

Clip-based road-marking detection with LLM-guided driving prompts

Shaofan Sheng^{a,*}, Nicolette Formosa^b, Yuxiang Feng^a, Mohammed Quddus^a

^a Department of Civil and Environmental Engineering, Imperial College London, London, UK

^b National Highways, Birmingham, UK

ARTICLE INFO

Keywords:

Road marking detection
RoadGPT
Decision-making
Autonomous vehicles
Visual-Language Models (VLMs)

ABSTRACT

The advancement of artificial intelligence (AI) has significantly improved the perception and decision-making abilities of autonomous vehicles (AVs), yet real-time and accurate road marking detection remains difficult under faded markings, nighttime scenes and adverse road-weather conditions. This paper presents RoadGPT, a vision-language pipeline that couples a CLIP-based detector (RoadCLIP) with a planner-facing, LLM-generated advisory layer. The model is trained and evaluated on 3,696 road marking images covering 14 UK Highway Code classes, comprising 2,576 real images (69.7 %) from Google Street View and 1,120 text-to-image synthetic images (30.3 %) that broaden rare appearances and degraded conditions. We use 2,968 images for training (2,072 real, 896 virtual) and 728 images for testing (504 real, 224 virtual), keeping the same 70:30 real-virtual ratio in both splits. On this test set, RoadCLIP attains 98.5 % precision, 97.9 % recall and 98.2 % F1 for non-lane markings, while lane-marking subclasses reach up to 89.8 % F1. The advisory layer transforms recognised markings into structured driving prompts and is assessed via semantic similarity using Sentence-BERT (all-mpnet-base-v2) cosine scores against Highway Code-based references, achieving 89.3 % similarity, alongside an external LLM-as-judge rating of 4.68/5 for accuracy, completeness, and concise effectiveness. The full camera-to-advisory path runs in real time at 135 FPS (batch size 1, 224 × 224) on an RTX 4070 under a unified timing protocol. A remaining limitation is that visually similar lane-marking classes and extreme low-light scenes still reduce discriminability compared with symbol-like, non-lane markings.

1. Introduction

The fast development of artificial intelligence (AI) has greatly enhanced how autonomous driving systems perceive their environment and make decisions, paving the way toward Level 5 autonomous driving. Autonomous vehicles, equipped with advanced technologies, can enhance road safety, efficient driving, and improved accessibility (Bimbraw K, 2015). Among these technologies, the accurate detection of road markings plays a vital role in navigation and traffic law compliance (Morrissett & Abdelwahed, 2020).

Road markings like lane lines, pedestrian crossings, and traffic symbols carry essential cues that help AVs stay in their lanes and adapt to changing road scenarios. As a result, the effectiveness and safety of autonomous driving systems are deeply reliant on how precisely these markings are detected.

Conventional road marking detection methods typically rely on feature extraction using machine learning algorithms, involving manual identification of colour, texture, and shape features (Bailo et al., 2017;

Chen et al., 2015; Greenhalgh & Mirmehdi, 2015; Gupta & Choudhary, 2018; Wu et al., 2014). However, the effectiveness of these methods depends significantly on feature extraction quality, which can be labour-intensive and limited in generalisation across different environments.

The rise of deep learning has notably advanced road marking detection techniques due to its capacity to handle complex data in diverse scenarios (Jayasinghe et al., 2022; Lee et al., 2017; Liu et al., 2017; Ye et al., 2020). Recently, transformer-based models, which process data in parallel and capture global dependencies, have shown promising results in various computer vision tasks (Carion et al., 2020; Dosovitskiy et al., 2020; Vaswani et al., 2017). For instance, Yang et al. (2022) successfully utilised transformers for lane detection in challenging scenarios.

Despite their performance, deep learning methods present challenges including high computational demands and the need for extensive annotated datasets (Zhao et al., 2019). Vision-Language models (VLMs), particularly CLIP (Contrastive Language-Image Pretraining), offer new opportunities by effectively integrating image and text data, enabling

* Corresponding author.

E-mail addresses: shaofan.sheng21@imperial.ac.uk (S. Sheng), nicolette.formosa@nationalhighways.co.uk (N. Formosa), y.feng19@imperial.ac.uk (Y. Feng), m.quddus@imperial.ac.uk (M. Quddus).

<https://doi.org/10.1016/j.jag.2025.105012>

Received 25 May 2025; Received in revised form 12 November 2025; Accepted 6 December 2025

Available online 8 December 2025

1569-8432/© 2025 Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

complex semantic understanding and cross-modal learning (Radford et al., 2021).

The limited availability of publicly datasets further challenges road marking detection research. Existing datasets, such as TRoM (Liu et al., 2017), VPGNet (Lee et al., 2017), and CeyMo (Jayasinghe et al., 2022), often have limited generalisation due to varying road marking standards across regions. Synthetic datasets which are based on diffusion models like DALL-E 3 (Betker et al., 2023) gives a viable solution, enhancing data diversity and adaptability.

To address these challenges, this paper aims to build and evaluate a vision–language pipeline, RoadGPT, that uses a CLIP-based detector (RoadCLIP) to recognise UK road markings and an LLM-generated advisory layer to express them as planner-facing driving prompts. Based on this motivation, we ask the following research question: Does injecting text-conditioned attention together with a mixed real–virtual training regime improve the detection and description of non-lane road markings under degraded conditions (low light, fading, partial occlusion)? To make this question testable on our dataset and protocol, we formulate three hypotheses: (H1) text-conditioned attention improves non-lane marking detection in low-light and faded scenes compared with an otherwise equivalent vision-only baseline — that is, we test whether text-conditioned attention improves non-lane detection in low light and under fading, as suggested by the reviewer; (H2) keeping a 70:30 ratio of real to synthetic road-marking images in both training and testing improves cross-condition generalisation over using real-only data; and (H3) LLM-guided driving prompts generated from CLIP-based detections achieve high semantic alignment (target $\geq 85\%$) with Highway-Code-based reference descriptions, so they can be safely used in advisory-only driving stacks.

Key contributions of this paper are summarised as follows:

- Text-conditioned lightweight attention for road markings. We introduce a text-conditioned multi-head attention on top of CLIP that encodes an explicit cross-modal inductive bias—text queries attend to image tokens—specialising elongated/glyph-like and composite layouts with negligible overhead; the perception module runs at ~ 214 FPS (BS = 1, 224×224), preserving low-latency suitability for AV deployment.
- Open-vocabulary CLIP adaptation with label-side edits. We provide a domain adaptation recipe that is, to our knowledge, among the first systematic uses of CLIP for road-marking semantics and decision support: a two-label supervision regime (geometry-aware Label 2 cosine pre-training; canonical Label 1 cross-entropy fine-tuning) plus text-side label edits (synonyms, regional variants, composites) enabling cross-region portability without retraining.
- Curated data with AI-assisted augmentation. We apply the method to a curated dataset mixing real images with text-to-image-generated markings to cover rare or degraded cases (e.g., adverse weather, nighttime, faded), offering a reproducible augmentation pipeline for robustness.
- RoadGPT: planner-consumable advisory interface. We build an end-to-end from camera input through RoadCLIP to advisory pipeline that converts recognised markings into schema-validated, confidence/ODD checks planner advisories (speed caps, no-stop polygons, lane-occupancy penalties, yield flags) with deterministic decoding; prompts reach 89.3 % semantic similarity and LLM-J = 4.68/5, and the full path sustains ~ 135 FPS under a unified timing protocol.

2. Literature review

This section reviews current road marking detection methods, the integration of vision-language models in computer vision tasks and autonomous driving, and the datasets available for road marking detection. It shows the evolution of techniques, the advantages and limitations of existing methods, and the relevant models in road marking detection.

3. Road marking detection methods

Current detection methods rely primarily on machine learning and deep learning. These techniques aim to improve the accuracy and efficiency of detecting various types of road markings under different conditions.

In (Greenhalgh & Mirmehdi, 2015), maximally stable extremal regions (MSER) (Matas et al., 2004) are used to detect the candidates for road markings. Symbol-based road markings were classified and recognised combined with histogram of oriented gradient (HOG) features and support vector machines (SVM). For text-based regions, the optical character recognition (OCR) algorithm (Memon et al., 2020) is used to obtain the meaning of road markings. Similarly, MSER was also used in (Bailo et al., 2017) to obtain candidate regions which are merged using density-based clustering and PCANet (Chan et al., 2015) in combination with SVM is used for classification. Binarized Normed Gradients (BING) (Cheng et al., 2014) methods are also utilised for the initial detection of potential road markings in Chen et al. (2015), followed by the use of SVM for classification.

In Liu et al. (2017), advanced deep learning techniques, specifically a model named RPP (ResNet (He et al., 2016) with Pyramid Pooling), which combines the strengths of ResNet architecture and pyramid pooling to detect road markings using the TRoM dataset. Zhang et al. (2018) enhanced the Fast R-CNN framework for road marking detection while Ye et al. (2020) proposed a two-stage road marking detection method using YOLOv2 for initial detection and a custom lightweight, transformation-invariant classification network (RM-Net) for reclassification. Jayasinghe et al. (Jayasinghe et al., 2022) used two Mask R-CNN (He et al., 2017) based architectures and two SSD (W. Liu et al., 2016) based-models to do the instance segmentation and detection of road markings, enhancing the diversity and complexity of scenarios in road marking detection.

Attention mechanisms enhance accuracy by capturing detailed road marking features. Yu et al., 2021 introduced MarkCapsNet, a self-attention guided high-resolution capsule network for extracting road markings from aerial images. This approach enhances the accuracy of pixel-wise marking extraction and is robust to various types of road markings in different scenarios. Another method for road marking segmentation was proposed by Wu et al. (2022), which employs a multi-scale attention-based dilated CNN (MSA-DCNN) to process the dataset. The method utilises multiscale attention to combine weighting outputs of different scale inputs with dilated convolution to get information from spatial-context dimensions, demonstrating significant improvements over existing techniques in road marking segmentation.

Beyond earlier CNN-only or attention-augmented pipelines, recent studies push road-marking perception toward robust pixel-level modeling and at-scale inspection. Zhang et al. (2024) introduce a hybrid convolutional–transformer architecture (C-Transformer) for automated pixel-level pavement marking detection, explicitly combining self-attention with convolution to recover thin/elongated structures and reduce discontinuities. In parallel, Wu et al. (2024) present a Street View image-based inspection system that uses semantic segmentation with inverse perspective mapping and thresholding to quantify marking damage from videos captured by commodity cameras, underscoring practical data pathways for city-scale assessment. Complementing these task-specific advances, Bakirci (2025) systematically analyzes YOLOv8 variants for intelligent transportation, reporting explicit speed–accuracy trade-offs and compute/hardware profiles (e.g., nano/small variants for low-latency scenarios). Although centered on vehicles rather than markings, these findings reinforce the importance of lightweight, real-time design choices for embedded road-scene perception.

Our approach follows this deployment-oriented ethos—keeping the backbone intact and introducing a lightweight, text-conditioned attention head to align morphology with semantics—while advancing from detection to regulatory meaning and advisory prompts for decision

support.

4. Vision language models in computer vision tasks and autonomous driving

The combination of visual data and natural language (Lu et al., 2019) has enabled models to interpret images at an open-vocabulary and more human-like level, which is particularly attractive for road scenes where many elements are text-like or symbol-like rather than richly textured. Among existing vision-language models, CLIP (Radford et al., 2021) is especially influential because its contrastive pre-training provides strong transferable representations that work on downstream tasks without task-specific re-training. On top of CLIP, Lüddecke and Ecker (2022) introduced CLIPSeg, adding a transformer-based decoder to perform dense prediction from text or image prompts, and Zhong et al. (2022) showed that aligning image regions with text improves open-vocabulary detection.

Compared with transformer-based detectors trained from scratch or on purely visual corpora (e.g. DETR- or Mask2Former-style models), this CLIP-style transfer is better suited to road markings. Markings in the UK context are often low-contrast, elongated, partly faded, or captured at night, so their visual signal is weak, but their linguistic identity (“SLOW”, “BUS LANE”, “KEEP CLEAR”, “Give Way”) is strong and regulated. Text-conditioned attention from a CLIP-like model can therefore pull apart visually similar but semantically different markings and even support rare or newly appearing categories without collecting large amounts of degraded training images. This is the main reason we build on CLIP rather than training a detector from scratch on our relatively small, long-tailed dataset.

Building on vision-language foundations, recent work has started to use large language models (LLMs) to verbalise or evaluate driving decisions. Sha et al. (2023) proposed LanguageMPC, where an LLM interprets the scene and guides control, and Cui et al., 2023 introduced DriveLLM to improve robustness and interaction. Our work follows this line but uses the LLM layer mainly to turn CLIP-based road-marking detections into planner-facing driving prompts and to evaluate them via a language interface.

5. Datasets for road marking detection

Datasets serve as the foundation upon which models are trained, validated, and tested. These datasets provide a vast array of examples from which deep learning algorithms can learn patterns, features, and relationships within the data. When it comes to road markings, their diversity, different shapes, and susceptibility to environmental factors, it makes collecting and labeling road marking data more complex and time-consuming. Therefore, publicly available datasets are essential for road marking detection tasks.

T. Wu and Ranganathan (2012) presented the first publicly accessible dataset for road marking detection, featuring 1443 annotated images across 11 distinct classes of road markings. However, since their primary focus is on approaches based on image processing, it lacks distinct training and testing sets and does not specify a clear evaluation metric. The TRoM dataset (Liu et al., 2017) has 712 images including 19 distinct classes of lane and road markings. However, the design of this dataset’s annotation format and evaluation metric are specifically for approaches based on semantic segmentation, thereby restricting its applicability in algorithms which aimed at detecting non-lane road markings. On the other hand, VPGNet (Lee et al., 2017) is a substantial dataset of approximately 20,000 images, covering 17 different classes of lanes and road markings. It offers detailed annotations at both the pixel and grid levels for lanes and road markings, along with the lanes’ vanishing points. Despite its extensive coverage, the dataset only includes annotations for 8 specific road marking classes, while the remaining are labelled as “other road markings”. VPGNet has limited application in road marking detection in further studies since its aim is to enhance lane

detection through vanishing point guidance. The CeyMo dataset (Jayasinghe et al., 2022), a new benchmark for road marking detection, including 2,887 images, with a resolution of 1920×1080 . This dataset is segmented into a training set with 2,099 images and a testing set with 788 images. Within the test set, the 788 images are further categorised into six groups representing one normal and five challenging conditions: crowded environments, dazzle light, nighttime, rainy weather, and shadowed areas.

The acquisition and annotation of real-world data in computer vision tasks have challenges due to their time-consuming, labour-intensive, and error-prone nature. Recognising these challenges, researchers are increasingly exploring the use of virtual images as training and testing datasets to overcome these limitations. ParallelEye dataset introduced by Li et al. (2019) has contributed to traffic research. This fully virtual dataset has shown effectiveness in improving the accuracy of object detection and semantic segmentation algorithms. The study emphasises the significance of environmental conditions on detection performance.

In summary, to address these challenges, this paper modifies CLIP to autonomous driving tasks, which is not widely used in existing studies. By leveraging CLIP’s cross-modal learning capabilities, the proposed RoadCLIP method aims to reduce high computational resources and improve model robustness under different conditions. This study also utilises synthetic images to enlarge the dataset and large language models to facilitate decision-making process. In addition, the paper develops RoadGPT, combining RoadCLIP with GPT-4’s reasoning ability to provide concise but accurate driving prompts in real time, thereby enhancing the decision-making of autonomous driving systems.

6. Dataset description

This section introduces the data collection and composition, giving an overview of the real and virtual road marking images used in this paper. It describes the dataset’s structure, the data enhancement methods, and the synthetic data used to ensure the diversity of dataset. Notably, the data here was used to fine-tune CLIP for the task of road marking detection, which does not require a large number of new data for training due to the excellent transfer learning capability of the CLIP model (Radford et al., 2021).

7. Data collection and composition

A comprehensive road marking dataset was created for training and evaluation. This dataset includes a large amount of real and virtual road marking images, which ensure robustness and adaptability in different driving scenarios.

A total of 14 distinct road markings has been collected, according to the categorisation in the UK Highway Code (The Highway Code, 2015). This official document by the UK government provides guidance for road users. Our class taxonomy follows the UK Highway Code for concreteness; to mitigate regional bias, we map classes to international equivalents via text-side label edits (for example, ‘GIVE WAY’ / ‘YIELD’) and we report zero-shot cross-region results (AU/US). We nonetheless acknowledge that some region-specific priors may remain. The number of images per class is shown in Table 1. All the core 3,696 images in this study were collected in the United Kingdom and follow the UK Highway Code taxonomy. To examine how well the proposed model transfers to road markings that are visually similar but come from other traffic jurisdictions, we additionally built a small transferability set of 334 images from China mainland, Hong Kong, Japan and Australia. This cross-region set is used only in the transferability experiments reported later (Table 10); it is not mixed into the UK training or test splits. For non-UK images, we only selected markings that could be mapped to one of the 14 UK categories (for example, yellow box, bus lane legends, dashed centre lines), so that the model could be evaluated through text-side label edits without changing the core taxonomy.

The real-world dataset was obtained from Google Street View images

Table 1

Number of images for per-class.

Class	Total	Real	Virtual
Non-Lane Markings	Bus Lane	248	173
	Bus Stop	232	162
	White diagonal stripes	256	178
	Yellow box junction	280	195
	No Entry	236	164
	Slow	292	204
	Give way	284	198
	Keep Clear	260	181
	Cycle Lane	252	176
	Speed Limit sign	240	167
	Single solid	304	212
Lane Markings	Single dash	312	218
	Double solid	248	173
	One dash and one solid	252	175
	Total	3696	2576

(Google, 2015), containing 2576 images capturing various road markings under diverse conditions, including different dates, weather conditions and driving situations. The number of images for different conditions in the dataset was shown in Table 2.

Manual selection and classification ensured dataset accuracy. Nighttime scenes were identified using timestamps in Google Street View images. Faded road markings are categorized by inputting the image into HSV space calculating the average luminance ratio between the marking area and the background area. The average luminance ratio R is computed in grayscale space:

$$R = \frac{I_{\text{marking}}}{I_{\text{bg}}} \quad (1)$$

where I_{marking} is average luminance of markings and I_{bg} is average luminance of background.

To obtain the marking area, we use two slightly different procedures. For non-lane markings (Bus Lane, Yellow box junction, SLOW, Give Way, etc.), the marking region is obtained by adaptive thresholding in HSV on the full image followed by morphological closing, because each image contains a single dominant marking. For lane-type markings (single solid, single dash, double solid, and one dash and one solid), we reuse our previously trained UK lane-marking detector (Sheng et al., 2024) to produce a pixel-level lane mask that is more robust under low light and wet asphalt. The closer the value of R is to 1, the more faded the marker is, so we classify the markers as 25 % faded, 50 % faded and 75 % faded based on the value of R . For illustration:

- **KEEP CLEAR** on dry asphalt: $I_{\text{marking}} = 182$, $I_{\text{bg}} = 108$, $R = 1.68 \rightarrow 25\%$ faded.
- **BUS LANE** legend with visible wear: $I_{\text{marking}} = 141$, $I_{\text{bg}} = 104$, $R = 1.36 \rightarrow 50\%$ faded.
- **nighttime Single dash**: $I_{\text{marking}} = 118$, $I_{\text{bg}} = 111$, $R = 1.06 \rightarrow 75\%$ faded.

Adverse weather conditions were categorised by manually reviewing images which include wet roads, reduced visibility, and snow coverage. These approaches allow for accurate classification based on human judgment. Fig. 1 shows examples from the real-world road marking

Table 2

Number of images for each condition.

Scenario	No. of Images	Percentage
Normal	2240	60.61 %
Nighttime	420	11.36 %
25 % Faded condition	333	9.01 %
50 % Faded condition	289	7.82 %
75 % Faded condition	114	3.08 %
Adverse weather conditions	300	8.12 %

dataset.

Additionally, virtual road marking images were generated to enhance the diversity and coverage of the dataset. These virtual images were created using Text-to-Image models, such as DALL-E 3 from OpenAI (Betker et al., 2023). These models based on textual descriptions can create highly detailed and realistic images, making them suitable for road marking detection tasks. To measure whether the generated road marking images can well simulate the road markings of the real scene, the FID index was used to measure it. FID measures the difference in the distribution of the generated image set and the real image set in the Inception feature space, which reflects both the fidelity of the generated samples and the diversity of the samples (Heusel et al., 2017), and the lower the value of FID indicates that the generative model is more effective. The pre-trained Inception-V3 network (Szegedy et al., 2016) was used for FID feature extraction and calculated the average FID score for the virtual dataset to be 41.57 which we interpret as a moderate distributional alignment given our deliberate coverage of hard conditions (night scenes, adverse weather, and 25/50/75 % wear). We therefore use FID as a distributional indicator rather than a stand-alone proxy for photorealism. To disambiguate interpretation, we also computed a condition-matched FID restricted to Normal scenes using the same Inception-V3 features and preprocessing. This yields FID = 14.21, indicating good alignment when conditions are matched; the higher overall FID reflects our deliberate over-sampling of Nighttime/Adverse/Faded cases to broaden coverage. Each synthetic image is generated with a structured prompt composed of a fixed class token (14 UK Highway Code categories) and enumerated attributes: time-of-day {day, dusk, night}, weather/surface {clear, rain, fog, snow, wet asphalt}, wear level {0/25/50/75 %}, viewpoint/range {front or oblique; near or far}, and occlusion/background {light/heavy traffic, parked cars, pedestrians}. We use simple stratified sampling so that each class is rendered across a balanced subset of these conditions (with fixed random seeds for reproducibility). Basic quality control is applied before inclusion: a text-image consistency checks against the class description and a light morphology sanity check to filter implausible markings. In addition, we fix sampling steps and guidance scale per generation, assign a unique random seed per image, and include negative prompts to suppress common artefacts (e.g., text overlays or geometric warping). Generations violating structural integrity or resolution/cropping constraints are rejected, and we manually spot-check 10 % of accepted samples to monitor failure patterns.

This approach not only increased the diversity of the dataset but also addressed the scarcity of specific road marking types that are less common in real-world data. By using synthetic images, the model can be trained on a large range of scenarios, enhancing its performance accurately in real-world applications. Fig. 2 provides some examples of the virtual road marking images.

8. Dataset overview

The types of road markings present in the study are categorised as follows: Bus Lane, Bus Stop, White Diagonal Stripes, Yellow Box Junction, No Entry, Slow, Give Way, Keep Clear, Cycle Lane, Speed Limit Sign, Lane Markings – Single Solid, Lane Markings – Single Dash, Lane Markings – Double Solid, and Lane Markings – One Dash and One Solid. Table 3 provides comprehensive information on the dataset's composition and the distribution of training and test data. Notably, 70 % of the dataset comprises real road marking images, with the remaining 30 % being virtual road marking images. The same ratio between real and virtual images was kept when splitting training and testing dataset. 80 % of the total images are used for training and 20 % for testing. We reserve 10 % of the UK training split (randomly selected) as an internal validation set for threshold and prompt calibration; the fixed 728-image test set is never used for model or threshold selection.



Figure 1 Examples from the real-world dataset

Fig. 1. Examples from the real-world dataset.



Figure 2 Examples of the synthetic road marking images

Fig. 2. Examples of the synthetic road marking images.

Table 3

Distribution of images (Real vs Virtual).

	Training	Testing	Total
Real world images	2072	504	2576
Virtual images	896	224	1120
Total	2968	728	3696

9. Methodology

This section outlines the methodology used to develop RoadGPT, an innovative system designed to enhance road marking detection and decision-making in autonomous driving. The approach leverages the powerful capabilities of the CLIP model, enhanced with a custom multi-head attention mechanism and integrates GPT-4 for real-time context-aware driving prompts.

10. CLIP model

The CLIP model uses contrastive learning, employing two independent encoders to extract the features from images and texts respectively and maps them into the same high dimensional space. By maximising the similarity between correct image-text pairs (positive pairs) and

minimising it between the incorrect pairs, CLIP effectively learns close associations between images and text. This training strategy endows CLIP with powerful multimodal understanding, zero-shot learning, and cross-task generalisation capabilities.

The CLIP model is trained using the Information Noise Contrast Estimation (InfoNCE) loss function (Oord et al., 2018) given as Eq. (2):

$$L = -\log \frac{\exp\left(\frac{\text{sim}(u,v)}{\tau}\right)}{\sum_{j=1}^N \exp\left(\frac{\text{sim}(u,v_j)}{\tau}\right)} \quad (2)$$

where u and v are the feature representations of the image and the corresponding text, respectively, $\text{sim}(u,v)$ is a cosine similarity measure between these two representations, τ is a temperature parameter controlling the smoothing of the distribution, N is the number of negative samples, and j represents the text representations in the negative samples.

11. Road marking detection based on the CLIP model

RoadCLIP is based on a pre-trained CLIP model. CLIP model has powerful image and text feature extraction capabilities and can be a

good starting point to be fine-tuned for the road marking detection task.

To enhance accuracy, a custom multi-head attention mechanism was introduced (see Fig. 3 below) to overcome the limitations of the stand-alone CLIP model. This is termed as RoadCLIP. This allows for better integration of image features with textual descriptions of road markings. The multi-head attention layer refines image feature representation, aligning it more closely with textual descriptions relevant to road marking detection allowing the model to prioritise visual features relevant to the task.

12. Customised multi-head attention mechanism for CLIP

To augment the model's understanding of road markings, a custom multi-head attention layer is added on top of the pre-trained CLIP model. This additional layer focuses on making the model pay more attention to the road marking features in images according to the text description. By adding this customised attention mechanism, the goal is to make the model focus on the road marking features while ignoring the unrelated environments information even in complex road detection tasks.

The multi-head attention mechanism (see Fig. 3) involves the following steps:

1. **Linear Transformation:** Initially, the image and text features undergo linear transformation through different fully connected layers, resulting in three sets of vectors for Query (Q), Value (V), and Key (K). Linear Transformation (text-conditioned): Derive the key K from the class text embedding t (Label 2 during attention training; Label 1 during fine-tuning), and derive queries/values Q,V from the image token matrix X output by the CLIP image encoder; This makes the attention cross-modal (text-to-image) rather than standard self-attention (image-to-image).
2. **Division into Multiple Heads:** The Q, K, and V vectors are divided into multiple heads, enabling the model to focus on information in different subspaces simultaneously.
3. **Attention Score Computation:** For each head, the attention score is computed by calculating the dot product of the Q and K vectors, normalised by the softmax function to obtain the weight of each value vector:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3)$$

where d_k is the dimensionality of the key vectors.

1. **Output Vector Formation:** Using the weights obtained in the previous step, the V vectors are weighted and summed to derive the output vector for each head.

2. **Concatenation and Linear Transformation:** The output vectors of all the heads are concatenated and linearly transformed through a fully connected layer to obtain the final output feature:

$$\text{Output} = W_o \cdot \text{Concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_n) \quad (4)$$

where W_o is the weight matrix of the final fully connected layer, and head_n represents the output vector from the n th attention head. This output feature is weighted, giving more emphasis to the visual content that aligns with the provided text description compared to the original image feature.

Unlike standard self-attention—where queries, keys, and values (Q/K/V) are all derived from image tokens—our module performs text-conditioned cross-attention: the key is computed from the text embedding t of a road-marking class, while queries/values come from the image tokens X produced by the CLIP image encoder. Formally, given $X \in R^{N \times d}$ and $t \in R^d$, we set $Q = tW_Q$, $K = XW_K$, $V = XW_V$; compute $A = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)$, $Z = AV$; and fuse Z through a linear projection and

residual connection to obtain the enhanced image feature. Pre-training of this head maximises the cosine similarity (Eq. (5)) between the attended image feature and the detailed Label 2 description, and the unified model is then fine-tuned with cross-entropy (Eq. (6)). This design explicitly aligns visual evidence with the semantic description and explains the observed gains on elongated/glyph-like and composition-sensitive markings.

13. Training process

The training process and loss are shown in Fig. 3. During model construction, two sets of text labels, namely Label 1 and Label 2, are prepared. Label 1 is the name label of the road marking, while Label 2 provides a detailed description of the road marking pattern, colour, and composition. For instance, a “Give Way” sign is labeled as “It’s a Give Way sign.” in Label 1 and “Give Way sign consisting of hollow inverted triangles.” in Label 2. We use two label sets with disjoint roles. Label 1 is the canonical class name for training/evaluation with cross-entropy; it is kept concise and normalised (lowercasing, punctuation stripping, capped token length), and the CLIP text encoder is frozen with precomputed text embeddings. Label 2 is a structured description template (shape/colour/composition) used only to pre-train the lightweight attention head with a cosine objective; the template is grammar-constrained, excludes contextual or region-specific hints, and is length-capped. Thus, any benefit from Label 2 stems from geometric specificity rather than verbosity, while all reported F1 results use Label 1 at inference.

During the training of the multi-head attention layer, Label 2 was used. We first encode the road-marking image and its paired textual description with the OpenAI CLIP ViT-B/32 backbone (input size 224×224), so that both the image and the text are represented as tensors of

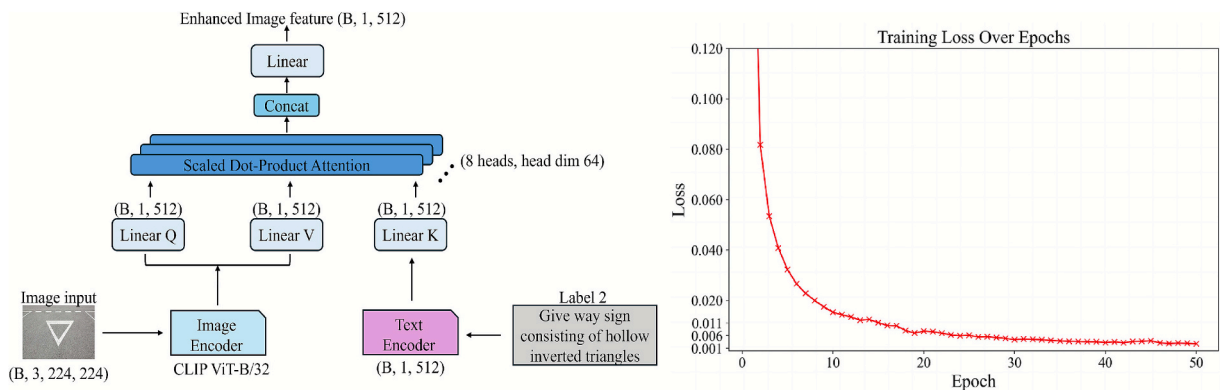


Fig. 3. Training process and Training Loss figure.

shape (B, 1, 512), where B is the batch size. On top of these CLIP features we add a text-conditioned multi-head attention block. Given the image feature (1 × 512) and the text feature (1 × 512), we apply an 8-head scaled dot-product attention with hidden size 512. The query and value are projected from the image feature, and the key is projected from the text feature; the eight heads (each of size 64) are concatenated and passed through a final linear layer back to 512 dimensions. In this implementation we do not apply layer normalisation before attention, but a pre-norm LayerNorm can be added if more training stability is needed.

This process includes putting image features and corresponding text descriptions into the attention layer and highlighting the relevant road-marking features based on the provided textual descriptions. The chosen loss function for training the attention layer is the cosine similarity loss function (Rahutomo et al., 2012), which measures the similarity between two vectors and is defined as Eq. (5):

$$L(x, y) = \begin{cases} 1 - \cos(x, y), & \text{if } y' = 1 \\ \max(0, \cos(x, y) - \text{margin}), & \text{if } y' = -1 \end{cases} \quad (5)$$

where, x and y are two vectors respectively, y' is a label indicating and $\cos(x, y)$ denotes the cosine similarity.

To optimise model performance, a grid search was conducted to fine-tune hyperparameters. The values of learning rate are set to be 0.0001, 0.001 and 0.01, batch size is 8, 16 and 32, and the number of epochs is 30, 50 and 100. After systematic evaluation, the optimal hyperparameter combinations were identified: an Adam optimiser with a learning rate of 0.001, number of epochs is 50, and the batch size is 16. This rigorous hyperparameter chosen process ensured effective training of the attention layer, improving the model's capability to accurately detect road markings in complex road environments.

While a joint multi-task loss is feasible (e.g., $L = \lambda L_{\cos} + (1 - \lambda) L_{ce}$), we adopt a staged schedule for stability and supervision fidelity under limited domain data: the text-conditioned head is first aligned to Label 2 with a cosine alignment loss on frozen (or lightly tuned) CLIP features, and only then do we fine-tune with Label 1 cross-entropy for discriminative accuracy. The advisory layer remains modular and advisory-only (no backpropagation into perception), keeping thresholds, verification, and failure modes explicit for safety. In our setting, staging reduced optimisation variance and mitigated backbone drift compared with a direct joint optimisation, while retaining the option to unify if future data/compute permit.

14. Integration and fine-tuning

After the training for the multi-head attention layer, it is integrated with the pre-trained CLIP model. The unified model is fine-tuned using Label 1 and training images. The fine-tuning process refines the model's performance to recognise and understand road markings when it comes to the specific task requirements.

After fine-tuning, images from the test dataset were used to evaluate the model, along with the corresponding Label 1 annotations. This evaluation assesses the performance of the model in recognizing and classifying road markings. Fig. 4 provides a visualization of the entire process, illustrating the comprehensive detection method of RoadCLIP.

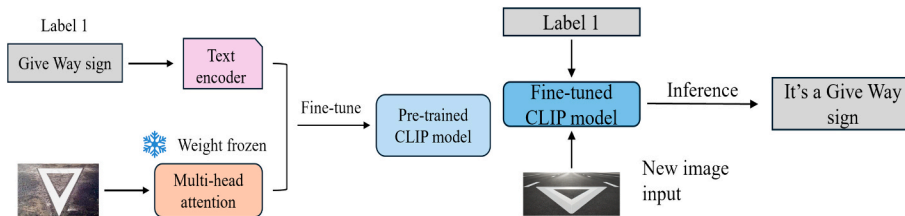


Fig. 4. Fine-tune and Inference of the RoadCLIP.

The top panel shows attention-head training with Label 2 and a cosine objective to emphasise elongated and glyph-like markings without disturbing the CLIP backbone. The left panel shows CLIP fine-tuning with Label 1 using cross-entropy to optimise discriminative performance. The right panel shows single-pass inference (from camera input through RoadCLIP to advisory prompts). This staging improves optimisation stability and preserves a safety-modular design, while inference remains a single pass. The cross-entropy loss function was used during the fine-tune process (Zhang and Sabuncu, 2018):

$$L(y, \hat{y}) = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^C y_{ij} \log(\hat{y}_{ij}) \quad (6)$$

where N is number of samples, C is number of classifications, y_{ij} is an indicator that if sample i belongs to class j , then $y_{ij} = 1$, otherwise 0. $\hat{y}_{ij}$ is the probability that the model predicts that sample i belongs to category j .

Our inference is unified (from camera input through RoadCLIP with the proposed attention head to advisory prompts). The training is staged by design for stability, supervision fidelity, and safety modularity. First, we align the attention head with text semantics using a cosine objective (Label 2) so that elongated, and glyph-like markings are emphasised without disturbing the generic CLIP backbone. Second, we fine-tune with cross-entropy (Label 1) to optimise discriminative performance. Third, the GPT-based prompt layer is kept advisory-only and decoupled from perception, so that confidence gates and planner costs/constraints remain explicit. A joint end-to-end formulation is possible by combining the two losses into a multi-task objective; we adopt the staged schedule to preserve backbone generality, reduce convergence sensitivity on limited domain data, and keep the advisory layer modular for safety.

The optimiser used during fine-tuning was stochastic gradient descents (SGD) with a learning rate of 0.00001, momentum of 0.9, and weight decay of 0.001. The Epoch number and batch size are 50 and 16, respectively. These training configurations contribute to refining the CLIP model, aligning its outputs more closely with the intended classification labels and enhancing its capability to accurately detect and classify road markings.

15. RoadGPT: Final pipeline based on RoadCLIP and GPT-4

By integrating RoadCLIP with GPT-4, we form a comprehensive end-to-end pipeline. It substantially improves the system's ability to interpret road information and to generate concise, context-aware driving prompts, thereby supporting more informed decision-making.

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. Fig. 5 illustrates the construction process of RoadGPT and the flowchart of the pipeline.

At runtime, RoadCLIP and the advisory layer run in a single pass: camera frames are processed to produce typed road-marking events {class, geometry, confidence, timestamp, which are converted into confidence-gated prompts and planner-readable flags (e.g., lane-occupancy penalty, no-stop polygon, speed cap). The module is

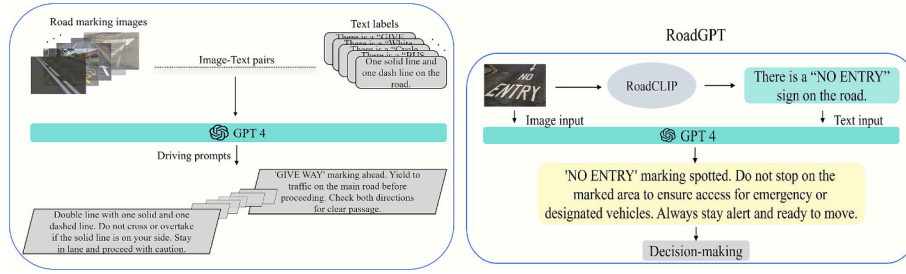


Fig. 5. Construction process of RoadGPT and Flowchart of the pipeline.

advisory-only and does not issue control commands; it integrates with the planner as costs/constraints alongside existing safety monitors.

By integrating RoadCLIP with GPT-4, a comprehensive end-to-end pipeline was established that substantially improves the autonomous driving system's ability to interpret road information and generate context-aware driving prompts, thereby supporting more informed 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. It reduces computational resource consumption and links the initially disparate perception and decision-making process for autonomous driving.

In a real vehicle, RoadCLIP runs at the semantic perception layer and RoadGPT serves as an advisory bridge to the behaviour planner. The module subscribes to the front-camera RGB stream after standard pre-processing and publishes typed road-marking events {class, geometry (polyline/ROI), confidence, and timestamp at N Hz. RoadGPT converts these events into confidence-aware prompts and planner-readable flags/costs, for example: BUS LANE ahead to add lane-occupancy penalty; KEEP CLEAR at junction to define a no-stop polygon; SLOW/SPEED LIMIT to cap the target speed supplied to the longitudinal controller. Advisories are gated by confidence (calibrated on the in-domain validation set) and simple operational-design-domain checks; thresholds are stricter at nighttime in view of the observed performance drop, and the system falls back to the baseline stack when confidence is low. The module is advisory-only—it never issues control commands—and integrates with the existing planner as costs/constraints alongside standard safety monitors. This keeps the backbone intact while enabling low-latency, text-conditioned semantic guidance. This design choice is consistent with lightweight detection studies that emphasise speed-accuracy trade-offs under embedded constraints; here, we extend perception outputs to planner-readable advisories while keeping the module lightweight and safety-gated.

All training was performed on an AMD Ryzen 5 3600 CPU and an NVIDIA RTX 4070 GPU with 32 GB RAM. The previously quoted speeds refer to training throughput (forward + backward) at 224×224 input resolution with batch size 16: attention-head training achieved 217.39 images/s, and CLIP fine-tuning achieved 177.28 images/s. For inference, we report batch size 1 latency at 224×224 (ms per frame and the equivalent FPS), measured with CUDA events after warm-up, including the CLIP image encoder forward and our attention head, and excluding data loading and GPT-4 prompt generation. Under this unified protocol, the perception module (RoadCLIP) runs at 4.67 ms per frame (~ 214 FPS), and the complete from camera input through RoadCLIP to advisory layer path (including schema validation and confidence/ODD checks) sustains 7.41 ms per frame (~ 135 FPS) on the RTX 4070. We use “throughput” only for training (images/s, e.g., 224×224 , batch size 16, forward + backward), and “latency” only for inference (ms per frame and FPS; 224×224 , batch size 1, forward only). Resolution and batch size are stated alongside every number, and timing is measured with CUDA events after warm-up, which makes the results directly comparable and reproducible. We considered a joint end-to-end formulation that combines the cosine alignment and cross-entropy terms. Given limited domain data and the need to keep the advisory layer modular for

safety, we opted for the staged schedule, which yielded stable optimisation while preserving backbone generality; inference remains a single-pass pipeline.

16. Results

16.1. Detection performance

RoadCLIP was applied on test data which includes various road marking images under different conditions. The model's ability to accurately detect and classify road markings in both real-world and synthetic images shows its robustness and generalisation capability. Fig. 6 shows the detection results, where the captions above the images indicate the most similar result chosen by the model with each road marking image from the provided labels. The successful recognition of both real-world and AI-generated images shows the efficacy of incorporating a mixed dataset during training, which emphasises the positive influence of synthetic data on model performance.

Precision, recall, and F1-score metrics evaluated RoadCLIP's performance. These metrics provide a comprehensive view of the model's accuracy and reliability in detecting various types of road markings. Table 4 shows the evaluation metrics for each of the 14 road markings and Table 5 gives the F1-score of markings in different conditions. Table 6 shows the comparison with state-of-arts. To make the error patterns explicit, we now additionally report the full 14×14 confusion matrix on the fixed 728-image UK test split (504 real, 224 virtual) (see Fig. 7). This matrix shows that the four lane-marking subclasses (single solid, single dash, double solid, and one dash and one solid) are indeed the main source of residual confusions, whereas non-lane markings almost never get misclassified, which is consistent with the 100 % or near-100 % precision/recall entries. Furthermore, to verify that the reported gains are not driven by synthetic images at evaluation time, we additionally report results on a UK real-only test subset (504 images), using the same model configuration, thresholds and hyper-parameters as in the main experiments. On this subset, the macro-averaged precision, recall and F1-score over the 14 classes are 93.6 %, 93.5 % and 93.5 %, respectively. These values are closely aligned with those obtained on the mixed real and synthetic test set, confirming that our conclusions do not rely on synthetic samples at evaluation time.

In addition, we sweep a similarity threshold over the CLIP scores and plot per-class PR curves. The full 14-class and four lane-marking PR curves are shown separately in Fig. 8. Non-lane classes stay close to the top-right corner because they rarely attract false positives, while the four lane-marking classes drop in precision at medium-high recall due to intra-lane confusions (“single dash” vs “one dash and one solid”, “single solid” vs “double solid”). Because Table 1 already lists the per-class image counts and Table 4 reports per-class precision, recall and F1, these two tables together provide the per-condition/per-class breakdown.

Besides, multiple ablation experiments were performed to demonstrate the superiority of using the enhanced image feature for fine-tuning. The overall average precision and recall of lane markings are 89.30 % and 90.30 % respectively, demonstrating superior performance

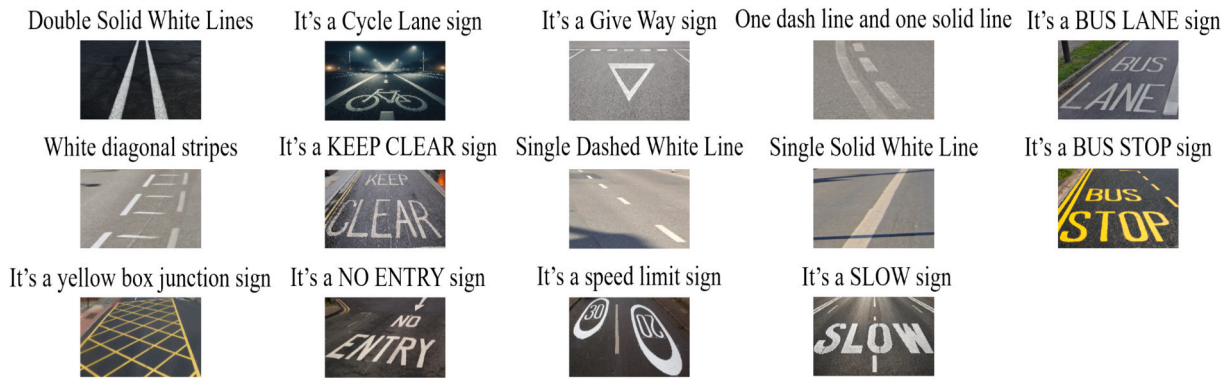


Fig. 6. Test results of road marking detection method.

Table 4

Evaluation metrics for each of the road markings.

Type of road marking		Precision	Recall	F1-score
Non-Lane Markings	Bus Lane	100.0 %	100.0 %	100.0 %
	Bus Stop	100.0 %	92.5 %	96.1 %
	White diagonal stripes	100.0 %	100.0 %	100.0 %
	Yellow box junction	100.0 %	100.0 %	100.0 %
	No Entry	97.6 %	100.0 %	98.8 %
	Slow	100.0 %	98.0 %	99.0 %
	Give Way	95.7 %	97.8 %	96.7 %
	Keep Clear	100.0 %	92.5 %	96.1 %
	Cycle Lane	98.0 %	98.0 %	98.0 %
	Speed Limit sign	93.8 %	100.0 %	96.8 %
Lane Markings	Single solid	80.6 %	96.7 %	87.9 %
	Single dash	100.0 %	85.7 %	92.3 %
	Double solid	93.0 %	88.9 %	90.9 %
	One dash and one solid	83.7 %	90.0 %	86.7 %

Table 5

F1-score in different conditions.

	Normal	Nighttime	Adverse weather	Faded condition
Non-Lane Markings	99.32 %	94.46 %	91.76 %	99.30 %
Lane Markings	91.44 %	85.81 %	89.89 %	82.92 %

compared with existing baseline methods. For non-lane markings, the model achieves even higher performance with an overall average precision of 98.50 %, recall of 97.90 % and F1-score of 98.20 %. These results highlight the significant advancements and improvements introduced by RoadCLIP, particularly in the accurate detection of non-lane road markings. Notably, while some comparison methods lack complete metric data, the approach in this work demonstrates a balanced and robust performance across all evaluated metrics. We quantify robustness for the paired, head-to-head ablation in Table 6 using three quantities on the same test split. First, $\Delta F1$ is the absolute difference in macro-F1 between methods. Second, the two-sided paired-bootstrap p-value estimates how likely it is for the observed $\Delta F1$ to be zero or negative due to sampling noise. Third, Cliff's δ is a non-parametric effect size that indicates how consistently one method wins over the other across images: values near zero mean no consistent advantage, whereas larger positive values indicate a consistent win rate. For Fine-tuned CLIP with Enhanced Image Feature versus Fine-tuned CLIP (Label 2) we obtain $\Delta F1 = +9.14$ percentage points, $p = 7.2 \times 10^{-4}$, and Cliff's $\delta = 0.41$ (medium), indicating a large absolute gain with a very low probability of arising by chance under paired resampling. We focus on this specific contrast because it is a like-for-like, fully paired ablation—same backbone, data split, resolution, optimiser, and

Table 6

Comparison with State-of-the-arts.

		Precision	Recall	F1-Score
VPGNet (Lee et al., 2017)		/	/	88.4 %
		67.6 %	/	/
		96.9 %	90.2 %	/
			%	
CeyMo (Jayasinghe et al., 2022)		/	/	90.6 %
		86.3 %	88.6 %	85.2 %
Road marking in (Gong et al., 2024)			%	
		93.6 %	88.4 %	90.9 %
Road marking in (Zhang et al., 2024)			%	%
		92.0 %	90.6 %	/
Pre-trained CLIP (with Label 1)	Lane Markings	47.3 %	32.9 %	31.1 %
	Non-Lane Markings	74.2 %	58.1 %	57.5 %
Pre-trained CLIP (with Label 2)	Lane Markings	76.0 %	74.3 %	75.2 %
	Non-Lane Markings	81.7 %	78.6 %	80.1 %
Fine-tuned CLIP (with Label 1)	Lane Markings	62.7 %	53.6 %	57.8 %
	Non-Lane Markings	90.3 %	85.4 %	87.8 %
Fine-tuned CLIP (with Label 2)	Lane Markings	83.7 %	86.9 %	85.3 %
	Non-Lane Markings	88.1 %	83.6 %	85.9 %
Fine-tuned CLIP with Enhanced Image Feature	Lane Markings	89.3 %	90.3 %	89.8 %
	Non-Lane Markings	98.5 %	97.9 %	98.2 %

schedule—so it isolates the effect of the proposed lightweight attention head and provides the most faithful statistical test of our claimed improvement. In addition, we repeated the same experiment on the same train/test split with three different random seeds (0, 1, 2), changing only the random initialisation, the shuffling order, and the stochastic augmentations. Across these three runs we obtained: baseline model (CLIP with Label 2): $F1 = 85.7 \pm 0.3$, attention-augmented model: $F1 = 95.6 \pm 0.4$, attention-layer gain: $\Delta F1 = (95.6 - 85.7) = +9.9 \pm 0.2$ percentage points. This confirms that the reported attention-layer gains are stable across runs and complements the earlier paired-bootstrap analysis. Table 7 gives the performance comparison with different state-of-the-arts VLMs. We include OWL-ViT v2 (Minderer et al., 2023) and GLIP (Li et al., 2022) as representative state-of-the-art open-vocabulary, text-conditioned detectors—covering zero-shot query-by-text (OWLv2) and language-grounded detection (GLIP)—to benchmark detector-centric pipelines against our semantic CLIP route. Under the unified setting (same split, 224×224 input, BS = 1 timing protocol),

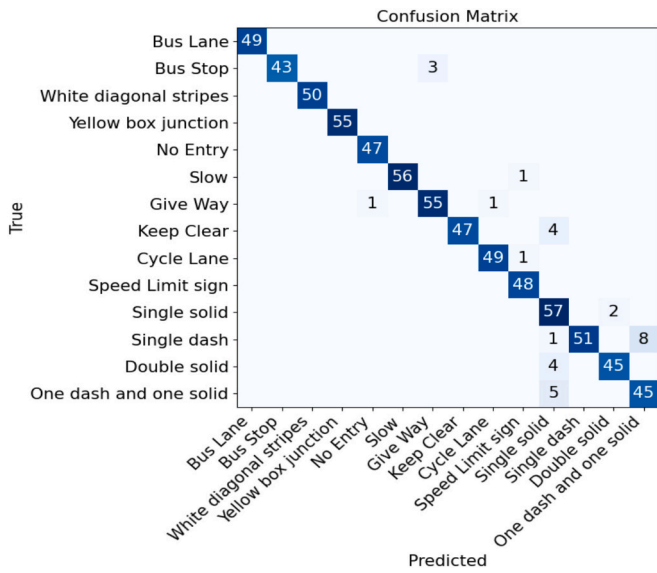


Fig. 7. Confusion matrix per-class.

fine-tuned CLIP achieves 85.6 % F1 at 214 FPS with ~ 151 M parameters, while fine-tuned OWL-ViT v2 yields 88.3 % F1 at 89 FPS with ~ 250 – 300 M parameters, and GLIP obtains 78.4 % F1 at 44 FPS with ~ 500 – 550 M parameters. This highlights a clear accuracy–latency–compute trade-off: OWLv2 offers slightly higher F1 but with $\sim 2.4 \times$ latency and $\sim 2 \times$ parameter cost relative to CLIP; GLIP is heavier and slower with lower F1. Our final choice—CLIP with a lightweight text-conditioned head plus an advisory-only GPT-4 layer—optimises the deployment Pareto (low latency/parameters with strong accuracy once the head is enabled; see Tables 4–6), preserves open-vocabulary label edits for cross-region transfer without retraining, and integrates naturally as planner costs/constraints behind confidence/ODD checks. OWLv2’s detector-centric, box-level interface is well suited to compact symbols but less aligned with elongated/glyph-like markings and planner-readable regions/constraints; GLIP’s cross-modal detection stack adds further overhead. In short, CLIP + GPT-4 provides the best fit for text-side adaptability + low-latency AV integration, while Table 6 quantifies how alternatives would shift accuracy and runtime under the same conditions.

To avoid misunderstanding, we clarify that the virtual percentages shown in Table 8 (25 %, 50 %, 75 %, 100 %) refer only to the composition of the training set. The test set is kept fixed throughout at 728 images, consisting of 504 real images and 224 virtual images, and no virtual image used for training is reused in the test set. We have also added a baseline row trained with real data only to show the starting point. From the table, it shows that the addition of the virtual dataset compensates to some extent for the lack of data volume and increases the effectiveness of model fine-tuning which means that the virtual road

marking images simulate real-world scenarios very well, thus improving the performance of the model. Despite the moderate overall FID driven by deliberate coverage of harder conditions, the virtual-ratio ablation in Table 6 shows consistent performance gains as synthetic coverage increases, confirming task utility. Combined with the condition-matched Normal FID of 14.21, this indicates that the synthetic images are well aligned under common conditions while intentionally diversifying toward harder cases that improve robustness.

17. The effect of different text labels on the model

Before the training of RoadCLIP, a pre-trained CLIP model was used to test the effect of different text labels on model performance. Fig. 9 shows how the similarity values change when different text labels are used with input images. The labels on the left image are from our Label 1, and the labels on the right image are from Label 2. The text description of the road markings in Label 2 is more similar to the road marking image because it contains more detailed information about the road marking’s colour and pattern. This detailed textual description helps the model better understand the relationship between the image

Table 7

Performance Comparison with Different State-of-the-arts VLMs.

	F1-Score	Parameters	Inference Time
Fine-tuned CLIP	85.6 %	~ 151 M	214 FPS
Fine-tuned OWL-ViT v2	88.3 %	~ 250 – 300 M	89 FPS
Fine-tuned GLIP	78.4 %	~ 500 – 550 M	44 FPS

Table 8

Results Comparison with different portions of virtual images in training.

		Precision	Recall	F1-Score
Fine-tuned CLIP with 0 % virtual (real-only training)	Lane Markings	78.4 %	74.1 %	76.2 %
	Non-Lane Markings	82.7 %	81.4 %	82.0 %
Fine-tuned CLIP with 25 % virtual dataset	Lane Markings	82.9 %	78.1 %	80.4 %
	Non-Lane Markings	85.3 %	83.4 %	84.3 %
Fine-tuned CLIP with 50 % virtual dataset	Lane Markings	85.6 %	82.1 %	83.8 %
	Non-Lane Markings	90.4 %	91.7 %	91.0 %
Fine-tuned CLIP with 75 % virtual dataset	Lane Markings	87.8 %	86.7 %	87.3 %
	Non-Lane Markings	97.2 %	98.1 %	97.7 %
Fine-tuned CLIP with 100 % virtual dataset	Lane Markings	89.3 %	90.3 %	89.8 %
	Non-Lane Markings	98.5 %	97.9 %	98.2 %

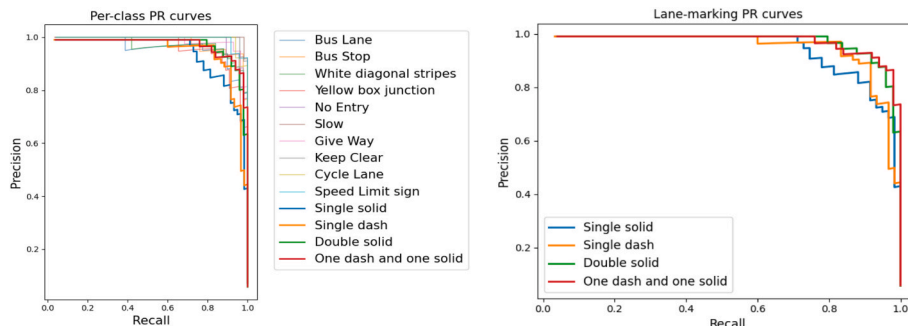


Fig. 8. 14-class and 4 lane-marking PR curves.

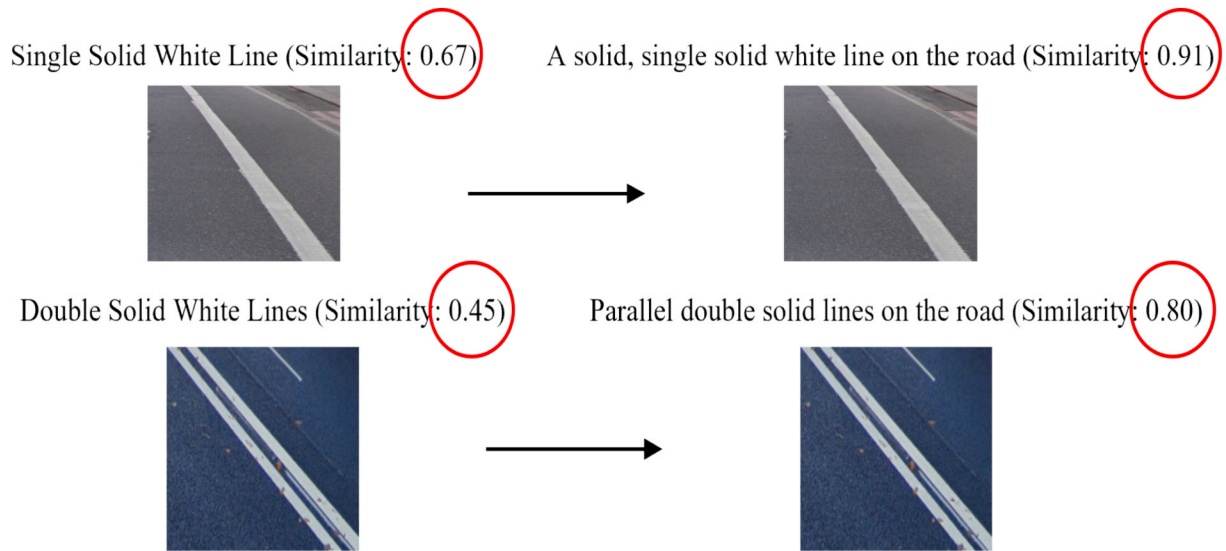


Figure 9 Impact of different text labels on the model

Fig. 9. Impact of different text labels on the model.

and the text, thereby improving its performance. This demonstrates that providing more detailed textual descriptions can improve the model's ability to learn and accurately identify road markings. Beyond higher similarity values, Label 2 introduces geometry-, colour-, and composition-level tokens that place the text embedding along attribute-sensitive directions in CLIP's joint space. In our text-conditioned cross-attention, these tokens act as query directions that pull attention mass onto image tokens expressing matching local cues (for example, hollow inverted triangles, grid-like meshes, stroke width and spacing), reducing attention entropy over background. This increases positive image-text alignment and widens the margin to confusable classes. The effect is strongest for composition-sensitive, elongated / glyph-like markings such as KEEP CLEAR, Yellow box and BUS LANE, and smaller for simple, uniform shapes. Because Label 2 is only used to pre-train the lightweight head (cosine objective) while inference uses concise Label 1, the improvement persists without relying on verbose descriptors at test time. To avoid length-induced bias, we standardise Label 1 as concise

canonical tokens and restrict Label 2 to a geometry-focused template with a capped length. The qualitative gain observed with Label 2 therefore reflects added geometric specificity for elongated / glyph-like markings rather than mere verbosity.

18. RoadGPT results

The integration of RoadCLIP with GPT-4 makes the system generate concise but accurate driving prompts in real time. Fig. 10 gives the results of the final pipeline, including all 14 road marking types. For each road marking, the pipeline output comprises detailed road marking information and corresponding driving prompts, seamlessly integrating perception and decision-making processes.

A reference text was created for each road marking based on its official description and the associated driving operation to evaluate the accuracy and effectiveness of the generated driving prompts. This reference text serves as the benchmark for assessing the generated

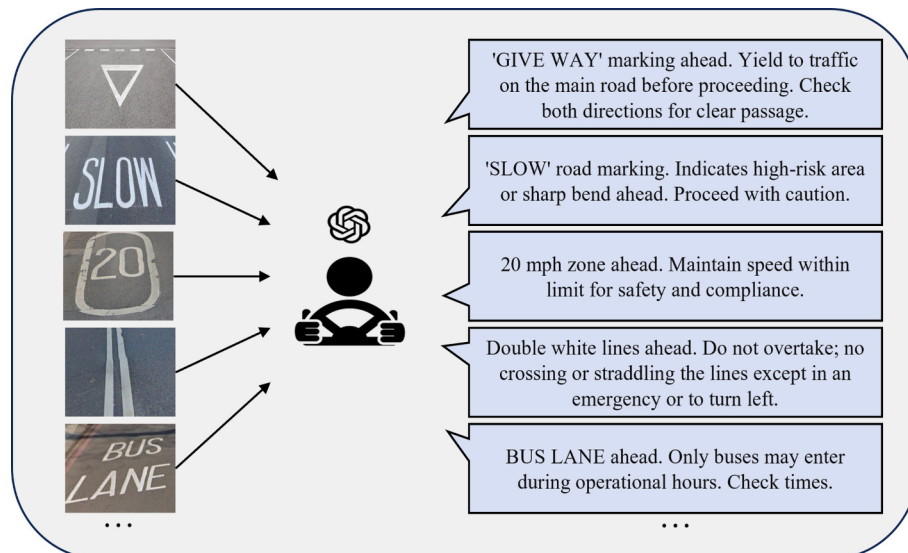


Fig. 10. Results of the final pipeline.

prompts. First the semantic similarity between the reference text and the generated driving prompts was calculated as a baseline measure. Subsequently, we manually defined key information slots (marking type, action, lane relation) based on the Highway Code. A reward mechanism is employed to increase the similarity value if the generated prompts contain these key information points. For reproducibility, semantic similarity is computed with Sentence-BERT (all-mpnet-base-v2) cosine similarity (Reimers and Gurevych, 2019) under a fixed normalisation (lowercasing, punctuation stripping, max length = 128). In addition, we introduce a secondary LLM-J (LLM-as-judge) score to target actionable prompt quality: an external instruction-tuned LLM: Llama-3-70B-Instruct (Grattafiori et al., 2024) rates each prompt on three 0–5 dimensions with anchored rubrics—Accuracy (whether the advised actions are correct), Completeness (whether required actions are covered), and Concise effectiveness (short, unambiguous, easy to parse). We choose Llama-3-70B-Instruct because it combines strong instruction-following with open-weight, offline reproducibility and, being independent from the GPT-4 generator, avoids self-evaluation bias. The judge returns a JSON record {accuracy, completeness, conciseness, unsafe, comment} with temperature = 0 and a fixed prompt template; we aggregate LLM-J = (Accuracy + Completeness + Conciseness) / 3 over the test set.

Table 9 presents the comprehensive semantic similarity and LLM-J evaluation results between the generated driving prompts and the reference text for each road marking. The overall average similarity of the generated driving prompts to their target road-marking descriptions reaches 89.31 %. Here similarity is computed as cosine similarity from Sentence-BERT (model: all-mpnet-base-v2) between the generated text and the Highway-Code-style reference description, after lower-casing, punctuation stripping and truncating to 128 tokens. Measured per class over the 14 markings, the similarities range from 72.40 % (“Single dash”) to 93.28 % (“White diagonal stripes”), with a standard deviation of about 4.9 percentage points. Beyond the mean semantic similarity of 89.31 %, Table 7 shows that the prompts also score highly on the LLM-J judge (0–5): the overall average is $\approx 4.68/5$, with non-lane markings averaging 4.80 and lane markings 4.38. Notably, clearly regulated markings such as Yellow box junction, No Entry, Slow, Give Way, Keep Clear all reach 5.0, indicating that the generated prompts are both accurate and complete from an actionability standpoint. In total, 12/14 classes achieve ≥ 4.5 on LLM-J; the only classes below 4.5 are the more context-dependent lane-boundary variants (Single dash: 4.25; One dash & one solid: 4.0), where conciseness vs. completeness can trade off in the judge rubric. Importantly, LLM-J also captures actionable quality even when wording diverges from the reference—for example, Keep Clear attains LLM-J = 5.0 despite a lower similarity (85.77 %), showing that the prompt still conveys the correct “no-stop polygon” advisory.

Table 9
Semantic similarity results for driving prompts.

Type of road marking		Similarity	LLM-J evaluation(0–5)
Non-Lane Markings	Bus Lane	88.53%	4.5
	Bus Stop	91.10%	4.75
	White diagonal stripes	93.28%	4.75
	Yellow box junction	85.99%	5.0
	No Entry	90.62%	5.0
	Slow	88.77%	5.0
	Give way	90.25%	5.0
	Keep Clear	85.77%	5.0
	Cycle Lane	92.19%	4.5
	Speed Limit sign	88.10%	4.5
	Single solid	91.30%	4.5
	Single dash	72.40%	4.25
	Double solid	90.04%	4.75
	One dash and one solid	86.80%	4.0
Lane Markings			

Together, the high similarity and strong LLM-J scores support that RoadGPT’s prompts are both semantically aligned and readily consumable as planner advisories, consistent with our advisory-only integration in the AV stack. To ensure actionability, all outputs pass the advisory JSON schema and are mapped to planner flags/costs as specified above; prompts failing schema/threshold checks are dropped. Thus, semantic similarity and LLM-J quantify prompt quality, while the schema and gates guarantee machine interpretability and safe consumption by the planner.

The results indicate that RoadGPT can effectively generate concise but accurate driving prompts. This high level of accuracy is important for real-time decision-making for autonomous vehicles, ensuring that vehicles can respond appropriately to various road scenarios.

A brief error analysis indicates that extreme low light is a major source of misrecognition; see the dedicated subsection “Limitations and Representative Failure Cases” for quantified trends and representative failure patterns.

19. Exploration of the transferability of models

In the road marking detection task, model transferability is crucial as it allows pre-trained models to adapt to different environments and changing road conditions. This capability is especially important for rapidly changing road environments, such as road marking differences in different countries, to make sure that the model can still accurately recognise and respond when faced with unseen markings. In this section, the transferability of the RoadCLIP is examined.

A comprehensive transferability test was conducted using 334 road marking images from different countries and regions such as China mainland, Hongkong, Japan and Australia. The types of road markings selected were those that were already included in the original dataset. In the cross-country transfer experiments, the model parameters were not re-trained or adapted. The CLIP encoder and the text-conditioned attention layer remained the same as in the UK experiments. The only adaptation was applied on the text side of the prompts. This is because several of the non-UK road networks in our test set (notably mainland China, Hong Kong and Japan) use yellow longitudinal markings in situations where the UK Highway Code uses white ones. If the original UK-style label (for example, “single dashed white lane marking on asphalt road”) is used unchanged, the similarity score for a yellow marking decreases and the prediction can drift to another lane-marking class. After inserting the correct colour token, the match recovers. Two examples of the colour-aware prompt variants we used are: UK: “single dashed white lane marking on asphalt road” to China / Hong Kong: “single dashed yellow lane marking on asphalt road”. UK: “double solid white centre line” to Japan / Australia: “double solid yellow centre line on asphalt”. No re-training, no re-ranking, and no country-specific fine-tuning were performed; only the textual labels were colour-specialised to reflect the local road-marking standard. Tables 10 and 11 already list, for each country or region, the number of test images and the corresponding F1-scores under different conditions. Together, these tables provide the quantitative, per-country view of transferability, while the paragraph above clarifies that the adaptation was achieved only through

Table 10
Number of images for each condition in dataset for testing transferability.

Region	Normal	Adverse weather	Nighttime	Faded condition
China mainland	41 (47.1 %)	18 (20.7 %)	17 (19.5 %)	11 (12.7 %)
Hongkong	40 (44.4 %)	16 (17.8 %)	20 (22.2 %)	14 (15.6 %)
Japan	36 (45.0 %)	18 (22.4 %)	11 (13.8 %)	15 (18.8 %)
Australia	34 (44.2 %)	17 (22.1 %)	14 (18.2 %)	12 (15.5 %)

Table 11
Accuracy and F1-score in different regions under different conditions.

	Normal (Acc/F1)	Adverse weather (Acc/ F1)	Nighttime (Acc/F1)	Faded condition (Acc/F1)
UK (in-domain)	97.61 %/95.38 %	91.46 %/90.83 %	92.07 %/90.14 %	95.84 %/91.11 %
China	93.22 %/90.77 %	87.16 %/85.41 %	86.79 %/82.93 %	90.31 %/86.69 %
mainland Hongkong	94.46 %/92.63 %	91.05 %/90.61 %	88.32 %/85.72 %	92.61 %/89.58 %
Japan	93.89 %/91.84 %	88.67 %/87.44 %	88.16 %/85.37 %	91.57 %/89.06 %
Australia	97.38 %/95.08 %	93.02 %/92.57 %	90.17 %/88.37 %	91.96 %/90.06 %

colour-aware prompting.

Cross-region tests are conducted in a zero-shot setting: model weights are unchanged and only text-side label edits (e.g., regional color terms) are applied. Thresholds and post-processing are fixed from the UK validation. We report macro-F1 per region and per scenario (Normal, Adverse, Nighttime, Faded); distribution details (region by scenario counts and shares) are provided in Table 10, and scenario-wise results are summarized in Table 11. Nighttime shows the largest relative drop versus the UK baseline (for example, 90.14 in the UK vs. 82.93/85.72/85.37/88.37 in CN/HK/JA/AU), while Adverse weather in AU (92.57) approaches or exceeds the UK in-domain value (90.83), indicating strong zero-shot robustness in some challenging conditions. These results provide initial cross-region evidence of transferability under zero-shot label-edit, with clear scenario-wise differences documented in Tables 10–11. While performance remains strong without retraining—particularly under Adverse conditions in AU—Nighttime is the main challenge across regions. We therefore treat cross-region deployment as feasible with text-only adaptation and leave further per-class/per-scenario expansion for future work. In deployment, these scenario-wise differences inform the confidence gates (e.g., stricter at nighttime), ensuring advisories are issued only under reliable conditions while the planner retains conservative defaults otherwise.

Fig. 11 shows the test results of road markings in different countries. For lane markings with different colours, the model showed flexibility by requiring only the addition of specific colour descriptions in the text labels without the need for re-training. This showcases the model's efficiency in adapting to minor variations in road marking appearances.

For other road markings, the model successfully identified the variations in patterns and colours, accurately recognising road markings from different countries. This success is attributed to the model's ability to generalise from the learned feature information of the road marking images and their associated text descriptors. The ability to accurately detect and classify road markings across different regions highlights the robustness and adaptability of the RoadCLIP model.

The transferability of the RoadGPT model was tested using data from different regions and conditions. The model demonstrated a high degree of generalisation, accurately detecting and interpreting road markings from various countries without additional retraining. This capability underscores the robustness of the model in diverse real-world applications. By maintaining high accuracy and reliability across different environments, RoadCLIP and RoadGPT prove to be highly effective solutions for global deployment in autonomous driving systems.

20. Limitations and representative failure cases

We first summarise the quantitative trends already evident in our results. On the in-domain UK set, lane-marking F1 drops from 91.44 % (Normal) to 85.81 % (Nighttime) and 82.92 % (Faded), while non-lane F1 decreases from 99.32 % (Normal) to 91.76 % (Adverse) (Table 4). Cross-region, the largest relative drop consistently appears at Nighttime (e.g., UK 90.14 % vs. 82.93 %/85.72 %/85.37 %/88.37 % in CN/HK/JA/AU), whereas some Adverse-weather cases remain strong (e.g., AU 92.57 % vs. UK 90.83 %) (Table 9). These trends indicate that low illumination and heavy wear are the primary challenges, followed by weather-induced specular reflections.

Representative failure patterns observed in qualitative inspection include:

- Nighttime/low light: missed detections or class confusions for thin, glyph-like markings when contrast is low.
- Heavy wear (e.g., 75 %faded): reduced luminance ratio between marking and background leading to under-segmentation and ambiguity between lane types.
- Adverse weather with wet asphalt: specular highlights and water streaks form lane-like structures, inducing false positives or boundary leaks.
- Cross-region colour/style shift: yellow vs. white lanes or typography variants in text markings can trigger misclassification unless colour terms are edited on the text side.

These failure types align with the scenario-wise performance gaps reported in Tables 4 and 9.

Limitations therefore centre on low-illumination robustness, high-wear degradations, and separability among visually similar lane types, compounded by data sparsity for rare/extreme conditions. Practical mitigations include targeted augmentation for low-light/high-wear scenes, a light shallow-feature fusion to emphasise thin structures, and simple temporal smoothing to stabilise predictions. We leave a per-class, per-scenario expansion of the cross-region set for future updates, keeping the core zero-shot, label-edit deployment flow unchanged.



Fig. 11. Test results of road markings in different regions.

21. Conclusion

This paper developed a road marking detection and prompting framework, RoadGPT, which couples a CLIP-based visual-textual module (RoadCLIP) with GPT-4-driven driving prompts. By exploiting CLIP's ability to align images with natural-language descriptions, RoadCLIP alleviates the usual problems of hand-crafted feature design, limited annotated road-marking data, and the difficulty of learning fine-grained correspondence between markings and textual labels. The inclusion of synthetic road-marking images further enlarges the training distribution and improves robustness, so that the overall pipeline can generate real-time, scene-aware prompts to support autonomous driving decisions.

At the same time, several limitations remain and were observed consistently in our experiments. Lane-type markings are still less reliable than symbol-like non-lane markings, because the four lane subclasses are very close to one another in shape and colour, which makes CLIP scores easier to confuse. Performance can also drop in rare or extreme conditions — heavy rain, dense fog, nighttime capture, or severely faded paint — because such cases are still sparsely represented in the current collection. In addition, although virtual images clearly help the model learn more diverse layouts, they cannot fully reproduce the clutter, surface wear, shadows and lighting variation seen in large-scale real deployments, so a synthetic-to-real gap may persist for the rarest markings.

Going forward, we will make the evaluation and deployment path more concrete. On the reasoning side, we will place fixed, rule-based driving instructions derived from UK Highway Code templates next to LLM-generated prompts that are conditioned on exactly the same SAM2/CLIP detections; running this comparison on both the 14-class UK set and the 334 non-UK images collected for transferability will make it possible to isolate how much benefit actually comes from the language model rather than from perception. On the control side, we will embed the current perception-prompting stack in a closed-loop environment (for example, CARLA scenarios or nuScenes/DADA-2000 replays) to test whether different textual prompts truly lead to different manoeuvre or warning decisions (lane-keeping, give-way, temporary roadworks handling), instead of only improving semantic-similarity scores. On the deployment side, we will add a lightweight safety layer that rejects low-confidence or colour-inconsistent CLIP matches, logs LLM outputs that conflict with traffic rules, and allows a simple human-in-the-loop review whenever a novel or unseen marking is detected.

These steps will be carried out together with the improvements already planned in this work: a hybrid real-synthetic curriculum that first pre-trains the attention head on condition-balanced synthetic images and then fine-tunes on real data with class/condition-weighted sampling, following the strategy of Song et al. (2023); lightweight temporal modelling over short frame windows — either ConvLSTM (Shi et al., 2015) or temporal attention on CLIP features in the spirit of TimeSformer (Bertasius et al., 2021) — to stabilise lane-marking continuity under low light and high wear within a strict batch-size-1 latency budget; and broader cross-region evaluation with increased Night/Faded coverage and an explicit protocol for zero-shot text-side label edits. Taking together, these directions keep the conclusion honest — lane markings and extreme conditions are still the hardest — but provide a clear, forward-looking route to closed-loop validation and safer deployment.

Author contributions

The authors confirm contribution to the paper as follows: study conception and design: S. Sheng, N. Formosa, M. Quddus; data collection: N. Formosa, M. Quddus; analysis and interpretation of results: S. Sheng, N. Formosa, Y. Feng, M. Quddus; draft manuscript preparation: S. Sheng, N. Formosa, Y. Feng, M. Quddus. All authors reviewed the results and approved the final version of the manuscript.

CRedit authorship contribution statement

Shaofan Sheng: Writing – original draft, Visualization, Methodology, Investigation, Data curation, Conceptualization. **Nicolette Formosa:** Writing – review & editing, Supervision, Conceptualization. **Yuxiang Feng:** Writing – review & editing, Supervision, Formal analysis. **Mohammed Quddus:** Writing – review & editing, Writing – original draft, Supervision, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

- Bailo, O., Lee, S., Rameau, F., Yoon, J.S., Kweon, I.S. Robust road marking detection and recognition using density-based grouping and machine learning techniques. In: Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), pp. 760–768. IEEE (2017).
- Bakirci, M., 2025. Advanced aerial monitoring and vehicle classification for intelligent transportation systems with YOLOv8 variants. *J. Netw. Comput. Appl.* 237, 104134.
- Bertasius, G., Wang, H. and Torresani, L., 2021, July. Is space-time attention all you need for video understanding?. In *ICML* (Vol. 2, No. 3, p. 4).
- Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., Manassra, W., 2023. Improving image generation with better captions. *Comput. Sci.* 2 (3), 8. <https://cdn.openai.com/papers/dall-e-3.pdf>.
- Bimbraw, K. Autonomous cars: Past, present and future—a review of the developments in the last century, the present scenario and the expected future of autonomous vehicle technology. In: Proc. Int. Conf. Informat. Control Autom. Robot. (ICINCO), vol. 1, pp. 191–198. IEEE (2015).
- Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S. End-to-end object detection with transformers. In: Proc. Eur. Conf. Comput. Vis. (ECCV), pp. 213–229. Springer (2020).
- Chan, T.H., Jia, K., Gao, S., Lu, J., Zeng, Z., Ma, Y., 2015. PCANet: a simple deep learning baseline for image classification. *IEEE Trans. Image Process.* 24 (12), 5017–5032.
- Chen, T., Chen, Z., Shi, Q., Huang, X. Road marking detection and classification using machine learning algorithms. In: Proc. IEEE Intell. Vehicles Symp. (IV), pp. 617–621. IEEE (2015).
- Cheng, M.M., Zhang, Z., Lin, W.Y., Torr, P. BING: Binarized normed gradients for objectness estimation at 300 fps. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 3286–3293 (2014).
- Cui, Y., Huang, S., Zhong, J., Liu, Z., Wang, Y., Sun, C., Li, B., Wang, X., Khajepour, A., 2023. Drivellm: charting the path toward full autonomous driving with large language models. *IEEE Trans. Intell. Vehicles*.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16×16 words: Transformers for image recognition at scale. *arXiv Prepr. arXiv:2010.11929* (2020).
- Gong, Y., Zhang, X., Feng, J., He, X., Zhang, D. LiDAR-based HD Map Localization using Semantic Generalized ICP with Road Marking Detection. In: Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), pp. 3379–3386. IEEE (2024).
- Google. Discover Street View and contribute your own imagery to Google Maps. <https://www.google.com/streetview/>.
- Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A. and Yang, A., 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Greenhalgh, J., Mirmehdi, M. Automatic detection and recognition of symbols and text on the road surface. In: Pattern Recognit. Appl. Methods, ICPRAM 2015, Lisbon, Portugal, Jan. 10–12, 2015. Revised Sel. Pap., vol. 4, pp. 124–140. Springer (2015).
- Gupta, A., Choudhary, A., 2018. A framework for camera-based real-time lane and road surface marking detection and recognition. *IEEE Trans. Intell. Vehicles* 3 (4), 476–485.
- He, K., Zhang, X., Ren, S., Sun, J. Deep residual learning for image recognition. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 770–778 (2016).
- He, K., Gkioxari, G., Dollár, P., Girshick, R. Mask R-CNN. In: Proc. IEEE Int. Conf. Comput. Vis. (ICCV), pp. 2961–2969 (2017).
- Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S., 2017. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Adv. Neural Inf. Process. Syst.* 30.
- The Highway Code – Road markings – Guidance – GOV.UK. <https://www.gov.uk/guidance/the-highway-code/road-markings>.
- Jayasinghe, O., Hemachandra, S., Annettigama, D., Kariyawasam, S., Rodrigo, R., Jayasekara, P. Ceymo: See more on roads—a novel benchmark dataset for road marking detection. In: Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV), pp. 3104–3113 (2022).

- Lee, S., Kim, J., Shin Yoon, J., Shin, S., Bailo, O., Kim, N., Lee, T.H., Seok Hong, H., Han, S.H., So Kweon, I. VPGNet: Vanishing point guided network for lane and road marking detection and recognition. In: Proc. IEEE Int. Conf. Comput. Vis. (ICCV), pp. 1947–1955 (2017).
- Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., Chang, K.W., 2022. Grounded language-image pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10965–10975.
- Li, X., Wang, K., Tian, Y., Yan, L., Deng, F., Wang, F.Y., 2019. The paralleleye dataset: a large collection of virtual images for traffic vision research. IEEE Trans. Intell. Transp. Syst. 20 (6), 2072–2084.
- Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C. SSD: Single shot multibox detector. In: Comput. Vis.—ECCV 2016, Amsterdam, Netherlands, Oct. 11–14, 2016, Proc., Pt. I, vol. 14, pp. 21–37. Springer (2016).
- Liu, X., Deng, Z., Lu, H., Cao, L. Benchmark for road marking detection: Dataset specification and performance baseline. In: Proc. IEEE Int. Conf. Intell. Transp. Syst. (ITSC), pp. 1–6. IEEE (2017).
- Lu, J., Batra, D., Parikh, D., Lee, S., 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Adv. Neural Inf. Process. Syst. 32.
- Lüddecke, T., Ecker, A. Image segmentation using text and image prompts. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 7086–7096 (2022).
- Manongga, W.E., Chen, R.C., 2025. Road marking sign damage detection and classification using deep learning and fuzzy method. IEEE Access.
- Matas, J., Chum, O., Urban, M., Pajdla, T., 2004. Robust wide-baseline stereo from maximally stable extremal regions. Image Vis. Comput. 22 (10), 761–767.
- Memon, J., Sami, M., Khan, R.A., Uddin, M., 2020. Handwritten optical character recognition (OCR): a comprehensive systematic literature review (SLR). IEEE Access 8, 142642–142668.
- Minderer, M., Gritsenko, A., Hounsby, N., 2023. Scaling open-vocabulary object detection. Adv. Neural Inf. Process. Syst. 36, 72983–73007.
- Morrisett, A., Abdelwahed, S. A review of non-lane road marking detection and recognition. In: Proc. IEEE Int. Conf. Intell. Transp. Syst. (ITSC), pp. 1–7. IEEE (2020).
- Oord, A. v. d., Li, Y., Vinyals, O. Representation learning with contrastive predictive coding. arXiv Prepr. arXiv:1807.03748 (2018).
- Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In: Proc. Int. Conf. Mach. Learn. (ICML), pp. 8748–8763. PMLR (2021).
- Rahutomo, F., Kitasuka, T., Aritsugi, M. Semantic cosine similarity. In: Proc. Int. Student Conf. Adv. Sci. Technol. (ICAST), vol. 4, p. 1 (2012).
- Reimers, N. and Gurevych, I., 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.
- Sha, H., Mu, Y., Jiang, Y., Chen, L., Xu, C., Luo, P., Li, S.E., Tomizuka, M., Zhan, W., Ding, M. LanguageMPC: Large language models as decision makers for autonomous driving. arXiv Prepr. arXiv:2310.03026 (2023).
- Sheng, S., Formosa, N., Hossain, M., Quddus, M., 2024. Advancements in lane marking detection: an extensive evaluation of current methods and future research direction. IEEE Trans. Intell. Vehic.
- Shi, X., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K. and Woo, W.C., 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. Advances in neural information processing systems, 28.
- Song, Z., He, Z., Li, X., Ma, Q., Ming, R., Mao, Z., Pei, H., Peng, L., Hu, J., Yao, D., Zhang, Y., 2023. Synthetic datasets for autonomous driving: a survey. IEEE Trans. Intell. Veh. 9 (1), 1847–1864.
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z. Rethinking the inception architecture for computer vision. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 2818–2826 (2016).
- Tabelini, L., Berriel, R., Paixao, T.M., Badue, C., De Souza, A.F., Oliveira-Santos, T. Keep your eyes on the lane: Real-time attention-guided lane detection. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 294–302 (2021).
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.
- Wu, J., Liu, W., Maruyama, Y., 2022. Automated road-marking segmentation via a multiscale attention-based dilated convolutional neural network using the road marking dataset. Remote Sens. 14 (18), 4508.
- Wu, J., Liu, W., Maruyama, Y., 2024. Street view image-based road marking inspection system using computer vision and deep learning techniques. Sensors 24 (23), 7724.
- Wu, P.C., Chang, C.Y., Lin, C.H., 2014. Lane-mark extraction for automobiles under complex conditions. Pattern Recogn. 47 (8), 2756–2767.
- Wu, T., Ranganathan, A. A practical system for road marking detection and recognition. In: Proc. IEEE Intell. Vehicles Symp. (IV), pp. 25–30. IEEE (2012).
- Yu, Y., Li, Y., Liu, C., Wang, J., Yu, C., Jiang, X., Wang, L., Liu, Z., Zhang, Y., 2021. Markcap-snet: road marking extraction from aerial images using self-attention guided capsule network. IEEE Geosci. Remote Sens. Lett. 19, 1–5.
- Yang, Y., Peng, H., Li, C., Zhang, W., Yang, K., 2022. LaneFormer: real-time lane extraction and detection via transformer. Appl. Sci. 12 (19), 9722.
- Ye, X.Y., Hong, D.S., Chen, H.H., Hsiao, P.Y., Fu, L.C., 2020. A two-stage real time YOLOv2-based road marking detector with lightweight spatial transformation invariant classification. Image Vis. Comput. 102, 103978.
- Zhang, H., He, A., Dong, Z., Zhang, A.A., Liu, Y., Zhan, Y., Wang, K.C., Lin, Z., 2024. Automated pixel-level pavement marking detection based on a convolutional transformer. Eng. Appl. Artif. Intel. 133, 108416.
- Zhang, X., Yuan, Y., Wang, Q. Roi-wise reverse reweighting network for road marking detection. In: Proc. Brit. Mach. Vis. Conf. (BMVC), p. 219 (2018).
- Zhang, Z., Sabuncu, M., 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. Adv. Neural Inf. Process. Syst. 31.
- Zhao, Z.Q., Zheng, P., Xu, S.T., Wu, X., 2019. Object detection with deep learning: a review. IEEE Trans. Neural Netw. Learn. Syst. 30 (11), 3212–3232.
- Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., et al. RegionCLIP: Region-based language-image pretraining. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 16793–16803 (2022).