Commun Transp Res

Research Article

Distilling Vision-Language Models for Explainable Vehicle Collision Prediction

Ruici Zhang✉, Yuxiang Feng, Jose Escribano, Mohammed Quddus

Cite this article: https://doi.org/10.26599/COMMTR.2026.9640047

ABSTRACT: Vision-based collision warning systems are increasingly recognised as a promising countermeasure against traffic collisions. However, their development is constrained by the limited explainability of deep learning-based collision prediction models. Vision-language models (VLMs), with inherent self-explainability, offer a promising solution, yet two critical challenges persist: (1) how to adapt general-purpose VLMs to the task-specific domain of collision prediction, and (2) how to reduce their high inference latency. To tackle these challenges, this paper develops a novel approach that distils VLMs for explainable vehicle collision prediction. Specifically, a two-step chain-of-thought (CoT) prompting strategy was designed to guide VLMs to first analyse the driving scenario and then predict potential collisions, providing explicit rationales. The VLMs were fine-tuned on both pre-collision and normal driving video clips with descriptive annotations. To reduce latency, a feature-based knowledge distillation approach was introduced to transfer knowledge from the fine-tuned VLM to a smaller one through hidden-state supervision. Experimental results demonstrate that fine-tuning improves prediction accuracy by over 20%, while CoT prompting further enhances both performance and explainability. The distilled VLM reduces inference latency by more than 50% with less than a 5% performance degradation, highlighting its potential to advance vision-based collision warning systems and proactive traffic safety.

KEYWORDS: explainable vehicle collision prediction; vision-language model; knowledge distillation; chain-of-thought prompting; inference latency; vision-based collision warning system

1 Introduction

Road traffic collisions are a major global threat to traffic safety. As reported by the World Health Organisation (WHO), they account for approximately 1.19 million deaths annually and are the leading cause of death among people aged 5 to 29 (WHO, 2023). Advanced driver assistance systems (ADAS) are widely recognised by governments and industry as an important approach to reducing traffic collisions and alleviating their impacts (Yu et al., 2022). The core idea of ADAS lies in using collision warning systems to assess the likelihood of a vehicle being involved in a collision imminently (Formosa et al., 2022). When the vehicle is anticipated to be in a collision-prone state, proactive warning or interventions can then be implemented to help prevent the collision.

Traditional vehicle collision prediction methods can be mainly divided into two approaches: (1) safety indicator-based and (2) trajectory prediction-based methods. The first approach calculates safety indicators such as time headway (THW) and time-to-collision (TTC) based on vehicle kinematic data (McLaughlin et al., 2008; Vogel, 2003). When the values of these safety indicators fall below pre-defined thresholds, the vehicle is considered to be in a collision-prone state, accompanied by the triggering of a collision warning (Zhu et al., 2020). Such an approach struggles to account for road user behaviours and adapt to complex driving scenarios (Yu et al., 2022).

To overcome this issue, the trajectory prediction-based approach estimates the future trajectories of vehicles and other traffic participants based on historical motion patterns and then infers the collision occurrence probability (Lyu et al., 2020; Wu et al., 2024). However, trajectory data fails to fully convey the rich contextual and visual details in driving environments (e.g., occlusions and visual blind spots). Moreover, these methods often rely on costly sensing equipment (e.g., radar), which further adds to implementation costs and limits scalability (Yu and Ai, 2022).

Given the significant advancements in computer vision and deep learning technologies, vision-based collision warning algorithms have attracted considerable interest within the community, including industry leaders such as Tesla (Ingle and Phute, 2016). Specifically, this approach leverages deep learning techniques to map the relationship between sequential driving frames and potential collision risks (Shi and Guo, 2024; Yu et al., 2020). These deep learning models can produce relatively accurate collision prediction outcomes, but their black-box nature makes it challenging to explain the predictions (Shwartz-Ziv and Tishby, 2017). As a result, it is difficult to determine whether the models rely on incorrect or unreasonable features, which raises concerns about their reliability in real-world applications (Guan et al., 2025).

Recent breakthroughs in vision-language models (VLMs) offer significant potential to enhance explainability in vehicle collision prediction. As a multimodal extension of large language models (LLMs), VLMs combine natural language processing with visual understanding (Chen et al., 2024), and have demonstrated excellent

Centre for Transport Engineering and Modelling, Department of Civil and Environmental Engineering, Imperial College London, London SW7 2AZ, UK.

✉ Corresponding author. E-mail: **ruici.zhang24@imperial.ac.uk**

Received: February 4, 2026; Revised: June 21, 2026; Accepted: August 13, 2026

capabilities in knowledge reasoning and video analysis (Shi et al., 2025; Zhang et al., 2025). A key advantage of VLMs is their capability to generate explicit explanations of their reasoning process, which is essential for prediction and decision-making tasks. This self-explainability has been validated in various transport-related applications, such as autonomous driving decision-making (Liu et al., 2025b) and pedestrian behaviour prediction (Munir et al., 2025). However, the utilisation of VLMs for explainable vehicle collision prediction remains largely unexplored.

Furthermore, applying VLMs to predict traffic collisions faces several critical challenges. First, most existing open-source VLMs in the literature are designed for general-purpose applications and lack collision-related domain knowledge (Zhang et al., 2025). Previous studies have shown that directly using general-purpose VLMs for traffic collision analysis does not yield satisfactory performance (Shi et al., 2025), making it essential to customise VLMs for the specific task of collision prediction. Moreover, given that VLM performance is often positively correlated with their model size, achieving excellent results typically requires large-scale models, commonly exceeding 7 billion parameters (Tian et al., 2026). Such massive models incur substantial computational overhead and high inference latency, requiring prohibitively expensive computational resources for deployment on vehicle servers.

This study therefore develops, to the best of our knowledge, the first unified framework integrating VLMs and knowledge distillation for explainable and computationally efficient vehicle collision prediction. Ego-view driving videos, which serve as the primary data source for vehicle collision analysis (Fang et al., 2024), are employed for empirical analysis. The main contributions of this study include:

(1) Customised VLMs for explainable vehicle collision prediction are developed through fine-tuning general-purpose VLMs on real-world driving video clips and corresponding descriptive annotations. A two-step chain-of-thought (CoT) prompting method is further designed to guide the VLMs to analyse the driving scenario before predicting potential collisions, with the analysis serving as justification to improve explainability.

(2) A feature-based knowledge distillation approach is proposed to address the latency challenge of VLMs by transferring knowledge from the customised VLM (i.e., a teacher model) to a smaller, more efficient one (i.e., a student model). Unlike traditional fine-tuning that relies solely on ground truth supervision, this method additionally incorporates the hidden states of the teacher's transformer blocks, enabling the student model to better capture how the teacher represents and embeds the input.

(3) Extensive experiments under various prediction horizon settings demonstrate the superiority of the proposed approach. The customised VLM with CoT prompt achieves an average of 84.7% accuracy, 83.7% precision, and 86.8% recall, consistently outperforming all deep learning baselines. The distilled VLM achieves a latency reduction exceeding 50% while retaining over 95% of the original prediction performance.

The remainder of this paper is organised as follows: Section 2 reviews previous work related to this study. Section 3 presents the overall research framework and methodology. Section 4 describes the process of constructing the datasets used for empirical analysis. Experiments and results are given in Section 5. Section 6 discusses the practical implementation of the proposed approach, the limitations of the current study, and directions for future research. Finally, conclusions are provided in Section 7.

2 Related work

2.1. Explainable deep learning

Explainable deep learning techniques, which seek to reveal the internal decision-making process of deep learning models (LeCun et al., 2015), can generally be divided into two approaches: (1) feature attribution-based and (2) visualisation-based methods. The first approach aims to quantify the contribution of each input variable to the model's output (Heskes et al., 2020), with SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017) being a representative method. Such an approach is primarily designed for structured data (e.g., tabular inputs), and fails to scale to unstructured data (e.g., images and videos).

Visualisation-based methods explain model predictions by highlighting the input features most relevant to the decision (Zhou et al., 2016). A common form of such methods is saliency maps or heatmaps (e.g., attention maps) (An and Joe, 2022), where a coloured overlay is placed on the original input to indicate the high relevance areas. However, these visualisations still struggle to clearly illustrate the specific relationship between the highlighted areas and the model's output.

VLMs, a specialised form of multimodal LLMs, achieve self-explainability by combining visual feature extraction with natural language generation in a unified framework (Lin et al., 2024). Specifically, a visual encoder first extracts semantic information from visual inputs, which is then linked to a language model via vision-language alignment techniques (e.g., cross-attention) (Lin et al., 2024). This design enables VLMs to not only provide outcomes based on visual inputs but also generate natural-language explanations describing the reasoning behind their output (Guo et al., 2024).

2.2. Vision-language models

VLMs can be broadly categorised into two types: (1) Image-based LLMs (Image-LLMs) and (2) Video-based LLMs (Video-LLMs). Compared to Image-LLMs, which employ image encoders to convert image inputs into static feature representations (Huang et al., 2025), Video-LLMs leverage video encoders to capture temporal dynamics, enabling comprehensive video understanding and analysis (Lin et al., 2024). However, the majority of Video-LLMs still face challenges in modelling complex spatial-temporal interactions. To address this, VideoLLaMA2 (Cheng et al., 2024) introduces additional deep learning layers specifically designed for spatial-temporal interaction and aggregation, achieving superior performance across various video-language tasks. Accordingly, this study adopts VideoLLaMA2 for driving video analysis.

The development of VLMs typically involves two main stages: pre-training and fine-tuning. The pre-training stage aims to establish a general-purpose VLM equipped with broad linguistic and world knowledge based on large-scale unstructured data (Li et al., 2023a). Then in the fine-tuning stage, the pre-trained model is adapted to a specific task or domain to improve its task-specific performance (Ding et al., 2023). Due to the high computational cost of pre-training, most downstream tasks now adopt existing pre-trained models and fine-

Table 1. Comparison of this study with representative vision-based collision prediction studies.

Studies	Primary prediction backbone	Does the study consider prediction explainability?	Does the study account for inference latency?*
Yu and Ai (2022)	CNN + LSTM	✗	-
Wang et al. (2023)	Graph neural networks	✗	-
Shi and Guo (2024)	Attention mechanisms	✗	-
Liao et al. (2025)	Attention mechanisms	✓ (rationale generation by a general-purpose VLM)	✗
Guan et al. (2025)	Attention mechanisms	✓ (rationale generation by a general-purpose VLM)	✗
Wang et al. (2026a)	VLM-based input extraction + LSTM	✗	✗
This study	**Task-specific fine-tuned VLM**	✓ (**CoT-generated rationale**)	✓ (**feature-based knowledge distillation**)

*Note: In this column, "–" denotes non-VLM-based studies which do not face the same VLM-induced inference latency bottleneck.

tune them using smaller, customised datasets. Moreover, the CoT (Wei et al., 2022) prompting technique is commonly used to enhance the model's reasoning capabilities by guiding it through a step-by-step thought process.

The applications of VLMs in traffic collision analysis have grown significantly, including collision detection (Shao et al., 2024), classification (Zhang et al., 2025), and scenario description (Guan et al., 2025; Shi et al., 2025). However, few studies have directly employed VLMs for vehicle collision prediction with explicit rationales.

2.3. Knowledge distillation for VLMs

Latency is a critical factor in vehicle collision prediction models, as it determines whether collision warnings can be issued timely. However, VLMs often face latency issues due to their massive scale, typically involving billions of parameters and substantial computational overhead. Model compression techniques such as pruning (He et al., 2024) and quantisation (Liu et al., 2025c) can reduce model size but inevitably lead to performance degradation, especially when aggressive compression is applied. Knowledge distillation (Hinton et al., 2015), which transfers knowledge from a large, powerful teacher model to a smaller, more efficient student model, has emerged as a promising solution. The core idea is to guide the student to replicate the teacher's behaviour and internal representations, allowing it to achieve comparable performance with significantly reduced latency.

Existing studies on knowledge distillation for VLMs can be mainly divided into three categories: (1) instruction-based distillation, (2) logit-based distillation, and (3) feature-based distillation. In instruction-based distillation, the instruction–response pairs generated by the teacher are used to supervise the student (Cai et al., 2025; Zhang et al., 2024b). This approach enables the student to learn high-level instruction-following abilities, but the learning process is indirect as the student mimics the teacher's linguistic output rather than the underlying reasoning process.

Compared to instruction-based distillation, logit-based distillation provides a more fine-grained and direct learning signal through aligning the output logits (i.e., the token probability distributions over the whole vocabulary) of the student and teacher models (Li et al., 2025; Sun et al., 2024). Such alignment still occurs at the output level, without explicitly capturing the internal representation space, and its performance can be further constrained by differences in vocabularies between the teacher and student.

To overcome these issues, feature-based distillation focuses on aligning intermediate representations (e.g., hidden states of the intermediate layers) between the teacher and student (Bing et al., 2025; Li et al., 2023b). By incorporating feature-level supervision, this approach enables the student model to learn richer multimodal associations and reasoning patterns rather than merely imitating the teacher's outputs.

2.4. Research positioning

Following the above review, Table 1 further positions the proposed framework against six recent representative studies on vision-based vehicle collision prediction and highlights its distinct contributions. Unlike conventional deep learning methods with limited explainability, the proposed framework employs a task-specifically fine-tuned VLM with CoT prompting to generate explicit textual rationales for collision predictions. It also differs from recent VLM-based studies in two key respects: (i) the VLM serves directly as the primary prediction backbone, integrating collision prediction and explanation generation within a unified framework, rather than being used only to justify or interpret the outputs of another prediction model; and (ii) feature-based knowledge distillation is introduced to reduce computational demand and support practical latency requirements. Together, these designs enable the framework to jointly address prediction performance, explainability and deployment feasibility, which have rarely been considered together in existing studies.

3 Methodology

3.1. Overall framework

The overall framework of this study is illustrated in Fig. 1, consisting of four stages: (1) data extraction, (2) data annotation, (3) VLM fine-tuning, and (4) VLM knowledge distillation. In the data extraction stage, pre-collision and normal driving video clips are extracted to represent collision-prone and collision-free scenarios respectively. In the data annotation stage, the extracted clips are annotated with corresponding scenario descriptions using an LLM-based framework. In the VLM fine-tuning stage, a general-purpose VLM is fine-tuned using both pre-collision and normal driving clips with their annotated descriptions, resulting in a customised VLM tailored for vehicle collision prediction (referred to as CustVLM). Finally, in the VLM knowledge distillation stage, the knowledge from CustVLM is transferred to a smaller VLM to enhance computational efficiency while maintaining comparable

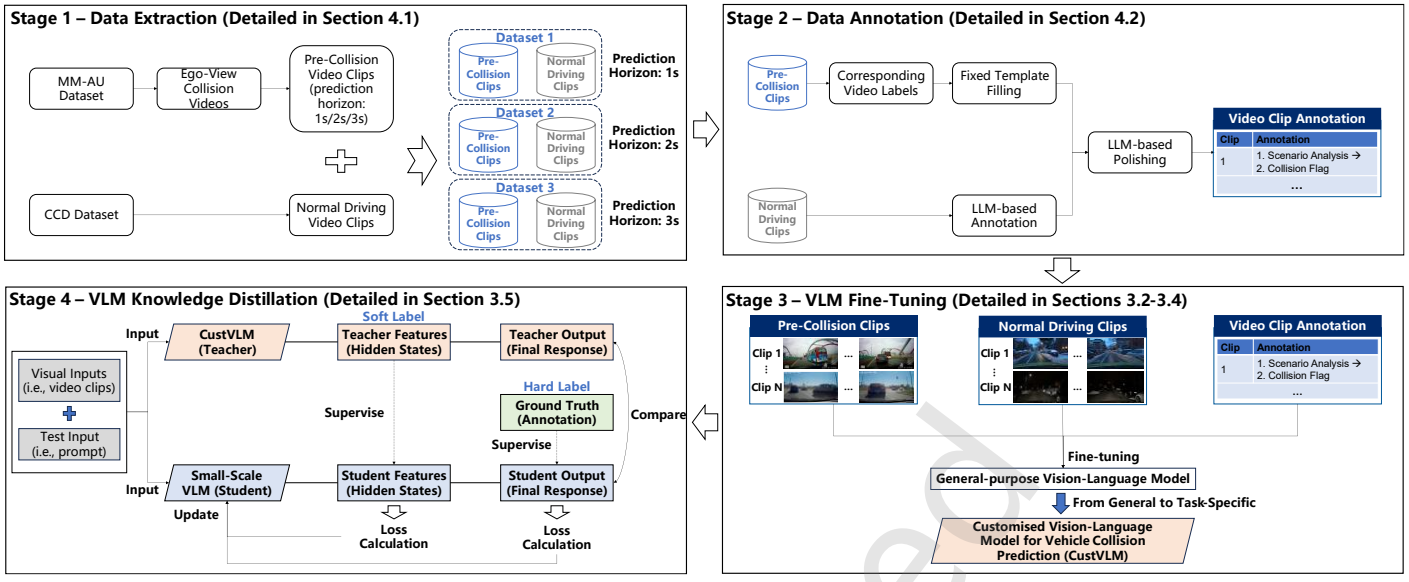

Fig. 1. Overall framework

prediction performance. After distillation, the resulting compact VLM can be independently employed for explainable vehicle collision prediction.

This section mainly focuses on the key methods used in stages 3 and 4, while stages 1 and 2 will be detailed in the subsequent Data Preparation section.

3.2. Vision-language model for driving video analysis

This study employs VLMs for scenario analysis-based collision prediction, with VideoLLaMA2 (Cheng et al., 2024) specifically chosen for implementation, given its strong performance across various vision-language tasks.

The overall architecture of VideoLLaMA2 is illustrated in Fig. 2. Visual inputs (i.e., driving video clips) are first processed by a visual encoder, and the extracted features are then passed to a spatial-temporal convolution connector (STC Connector) designed for enhancing spatial-temporal feature mining. A projector is utilised to map the extracted visual semantic features into the input space of the LLM, which corresponds to the embedding space of the textual modality. Finally, the LLM jointly processes the text prompt and the projected visual features to generate the final response.

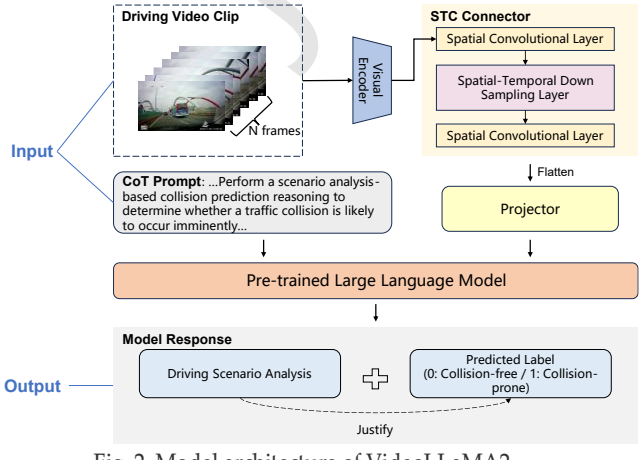

Fig. 2. Model architecture of VideoLLaMA2

3.3. Two-step chain-of-thought (CoT) prompting method

To guide VLMs in scenario analysis-based collision prediction, a two-step CoT prompting method is proposed as follows:

(1) Step 1: Scenario Analysis - Analyse the current driving scenario in detail. Identify and describe any potential sources of collision risk, whether involving the ego vehicle or any other vehicles in the scenario.

(2) Step 2: Collision Prediction - Based on your analysis in step 1, assess whether a vehicle-to-vehicle collision is likely to occur imminently. This includes potential collisions involving the ego vehicle or only between other vehicles in the scenario.

3.4. Low-rank adaptation (LoRA) for VLM fine-tuning

Traditional full fine-tuning method (Ding et al., 2023) directly optimises the whole weight matrix $W \in \mathbb{R}^{d\times k}$ using Eq. (1), where L denotes the loss function and η is the learning rate. Since all elements of W are involved in training, the number of trainable parameters scales as $\mathcal{O}(dk)$. This large number of parameter updates leads to significant computational overhead, making the fine-tuning process costly.

$$W \leftarrow W + \Delta W = W - \eta \frac{\partial L}{\partial W} \tag{1}$$

Rather than updating the entire weight matrix $W \in \mathbb{R}^{d\times k}$, LoRA (Hu et al., 2022) freezes the pre-trained parameter matrix $W_0 \in \mathbb{R}^{d\times k}$ and represents the weight change $\Delta W \in \mathbb{R}^{d\times k}$ as the product of two smaller matrices: $A \in \mathbb{R}^{r\times k}$ and $B \in \mathbb{R}^{d\times r}$, where the rank r is much smaller than $min(d,k)$. Both A and B are trainable. The fine-tuned weight matrix is then computed using Eq. (2).

$$W = W_0 + \Delta W = W_0 + BA \tag{2}$$

By factorising ΔW into the smaller matrices A and B, the number of trainable parameters is reduced from $\mathcal{O}(dk)$ to $\mathcal{O}(r(d+k))$. Given that the rank r is much smaller than $min(d,k)$, the number of trainable parameters in LoRA (i.e., $\mathcal{O}(r(d+k))$) is significantly lower than that of full fine-tuning (i.e., $\mathcal{O}(dk)$), resulting in substantially reduced computational costs and improved fine-tuning efficiency.

3.5. Feature-based knowledge distillation

The proposed feature-based knowledge distillation framework in this study is illustrated in Fig. 3. During the distillation process, the

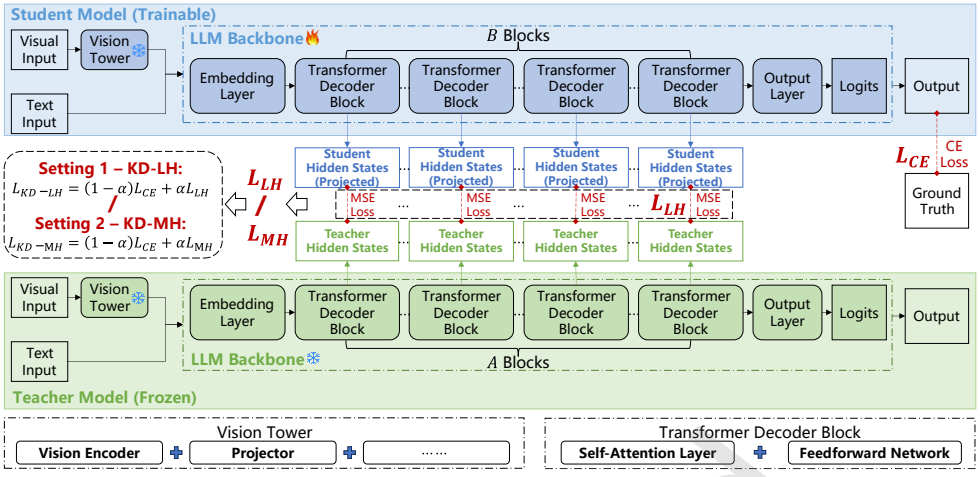

Fig. 3. Feature-based knowledge distillation framework

teacher model f_T is kept frozen, and only the student model f_S, specifically, its LLM backbone is optimised.

Assume that the training dataset $\{x_i, y_i\}_{i=1}^N$ contains N training samples, where x_i and y_i represent the i^{th} input and the corresponding label. The basic fine-tuning process for the student model is performed using the cross entropy (CE) loss function, as defined in Eq. (3).

$$L_{CE} = -\frac{1}{N}\sum_{i=1}^{N} log p_s(y_i|x_i) \tag{3}$$

$$p_s(y_i|x_i) = softmax\left(z_s(x_i)\right) \tag{4}$$

Where $p_s(y_i|x_i)$ represents the predicted probability assigned by the student model to label y_i conditioned on input x_i, which is obtained by applying the Softmax function to the output logits $z_s(x_i)$ of the student model, as shown in Eq. (4).

In addition to using the ground truth (commonly referred to as hard labels) as supervision during fine-tuning, the core idea of knowledge distillation lies in introducing the features of the teacher model (commonly referred to as soft labels) to guide the student model. These soft labels contain richer information than the hard labels, revealing how the teacher represents and embeds the input. Thus, soft labels enable the student to more effectively replicate the teacher's behaviour (Zhang et al., 2024a). Typical knowledge distillation studies commonly use the teacher's logits to supervise the student (Li et al., 2024). However, the performance of such an approach is limited when the vocabularies of the teacher and student models are inconsistent. To address this issue, the hidden states from the last transformer block of the teacher model are used as supervision signals (referred to as KD-LH in this study for convenience), since these hidden states are highly correlated with the logits. The student's and teacher's hidden states in the last transformer block can then be aligned using Eq. (5).

$$L_{LH} = \frac{1}{N}\sum_{i=1}^{N} MSE\left(s_{i,B}W, t_{i,A}\right) \tag{5}$$

$$MSE(q,v) = \frac{1}{HG}\sum_{h=1}^{H}\sum_{g=1}^{G}(q_{h,g} - v_{h,g})^2, q, v \in \mathbb{R}^{H\times G} \tag{6}$$

$$t_i = [t_{i,1}, \cdots, t_{i,a}, \cdots, t_{i,A}] = f_T(x_i) \tag{7}$$

$$s_i = [s_{i,1}, \cdots, s_{i,b}, \cdots, s_{i,B}] = f_S(x_i) \tag{8}$$

Where MSE denotes the mean squared error function given in Eq. (6) exemplified by the computation between two vectors q and v in $\mathbb{R}^{H\times G}$, W represents a learnable linear transformation that projects the student hidden embeddings into the same dimension as teacher hidden embeddings, A and B denote the numbers of transformer blocks in the teacher and student models, $t_{i,A}$ and $s_{i,B}$ are the hidden states extracted from the teacher's and student's last transformer blocks respectively, as defined in Eqs. (7) and (8). Here, $t_{i,a}$ and $s_{i,b}$ indicate the hidden states obtained from the a^{th} ($a \in [1, A]$) layer of the teacher f_T and the b^{th} ($b \in [1, B]$) layer of the student f_S, respectively. The final objective function of KD-LH is given in Eq. (9), where $\alpha \in (0, 1)$ is a weight hyperparameter that balances the L_{CE} and L_{LH}.

$$L_{KD\text{-}LH} = (1-\alpha)L_{CE} + \alpha L_{LH} \tag{9}$$

In KD-LH, only the hidden states from the last layers of the teacher and student models are utilised and aligned. Aligning multiple transformer decoder blocks may further improve knowledge distillation performance, as this allows the student to better learn how the teacher represents and embeds the input. In this study, the teacher and student models are assumed to have the same number of transformer blocks, denoted as K (i.e., $A = B = K$). Accordingly, all transformer blocks can be aligned using Eq. (10), where W_k denotes a trainable linear transformation that maps the student's hidden embeddings from the k^{th} ($k \in [1, K]$) layer into the same dimensional space as those of the teacher's k^{th} layer. This method is referred to as KD-MH for convenience in this study, and the objective function is given in Eq. (11). In cases where the teacher and student models have different numbers of transformer blocks, a subset of layers can be selected for alignment.

$$L_{MH} = \frac{1}{NK}\sum_{i=1}^{N}\sum_{k=1}^{K} MSE\left(s_{i,k}W_k, t_{i,k}\right) \tag{10}$$

$$L_{KD\text{-}MH} = (1-\alpha)L_{CE} + \alpha L_{MH} \tag{11}$$

3.6. Evaluation metrics

3.6.1 Prediction performance

Vehicle collision prediction can be formulated as a binary classification problem in this study. Ground truth labels are assigned as 1 (collision-prone) for pre-collision clips and 0 (collision-free) for normal driving clips. Accuracy, precision, and recall are adopted as evaluation metrics as they are standard measures for binary classification tasks. Higher values of these metrics indicate better prediction performance.

Given an evaluation dataset $\{x_i, y_i\}_{i=1}^N$ containing N samples, where $y_i \in \{0,1\}$ denotes the ground truth label and $\hat{y}_i \in \{0,1\}$ represents the

predicted label for sample x_i. The values of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) can be derived according to Eqs. (12)-(15). The indicator function $\mathbf{1}(\cdot)$ takes the value 1 when the stated condition is true and 0 otherwise. Based on these definitions, the three metrics are computed using Eqs. (16)-(18).

$$TP = \sum_{i=1}^{N} \mathbf{1}(\hat{y}_i = 1 \wedge y_i = 1) \tag{12}$$

$$TN = \sum_{i=1}^{N} \mathbf{1}(\hat{y}_i = 0 \wedge y_i = 0) \tag{13}$$

$$FP = \sum_{i=1}^{N} \mathbf{1}(\hat{y}_i = 1 \wedge y_i = 0) \tag{14}$$

$$FN = \sum_{i=1}^{N} \mathbf{1}(\hat{y}_i = 0 \wedge y_i = 1) \tag{15}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{\sum_{i=1}^{N} \mathbf{1}(\hat{y}_i = y_i)}{N} \tag{16}$$

$$Precision = \frac{TP}{TP + FP} = \frac{\sum_{i=1}^{N} \mathbf{1}(\hat{y}_i = 1 \wedge y_i = 1)}{\sum_{i=1}^{N} \mathbf{1}(\hat{y}_i = 1)} \tag{17}$$

$$Recall = \frac{TP}{TP + FN} = \frac{\sum_{i=1}^{N} \mathbf{1}(\hat{y}_i = 1 \wedge y_i = 1)}{\sum_{i=1}^{N} \mathbf{1}(y_i = 1)} \tag{18}$$

3.6.2 Computational efficiency

Following existing literature (Fang et al., 2021; Liu et al., 2025a), three metrics are selected to assess the model's computational efficiency, with lower values indicating better efficiency.

(1) Latency: Average inference time per sample on the testing dataset, indicating the time required per input.

(2) Params: Total number of parameters, reflecting the model size.

(3) FLOPs: Number of floating-point operations, measuring the amount of arithmetic computation involved in a forward pass.

4 Data preparation

4.1. Data extraction

This study develops vehicle collision prediction models that classify current driving conditions as either collision-prone or collision-free. Thus, both types of scenarios are required for model development. Collision-prone scenarios are defined as driving segments that precede a collision, maintaining a certain time gap from the actual event (Gao et al., 2023). These segments are extracted as pre-collision video clips from collision videos, while normal driving video clips serve as representations of collision-free scenarios.

Specifically, the collision-prone scenarios are sourced from the MM-AU dataset presented in Fang et al. (2024), which is designed for multi-modal collision video understanding and contains a large collection of real-world ego-view collision videos with rich annotations. Normal driving video clips from the CCD dataset (Bao et al., 2020) are employed to represent collision-free scenarios. The collision videos from CCD are not used because they are limited to 5-second clips, making the pre-collision segments too short (always less than 5 seconds) for pre-collision clip extraction.

Although the pre-collision clips and normal driving clips were obtained from two different datasets, MM-AU and CCD, respectively, both datasets consist of real-world ego-view dashcam videos. In addition, both data sources cover diverse camera mounting positions and geographic conditions, which supports their comparability in terms of visual perspective and driving environment. A preliminary scenario-level comparability check further confirmed that the clips from the two sources cover broadly similar real-world driving conditions, including daytime and nighttime scenes, different weather conditions, various road environments, and multiple road types.

The process of pre-collision video clip extraction is illustrated in Fig. 4, using a single collision video as an example. Collision videos in the MM-AU dataset vary in duration, with the majority ranging from 5 to 15 seconds. Assuming the collision occurs at time t in the video, and a time gap of k seconds is reserved, with the length of the extracted pre-collision video clip being a seconds, then the pre-collision clip corresponds to the time interval from $t-k-a$ to $t-k$. The time gap k can be viewed as the prediction horizon, indicating that the model aims to predict the collision k seconds in advance. Referring to existing literature (Liao et al., 2024; Zhang et al., 2025), the data window for the pre-collision clip is set to 5 seconds. Three prediction horizons are considered: 1, 2, and 3 seconds. To ensure sufficient pre-collision context, only collision videos longer than 8 seconds are used for clip extraction. For each prediction horizon, a total of 1,500 pre-collision video clips are extracted.

Table 2. Dataset summary.

Data-set	Total Samples	Pre-collision Clips	Normal Clips	Prediction Horizon	Data Window
1	3,000	1,500	1,500	1 s	5 s
2	3,000	1,500	1,500	2 s	5 s
3	3,000	1,500	1,500	3 s	5 s

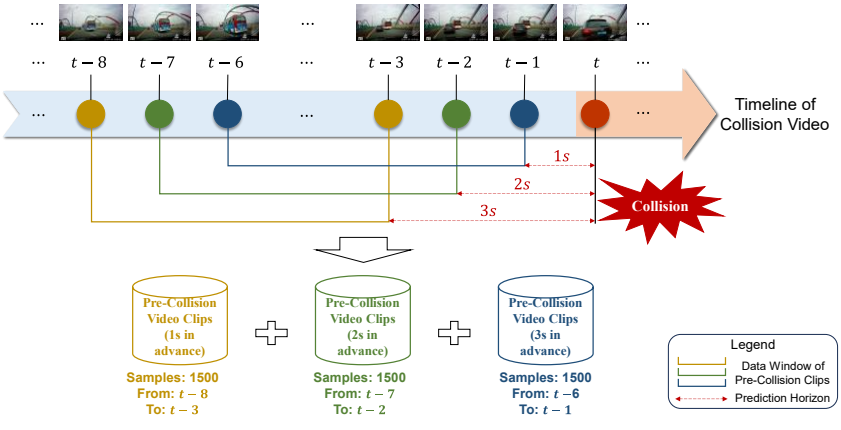

Fig. 4. Pre-collision video clip extraction diagram

To avoid the issue of imbalanced learning, an equal number of normal driving video clips were randomly sampled from the CCD dataset. These clips are provided as 5-second segments by default, matching the length of the extracted pre-collision clips. All clips used for model training and testing were processed consistently. Specifically, the inputs were standardised into 5-second video clips sampled at 10 Hz and subsequently processed according to the standard input configuration of the adopted VLM. This unified preprocessing helps reduce the influence of low-level dataset-specific differences and mitigates the potential concern of dataset-level domain shift. Table 2 summarises the three balanced modelling datasets. The scale of the developed datasets is comparable to that of existing studies (Liao et al., 2024; Yu et al., 2022).

Both the extracted pre-collision and normal driving clips encompass a broad diversity of scenarios, including different environments (e.g., urban and rural), lighting conditions (daytime and nighttime), road types (intersections, arterials, etc.), and weather conditions (sunny, cloudy, rainy, snowy). Originating from multiple countries and varied geographic regions, these videos represent a wide range of global driving scenarios. Notably, the collision videos from which the pre-collision clips were extracted include both ego-involved collisions and collisions involving only other vehicles, where the ego vehicle is not directly involved. Accordingly, the extracted pre-collision clips are used to represent scene-level collision risk in ego-view driving videos, as safety-critical events may occur either in direct relation to the ego vehicle or within the surrounding visible traffic environment. Non-ego collisions are also retained as collision-prone scenarios because they may still influence subsequent driving conditions by triggering sudden braking, evasive manoeuvres, lane blockages, or secondary risks (Wang et al., 2026b; Wu et al., 2025).

4.2. Data annotation

For the normal driving video clips, this study employs LLaMA3-8B, the 8B-parameter instruction-tuned version of LLaMA3 (Dubey et al., 2024), to directly generate descriptions of safe driving scenarios without potential collision risks. These descriptions are then self-refined by the same model to ensure coherence and to avoid repetitive structures or expressions across clips. For the pre-collision clips, the annotation process is illustrated in Fig. 5. Specifically, two labels from the corresponding collision videos in the MM-AU dataset: collision reason and ego involvement flag, are used to fill in a fixed template and generate initial descriptions. These initial descriptions are then polished and refined using LLaMA3-8B to enhance coherence and diversity, ensuring that each clip is paired with a unique and contextually appropriate description.

Notably, LLaMA3-8B was not used to generate annotations in a fully open-ended manner. Instead, the generated descriptions were grounded in available label information and pre-defined annotation templates. For pre-collision clips, the key semantic information was derived from the verified MM-AU labels, including the collision reason and ego-involvement flag, and was first organised using a fixed template. LLaMA3-8B was then used to convert these label-constrained templates into fluent and diverse scenario descriptions. For normal driving clips, the model was instructed to generate descriptions under the pre-defined collision-free label. In this way, the annotation process was guided by existing labels and task definitions rather than relying on free-form LLM inference. Manual verification of a stratified sample of 800 annotations, comprising 400 pre-collision clips and 400 normal-driving clips, identified no critical factual inconsistencies with the corresponding videos or source labels, confirming the reliability and satisfactory quality of the LLM-assisted automatic annotations. Such sampling-based manual verification is consistent with prior LLM-assisted annotation studies and provides a scalable quality-control strategy for large-scale annotation (Zhou et al., 2025). Representative examples of the annotated data are shown later in Fig. 8.

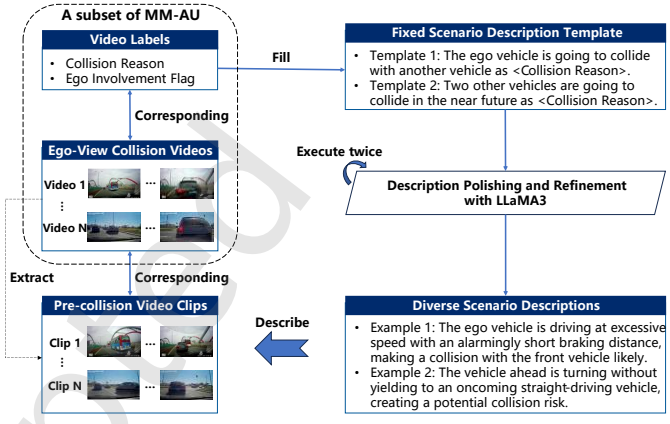

Fig. 5. Pre-collision video clip annotation process

5 Experiments and results

This section first customises general-purpose VLMs for vehicle collision prediction through fine-tuning and then compresses them via knowledge distillation. Experiments are conducted on the three developed datasets. To avoid data leakage, the train-test split was performed at the original-video level. Specifically, clips derived from the same original video were assigned exclusively to either the training set or the testing set, ensuring no overlap in source videos between the two sets. For each dataset, the training and testing subsets followed a 4:1 ratio, containing 2,400 and 600 samples respectively, with an equal proportion of pre-collision and normal driving clips in each subset. All experiments are conducted on a desktop workstation equipped with an Intel Xeon W7-3455 CPU, two NVIDIA RTX A6000 GPUs, and 256 GB RAM. Finally, explainable prediction cases are also presented.

5.1. VLM fine-tuning

This subsection aims to adapt general-purpose VLMs into customised models through fine-tuning on real-world driving clips. The developed models are evaluated solely on prediction performance metrics, as computational efficiency is not considered at this stage.

5.1.1 Experimental setup

VideoLLaMA2 is fine-tuned via LoRA on the three constructed datasets to obtain customised VLMs. In addition, four benchmark models are considered:

(1) CNN-LSTM: Based on Yu et al. (2020), this integrated model extracts spatial features from individual video frames using a CNN, and then captures temporal dependencies across frames via an LSTM to predict collision risks.

(2) Vision Transformer (ViT): Following Kang et al. (2022), ViT is used to split driving images into patches, which are then processed through transformer layers to output a collision risk score.

Table 3. Model evaluation results. The best and second-best performances on each dataset are denoted by bold and underlined, respectively. Legend: DL - deep learning, OrigVLM-Direct - original VLM with direct prompt, OrigVLM-CoT - original VLM with CoT prompt, CustVLM-Direct - customised VLM with direct prompt, CustVLM-CoT - customised VLM with CoT prompt.

Dataset / Prediction Horizon (s)	Model		Accuracy (%)	Precision (%)	Recall (%)
1	Baseline DL	CNN-LSTM	78.67	76.88	82.00
		ViT	82.50	81.76	83.67
		SwinT	82.67	81.82	84.00
	Baseline VLM	OrigVLM-Direct	68.33	**91.67**	40.33
		OrigVLM-CoT	58.67	70.97	29.33
	Ours	CustVLM-Direct	<u>86.17</u>	82.39	**92.00**
		CustVLM-CoT	**87.17**	<u>87.29</u>	<u>87.00</u>
2	Baseline DL	CNN-LSTM	78.33	<u>83.20</u>	71.00
		ViT	80.33	77.41	<u>85.67</u>
		SwinT	81.67	81.88	81.33
	Baseline VLM	OrigVLM-Direct	67.33	64.13	78.67
		OrigVLM-CoT	48.67	49.22	83.67
	Ours	CustVLM-Direct	<u>83.67</u>	**86.07**	80.33
		CustVLM-CoT	**85.33**	79.44	**95.33**
3	Baseline DL	CNN-LSTM	76.50	81.42	68.67
		ViT	75.67	74.84	77.33
		SwinT	<u>79.67</u>	79.08	<u>80.67</u>
	Baseline VLM	OrigVLM-Direct	65.67	**89.17**	35.67
		OrigVLM-CoT	61.33	83.33	28.33
	Ours	CustVLM-Direct	<u>79.67</u>	78.16	**82.33**
		CustVLM-CoT	**81.67**	<u>84.17</u>	78.00

(3) Swin Transformer (SwinT): Developed by Liu et al. (2021), SwinT extends ViT with a hierarchical structure and shifted window mechanism, enabling more effective capture of both local and global image features for prediction tasks.

(4) Original VLM: The VideoLLaMA2 model is directly applied using the parameters provided in Cheng et al. (2024), without additional fine-tuning.

To ensure model convergence, the CNN-LSTM, ViT and SwinT models were trained for 20 epochs on all datasets with a learning rate of 3e-4, while customised VLMs were fine-tuned for 30 epochs with a learning rate of 2e-5.

During inference, both the original and customised VLMs were tested with two prompts: (1) direct prompt, which directly asks the model to determine whether a collision is likely to occur imminently, and (2) CoT prompt, which guides the model to first analyse the current driving scenario and then make a decision based on the analysis, as given in Section 3.3. For both prompt settings, the model received only the input driving video clip and the corresponding task prompt as input. No generated description or other annotation metadata was provided during inference.

5.1.2 Evaluation results

The evaluation results of all models are summarised in Table 3. Across all datasets, the models exhibit similar performance patterns: the original VLMs perform the worst, while the customised VLMs (i.e., CustVLM) achieve the best results, outperforming SwinT, which in turn surpasses ViT and CNN-LSTM. For the customised VLMs, the CoT prompt leads to better performance compared to the direct prompt. Additionally, the models generally perform better on Dataset 1 than on Datasets 2 and 3, suggesting that reducing the time gap between pre-collision segments and collision events can improve model performance to some extent.

Table 4. Average performance of all models across datasets. Bold and underlined indicate top and second-best results, respectively. This table uses the same legend as Table 3.

Model		Accuracy (%)	Precision (%)	Recall (%)
Baseline DL	CNN-LSTM	77.83	80.50	73.89
	ViT	79.50	78.00	82.22
	SwinT	81.33	80.93	82.00
Baseline VLM	OrigVLM-Direct	67.11	81.65	51.56
	OrigVLM-CoT	56.22	67.84	47.11
Ours	CustVLM-Direct	<u>83.17</u>	<u>82.21</u>	<u>84.89</u>
	CustVLM-CoT	**84.72**	**83.64**	**86.78**

The average performance of all models across the three datasets is summarised in Table 4. CustVLM-CoT achieves 84.72% accuracy, 83.64% precision, and 86.78% recall, showing improvements of 1.86%, 1.74%, and 2.23% over the customised VLM with direct prompt (CustVLM-Direct), and gains of 4.17%, 3.35%, and 5.83% compared to the best-performing deep learning baseline (SwinT). Furthermore, compared to the original VLM with direct prompt (OrigVLM-Direct), CustVLM-CoT achieves substantial improvements in both accuracy and recall, demonstrating that fine-tuning can greatly enhance the performance of general-purpose VLMs.

5.1.3 Sensitivity analysis

In this subsection, the effects of two key factors are analysed: (1)

collision involvement scope and (2) fine-tuning sample size. Dataset 1 (with a 1-second prediction horizon) is used as the data source, and CustVLM-CoT is selected as the evaluation model.

(1) Collision Involvement Scope: In the main experiments, both ego-involved collisions and collisions involving only surrounding vehicles are treated as collision-prone events. To further examine the effect of collision involvement scope, an additional ego-focused experiment was conducted using Dataset 1. In this setting, pre-collision clips involving only surrounding vehicles were re-labelled as collision-free, while ego-involved pre-collision clips remained labelled as collision-prone. During both fine-tuning and inference, CustVLM-CoT was instructed to predict only whether the ego vehicle is likely to be involved in a near-future collision, with all other experimental settings kept unchanged. As shown in Fig. 6, under this ego-focused setting, the model achieved improved performance: accuracy increased from 87.17% to 91.67%, recall increased from 87.00% to 97.67%, and precision remained nearly unchanged. This result indicates that the model remains effective when the prediction target is restricted to ego-involved collision risks.

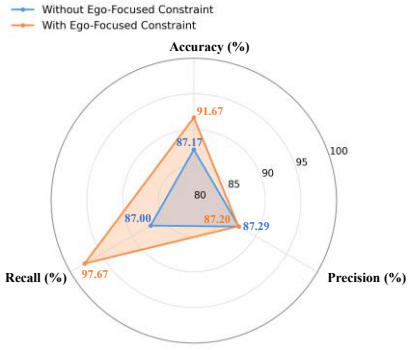

Fig. 6. Model performance comparison (with and without ego-focused constraint)

(2) Fine-Tuning Sample Size: Based on Dataset 1, a fixed testing set of 600 samples was used, while the training set size for fine-tuning was varied. Specifically, in addition to the original 2,400 fine-tuning samples, experiments were conducted using 1,000, 1,500, 2,000, 2,200, 2,600, and 3,000 samples to investigate the effect of fine-tuning sample size on model performance (using accuracy as the metric).

The accuracy differences of CustVLM-CoT under these settings, relative to its performance in the original setting (2,400 samples), are shown in Fig. 7. It can be observed that the reduction of the fine-tuning sample size to 1,000 or 1,500 leads to a noticeable decline in model performance, while the performance stabilises when the fine-tuning sample size exceeds 2,000, with the accuracy difference approaching zero.

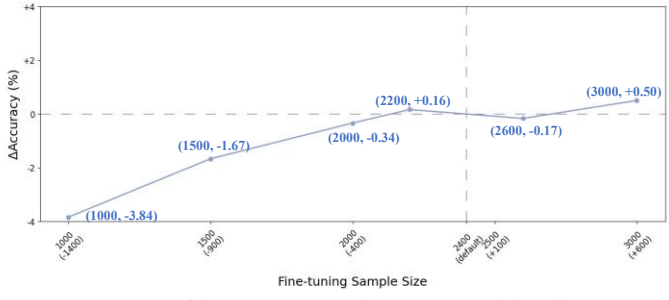

Fig. 7. Impact of fine-tuning sample size on model performance

5.2. VLM knowledge distillation

Although CustVLM-CoT demonstrates superior performance in vehicle collision prediction, its large model size leads to high computational cost, limiting real-world deployment under constrained resources. In this subsection, CustVLM-CoT is employed as the teacher model to distil knowledge into a smaller, more efficient VLM (i.e., student model). The selection of a student model involves a trade-off between performance and efficiency, as performance generally increases with model size, whereas efficiency decreases. According to the literature, the student model size typically ranges from 20% to 50% of that of the teacher model (Yang et al., 2024). Based on this guideline, Qwen2-VL-2B (Wang et al., 2024) is chosen as the student model due to its compact size and strong benchmark performance. Both prediction performance and computational efficiency metrics are used for model evaluation in this subsection.

5.2.1 Experimental setup

Five models are developed and compared on each dataset, as presented in Table 5. S is the vanilla Qwen2-VL-2B with publicly available parameters from Wang et al. (2024), without additional fine-tuning. S-FT is obtained through LoRA fine-tuning using ground-truth labels as supervision signals. S-KD-LH and S-KD-MH further incorporate the hidden states from the teacher model (T) as additional supervision. Notably, for each dataset, T employs the parameters of CustVLM-CoT fine-tuned on that specific dataset, obtained from Section 5.1.

Table 5. Model settings.

Type	VLM	Supervision Signal	Objective Function	Model Name
Teacher	CustVLM	-	-	T
Student	Qwen2-VL-2B	-	-	S
		Ground Truth	Eq. (3)	S-FT
		Ground Truth + Last-Layer Teacher Hidden States	Eq. (9)	S-KD-LH
		Ground Truth + Multi-Layer Teacher Hidden States	Eq. (11)	S-KD-MH

To ensure model convergence, both S-KD-LH and S-KD-MH were trained for 25 epochs on all datasets with a learning rate of 1e-5, whereas S-FT was fine-tuned for 20 epochs using the same learning rate. The weight hyperparameter α is selected as 0.5 to balance different loss terms in the knowledge distillation process. All models adopt the CoT prompt defined in Section 3.3, as it has been demonstrated in Section 5.1 to enhance both model performance and explainability.

5.2.2 Evaluation results

Table 6 summarises the model evaluation results across datasets. Regarding the prediction performance, models trained on Dataset 1 achieve the highest performance, followed by those on Dataset 2, with Dataset 3 showing the lowest performance. Across all datasets, the teacher model (T) consistently achieves the best performance, while the vanilla student model (S) performs the worst. S-FT significantly outperforms S, indicating that task-specific fine-tuning substantially enhances the VLM's performance. Both S-KD-LH and S-KD-MH further outperform S-FT, demonstrating that knowledge distillation via

Table 6. Model evaluation results. The best and second-best prediction performances on each dataset are denoted by bold and underlined, respectively.

Dataset / Prediction Horizon (s)	Model	Prediction Performance			Computational Efficiency		
		Accuracy (%)	Precision (%)	Recall (%)	Latency (s)	Params (B)	FLOPs (G)
1	T	**87.17**	**87.29**	**87.00**	2.76	8.421	1810
	S	57.17	54.49	**87.00**	1.06	2.209	395
	S-FT	77.50	78.55	75.67	1.11	2.213	396
	S-KD-LH	80.67	80.07	81.67	1.06	2.213	396
	S-KD-MH	<u>84.33</u>	<u>83.44</u>	<u>85.67</u>	1.21	2.213	396
2	T	**85.33**	<u>79.44</u>	**95.33**	2.89	8.421	1810
	S	59.67	56.42	<u>85.00</u>	1.01	2.209	395
	S-FT	75.83	75.75	76.00	1.14	2.213	396
	S-KD-LH	80.33	78.80	83.00	1.05	2.213	396
	S-KD-MH	<u>82.83</u>	**81.88**	84.33	1.20	2.213	396
3	T	**81.67**	**84.17**	<u>78.00</u>	2.68	8.421	1810
	S	59.83	56.46	**86.00**	1.00	2.209	395
	S-FT	71.83	71.76	72.00	1.11	2.213	396
	S-KD-LH	74.33	73.70	75.67	1.22	2.213	396
	S-KD-MH	<u>80.17</u>	<u>81.75</u>	77.67	1.19	2.213	396

Table 7. Average performance of all models across datasets. Bold and underlined indicate top and second-best prediction performance, respectively. The numbers in parentheses show the performance percentage change relative to the teacher T.

Model	Prediction Performance			Computational Efficiency		
	Accuracy (%)	Precision (%)	Recall (%)	Latency (s)	Params (B)	FLOPs (G)
T	**84.72**	**83.64**	**86.78**	2.78	8.421	1810
S	58.89 (-30.49%)	55.79 (-33.30%)	<u>86.00</u> (-0.90%)	1.02 (-63.15%)	2.209 (-73.77%)	395 (-78.18%)
S-FT	75.06 (-11.41%)	75.35 (-9.91%)	74.56 (-14.08%)	1.12 (-59.66%)	2.213 (-73.72%)	396 (-78.12%)
S-KD-LH	78.44 (-7.41%)	77.52 (-7.31%)	80.11 (-7.68%)	1.11 (-60.02%)	2.213 (-73.72%)	396 (-78.12%)
S-KD-MH	<u>82.44</u> (-2.69%)	<u>82.36</u> (-1.53%)	82.56 (-4.87%)	1.20 (-56.78%)	2.213 (-73.72%)	396 (-78.12%)

hidden-state supervision provides additional performance gains. Moreover, S-KD-MH outperforms S-KD-LH, suggesting that aligning multiple transformer blocks yields better results than aligning only the last block, as S-KD-MH utilises hidden states from all transformer blocks of the teacher model to supervise the student model, whereas S-KD-LH relies solely on the last block. Regarding the computational efficiency, all student series models (S, S-FT, S-KD-LH and S-KD-MH) achieve significant reductions across metrics and datasets.

The average performance of all models across the three datasets is shown in Table 7. Compared to the teacher model, the best-performing distilled student model, S-KD-MH, achieves a reduction in prediction performance of less than 5%, while largely improving computational efficiency. Specifically, it reduces Params and FLOPs to approximately one-fourth of the teacher model and cuts latency by more than half. Compared with S-KD-LH, S-KD-MH yields additional gains of 5.10%, 6.24%, and 3.05% in accuracy, precision, and recall, respectively.

5.2.3 Ablation study

To further analyse the contribution of feature distillation at different transformer depths, a stage-wise ablation study was conducted on Dataset 1. Both the teacher and student VLMs contain 28 transformer blocks, which were divided into four consecutive stages: early stage (Blocks 1-7), lower-middle stage (Blocks 8-14), upper-middle stage (Blocks 15-21), and late stage (Blocks 22-28). Based on this partition, four ablation variants were developed by removing the feature distillation loss from one stage at a time while keeping the remaining stages and all other experimental settings unchanged. These variants were compared with S-KD-LH, which uses only the last transformer block, and S-KD-MH, which uses all transformer blocks.

Table 8. Stage-wise ablation results on Dataset 1.

Model	Distilled Blocks	Accuracy (%)	Precision (%)	Recall (%)
S-KD-LH	28 (last block)	80.67	80.07	81.67
S-KD-MH	1-28 (all blocks)	84.33	83.44	85.67
w/o early stage	8-28	83.67	83.01	84.67
w/o lower-middle stage	1-7, 15-28	83.17	83.96	82.00
w/o upper-middle stage	1-14, 22-28	82.00	81.79	82.33
w/o late stage	1-21	81.00	81.42	80.33

As shown in Table 8, S-KD-MH achieves the best overall performance, confirming the benefit of using multi-stage hidden-state

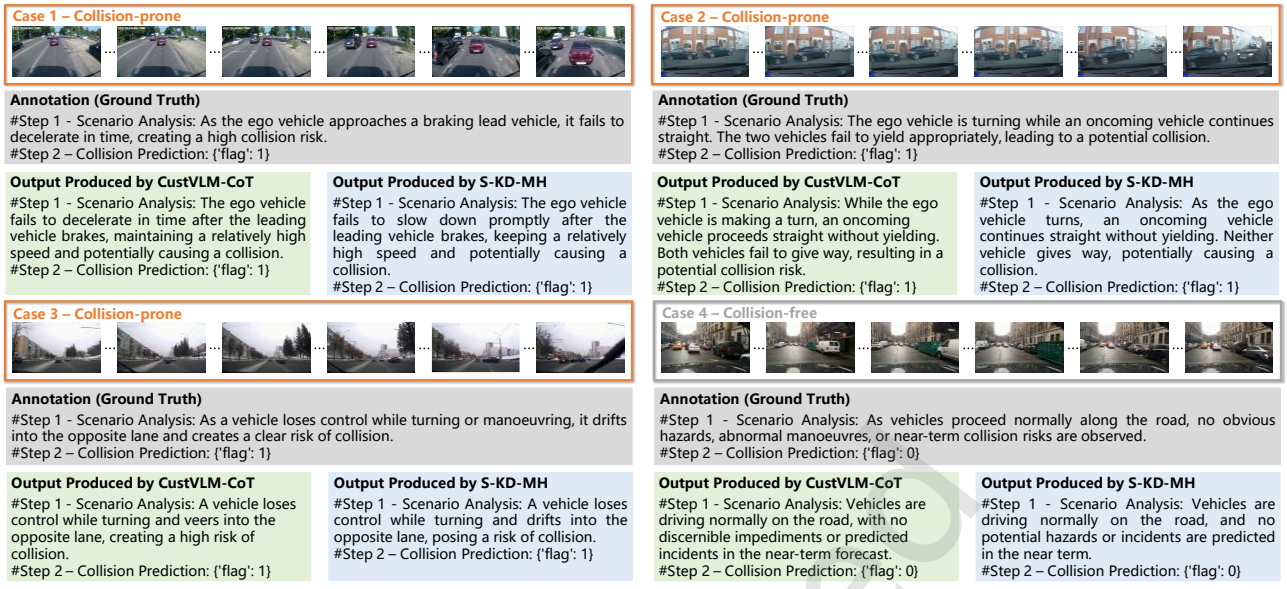

Fig. 8. Explainable prediction cases.

supervision. Removing any stage reduces performance, indicating that hidden states from transformer blocks at different depths provide complementary supervision. The performance degradation is smallest when the early stage is removed and largest when the late stage is removed, suggesting that later transformer blocks contribute more strongly to collision-risk prediction. Nevertheless, the multi-stage ablation variants generally outperform S-KD-LH, further demonstrating the value of incorporating hidden-state supervision from transformer blocks at different depths.

5.3. Explainable prediction cases

Beyond achieving strong prediction performance, the developed models also demonstrate a key advantage in explainability. Fig. 8 presents four examples of model outputs on randomly selected video clips from Dataset 1. Results are shown for both CustVLM-CoT and S-KD-MH, the best-performing customised and distilled VLMs developed in Sections 5.1 and 5.2, respectively.

As observed, for pre-collision clips, the models effectively identify and explain the potential sources of risk, which in turn support their decisions to classify the scenario as collision-prone. This scenario analysis-based collision prediction strategy allows for evaluating whether the model has learned to make decisions based on correct and relevant factors. Moreover, the outputs of S-KD-MH (i.e., the student model) closely resemble those of CustVLM-CoT (i.e., the teacher model) in both tone and content, further confirming the effectiveness of the proposed knowledge distillation approach.

6 Discussions

6.1. Model implementation

The explainable vehicle collision prediction model developed in this study is expected to be integrated into a vision-based collision warning system. During vehicle operation, the onboard dash camera continuously records driving video, from which 5-second clips are fed into the proposed model. The model then provides an analysis of the current driving scenario and determines whether it is collision-prone or collision-free by assessing the likelihood of a near-future collision. When the current scenario is determined to be collision-prone, the system is automatically triggered to issue a warning message to the driver. Notably, the scenario analysis generated by the model can support the design of more informative and context-aware warning messages, thereby enhancing user trust in ADAS.

By default, the proposed framework predicts both ego-involved and non-ego-involved collision risks, as both are treated as collision-prone scenarios in the constructed dataset. In practical ADAS implementation, the generated contextual descriptions can help identify whether the predicted risk directly involves the ego vehicle and support different warning strategies, such as urgent warnings for ego-involved risks and situational-awareness alerts for risks that do not directly involve the ego vehicle. The warning scope can also be adjusted according to specific application needs, for example by adapting the framework to predict only ego-involved collision risks.

In addition, inference latency should be carefully considered in practical applications given the high computational costs of VLMs. In this study, latency (reflected by the average inference time) was measured on our workstation, and servers with stronger computing resources would naturally achieve shorter inference time. Nevertheless, the proposed feature-based knowledge distillation approach would improve the computational efficiency of VLMs with marginal performance loss under any computing resource conditions. In other words, it can cut down on the computational resource demand for any given inference latency requirement. Thus, this approach can promote the practical deployment of VLMs in vehicle collision prediction by reducing the associated computational and implementation costs.

6.2. Limitations and future directions

This study is among the first to apply VLMs for explainable vehicle collision prediction and addresses two key challenges: (1) adapting general-purpose VLMs to the specific task of vehicle collision prediction, and (2) reducing VLM inference latency to mitigate the excessive computational demands. However, several limitations remain in the current study.

From the data perspective, the annotations for pre-collision clips in this study are primarily based on the ego involvement flag and

simplified collision reasons from the MM-AU dataset (Fang et al., 2024). Incorporating kinematic information, such as the speed, acceleration, and positions of traffic participants, may allow VLMs to generate more informative scenario descriptions and in turn make more accurate predictions. However, this is not feasible in the current study due to the lack of publicly available collision video datasets with corresponding kinematic data.

From the methodological perspective, this study primarily relied on the VLM's self-reported justifications to explain its decision-making process. Future work could further explore methods for interpreting the model's internal reasoning mechanisms. In addition, as this study focuses on developing a specialised VLM for vehicle collision prediction, potential changes in its performance on general-purpose benchmarks after task-specific fine-tuning were not evaluated. The potential trade-off between task specialisation and the preservation of broader model capabilities could be explored in future work.

7 Conclusions

This study advances the development of reliable vision-based collision warning systems by distilling VLMs for explainable vehicle collision prediction. Pre-collision and normal driving clips were extracted and annotated to construct three modelling datasets with prediction horizons of 1, 2, and 3 seconds. General-purpose VLMs were fine-tuned on these datasets using LoRA to obtain customised models, which were subsequently compressed into smaller and more efficient VLMs with comparable performance through feature-based knowledge distillation. A two-step CoT prompt was designed to guide VLMs to perform scenario analysis-based collision prediction, where the scenario analysis provided a clear rationale for the prediction outcomes.

Results from extensive experiments demonstrate that fine-tuning substantially enhances the performance of VLMs on specific tasks, while the proposed CoT prompting strategy provides additional performance gains. CustVLM-CoT consistently achieves the best results across all datasets, with average accuracy, precision, and recall of 84.7%, 83.7%, and 86.8%, surpassing the best-performing deep learning baseline (SwinT) by 4.2%, 3.4%, and 5.8%. Furthermore, the distilled VLM, with CustVLM-CoT serving as the teacher model, achieves a latency reduction of more than 50% while preserving over 95% of the original prediction performance.

This study is among the first to develop a unified framework that systematically integrates VLMs and knowledge distillation for explainable vehicle collision prediction under the computational constraints of real-world deployment. The proposed approach provides reliable collision predictions together with analyses of potential risk sources in the current driving scenario, thereby offering explicit rationales for its decisions. The feature-based knowledge distillation framework improves deployment feasibility by reducing computational resource requirements and inference latency while retaining most of the teacher model's prediction performance. Moreover, the framework is designed to be applicable beyond the specific teacher and student models examined in this study. The resulting model has the potential to support future integration into ADAS, thereby enhancing user trust and contributing to proactive traffic safety.

Author contributions

Ruici Zhang: Writing – original draft, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Yuxiang Feng:** Writing – review & editing, Methodology, Conceptualization. **Jose Escribano:** Writing – review & editing, Methodology, Conceptualization, Supervision. **Mohammed Quddus:** Writing – review & editing, Methodology, Conceptualization, Supervision.

Replication and data sharing

The data and code used in this study are available at https://doi.org/10.26599/ETSD.2026.9190034.

Acknowledgements

This work was funded by the President's PhD Scholarships of Imperial College London. The authors are grateful to Prof. Rongjie Yu (Tongji University) for his valuable help during the revision of this manuscript.

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

References

An, J., Joe, I., 2022. Attention map-guided visual explanations for deep neural networks. *Applied Sciences*, 12, 3846.

Bao, W., Yu, Q., Kong, Y., 2020. Uncertainty-based traffic accident anticipation with spatio-temporal relational learning. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 2682-2690.

Bing, Z., Li, L., Liang, J., 2025. Optimizing Knowledge Distillation in Transformers: Enabling Multi-Head Attention without Alignment Barriers. *arXiv preprint arXiv:2502.07436*.

Cai, Y., Zhang, J., He, H., He, X., Tong, A., Gan, Z., et al., 2025. Llava-kd: A framework of distilling multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 239-249.

Chen, L., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., et al., 2024. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37, 27056–27087.

Cheng, Z., Leng, S., Zhang, H., Xin, Y., Li, X., Chen, G., et al., 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*.

Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., et al., 2023. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature machine intelligence*, 5, 220–235.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., et al., 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Fang, J., Li, L., Zhou, J., Xiao, J., Yu, H., Lv, C., et al., 2024. Abductive ego-view accident video understanding for safe driving perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22030-22040.

Fang, Z., Wang, J., Hu, X., Wang, L., Yang, Y., Liu, Z., 2021. Compressing visual-linguistic model via knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1428-1438.

Formosa, N., Quddus, M., Ison, S., Timmis, A., 2022. A new modeling approach for predicting vehicle-based safety threats. *IEEE transactions on intelligent transportation systems*, 23(10), 18175-18185.

Guan, Y., Liao, H., Wang, C., Wang, B., Zhang, J., Hu, J., Li, Z., 2025. Domain-enhanced dual-branch model for efficient and interpretable accident anticipation. *Communications in Transportation Research*, 5, 100214.

Guo, X., Zhang, Q., Jiang, J., Peng, M., Zhu, M., Yang, H.F., 2024. Towards explainable traffic flow prediction with large language models. *Communications in Transportation Research*, 4, 100150.

Gao, Z., Wang, L., Xu, J., Yu, R., Wang, L., 2023. Hazardous driving scenario identification with limited training samples. In *International Conference on Neural Information Processing*, pp. 203-214.

He, S., Li, A., Chen, T., 2024. Rethinking pruning for vision-language models: Strategies for effective sparsity and performance restoration. *arXiv preprint arXiv:2404.02424*.

Heskes, T., Sijben, E., Bucur, I.G., Claassen, T., 2020. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. *Advances in neural information processing systems*, 33, 4778-4789.

Hinton, G., Vinyals, O., Dean, J., 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., et al., 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2), 3.

Huang, S., Zhang, H., Zhong, L., Chen, H., Gao, Y., Hu, Y., Qin, Z., 2025. From image to video, what do we need in multimodal llms?. *arXiv preprint arXiv:2404.11865*.

Ingle, S., Phute, M., 2016. Tesla autopilot: semi autonomous driving, an uptick for future autonomy. *International Research Journal of Engineering and Technology*, 3, 369–372.

Kang, M., Lee, W., Hwang, K., Yoon, Y., 2022. Vision transformer for detecting critical situations and extracting functional scenario for automated vehicle safety assessment. *Sustainability*, 14, 9680.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. *Nature* 521, 436–444.

Li, J., Li, D., Savarese, S., Hoi, S., 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730-19742.

Li, M., Zhou, F., Song, X., 2025. Bild: Bi-directional logits difference loss for large language model distillation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 1168-1182.

Li, X., Fang, Y., Liu, M., Ling, Z., Tu, Z., Su, H., 2023b. Distilling large vision-language model with out-of-distribution generalizability. in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2492-2503.

Li, Y., Li, P., Yan, D., Liu, Y., Liu, Z., 2024. Deep knowledge distillation: A self-mutual learning framework for traffic prediction. *Expert Systems with Applications*, 252, 124138.

Liao, H., Li, Y., Li, Z., Bian, Z., Lee, J., Cui, Z., et al., 2024. Real-time accident anticipation for autonomous driving through monocular depth-enhanced 3D modeling. *Accident Analysis & Prevention*, 207, 107760.

Liao, H., Rao, B., Sun, H., Wang, C., Chang, Q., Li, S. E., et al., 2025. Chain-of-Thought Guided Multimodal Large Language Models for Scene-Aware Accident Anticipation in Autonomous Driving. *IEEE Transactions on Intelligent Transportation Systems*.

Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., et al., 2024. Video-llava: Learning united visual representation by alignment before projection. in: *Proceedings of the conference on empirical methods in natural language processing*, pp. 5971-5984.

Liu, L., Wang, P., Wu, G., Jiang, J., Yang, H., 2025a. Towards optimal mixture of experts system for 3d object detection: A game of accuracy, efficiency and adaptivity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Liu, W., Zhang, J., Zheng, B., Hu, Y., Lin, Y., Zeng, Z., 2025b. X-driver: Explainable autonomous driving with vision-language models. *arXiv preprint arXiv:2505.05098*.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 10012–10022.

Liu, Z., Zhao, C., Fedorov, I., Soran, B., Choudhary, D., Krishnamoorthi, R., et al., 2025c. Spinquant: Llm quantization with learned rotations. *arXiv preprint arXiv:2405.16406*.

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Lyu, N., Wen, J., Duan, Z., Wu, C., 2020. Vehicle trajectory prediction and cut-in collision warning model in a connected vehicle environment. *IEEE Transactions on Intelligent Transportation Systems*, 23, 966–981.

McLaughlin, S.B., Hankey, J.M., Dingus, T.A., 2008. A method for evaluating collision avoidance systems using naturalistic driving data. *Accident Analysis & Prevention*, 40, 8–16.

Munir, F., Azam, S., Mihaylova, T., Kyrki, V., Kucner, T.P., 2025. Pedestrian vision language model for intentions prediction. *IEEE Open Journal of Intelligent Transportation Systems*.

Shao, Y., Cai, H., Long, X., Lang, W., Wang, Z., Wu, H., et al., 2024. Accidentblip2: Accident detection with multi-view motionblip2. *CoRR*.

Shi, L., Guo, F., 2024. Two-stream video-based deep learning model for crashes and near-crashes. *Transportation Research Part C: Emerging Technologies*, 166, 104794.

Shi, L., Jiang, B., Zeng, T., Guo, F., 2025. ScVLM: Enhancing Vision-Language Model for Safety-Critical Event Understanding, in: *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 1061–1071.

Shwartz-Ziv, R., Tishby, N., 2017. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*.

Sun, S., Ren, W., Li, J., Wang, R., Cao, X., 2024. Logit standardization in knowledge distillation, In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15731–15740.

Tian, H., Reddy, K., Feng, Y., Quddus, M., Demiris, Y., Angeloudis, P., 2026. Large (Vision) Language Models for Autonomous Vehicles: Current Trends and Future Directions. *IEEE Transactions on Intelligent Transportation Systems*, 27, 187–210.

Vogel, K., 2003. A comparison of headway and time to collision as safety indicators. *Accident Analysis & Prevention*, 35, 427–433.

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., et al., 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wang, J., Zhang, W., Fu, T., Shangguan, Q., 2026a. DRPVLM: A generative multimodal large language model for real-time driving risk prediction. *Accident Analysis & Prevention*, 231, 108505.

Wang, P., Wu, G., Zhao, Y., Lin, Y., Yang, H.F., 2026b. LLM agents can partially mimic human driving behaviors and decision values but better align with aggressive profiles. *Transportation Research Part C: Emerging Technologies*, 186, 105606.

Wang, T., Chen, K., Chen, G., Li, B., Li, Z., Liu, Z., Jiang, C., 2023. GSC: A Graph and Spatio-Temporal Continuity Based Framework for Accident Anticipation. *IEEE Transactions on Intelligent Vehicles* 9, 2249–2261.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., et al., 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35, 24824-24837.

World Health Organisation, 2023. Global status report on road safety 2023 [WWW Document]. URL https://www.who.int/teams/social-determinants-of-health/safety-and-mobility/global-status-report-on-road-safety-2023 (accessed 7.15.25).

Wu, G., Su, B., Zhao, Y., Wang, P., Lin, Y., Yang, H. F., 2025. Towards physics-informed spatial intelligence with human priors: An autonomous driving pilot study. *Advances in Neural Information Processing Systems*.

Wu, X., Yan, L., Li, H., Su, C., 2024. Forward collision warning system using multi-modal trajectory prediction of the intelligent vehicle. *Proceedings of the Institution of Mechanical Engineers, Part D:*

Journal of Automobile Engineering, 238(2-3), 358-373.

Yang, C., Zhu, Y., Lu, W., Wang, Y., Chen, Q., Gao, C., et al., 2024. Survey on knowledge distillation for large language models: methods, evaluation, and application. *ACM Transactions on Intelligent Systems and Technology*, 16(6), 1-27.

Yu, R., Ai, H., 2022. Vehicle forward collision warning based upon low frequency video data: A hybrid deep learning modeling approach, in: *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 59–64.

Yu, R., Ai, H., Gao, Z., 2020. Identifying high risk driving scenarios utilizing a CNN-LSTM analysis approach, in: *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1–6.

Yu, R., Zhang, R., Ai, H., Wang, L., Zou, Z., 2022. Personalized driving assistance algorithms: Case study of federated learning based forward collision warning. *Accident Analysis & Prevention*, 168, 106609.

Zhang, J., Xie, C., Cai, H., Shen, W., Yang, R., 2024a. Knowledge distillation-based spatio-temporal mlp model for real-time traffic flow prediction. *IEEE Transactions on Intelligent Transportation Systems*, 25, 18122–18135.

Zhang, R., Wang, B., Zhang, J., Bian, Z., Feng, C., Ozbay, K., 2025. When language and vision meet road safety: leveraging multimodal large language models for video-based traffic accident analysis. *Accident Analysis & Prevention*, 219, 108077.

Zhang, Z., Meyer, G.P., Lu, Z., Shrivastava, A., Ravichandran, A., Wolff, E.M., 2024b. Vlm-kd: Knowledge distillation from vlm for long-tail visual recognition. *arXiv preprint arXiv:2408.16930*.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning deep features for discriminative localization, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2921–2929.

Zhou, Y., Bai, L., Cai, S., Deng, B., Xu, X., Shen, H. T., 2025. TAU-106k: A new dataset for comprehensive understanding of traffic accident. In *The Thirteenth International Conference on Learning Representations*.

Zhu, M., Wang, X., Hu, J., 2020. Impact on car following behavior of a forward collision warning system with headway monitoring. *Transportation research part C: emerging technologies*, 111, 226–244.

Author biography

Ruici Zhang received the B.Eng. and M.Eng. degrees in Transportation Engineering from Tongji University, China, in 2021 and 2024, respectively. He is currently a Ph.D. student with the Department of Civil and Environmental Engineering, Imperial College London, U.K. His research interests include traffic safety and AI.

Yuxiang Feng received the M.Sc. degree in mechatronics and the Ph.D. degree in automotive engineering from the University of Bath. He is a Research Fellow and Lab Manager in Department of Civil and Environmental Engineering, Imperial College London. His main research interests include environment perception, sensor fusion, and artificial intelligence for robotics and autonomous vehicles.

Jose Escribano received the Ph.D. degree in March 2021. He is an Assistant Professor with the Centre for Transport Engineering and Modelling, Imperial College London. His research focuses on collaborative vehicle control, optimisation of last-mile logistics, urban air mobility, and machine learning and game theoretical models.

Mohammed Quddus received the B.Sc. degree in civil engineering from Bangladesh University of Engineering and Technology in 1998, the master's degree in transportation engineering from the National University of Singapore in 2001, and the Ph.D. degree from Imperial College London in 2006. He is currently a Chair Professor with the Centre for Transport Engineering and Modelling, Imperial College London. His research interests include connected and autonomous vehicles, AI, and statistical modelling.