-4 (Generative Pre-trained Transformer), an LLM-based language model, for question answering (QA) of road scenarios based on model's output. This pipeline enhances the autonomous driving system's ability to perceive road surfaces and aids in the decision-making process while driving. The key contributions of this paper are outlined as follows: (1) Development of a new road marking detection model based on VLM, offering reduced training costs compared to traditional methods, (2) Construction of an end-to-end pipeline to enhance the autonomous driving system's comprehension of road information and generation of driving prompts to support decision-making processes.

2 Methodology

2.1 Road marking detection model

The Contrastive Language-Image Pretraining (CLIP) model (Radford et al., 2021), a typical representation of VLMs, is employed as the foundation, leveraging its powerful image and text feature extraction capabilities. Hence, the name of the model is given as *RoadCLIP*. However, recognising the need to further enhance the accuracy of detection, a custom multi-head attention mechanism is introduced. This addition addresses the limitations of the standalone CLIP model by facilitating a more effective integration of image features with textual descriptions of road markings. Consequently, the model can discern and prioritise visual features relevant to the detection task. Figure 1 provides a visual representation of the methodology, illustrating the comprehensive detection method employed by RoadCLIP.

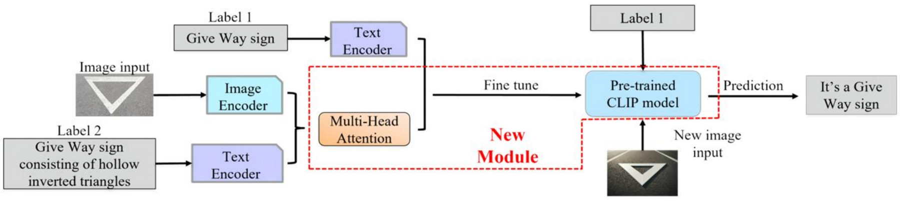


Figure 1 – *Flowchart of the RoadCLIP for comprehensive road marking detection.*

During the model construction, two sets of text labels, namely Label 1 and Label 2, were prepared. Label 1 served as the name label of the road marking, while Label 2 provided a detailed description of the road marking pattern, colour, and composition. During the training of the multi-head attention layer, Label 2 was used. This process involved feeding image features and corresponding text descriptions into the attention layer, adapting, and optimising the image feature representation based on the provided textual descriptions. Following the completion of training for the multi-head attention layer, the next step involved integrating this enhanced layer with the pre-trained CLIP model. This unified model is then fine-tuned using Label 1 and the training set images. The fine-tuning process refines the model's ability to recognise and understand road markings based on the specific task requirements. In training and testing the road detection model, a comprehensive set total of 14 distinct road markings was collected, including 644 real world images from Google Street View and 280 virtual road marking images generated based on the existing Text-to-Image models.

2.2 RoadGPT: Question Answering (QA) Systems based on GPT-4

By merging RoadCLIP, and the QA function of GPT-4, RoadGPT is created. This aims to deepen the understanding and analysis of road scenarios. By using the robust reasoning capabilities of LLMs, RoadGPT generates valuable driving prompts to enhance the decision-making process. Figure 2 illustrates the construction process of RoadGPT and the flowchart of the pipeline.

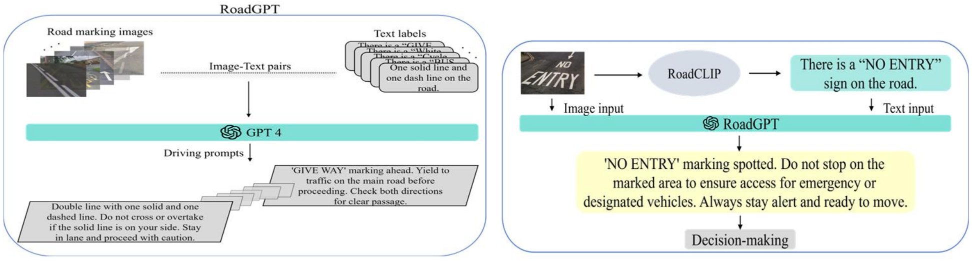
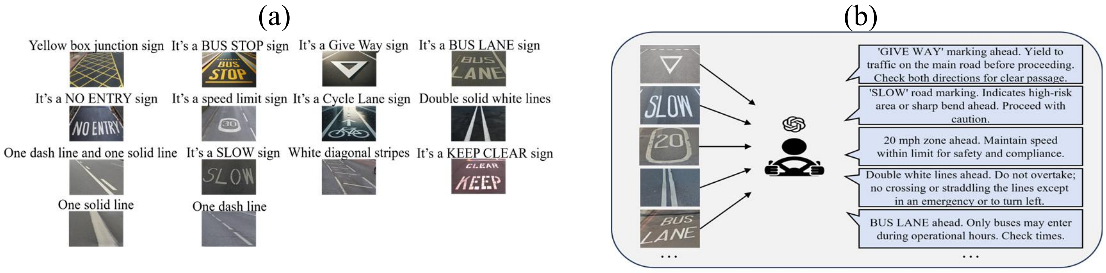

RoadGPT
Road marking images
Text labels
There is a "GIVE
There is a "White
There is a "Cycle
There is a "BUS
One solid line and one dash line on the road.
Image-Text pairs
GPT 4
Driving prompts
'GIVE WAY' marking ahead. Yield to traffic on the main road before proceeding. Check both directions for clear passage.
Double line with one solid and one dashed line. Do not cross or overtake if the solid line is on your side. Stay in lane and proceed with caution.
NO ENTRY
RoadCLIP
There is a "NO ENTRY" sign on the road.
Image input
Text input
RoadGPT
'NO ENTRY' marking spotted. Do not stop on the marked area to ensure access for emergency or designated vehicles. Always stay alert and ready to move.
Decision-making

Figure 2 – *Construction process of RoadGPT and Flowchart of the pipeline*

The official OpenAI API of GPT is used to construct the RoadGPT. During the fine-tuning, the Image-Text pairs from the RoadCLIP serve as inputs. By controlling the content and word count of the output, GPT-4 delivers concise yet accurate driving prompts relating to driving behaviour. By combining RoadCLIP and RoadGPT, a comprehensive end-to-end pipeline is constructed to enhance the autonomous driving system's comprehension of road information and generation of driving prompts to support decision-making processes. This pipeline streamlines the process requiring only the input of the road marking image to obtain relevant marking information and driving prompts. This reduces computational resource consumption and links the initially disparate perception and decision-making process in autonomous driving.

3 Results

RoadCLIP was applied to the test data using both real-world and AI road marking images. Figure 3 (a) shows the detection results, showcasing the model's proficiency. The captions above the images indicate the most similar result associated by the model with each road marking image from the provided labels. Moreover, all 14 road marking images were subjected to the end-to-end pipeline for testing and the results are shown in Figure 3 (b).



'GIVE WAY' marking ahead. Yield to traffic on the main road before proceeding. Check both directions for clear passage.

'SLOW' road marking. Indicates high-risk area or sharp bend ahead. Proceed with caution.

20 mph zone ahead. Maintain speed within limit for safety and compliance.

Double white lines ahead. Do not overtake; no crossing or straddling the lines except in an emergency or to turn left.

BUS LANE ahead. Only buses may enter during operational hours. Check times.

Figure 3 – *(a) Test results of RoadCLIP and (b) results of the end-to-end pipeline*

For each road marking, the pipeline output comprises road marking information plus corresponding driving prompts. This seamless integration of perception and decision-making processes in the autonomous driving system enables rapid understanding of road markings and timely decisions about driving behaviour. To assess the accuracy and effectiveness of the generated driving prompts a reference text is established for each road marking based on its official description and the associated driving operation. This reference text serves as the standard for evaluating the generated prompts. First the semantic similarity between the reference text and the generated driving prompts is calculated as a baseline measure. Subsequently, a manual fine-grained analysis is conducted, focusing on specific driving operations as the key information points. A reward mechanism is employed to increase the similarity value if the generated prompts contain these key information points. Table 1 presents the detection rates and the values of the comprehensive semantic similarity between the generated driving prompts and the reference text for each road marking.

Table 1 – *Detection rates and semantic similarity results for driving prompts*

Type of road marking		**Detection rate**	**Similarity**
Bus Lane		94.1%	88.53%
Bus Stop		88.6%	91.10%
White diagonal stripes		100%	93.28%
Yellow box junction		89.5%	85.99%
No Entry		83.3%	90.62%
Slow		88.6%	88.77%
Give way		96.7%	90.25%
Keep Clear		85.7%	85.77%
Cycle Lane		100%	92.19%
Speed Limit sign		90%	88.10%
Lane Markings	Single solid	90.33%	91.3%
	Single dash	91.29%	72.4%
	Double solid	91.11%	90%
	One dash and one solid	83.10%	86.8%

The overall average detection rate of road markings achieves 89.79% surpassing the average accuracy reported in (Liu et al., 2017) standing at 67.56%. It is worth noting that images captured during nighttime, rainy weather and faded road markings are the most challenging to deal with. Despite these adverse conditions, the new developed model demonstrates robust performance, achieving consistently high detection rates for each road marking. This resilience highlights the model's ability to reliably identify road markings in real-world scenarios, contributing to enhanced safety and efficiency in road infrastructure management. Furthermore, the overall average similarity of road markings can reach 89.31%, indicating the accuracy and effectiveness of the generated driving prompts. This comprehensive evaluation approach ensures that the prompts adequately convey the necessary driving instructions, further improving the system's utility in facilitating safe and informed decisions.

4 Discussion

In this paper, RoadCLIP, a VLM-based road marking detection model, and RoadGPT, an LLM-based road marking QA model, were developed. In addition, an end-to-end pipeline for road marking detection and driving prompt generation was proposed. These innovations present a significant advancement in the field of autonomous driving technology. Compared to traditional detection methods, RoadCLIP can detect road markings at 89.79% of average detection rate with reduced computational resources and training data. This efficiency is crucial for the practical implementation of autonomous driving systems, where resource constraints are often a concern. Furthermore, RoadCLIP's proficiency in recognising both real-world and AI-generated road marking images underscores its versatility and robustness. The integration of RoadGPT and the pipeline represents a new approach to connecting the perception and decision-making process of autonomous driving. By leveraging the reasoning capabilities of LLMs, RoadGPT generates driving prompts that aid in decision-making. The overall driving prompts generated by this pipeline reach an average similarity of 89.31% for each road marking. This seamless integration enhances the vehicle's understanding of driving scenarios, facilitating quicker and more informed decision-making—an essential aspect of autonomous driving. However, the rationality, accuracy, and real-time performance of the generation of LLM-based driving prompts need to be further explored, which may imply the consumption of more computational resources. In the future, for autonomous driving, the application of VLMs and LLMs should be further explored to enhance the vehicle's ability to understand the driving scenarios with these models, and to closely link the perception, planning and decision-making processes.

5 REFERENCES

1. Jayasinghe, O., Hemachandra, S., Anhettigama, D., Kariyawasam, S., Rodrigo, R. and Jayasekara, P., 2022. Ceymo: See more on roads-a novel benchmark dataset for road marking detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 3104-3113).
2. Liu, X., Deng, Z., Lu, H. and Cao, L., 2017, October. Benchmark for road marking detection: Dataset specification and performance baseline. In 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC) (pp. 1-6). IEEE.
3. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. and Krueger, G., 2021, July. Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PMLR.
4. Sha, H., Mu, Y., Jiang, Y., Chen, L., Xu, C., Luo, P., Li, S.E., Tomizuka, M., Zhan, W. and Ding, M., 2023. Languagempc: Large language models as decision makers for autonomous driving. arXiv preprint arXiv:2310.03026.
5. Zhang, J., Huang, J., Jin, S. and Lu, S., 2024. Vision-language models for vision tasks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence.