scientific reports

OPEN

Improving intelligent perception and decision optimization of pedestrian crossing scenarios in autonomous driving environments through large visual language models

Xiao Teng[1,2], Lin Huang[3], Zhenjiang Shen[3,4]✉ & Wankai Li[1]

This study leverages large Visual Language Models (VLM) to develop an intelligent pedestrian crossing scenario system within autonomous driving environments. By establishing standardized checklists and prompts, the system minimizes the risks of misjudgment and omission through multimodal data processing. It offers data-driven decision-making support, presenting an innovative approach to integrating autonomous driving technology with intelligent transportation systems. The study begins by classifying pedestrian crossing scenarios based on international autonomous driving standards, distinguishing between pedestrian crossings and autonomous vehicle crossings, as well as dynamic and static entities. Next, standardized prompts derived from these standards are fed into the VLM, generating structured scenario checklists of dynamic and static entities, outputted in JSON format. This systematic identification and processing of entities—such as pedestrians, vehicles, and traffic facilities—enables the construction of structured data representations for complex traffic scenarios. Building on this foundation, the VLM analyzes scenario data to predict collision risks by modeling the behaviors of both pedestrians and vehicles, supporting real-time decision-making for autonomous vehicles and road users. Furthermore, the VLM processes scene data to anticipate potential conflicts and provide actionable safety recommendations, enhancing the overall security of all traffic participants. The system achieved a perception accuracy of 93.05%, with risk prediction consistency and decision-making rule consistency rates of 85.91% and 87.72% respectively. By constructing a VLM-based intelligent pedestrian crossing perception system, this study offers a novel technical framework for improving perception, prediction, and decision-making in autonomous driving. Unlike traditional rule-based and deep learning approaches, which struggle with complex pedestrian behaviors and dynamic environments, our method integrates visual perception with reasoning capabilities, enabling structured, standardized, and explainable decision-making in pedestrian crossing scenarios.

Keywords Visual language models, Scenario classification, Standardized checklists, Prompt words, Collision risk prediction, Action decision-making

Pedestrian crossings are critical facilities for ensuring pedestrian safety, playing a key role in improving traffic efficiency, reducing accident risks, and protecting the lives of pedestrians[1,2]. With the rapid advancement of autonomous driving technology, precise perception of pedestrian crossings and their surrounding environments has become especially important[3,4]. This is not only crucial for the safe operation of autonomous vehicles but also directly affects the safety of pedestrians crossing the road[5]. However, due to the complexity of traffic

[1]Faculty of Transdisciplinary Sciences, Institute of Philosophy in Interdisciplinary Sciences, Kanazawa University, Kanazawa 920-1192, Japan. [2]China Youke Communication Technology Co.,Ltd, Fujian, China. [3]Graduate School of Natural Science & technology, Kanazawa University, Kanazawa 920-1192, Japan. [4]International Joint Laboratory of Spatial Planning and Sustainable Development (FZUKU-LAB SPSD), Fuzhou University, Fuzhou 350025, China. ✉email: shenzhe@se.kanazawa-u.ac.jp

environments, diverse observation perspectives, and the multiplicity of behavioral entities, traditional data-driven and rule-based approaches face numerous challenges when perceiving complex scenarios. A lack of accurate perception can result in pedestrian injuries or fatalities, legal disputes, public trust crises in autonomous driving technology, and decreased system performance[6,7]. Therefore, enhancing the perception capabilities of autonomous driving systems is essential for ensuring their safety and efficiency. In this context, the establishment and implementation of standardized output checklists[8] are also necessary. These checklists not only help the system achieve consistency and predictability in perception[9,10], prediction[11], and decision-making[12] but also improve system reliability, providing clear foundations for fault diagnosis and problem resolution[8]. To address these challenges, large visual language models (VLM) offer an innovative solution[13,14]. By integrating the reasoning, common sense, and generalization capabilities of large language models (LLM) with visual perception, VLM breaks the boundaries between text and vision[15]. This enables systems to better perceive and understand complex pedestrian crossing scenarios, as well as to possess powerful reasoning, comprehension, and zero-shot learning capabilities. This innovation presents a new approach to overcoming the perception difficulties in autonomous driving and is expected to promote the further development of autonomous driving technology while enhancing the safety and reliability of intelligent transportation systems[16].

From the research area, in current autonomous driving research, most studies collect data from the perspective of the autonomous vehicle[17–20], or by integrating data from various vehicle-mounted sensors to improve positioning accuracy[21,22]. however, roadside traffic equipment is also critically important[23,24]. Alternatively, these studies tend to focus solely on the condition of the roads, rarely dedicating attention to the specific scenario of pedestrian crossings. However, pedestrian crossings are the locations where autonomous vehicles and pedestrians interact most frequently, and the traffic situations are complex[25–27]. Conducting research by acquiring data from the perspective of this scenario holds significant value.

The recognition and decision-making problem of pedestrian crossing scenes is one of the important research contents in the field of autonomous driving. Previous research results can be divided into two categories: one is the rule-based decision-making method, which relies on pre-set traffic rules to respond to pedestrian crossing behavior[28,29]. The system automatically takes corresponding decision measures when the pedestrian meets specific trigger conditions. This type of method has clear logic and reliable decision-making. Still, it lacks full consideration of the complexity of the real environment, especially in situations where pedestrian behavior is highly random and uncertain, it is easy to be slow to react or make mistakes. In addition, this type of research usually simplifies the scene into a binary classification problem, and fails to fully consider the various subtle states and complex behaviors of pedestrians crossing the road. The other type of research is dominated by deep learning methods of visual perception, among which the application of convolutional neural networks (CNNs) has greatly promoted the recognition, tracking and spatial positioning capabilities of autonomous driving systems for targets in traffic scenes[30,31]. This method usually focuses on the accurate detection and continuous tracking of specific targets such as pedestrians, vehicles, and traffic facilities, and predicts their dynamic behaviors[32]. However, existing deep learning methods still have obvious limitations. Most of their model training and applications focus on the recognition of target positions and motion trajectories, and lack the understanding of the deep semantic information behind pedestrian behavior. Although the model can capture the position and movement trajectory of pedestrians, it is difficult to accurately analyze the pedestrian's psychological state, behavioral intentions, and the interaction with other environmental elements, making it difficult to effectively respond to more complex traffic scenarios or abnormal events[33].

Through existing research, we noticed that previous studies have not systematically classified pedestrian crossing scenarios. Existing studies usually regard "pedestrian crossing the street" as a single behavior and do not further divide different scenarios. This simplified treatment of the scene weakens the adaptability of the model when facing complex real-life situations, and cannot provide sufficiently detailed information support for subsequent prediction and decision-making links. More importantly, existing studies have not built an entity recognition list based on visualization targets and autonomous driving standards for prediction and simulation. Although some studies focus on elements such as pedestrians, vehicles, and traffic lights, they are often scattered target recognition tasks and have not formed a unified and standardized scene entity output format. In addition, existing studies generally lack in-depth integration with autonomous driving safety standards, and do not convert perception results into structured data tables or standardized output formats, which makes it difficult for the system to directly call these perception results for reasoning in subsequent decision-making links.

Therefore, this study proposes three key innovations on this basis: First, based on the existing national standard classification documents, we established a scene classification framework for pedestrian crossings, refined the various scenarios of pedestrian crossing behavior, and further combined weather conditions and traffic signal status for detailed research. This classification method can provide richer input for VLM, which helps the system to accurately infer the behavioral intentions of pedestrians and vehicles, thereby achieving more reasonable decisions. Second, based on the scene classification and standard documents, we constructed a scene entity recognition list based on VLM and autonomous driving standards. Through custom prompts, the system can identify key entities such as pedestrians, vehicles, traffic signs, road conditions, etc. in the scene, and output them in a structured JSON format. This standardized approach not only makes the perception results more operational, but also provides clear input data for the prediction and decision-making links, so that the entire process from perception to decision-making forms a coherent information chain. At the same time, based on the visualization target, the visualization content is perceived-predicted-decision-made. Third, this study focuses on pedestrian crossings as the research area, narrowing the research scope, which helps to explore the detailed characteristics of pedestrian crossing behavior more thoroughly and improves the specificity and accuracy of scene classification and entity recognition. Additionally, this focused approach reduces environmental interference, enhances data quality, and provides more precise support for predicting and making decisions about pedestrian crossing behavior. At the same time, this research framework is scalable, laying a solid foundation for future expansion into more complex traffic environments. Unlike traditional rule-based or deep learning approaches, our method leverages the reasoning and zero-shot learning capabilities of VLM to achieve structured perception and robust decision-making in pedestrian crossing scenarios. Meanwhile, this study uniquely adopts a third-party perspective—distinct from those of pedestrians and vehicles—by utilizing roadside camera data. In the context of autonomous driving, this approach complements traditional vehicle-centric research and enables a more comprehensive perception of the environment, particularly in pedestrian crossing scenarios that are often overlooked in previous studies.

The remainder of this paper is organized as follows: Section 2 introduces the research methodology. Section 3 focuses on scenario construction and standardized decomposition for pedestrian crossings, including functional classification and structured representation of crossing environments. Section 4 presents the perception, prediction, and decision-making processes based on VLM, covering scenario description, risk prediction, and autonomous decision-making. Section 5 reports the quantitative experimental results and provides analysis. Finally, Section 6 concludes the study and highlights future research directions.

Methodology

This paper leverages the performance of VLM models in complex tasks such as scenario understanding and causal reasoning, focusing on the study of pedestrian crossings in autonomous driving scenarios, and providing roadside support for autonomous driving environments. In this study, we employed GPT-4o (with Vision) developed by OpenAI as the core VLM. The model accepts visual inputs and prompts, and produces structured JSON outputs as well as natural language scene descriptions. Its multimodal capabilities are leveraged through API-based access under the ChatGPT-4o framework. By inputting both visual scenario images of pedestrian crossings and designed prompt information from a third-party perspective, independent of pedestrians and vehicles, the VLM generates a standardized JSON-format checklist[34]. This allows the VLM to perform perception, prediction, and decision-making within the traffic scenarios of pedestrian crossings in autonomous driving environments. Additionally, to comprehensively set up the pedestrian crossing scenarios and improve the interpretability of the results, ensuring that these scenarios apply to autonomous driving environments, the datasets used in this study were all collected through optical sensor simulations from the open-source autonomous driving simulator CARLA (version 0.9.15)[35], using the simulation map Town10HD_Opt. This paper presents application scenarios of VLM in pedestrian crossings within autonomous driving environments, aiming to support the development of pedestrian crossings in the context of autonomous driving.

Taking the visual scenario of the pedestrian crossing as input information, and the data is obtained by optical sensors in the Carla simulator. In actual scenarios, this data usually comes from cameras installed at the pedestrian crossing. After obtaining the scenario information, the prompt words are input at the same time to make the VLM output content. In addition, in order to make the content output by VLM more targeted, this paper imposes certain restrictions on the prompt words. In order to standardize the output, a standardized checklist example is established to convert the content into a JSON format checklist. Through these operations, a smart scenario system is built with the help of VLM to perceive, predict and make decisions on the pedestrian crossing scenario. Perception capabilities include identifying and understanding key elements in the environment, such as pedestrians, traffic lights, road signs and other vehicles. Through this information, a comprehensive understanding of the current traffic situation is formed. Prediction capabilities involve inferring the behavior of autonomous vehicles and pedestrians in the scenario to predict the degree of collision risk between pedestrians and autonomous vehicles. Decision-making capabilities perform corresponding operations based on the prediction results. When it is predicted that a pedestrian is about to enter the pedestrian crossing, it is necessary to determine whether to slow down or stop to avoid a collision. The aim is to generate a standardized checklist of pedestrian crossings in an autonomous driving environment through VLM and realize smart scenario construction, improve traffic safety levels and reduce the incidence of traffic accidents. In many real-world scenarios where real-time communication is limited, components of the VLM model can be deployed on edge computing units integrated with roadside cameras to handle local perception and risk prediction. Critical safety information can then be synchronized with vehicles through low-frequency communication. This setup helps to mitigate the challenge of limited infrastructure-to-vehicle communication in real-world environments.

The research approach of this paper is illustrated in Fig. 1. Firstly, this study constructs a smart pedestrian crossing scenario for autonomous driving within the Carla simulator. Based on the differences between the two main entities—autonomous vehicles and pedestrians—the scenario is categorized into pedestrian-crossing and vehicle-crossing scenarios, each of which is analyzed. In conjunction with autonomous driving road, vehicle, and safety standards listed in Table 1[36–38], the visual elements within these scenarios are examined, encompassing traffic entities such as vehicles and pedestrians, pets accompanying pedestrians, and other scenario entities like pedestrian crossings, traffic lights, basic environmental conditions, and additional factors. Secondly, using these scenario elements, the study designs and defines prompt terms and example output formats for each element, directing the system to output the perception content in JSON format. Then, with the prompt terms and example output texts as a basis, visual scenario data of pedestrian crossings captured by optical sensors within the Carla simulator is used as input for the VLM. This data undergoes perception processing to generate a standardized output checklist of attributes for each scenario entity in JSON format. Finally, based on the output content, the smart pedestrian crossing scenario within the VLM framework predicts collision risks and makes decisions, forecasting the probability of collisions between traffic entities. The decision-making process is driven by these predictions, providing each entity with action decisions that are executed to avoid or mitigate collision risks.

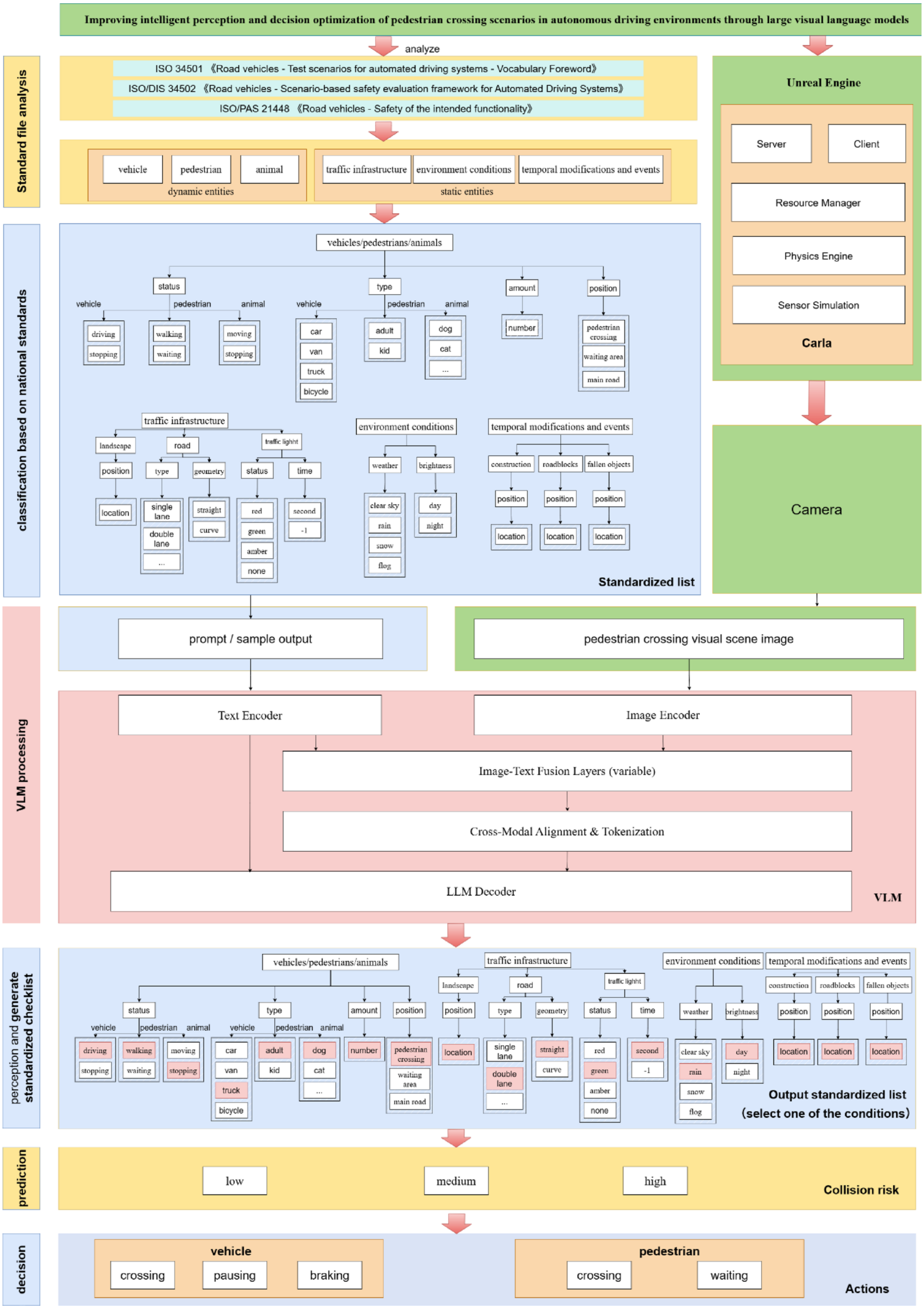


Fig. 1. Research approach.

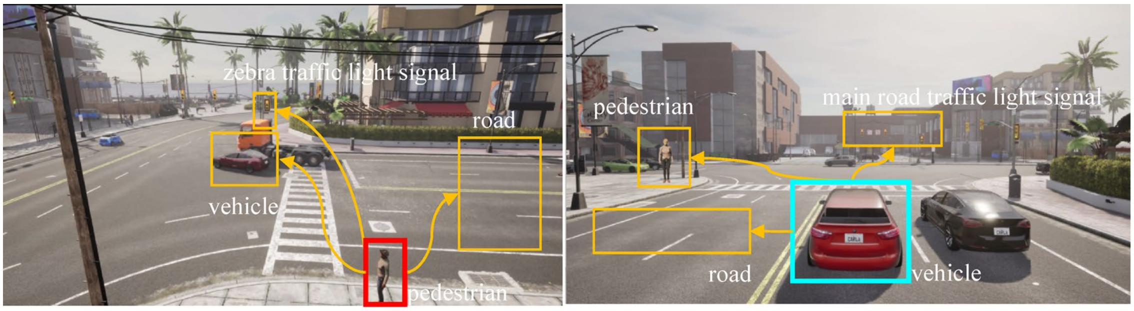
Scenario construction and standardized decomposition for pedestrian crossings

Scenario construction and functional classification for pedestrian crossings

This study constructs a smart pedestrian crossing scenario for autonomous driving using the Carla simulator, where various vehicles and pedestrians are added to the simulated Carla environment. Vehicles are set to operate

Standard	Elements
ISO 34,501 《Road vehicles - Test scenarios for automated driving systems - Vocabulary Foreword》	Scenario, static entities, dynamic entities, vehicle, pedestrian, road model, functional scenario, main road traffic light signal, zebra traffic light signal, (amber, red, green), perception, temporal modifications and events, fallen objects, temporal installations, animals, environmental conditions, road structure, road shape, functional concept, weather conditions, (clear sky, rain, fog, snow), shape of road/lane, (straight, curve), action, event
ISO/DIS 34,502 《Road vehicles - Scenario-based safety evaluation framework for Automated Driving Systems》	
ISO/PAS 21,448 《Road vehicles - Safety of the intended functionality》	

Table 1. Elements from autonomous driving road, vehicle, and safety standards.



(a)Pedestrian Crossing at a Pedestrian Crossing Scenario (b)Autonomous Vehicle Crossing at a Pedestrian Crossing Scenario

Fig. 2. Pedestrian crossing functional scenario classification.

in autonomous driving mode, creating a simulated scenario featuring autonomous vehicles and pedestrians. In an autonomous driving environment, ensuring that pedestrians and vehicles safely and efficiently navigate pedestrian crossings is a crucial goal for implementing an intelligent transportation system. In this study, functional scenarios are classified based on the primary dynamic entities involved—pedestrians and autonomous vehicles—aligned with ISO 34,501 and ISO/DIS 34,502 standards. Given the differing characteristics and impacts of autonomous vehicles and pedestrians as they cross the pedestrian crossing, this paper classifies the functional scenarios into two types: pedestrian-centered scenarios, which include pedestrians as the primary subject along with other pedestrian crossing elements, and vehicle-centered scenarios, where autonomous vehicles serve as the primary subject alongside other pedestrian crossing elements. This classification aligns with the ISO 34,501 and ISO/DIS 34,502 standards, emphasizing functional safety and situational awareness in autonomous driving systems within complex environments. By clearly distinguishing between pedestrian-centered and vehicle-centered scenarios, it helps design precise test cases to validate the system's responsiveness in various situations. The core purpose is to realistically simulate real-world traffic scenarios, enabling system to make optimal decisions based on different dynamic entities, thereby enhancing overall traffic safety and efficiency.

Figure 2a illustrates the scenario of pedestrians crossing a pedestrian crossing, with pedestrians as the primary element. In contrast, Fig. 2b presents the scenario of an autonomous vehicle crossing a pedestrian crossing, where the autonomous vehicle is the primary element. Since these two scenarios have different primary elements, the associated reference scenario elements also vary. In the pedestrian-crossing scenario, the system needs to focus on pedestrian intentions and account for elements affecting pedestrian crossing, such as zebra traffic light signals, vehicles on both sides of the road, and surrounding environmental factors. Conversely, in the autonomous vehicle-crossing scenario, the system must consider main road traffic light signal, pedestrians in the waiting zones on either side of the pedestrian crossing, and other surrounding environmental factors.

Classifying scenarios based on primary elements (pedestrians and autonomous vehicles) helps the system more accurately identify key elements and decision-making requirements specific to each scenario, simplifying the perception and decision-making processes. This approach enhances the system's adaptability and safety in complex traffic environments. This classification strategy provides a theoretical basis and practical guidance for autonomous driving systems' efficient and reliable operation within dynamic traffic settings.

Standardized decomposition of pedestrian crossing environment scenarios

In a pedestrian crossing environment, environmental perception is crucial for both autonomous vehicles and pedestrians, as understanding the scenario is a fundamental prerequisite for achieving safe and efficient autonomous driving. This section analyzes the visual scenarios of pedestrian crossings, utilizing VLM to achieve perception within these scenarios and generate standardized JSON-format output checklists. This serves as a reference for the future development of intelligent transportation systems, the JSON-structured output list designed in this study accounts for uncertainties in future Vehicle-to-Infrastructure (V2I) communication. By adopting a unified intermediate data format, it facilitates efficient compression and reliable transmission of information under bandwidth limitations or unstable network conditions. The classification standards adopted in this study are based on international standard documents, formulated to achieve the industry's development goals.

The structured output of JSON relies on the VLM's ability to interpret prompts accurately, while VLM are susceptible to hallucinations, such as misclassifying static and dynamic vehicle states or incorrectly interpreting pedestrian traffic light signals. These errors can propagate and adversely affect collision risk prediction. To address this issue, this paper incorporates content from traffic standard documents and employs specifically designed prompts, integrating them with visual scene information to impose logical constraints on the VLM's predictions. This approach mitigates hallucinations in traffic perception, enhances the accuracy of structured outputs, and improves the reliability of collision risk prediction.

By analyzing and structuring each entity within the constructed scenario, in alignment with standard documentation, a deeper understanding of each element's role is achieved. This facilitates a more targeted evaluation of the interactions among pedestrians, vehicles, and other scenario elements, enhancing the purposefulness of scenario perception. The scenario of factors segmentation beneath the pedestrian crossing is systematically divided as illustrated in Fig. 3.

The diagram primarily includes the following scenario of factors:

Dynamic Entities:

(a) Vehicles: This includes the number, type, status, and position of vehicles in the current scene.
(b) Pedestrians: This includes the number, type, status, and position of pedestrians in the current scene.
(c) Animals: This includes the number, type, status, and position of animals in the current scene.

Static Entities:

(d) Traffic Infrastructure: The traffic light system is an important part of urban traffic management, divided into main road traffic light signals and zebra traffic light signals, used to manage motor vehicle and pedestrian traffic flow, respectively. The attributes of traffic lights include color and the remaining time for the current state.
(e) Environmental Conditions: These include weather and brightness in the pedestrian crossing scene. Weather information consists of conditions such as sunny, rainy, snowy, or foggy, while brightness indicates the time of day (daytime or nighttime). These natural factors significantly impact both autonomous vehicles and roadside equipment and cannot be ignored.
(f) Temporal Modifications and Events: This includes sudden scenarios and factors that may be overlooked, such as temporary installations or fallen objects on the road.

As illustrated in Fig. 3, a pedestrian crossing scenario was created in the Carla simulator. The scenario is set in rainy weather with low visibility and wet road conditions. Traffic participants include autonomous vehicles, non-motorized vehicles, and pedestrians. The scenario also features buildings, palm trees, and traffic infrastructure like traffic lights. Based on the above scenario divisions, the diagram labels the content of each scenario element.

In the current research paradigm, prompts have become the mainstream approach for applying LLM to specific natural language processing tasks. Therefore, based on the above scenario classification, it is necessary to design prompts that align with the scenario analysis. The purpose of prompt design is to guide the expected output in text generation tasks, ensuring clarity, standardization, and thorough explanations to maximize the model's reasoning efficiency and output quality in pedestrian crossing scenarios. In this context, prompts should explicitly and concisely describe the desired output content and provide necessary contextual information to enable the VLM to accurately interpret and process the input.

Perception, prediction and decision making based on VLM

Perception and scenario description of pedestrian crossings using VLM

The prompts in this study specifically require output in a standardized JSON format, strictly adhering to the defined fields and data type specifications for pedestrian crossing scenario entities as provided in the design. This ensures consistency and readability in the output. The use of standardized JSON-format lists also anticipates the challenge of limited communication between infrastructure and vehicles in real-world deployments, as noted in previous studies. To address this, a potential future direction is the development of a vehicle-infrastructure cooperative inference mechanism, in which both vehicle-side and infrastructure-side systems run simplified or lightweight VLM modules. These modules can exchange structured information—such as standardized JSON-format outputs—through predefined communication protocols, thereby enabling collaborative perception and enhancing the system's robustness and flexibility under constrained communication conditions. Table 2 outlines the factors and data types involved in a pedestrian crossing scenario. It provides a comprehensive breakdown of the fields, including field names, data types, and descriptions of scenario entities. The table clearly defines the required format and scope to guarantee high precision and ensure that the output adheres to model-generated requirements. Additionally, to further enhance the accuracy and standardization of the output, the prompt provides an example output, demonstrating how to convert scenario elements into a standardized JSON format checklist.

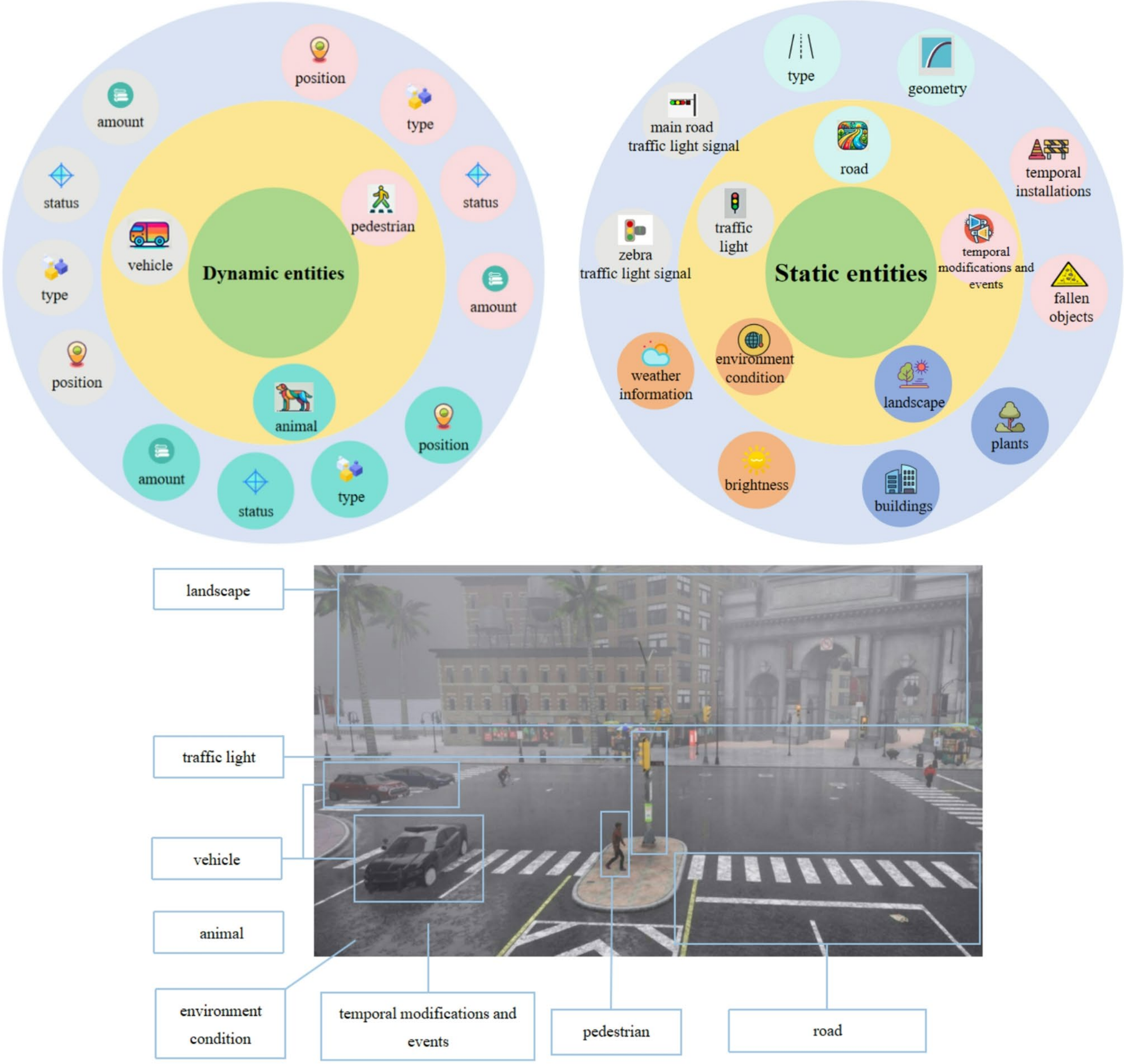


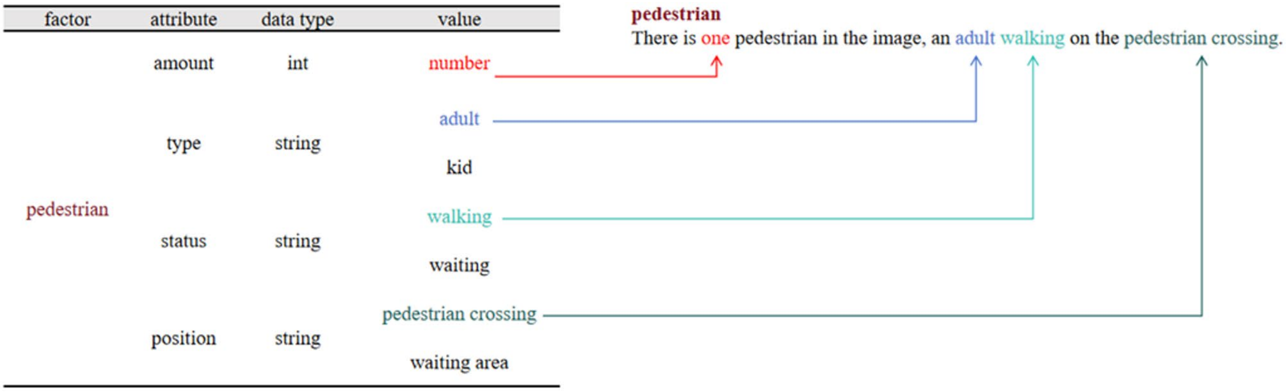
Fig. 3. The scenario of factors segmentation beneath the pedestrian crossing.

The prompt design for perception in this study, shown in Appendix A (Fig. A1), ensures that the output format is precise and consistent, laying a solid data foundation for further research and applications.

Using Fig. 3 as the input scenario, the perception process based on the designed prompts generated the corresponding output results, as shown in Appendix D (Fig. D 1). The output results indicate that the system has systematically detailed various types of information within the scenario, including vehicles, pedestrians, animals, landscapes, roads, traffic signals, environmental conditions, as well as temporary changes and events. By comparing with Fig. 3, in terms of quantity, the system identified 3 cars, 1 bicycle, 2 pedestrians, and 0 animals, demonstrating its accurate recognition of the numbers of cars, pedestrians, and animals in the scenario. For spatial awareness of entities, the system output shows vehicles driving on the main road and pedestrians on the pedestrian crossing. Regarding vehicle status, multiple experiments reveal that the system may misinterpret the dynamic or static state of vehicles, primarily due to specific scenarios involving pedestrian crossings. Consequently, when visual scenarios of vehicles across several consecutive time points are inputted, the system can rectify previous perceptions and output correct results. For pedestrian status, due to the distinctive behavioral features of pedestrians in motion versus at rest, the system almost always accurately identifies whether a pedestrian is moving. In terms of landscape perception, the system can recognize greenery, buildings, and water towers in the scene. For road perception, the system understands the multi-lane configuration and geometric conditions of the road. Regarding traffic signals, with multiple intersections and traffic lights on the scenario, the system identified the main road traffic light as red but failed to recognize the zebra traffic

Factor	Attribute	Data type	Value	Explanation
Vehicle	Position	String	Pedestrian crossing	Vehicle position in the image, pedestrian crossing, waiting area
			Waiting area	
			Main road	
	Status	String	Driving	Vehicle status in the image, driving, stopping
			Stopping	
	Type	String	Truck	the vehicle type in the image, car, van, truck, …
			Bus,	
			Bike	
			…	
	Amount	int	Number	Total number of vehicles in the image, including non-motor vehicles
Pedestrian	Type	string	Adult	Types of pedestrians in the image, adult, kid
			Kid	
	Status	string	Walking	Pedestrian status in the image, walking, waiting
			Waiting	
	Position	string	Pedestrian crossing	Pedestrian position in the image
			Waiting area	
	Amount	int	Number	total number of pedestrians in the image
Animal	Type	string	Cat	Name of the animal in the image, cat, dog, …
			Dog	
			…	
	Status	string	Moving	Animal status in the image, moving, stopping
			Stopping	
	Position	string	Position	The location of the animals in the image
	Amount	int	Number	Total number of animals in the image
Landscape	facility name	string	Plant	The names of facilities in the scene, green plants, buildings, etc.
			Building	
			…	
	Position	string	Position	The location of the facilities in the image
Road	Type	string	Single lane	The lanes of the road
			Multiple lanes	
	Geometry	string	Straight	The geometry of the road, straight, curve
			Curve	
Traffic light	main road traffic light signal	string	Red	manages the right of way for vehicles on main roads
			Green	
			Amber	
	zebra traffic light signal	string	Red	Mainly for pedestrians, but also controls approaching vehicles, requiring them to stop when the pedestrian signal is green.
			Green	
			Amber	
	Time	int	Number	The remaining time displayed by the traffic light, if there is no perception, it will display – 1
	None	string	None	If not perceived, then output NONE.
Environment condition	Weather	string	Clear sky	Weather conditions in images, clear sky, rain, snow, flog
			Rain	
			Snow	
			Flog	
	brightness	string	Day	Light intensity in the image, day, night
			Night	
Temporal modifications and events	name	string	Construction	Temporal modifications and events' name
			Roadblocks	
			Fallen objects	
	position	string	Position	The location of the construction observed in the image. If there is none, output NONE

Table 2. The factors and data types of pedestrian crossing scenario.



factor	attribute	data type	value
pedestrian	amount	int	number
	type	string	adult kid
	status	string	walking waiting
	position	string	pedestrian crossing waiting area

pedestrian

There is one pedestrian in the image, an adult walking on the pedestrian crossing.

Fig. 4. Attributes of dynamic entities and their natural language descriptions.

Functional scenario: Pedestrian Crossing at a Pedestrian Crossing Scenario
Dynamic entities
Pedestrian One adult pedestrian is waiting on the sidewalk.
Vehicle There are two vehicles, a truck and a car, stopped at a pedestrian crossing, and another car driving on the main road. **Animal** There are no animals in the scene.
Static entities
Landscape There are palm trees lining the side of the road and buildings in the background. **Road** The road is a straight, multi-lane avenue. **Zebra traffic light signal** The traffic light for the pedestrian crossing is red. **Environment condition** The weather is clear and it is daytime. **Temporal modifications and events** There are no construction sites, roadblocks, or fallen objects in the scene.

Table 3. The attributes of these relevant entity elements.

light signal. This ambiguity about the scenario's traffic lights indicates that input data must correspond traffic lights to intersections individually, avoiding unclear situations. Environmental conditions were recognized as daytime with rainy weather. For construction activities, roadblocks, or fallen objects not present in the scenario, the system outputs results based on actual conditions. These results demonstrate that, through well-designed prompts, the system can accurately analyze complex scenarios, providing fine-grained descriptions of various types of objects and environmental factors.

In pedestrian crossing scenario perception, the standardized output checklist provides essential reference data for analyzing the risk levels of different functional scenarios. As this study primarily explores the perception capabilities of large models in visual scenarios, these functional scenarios are conceptually described in natural language, excluding specific physical quantities. Based on the conditions of the standard list and the scene information obtained by the camera, this paper generates or selects (chooses one) the correct state that matches the current state from the manual standard list through VLM to describe the scene. Figure 4 illustrates the relationship between the attributes of the dynamic entity, the pedestrian, and corresponding descriptive terms in natural language. By outputting sentences that naturally describe these attributes through the large model, a scenario description is formed. In the prompt design of Sect. 3.1, additional descriptive attributes are included to allow the model to generate natural language descriptions based on specific attributes.

Taking the pedestrian crossing scenario depicted in Fig. 2a as an example, the perceived entity elements, following the perception process, are structured into a natural language description. The attributes of these relevant entity elements are detailed in Table 3.

The standardized checklist represents the state attributes of various entities on the pedestrian crossing, and translating these attribute values into natural language descriptions better aligns with the input requirements of LLM. The structured data in the standardized checklist ensures information completeness and consistency, while natural language descriptions enhance the model's effectiveness in understanding, reasoning, and generating language. Combining these approaches facilitates efficient data representation and model interaction: the standardized checklist delivers precise and systematic semantic information, minimizing redundancy and potential misinterpretation, whereas natural language descriptions present information in a more human-readable format, improving the quality of language generation by the LLM when handling complex scenarios.

This combination ensures both scientific and objective data handling and enhances the flexibility and adaptability of language models in application scenarios. Through the use of both the standardized checklist and natural language descriptions, this study integrates VLM with multimodal information (such as traffic signs and signals) to infer high-level semantics and interpret traffic regulations. This approach aids in predicting collision risks in complex traffic scenarios and supports decision-making for autonomous vehicles and pedestrians.

Prediction and decision-making process based on perception results

The process of the intelligent pedestrian crossing scenario system in this study is illustrated in Fig. 5. Unlike conventional models that rely purely on feature extraction, the VLM not only performs object detection but also leverages semantic associations to carry out causal reasoning. For example, when a pedestrian steps into the crosswalk while the traffic light is red, the VLM integrates visual and contextual prompt information to infer the likely violation behavior and the associated collision risk. The entire workflow is based on the VLM to perceive, interpret, and process information within the scenario, predict outcomes, and ultimately make decisions. First, an image of the pedestrian crossing in the autonomous driving environment is input into the system, along with *Prompts_1* for scenario perception, producing the perception results. Next, the perception results are used as prior information and input with *Prompts_2* for scenario prediction, yielding the prediction results. Finally, these results are combined with *Prompts_3* for scenario decision-making to make optimal decisions about each entity, to avoid or reduce collision risks. Meanwhile, The factors attributes and data types for scenario prediction and decision-making is shown in Table 4.

This process leverages visual scenario information and precise prompts, allowing the VLM to progressively reason through perception, prediction, and decision-making in the scenario, ultimately achieving comprehensive understanding and making rational decisions for the scenario.

Collision risk prediction through VLM scenario analysis

The pedestrian crossing is a crucial pathway for pedestrians, where the likelihood of interactions between pedestrians and vehicles is higher than on other road segments, necessitating increased attention. When there is a potential collision risk between pedestrians and autonomous vehicles, pedestrians and drivers rely on their experience and common sense to make accurate predictions and decisions to avoid accidents. However, in an autonomous driving environment, vehicles must make predictions and decisions autonomously through driving algorithms. Traditional data-driven methods often lack the experience and common sense necessary for nuanced judgment, limiting their prediction accuracy and reliability.

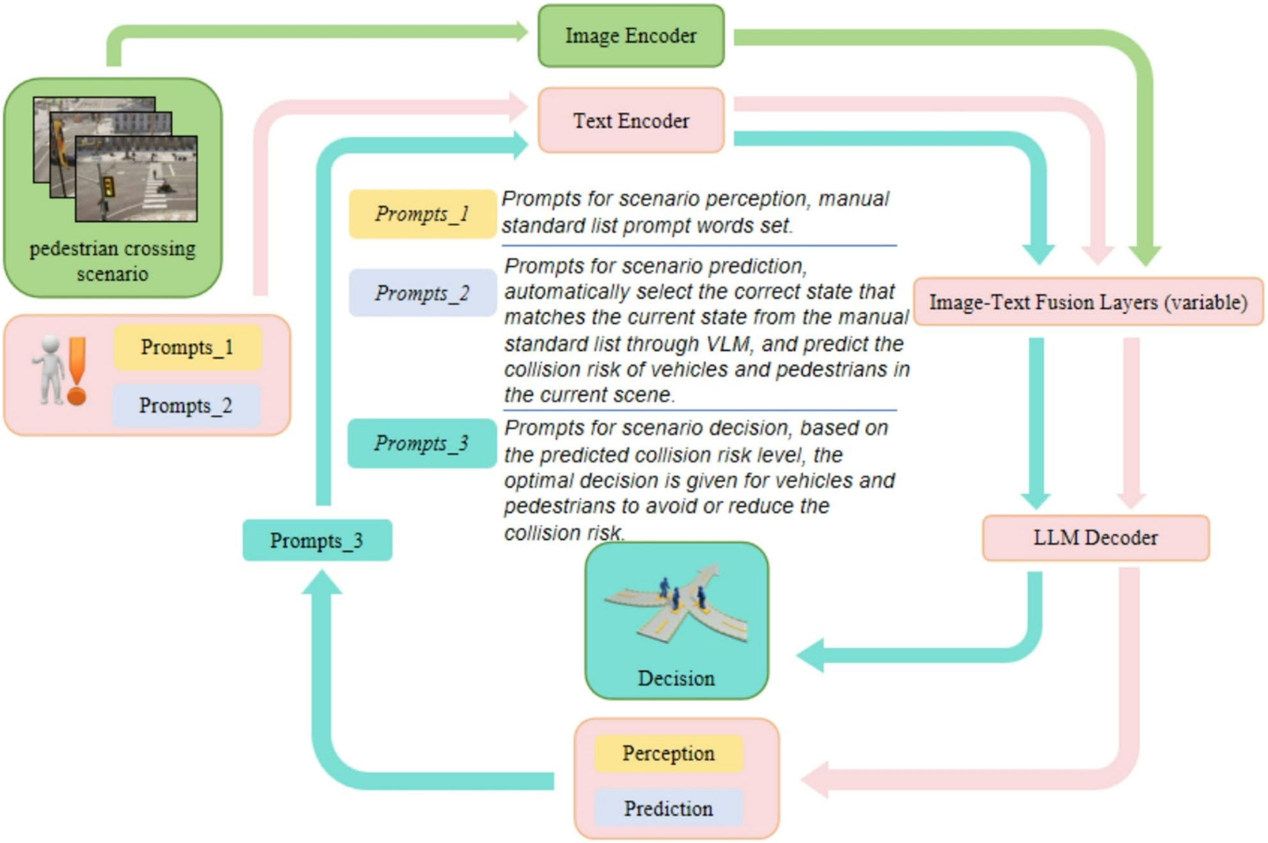


Image Encoder
Text Encoder
pedestrian crossing scenario
Prompts_1
Prompts_2
Prompts_3
Prompts for scenario perception, manual standard list prompt words set.
Prompts for scenario prediction, automatically select the correct state that matches the current state from the manual standard list through VLM, and predict the collision risk of vehicles and pedestrians in the current scene.
Prompts for scenario decision, based on the predicted collision risk level, the optimal decision is given for vehicles and pedestrians to avoid or reduce the collision risk.
Image-Text Fusion Layers (variable)
LLM Decoder
Decision
Perception
Prediction

Fig. 5. Structured decision pipeline for pedestrian crossings in system.

Factor	Atribute	Data type	Value	Explanation
Vehicle	id	int	Number	The id of the vehicles
	type	string	Truck	The type of the vehicle
			Bus	
			Bike	
			…	
	prediction	string	Low	Predict the possible actions of the vehicle and whether a collision will occur
			Medium	
			High	
	reason	string	Text	Provide the reason for the prediction
	decision	string	Crossing	Decide on the autonomous vehicle's next action
			Pausing	
			Braking	
	reason	string	Text	The reason for the decision
Pedestrian	id	int	Number	The id of the pedestrian
	type	string	Adult	Types of pedestrians in the image, adult, kid
			Kid	
	prediction	string	Low	Predict whether there is a risk of collision under the current state
			Medium	
			High	
	reason	string	Text	Provide the reason for the prediction
	decision	string	Crossing	Decide on the pedestrian's next action
			Waiting	
	reason	string	Text	The reason for the decision

Table 4. Factors attributes and data types for scenario prediction and decision-making.

To explore enhanced decision-making capabilities in complex traffic environments, this study leverages the VLM's capacity for common sense reasoning and zero-shot recognition to predict potential collisions between vehicles and pedestrians. Based on scenario perception and classification, this research employs large-scale models to develop an intelligent scenario system that predicts the next actions of pedestrians and vehicles, as well as potential collision risks. Through these predictions, the system provides decision-making guidance to both pedestrians and vehicles, enabling optimal choices in complex traffic environments to avoid dangerous situations or, when unavoidable, to minimize potential losses. This approach aims to enhance traffic safety and efficiency.

Table 4 outlines the factors attributes and data types for scenario prediction metrics based on functional scenario classifications. The primary focus is on predicting the actions of autonomous vehicles and pedestrians. The ***id*** and ***type*** attributes of vehicles and pedestrians correspond directly to those in Table 2 to uniquely identify each vehicle and pedestrian. The ***prediction*** attribute forecasts collision risk, outputting prediction results. The VLM links perceived scenario elements with natural language descriptions to establish a unified semantic understanding, mapping input information to risk levels ("low," "medium," or "high") through predefined risk classification rules and a domain-specific fine-tuning model, while providing predictive feedback. In order to realize the intelligent identification of the collision risk of traffic participants, this paper combines the VLM with the risk level mapping method of structured perception information. Specifically, we directly embed the structured JSON data output by the perception module (including the spatial position, motion state of pedestrians and vehicles, and environmental information such as signal light color and weather conditions) into Prompt, and perform semantic parsing and risk reasoning on traffic scenes based on the advantages of VLM understanding and common sense reasoning.

To improve the consistency and professionalism of the prediction, we introduced a set of explicit risk classification rules based on the summary of urban traffic regulations and behavior patterns in Prompt to fine-tune the model. These rules cover high-risk behaviors (such as illegal crossing, limited sight distance, sudden movement, etc.) and low-risk situations (such as orderly passage under signal control, etc.), as knowledge support for the reasoning process of embedding domain knowledge into the model. Different from the traditional symbolic logic form, these rules are integrated into Prompt through few-shot examples, thereby retaining the flexible semantic generalization ability of the large language model and realizing the standardized guidance of typical traffic scenarios.

In the prompt design, the task goal is explicitly set as: requiring the model to determine the collision risk level (low, medium, high) for each traffic participant and give the corresponding judgment basis. When generating the prediction results, the model will comprehensively consider the spatial relationship, dynamic interaction and rule compliance between traffic elements to make an interpretable risk assessment. For example, when a pedestrian is crossing the lane and there is a high-speed approaching vehicle nearby, the model tends to judge it as a high risk and point out triggering factors such as "crossing behavior" and "the distance of the oncoming vehicle is too close"; while for pedestrians in a safe waiting area, the environment is stable, and the signal indication is clear, it can be assessed as a low risk level.

The advantage of this method is that, on the one hand, it utilizes the powerful language understanding and reasoning ability of the large language model, can process multimodal input and output structured risk judgments; on the other hand, through guided prompts and domain knowledge injection, the consistency and interpretability of the model output are significantly enhanced. In addition, the mechanism has good scalability and migration capabilities, which is convenient for accessing other perception dimensions (such as visual anomaly detection, behavior prediction, etc.), and can adapt to traffic regulations in different cities to achieve cross-regional intelligent assessment of traffic safety.

By integrating visual and textual data to assess risk levels, this approach offers a more comprehensive analysis of complex scenarios. To better interpret prediction outcomes, an additional ***reason*** attribute has been added to explain the rationale behind each prediction result.

To effectively test the system's perception, prediction, and decision-making capabilities, this study conducts an application analysis based on a selected scenario, as shown in Fig. 6. The scenario depicts a typical urban traffic intersection, illustrating the dynamic interactions between vehicles, pedestrians, and road infrastructure. The road consists of two-way lanes, separated by a yellow double solid line prohibiting lane-crossing by vehicles. At the intersection, a broad white pedestrian crossing provides a safe passageway for pedestrians. The sidewalks are paved with neatly arranged bricks, and the roadside is equipped with billboards and traffic light poles, reflecting well-established infrastructure. A sunshade and a small stall further in the background indicate some level of commercial activity or pedestrian gathering in the area. In terms of pedestrian and autonomous vehicle movement, the pedestrian crossing signal is currently red, prohibiting pedestrians from crossing the road. However, a pedestrian is seen crossing on the pedestrian crossing, disregarding the traffic signal. This suggests that the pedestrian has failed to comply with traffic signals, creating a potential conflict risk with nearby moving vehicles. This scenario highlights a challenge in urban traffic management: even with robust traffic infrastructure and signal control systems, some pedestrians continue to violate traffic rules, potentially leading to accidents. The scenario also emphasizes the need to enhance traffic safety by combining technological solutions with behavioral guidance to reduce violations and ensure the safety of all road users.

The prompts for the prediction component are shown in Appendix B (Fig. B1), with output results displayed in Appendix D (Fig. D2). These results highlight the system's strong capability in traffic context perception and risk assessment, allowing it to make reasonable judgments based on the dynamic behaviors of vehicles and pedestrians. In the vehicle prediction segment, the system effectively identifies the risk level associated with each vehicle and provides rational explanations based on their relative positions on the road. For instance, the system accurately assesses that vehicles driving within normal parameters have a low risk, while a truck near the pedestrian crossing is flagged as a potential risk, indicating its capability to anticipate future dangers. Additionally, the system correctly evaluates the low-risk status of a vehicle following closely behind another, reflecting a solid understanding of dynamic relationships among different road users. In pedestrian prediction, the system not only discerns the current actions of pedestrians (e.g., crossing the road or waiting in a safe zone) but also assesses the risk level associated with these actions. For example, a pedestrian crossing through traffic prompts a "medium" risk alert, whereas one standing on the sidewalk is classified as "low" risk, demonstrating the system's sensitivity to pedestrian dynamics. This precise judgment helps minimize false alarms and avoid unnecessary risk warnings. Overall, the system shows promise in understanding vehicle-pedestrian interactions and assessing potential risks, underscoring its potential for practical application in complex traffic scenarios.

Autonomous decision-making for pedestrian crossings using VLM

In the intelligent pedestrian crossing scenario, the system captures perception information about the environment surrounding vehicles and pedestrians and uses this information to predict collision risks. To ensure data consistency and high quality, a standardized checklist based on these perception and prediction outputs provides a unified, reliable input source for the decision-making layer of the system. This standardized data format minimizes complexity due to inconsistent data structures, enhancing data processing efficiency and decision-making accuracy. Additionally, by supplying structured environmental data, the standardized checklist enables the autonomous driving system to better navigate complex traffic scenarios. This section leverages the information above to make decisions for pedestrians and vehicles in pedestrian crossing scenarios using a large model. These decisions aim to prevent or mitigate collision risks, thereby enhancing safety within pedestrian crossing scenarios. The factors attributes and data types for decision-making indicators are shown in Table 4. Similar to the prediction section, the decision-making process primarily focuses on providing recommendations for autonomous vehicles and pedestrians at pedestrian crossings. The ***decision*** attribute outputs the decision result, while the ***reason*** attribute explains the decision rationale.

For autonomous vehicles, decision outputs include:

Crossing: Continue driving and pass through the pedestrian crossing.

Pausing: Slow down and pause as the vehicle approaches the crossing or maintain a stopped state if already paused

Braking: Apply immediate braking measures to decelerate or come to a complete stop to prevent a collision.

For pedestrians, decision outputs include:

Crossing: If safe, the pedestrian proceeds quickly and safely across the crossing.

Waiting:: Wait for an appropriate moment to cross the pedestrian crossing.

This structured decision-making approach provides clarity and safety in complex pedestrian crossing environments.

The decision prompts, as shown in Appendix C (Fig. C1), and the decision output in Appendix D (Fig. D3) demonstrate the system's strong capability in traffic context analysis and decision support. By comprehensively

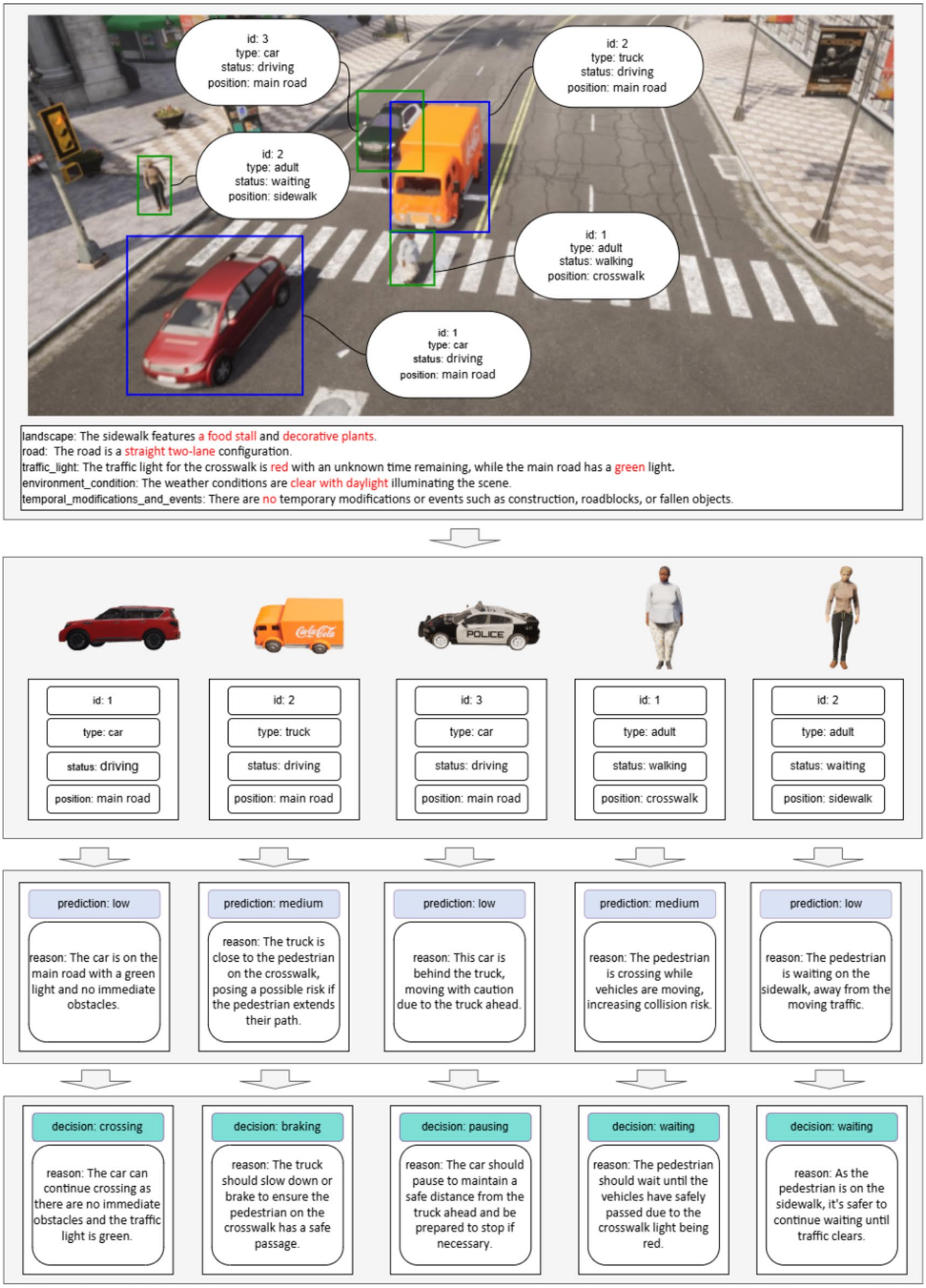

id: 3
type: car
status: driving
position: main road
id: 2
type: truck
status: driving
position: main road
id: 2
type: adult
status: waiting
position: sidewalk
id: 1
type: adult
status: walking
position: crosswalk
id: 1
type: car
status: driving
position: main road
landscape: The sidewalk features a food stall and decorative plants.
road: The road is a straight two-lane configuration.
traffic_light: The traffic light for the crosswalk is red with an unknown time remaining, while the main road has a green light.
environment_condition: The weather conditions are clear with daylight illuminating the scene.
temporal_modifications_and_events: There are no temporary modifications or events such as construction, roadblocks, or fallen objects.
POLICE
id: 1
type: car
status: driving
position: main road
id: 2
type: truck
status: driving
position: main road
id: 3
type: car
status: driving
position: main road
id: 1
type: adult
status: walking
position: crosswalk
id: 2
type: adult
status: waiting
position: sidewalk
prediction: low
reason: The car is on the main road with a green light and no immediate obstacles.
prediction: medium
reason: The truck is close to the pedestrian on the crosswalk, posing a possible risk if the pedestrian extends their path.
prediction: low
reason: This car is behind the truck, moving with caution due to the truck ahead.
prediction: medium
reason: The pedestrian is crossing while vehicles are moving, increasing collision risk.
prediction: low
reason: The pedestrian is waiting on the sidewalk, away from the moving traffic.
decision: crossing
reason: The car can continue crossing as there are no immediate obstacles and the traffic light is green.
decision: braking
reason: The truck should slow down or brake to ensure the pedestrian on the crosswalk has a safe passage.
decision: pausing
reason: The car should pause to maintain a safe distance from the truck ahead and be prepared to stop if necessary.
decision: waiting
reason: The pedestrian should wait until the vehicles have safely passed due to the crosswalk light being red.
decision: waiting
reason: As the pedestrian is on the sidewalk, it's safer to continue waiting until traffic clears.

Fig. 6. Perception-prediction-decision overall results.

understanding the behaviors of vehicles and pedestrians, the system provides each road user with reasonable action recommendations. For vehicles, the system makes clear judgments based on current road conditions and traffic light status. For example, it permits normal passage when the light is green and no obstacles are present, while requiring trucks to slow down near pedestrians to ensure safety. This reflects the system's sharp perception of priority in various traffic scenarios. Additionally, it advises following vehicles to maintain a safe distance and prepare to stop if necessary, showcasing precise control over dynamic traffic flow relationships. On the pedestrian side, the system prioritizes safety, encouraging pedestrians to cross at appropriate times or continue waiting. For instance, it suggests that a pedestrian attempting to cross the street waits until vehicles have safely passed to avoid potential conflicts. For pedestrians already in a safe area, the system determines that immediate action is unnecessary, thus minimizing unexpected risks with cautious decision-making. Overall, the system exhibits clear logic and sound decision-making, highlighting its high application value in pedestrian crossing and autonomous driving contexts.

Figure 6 illustrates the overall output results within the context of this study's scenario. It shows a comprehensive analysis of the behaviors of various participants in the traffic environment. By providing detailed decision recommendations, the system assists traffic participants in making safe decisions in complex traffic scenarios, thereby reducing potential risks. This decision-making and recommendation mechanism, based on dynamic information, not only significantly enhances the intelligence of autonomous driving systems but also greatly improves the overall safety of the roads. It offers better protection for all traffic participants, contributing to a safer and more efficient traffic environment.

Quantitative experimental results and analysis

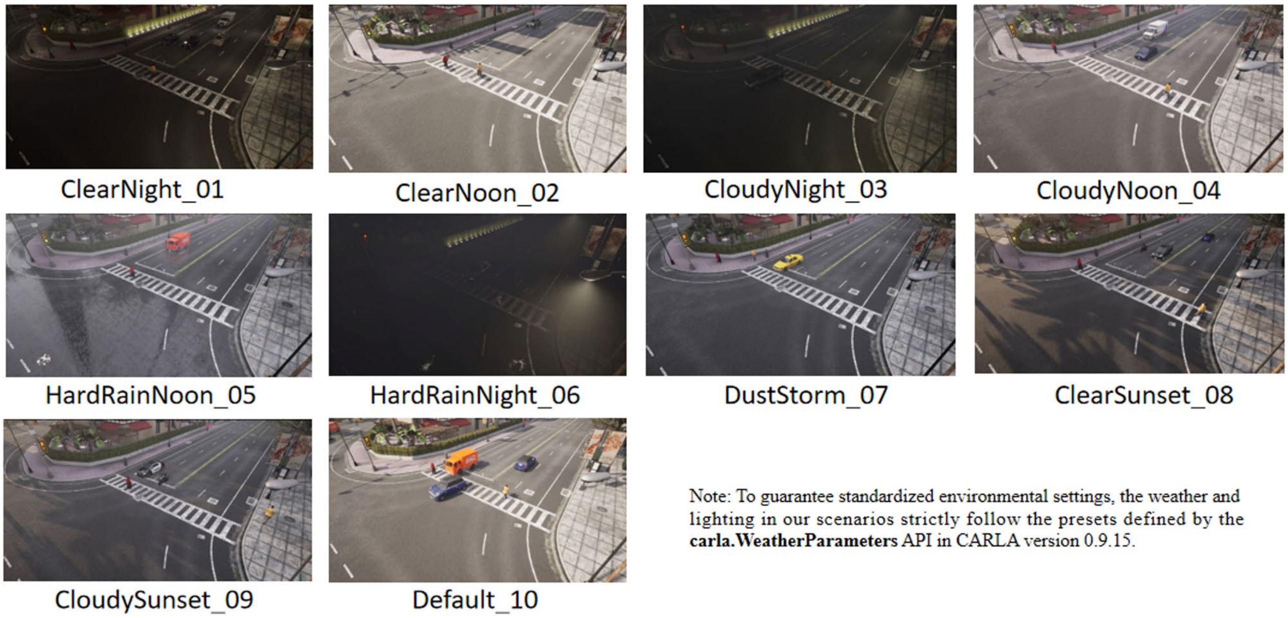
Experimental results and analysis

To evaluate the perception-prediction-decision-making capabilities of this system, this paper conducted results tests in 10 scenarios based on the CARLA simulation platform, covering a variety of weather conditions and combinations of lighting changes. This paper evaluates it through three quantitative indicators: (a) Perception Accuracy: This indicator is used to measure the VLM's ability to recognize key traffic scene elements. We compare the VLM recognition results with manual annotations to evaluate its recognition consistency of core elements such as vehicles, pedestrians, traffic lights, and road types. (b) Risk prediction consistency: To evaluate the risk assessment ability of VLM in dynamic traffic scenarios, the possible collision risks in each test scenario are graded and judged. The risk level output by VLM is compared with the level of rule judgment, and the consistency of risk level is used as the standard. (c) Decision-making rule consistency rate: This indicator is used to evaluate the rationality of the behavior decisions generated by VLM under traffic rules and safety priority strategies. We compare it with road traffic regulations and traffic safety principles.

The scene tested in this article is shown in Fig. 7. The scene's lighting and weather conditions refer to the carla. WeatherParameters interface in CARLA 0.9.15. The lighting conditions include noon, sunset, and night, and the weather conditions include sunny, cloudy, rainy, and sandstorm. The weather and lighting conditions used in the simulation scene are configured by referencing the carla.WeatherParameters interface in CARLA 0.9.15 to ensure consistency with the built-in environment presets.

The test results of this paper are shown in Table 5. By analyzing the results, it can be seen that the average performance of the system in the three main indicators is relatively stable overall, especially the Perception accuracy has reached 89.91%, which shows that the model has high reliability in scene perception and recognition. The Risk prediction consistency and Decision-making rule consistency rate are slightly lower, at 85.91% and 87.72% respectively. Although these two values are also at a relatively high level, they fluctuate greatly in specific scenarios, indicating that the stability of the system in different environments still has room for improvement.

In some specific environments, such as "ClearNoon_02" and "HardRainNoon_05", the system performed extremely well, with all three indicators reaching 97.5%. Through retrospective analysis of the scenes, it was found that the traffic environment of vehicles and pedestrians in the scenes was not very complex, which shows that the system can still maintain highly consistent and accurate performance when there is sufficient light during the



Note: To guarantee standardized environmental settings, the weather and lighting in our scenarios strictly follow the presets defined by the **carla.WeatherParameters** API in CARLA version 0.9.15.

Fig. 7. Simulation scenarios used in the test.

Scene Number	Perception accuracy	Risk prediction consistency	Decision-making rule consistency rate
ClearNight_01	94.59%	83.33%	88.89%
ClearNoon_02	97.5%	97.5%	97.5%
CloudyNight_03	92%	83.3%	83.33%
CloudyNoon_04	87.5%	87.5%	87.5%
HardRainNoon_05	97.5%	97.5%	97.5%
HardRainNight_06	90.0%	87.5%	75%
DustStorm_07	88%	97.5%	97.5%
ClearSunset_08	80.56%	66.67%	83.33%
CloudySunset_09	89.29%	75%	83.33%
Default_10	82.14%	83.33%	83.33%
Average	89.91%	85.91%	87.72%

Table 5. Factors attributes and data types for scenario prediction and decision-making.

Method	Perception accuracy	Risk prediction consistency	Decision-making rule consistency rate
Rule-based Method	/	93.12% (within predefined rules)	90.23% (within predefined rules)
Deep Learning Baseline	79.5% (YOLOv5)	77.70% (LSTM + Transformer)	76.00% (Learning Integration Strategies)
Our Method(VLM-based)	89.91%	85.91%	87.72%

Table 6. Comparison of experimental results of different methods.

day, or even in bad weather such as heavy rain, but the traffic environment is not complex. When in some scenes with complex or mixed lighting, such as "ClearSunset_08" and "CloudySunset_09", the performance declined significantly. The Perception accuracy of "ClearSunset_08" was only 80.56%, and the Risk prediction consistency was even lower at 66.67%. This shows that under low-light and strongly directional lighting conditions such as sunset and west-facing light, the model's perception and judgment capabilities will be disturbed, especially in predicting risks, the system's judgment may be biased or unstable. This type of problem deserves special attention, because the sunset period is very common in real driving scenarios, and the light conditions change rapidly, posing a challenge to the system. In contrast, the combination of night and extreme weather conditions, such as "HardRainNight_06", also exposed the shortcomings of the model. Although the perception accuracy remained at 90%, the consistency of the decision rules dropped significantly to 75%. This shows that even if the system can perceive the information, it may not be able to make a stable behavioral response in decision-making due to the high uncertainty of the input. This situation may be caused by the poor adaptability of the rule logic under night and rainy conditions. In the "DustStorm_07" scenario, although the perception accuracy was only 88%, the system was completely consistent (97.5%) in risk prediction and decision rules, which may indicate that even if the sensor signal is disturbed, the overall control strategy of the system is strengthened or specialized in this type of scenario, perhaps with the help of rule-driven or conservative strategies to maintain stability.

From the overall trend, the performance is best during clear daytime, followed by clear nights, and the most challenging is the edge period when the light changes dramatically or is disturbed, such as sunset. In addition, the consistency of risk prediction is generally higher than the perception accuracy and decision consistency, indicating that even when perception is limited, the system can still maintain a certain stability in risk judgment, but there is still room for further optimization of the decision-making process from prediction to action.

Comparative experiments with baseline methods

To better demonstrate the effectiveness and advantages of our proposed approach, we compare the results with both traditional rule-based methods and deep learning baselines. a): Traditional rule-based method (Rule-based): simulates the decision-making control process based on preset traffic rules and state trigger mechanism. This method is widely used in the early stage of autonomous driving system and has strong stability, but has limitations in dealing with complex dynamic environments. b) Deep Learning Baseline: YOLOv5 is used for visual target recognition, and LSTM model and Transformer model are used to predict pedestrian/vehicle trajectory and risk, and then the decision is made through learning fusion strategy. Table 6 shows the comparison of experimental results of different methods.

Although the rule-based method has shown extremely high accuracy in indicators such as risk prediction consistency and decision compliance rate, it is highly dependent on manually preset rules and lacks the ability to perceive and respond to dynamic changes and complex scenarios. Therefore, from a practical point of view, this method is not feasible for actual road deployment. Its application scenarios are limited to specific conditions with clear rules and controllable environment, and it is difficult to adapt to the complex and changeable traffic situations on real roads. In contrast, the multimodal understanding and causal reasoning capabilities of VLM can effectively improve the perception and decision-making quality of autonomous driving systems in complex scenarios. The VLM-based method in this paper is superior to the mainstream deep learning baseline method,

Aspect	Rule-based	Deep Learning Baseline	This Work (VLM-based)
Input Modality	Single rules or visual	Visual image/video	Multimodal visual + language
Semantic Understanding	Low struggles with context	Medium learns features, limited context	High understands objects, context, and semantics
Causal/Intent Reasoning	Almost none	Limited task-specific	Strong Explicit reasoning via language prompts
Generalization Capability	Weak poor transfer to unseen conditions	Moderate needs large labeled datasets for transfer	Strong pretrained on large-scale data, adapts well with few-shot tuning or prompt engineering
Interpretability	High transparent logic from rules	Low black-box models	High uses natural language explanations
Complex Scene Handling	Poor rule coverage is limited	Limited may miss contextual cues or rare cases	Strong models interaction and high-level dynamics
Data Needs	Low rule-based design	High large-scale labeled data	Moderate leverages pretraining + low-shot adaptation

Table 7. Systematic comparison of different methods along key dimensions.

which reflects the good balance between generalization, semantic understanding ability and safety of this method.

At the same time, we believe that the advantages of VLM are not only reflected in the experimental indicators, but also in the methodology. Table 7 shows a systematic comparison of the three methods in key dimensions. Compared with traditional methods, VLM shows significant advantages in perception, risk prediction and decision-making tasks in autonomous driving. Traditional methods usually rely on single-modal input, such as visual processing based on convolutional neural networks or rule-based systems. Such methods often find it difficult to fully capture semantic information and contextual dependencies in complex traffic environments. In contrast, VLM achieves multi-level understanding capabilities from target recognition to semantic reasoning by fusing images and text, allowing the model to not only "see" the scene, but also "understand" its deep meaning. This multimodal understanding is particularly critical in dynamic traffic scenarios. With the introduction of natural language, VLM shows a processing method that is closer to human cognition in risk identification and intention reasoning, and can model behavioral patterns and potential relationships in complex environments at a higher level, thereby providing more comprehensive and semantically rich support for downstream tasks.

The language generation capability of VLM enhances the interpretability of the system, making the model judgment process easier for humans to understand, thereby improving the transparency and credibility of the system. At the same time, since most VLMs are pre-trained on large-scale image-text data, they have been fine-tuned with a small number of samples or task-specific prompts, and have stronger generalization capabilities than traditional methods. When faced with unseen condition, they can significantly alleviate the "domain migration" problem of traditional models and maintain excellent performance under limited annotated data conditions.

Conclusion

This paper proposes and develops a smart scenario system based on VLM, aimed at enhancing intelligent perception and optimizing decision-making at pedestrian crossings in autonomous driving environments. Following autonomous driving road, vehicle, and safety standards, we conducted a thorough analysis of pedestrian crossing scenarios. By inputting customized prompts and sample outputs into the VLM, we created a standardized checklist of scenario entities, outputting them in JSON format. The system categorizes scenarios into two types: pedestrians crossing and autonomous vehicles passing through pedestrian crossings. The VLM facilitates perception, prediction, and decision-making, aiding in the development and optimization of intelligent scenarios.

Results indicate that the proposed smart scenario system, powered by VLM, uses a standardized checklist to systematically identify all entities in pedestrian crossing scenarios—such as pedestrians, vehicles, roads, and weather—offering a unified framework for intelligent perception. This approach enhances entity recognition and reduces misjudgments, particularly for tasks like counting pedestrians and classifying vehicle types, where traditional models struggle.

During perception, the system employs VLM with tailored prompts to identify all scenario details, even subtle ones. For example, in Fig. 3, the system equally highlights a water tower alongside other entities, showcasing its objective, comprehensive perception. This detailed awareness strengthens environmental modeling and supports more accurate intent prediction and decision-making.

In prediction, the system leverages VLM's reasoning to infer pedestrian and vehicle intentions by integrating scenario data with signals, signage, and environmental cues. It combines natural language and visual information, such as traffic lights, to enhance predictive accuracy in complex scenarios through multimodal analysis.

For decision-making, the system uses VLM to grasp high-level scenario semantics, aligning decisions with traffic rules and safety principles. It determines passage order and hazard responses by analyzing entities like traffic signals, vehicles, and pedestrians. VLM's reasoning also helps interpret traffic signs through natural language, guiding safe, regulation-compliant actions for both vehicles and pedestrians.

Although the proposed VLM-based system demonstrates zero-shot perception and reasoning capabilities, it does not inherently guarantee robust generalization to all unseen scenarios. This limitation arises from the data distribution gap between simulated and real-world environments, as well as the potential for misinterpretation of contextually ambiguous entities. Future work should incorporate few-shot learning and domain adaptation strategies to further enhance the system's generalizability.

While the use of roadside visual data proves effective in simulation, attention must still be paid to challenges in infrastructure-to-vehicle communication under real-world conditions. To mitigate this, we propose: (1) Edge Computing and On-site Deployment: In areas lacking real-time communication, parts of the VLM model can be deployed on edge computing units of roadside cameras to perform local perception and risk prediction. Key safety information can then be synchronized with vehicles via low-frequency communication. (2) Coordination Strategy between Vehicle-side and Infrastructure-side VLMs: A future direction could be the exploration of a "vehicle-infrastructure cooperative inference" mechanism, where both the vehicle and infrastructure sides run simplified or lightweight VLM modules. They can share structured information—such as standardized JSON-format lists—through predefined protocols to enable collaborative perception capabilities. (3) Standardized Interfaces and Intermediate Data Format Design: The JSON-structured output list designed in this study accounts for uncertainties in future V2I communication. By adopting a unified intermediate data format, it facilitates efficient compression and reliable transmission of information under bandwidth limitations or unstable network conditions.

While our VLM-based system exhibits superior perception and decision-making capabilities in pedestrian crossing scenarios, several limitations must be acknowledged. First, the high computational cost of VLM may pose challenges for real-time deployment, particularly in edge devices. Future work can explore model compression techniques such as quantization and knowledge distillation. Second, the reliance on simulated CARLA data limits the generalizability of our findings. A potential solution is to integrate real-world datasets such as the Waymo Open Dataset. Finally, although this study reduces the risk of VLM hallucinations by constructing a manually standardized checklist and incorporating visual information, VLM still tends to hallucinate in complex scenarios. Further work should explore hybrid approaches that incorporate rule-based constraints and knowledge graphs to mitigate misclassification risks. In recent years, artificial intelligence has driven autonomous driving technology forward. This study leverages VLM to address pedestrian crossing scenarios, fostering the integration of autonomous driving with intelligent transportation systems. Although VLM models are still evolving, this approach accelerates smart transportation development and advances intelligent mobility.

Data availability

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Received: 10 April 2025; Accepted: 4 August 2025
Published online: 25 August 2025

References

1. Guo, X., Teng, X. & Shen, Z. Smart pedestrian crossing design by using smart devices to improve pedestrian safety. *Rev. Adhes. Adhes.* **11**, 2023 (2023).
2. Budzynski, M., Guminska, L., Jamroz, K., Mackun, T. & Tomczuk, P. Effects of road infrastructure on pedestrian safety, IOP Conf. *Ser. Mater. Sci. Eng.* **603**, 042052. https://doi.org/10.1088/1757-899X/603/4/042052 (2019).
3. Rasouli, A. & Tsotsos, J. K. Autonomous vehicles that interact with pedestrians: A survey of theory and practice. *IEEE Trans. Intell. Transp. Syst.* **21**, 900–918. https://doi.org/10.1109/TITS.2019.2901817 (2020).
4. Rezwana, S. & Lownes, N. Interactions and behaviors of pedestrians with autonomous vehicles: A synthesis. *Future Transp.* **4**, 722–745. https://doi.org/10.3390/futuretransp4030034 (2024).
5. Van Brummelen, J., O'Brien, M., Gruyer, D. & Najjaran, H. Autonomous vehicle perception: the technology of today and tomorrow. *Transp. Res. Part. C Emerg. Technol.* **89**, 384–406. https://doi.org/10.1016/j.trc.2018.02.012 (2018).
6. Othman, K. Public acceptance and perception of autonomous vehicles: a comprehensive review. *AI Ethics*. **1**, 355–387. https://doi.org/10.1007/s43681-021-00041-8 (2021).
7. Sadaf, M. et al. Connected and automated vehicles: infrastructure, applications, security, critical challenges, and future aspects. *Technologies* **11**, 117. https://doi.org/10.3390/technologies11050117 (2023).
8. Atakishiyev, S., Salameh, M., Yao, H. & Goebel, R. Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions. *IEEE Access.* **12**, 101603–101625. https://doi.org/10.1109/ACCESS.2024.3431437 (2024).
9. Han, Y. et al. Collaborative perception in autonomous driving: methods, datasets, and challenges. *IEEE Intell. Transp. Syst. Mag.* **15**, 131–151. https://doi.org/10.1109/MITS.2023.3298534 (2023).
10. Li, J. et al. Domain Adaptation based Object Detection for Autonomous Driving in Foggy and Rainy (2024). https://doi.org/10.48550/arXiv.2307.09676
11. Ye, T. et al. FusionAD: Multi-modality Fusion for Prediction and Planning Tasks of Autonomous Driving (2023). https://doi.org/10.48550/arXiv.2308.01006
12. Hubmann, C., Becker, M., Althoff, D., Lenz, D. & Stiller, C. Decision making for autonomous driving considering interaction and uncertain prediction of surrounding vehicles. in *2017 IEEE Intell. Veh. Symp. IV, 2017:* 1671–1678. https://doi.org/10.1109/IVS.2017.7995949
13. OpenAI, GPTV System Card (2023).
14. Yang, Z. et al. The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision) (2023). http://arxiv.org/abs/2309.17421 (accessed May 23, 2024).
15. Tian, H. et al. Large (Vision) Language Models for Autonomous Vehicles: Current Trends and Future Directions (2024). https://doi.org/10.36227/techrxiv.172963218.80161917/v1
16. Zhou, X. & Knoll, A. C. GPT-4V as Traffic Assistant: An In-depth Look at Vision Language Model on Complex Traffic Events (2024). http://arxiv.org/abs/2402.02205 (accessed May 23, 2024).
17. Zhao, X. et al. Potential sources of sensor data anomalies for autonomous vehicles: an overview from road vehicle safety perspective. *Expert Syst. Appl.* **236**, 121358. https://doi.org/10.1016/j.eswa.2023.121358 (2024).
18. Wang, Y., Han, Z., Xing, Y., Xu, S. & Wang, J. A survey on datasets for the decision making of autonomous vehicles. *IEEE Intell. Transp. Syst. Mag.* **16**, 23–40. https://doi.org/10.1109/MITS.2023.3341952 (2024).

19. Galvão, L. G. & Huda, M. N. Pedestrian and vehicle behaviour prediction in autonomous vehicle system — A review. *Expert Syst. Appl.* **238**, 121983. https://doi.org/10.1016/j.eswa.2023.121983 (2024).
20. Cong, P. et al. A visual detection algorithm for autonomous driving road environment perception. *Eng. Appl. Artif. Intell.* **133**, 108034. https://doi.org/10.1016/j.engappai.2024.108034 (2024).
21. Teng, X., Shen, Z., Huang, L., Li, H. & Li, W. Multi-sensor fusion based wheeled robot research on indoor positioning method. *Results Eng.* **22**, 102268. https://doi.org/10.1016/j.rineng.2024.102268 (2024).
22. Huang, L. et al. *Int. Arch. Photogramm Remote Sens. Spat. Inf. Sci. XLVIII -3/W1-2022* 19–24. https://doi.org/10.5194/isprs-archives-XLVIII-3-W1-2022-19-2022. (2022).
23. Shen, Z., Teng, X., Zhang, Y., Fang, G. & Xu, W. Guidelines for installation of sensors in smart sensing platforms in underground spaces. *Sensors* **22**, 3215. https://doi.org/10.3390/s22093215 (2022).
24. Teng, X., Shen, Z., Ge, T. & Lei, R. A review of intelligent scenes design in underground pedestrian system, IOP conf. Ser. *Earth Environ. Sci.* **1157**, 012001. https://doi.org/10.1088/1755-1315/1157/1/012001 (2023).
25. Castañeda, K., Sánchez, O., Herrera, R. F., Pellicer, E. & Porras, H. BIM-based traffic analysis and simulation at road intersection design. *Autom. Constr.* **131**, 103911. https://doi.org/10.1016/j.autcon.2021.103911 (2021).
26. Zhao, X. et al. Crossing roads in a social context: how behaviors of others shape pedestrian interaction with automated vehicles. *Transp. Res. Part. F Traffic Psychol. Behav.* **102**, 88–106. https://doi.org/10.1016/j.trf.2024.02.008 (2024).
27. Izquierdo, R., Alonso, J., Benderius, O., Sotelo, M. Á. & Fernández, D. Llorca, pedestrian and passenger interaction with autonomous vehicles: field study in a crosswalk scenario. *Int. J. Human–Computer Interact.* 1–19. https://doi.org/10.1080/10447318.2024.2426856 (2024).
28. Tian, Z. et al. Evaluating scenario-based decision-making for interactive autonomous driving using rational criteria: A survey. *ArXiv Prepr*. ArXiv250101886 (2025).
29. Lu, Y., Ma, H., Smart, E. & Yu, H. Enhancing autonomous driving decision: A hybrid deep reinforcement Learning-Kinematic-Based autopilot framework for complex motorway scenes. *IEEE Trans. Intell. Transp. Syst.* (2025).
30. Ghintab, S. S. & Hassan, M. Y. CNN-based visual localization for autonomous vehicles under different weather conditions. *Eng. Technol. J.* **41**, 375–386 (2023).
31. Yu, H., Huo, S., Zhu, M., Gong, Y. & Xiang, Y. Machine learning-based vehicle intention trajectory recognition and prediction for autonomous driving. in *2024 7th Int. Conf. Adv. Algorithms Control Eng. ICAACE, IEEE* 771–775 (2024).
32. Thakur, A. & Mishra, S. K. An in-depth evaluation of deep learning-enabled adaptive approaches for detecting Obstacles using sensor-fused data in autonomous vehicles. *Eng. Appl. Artif. Intell.* **133**, 108550 (2024).
33. Rezwana, S. & Lownes, N. Interactions and behaviors of pedestrians with autonomous vehicles: A synthesis. *Future Transp.* **4**, 722–745 (2024).
34. Spivak, I., Krepych, S., Litvynchuk, M. & Spivak, S. Validation and data processing in JSON format. in *IEEE EUROCON 2021–19th int. Conf. Smart Technol.* 326–330. https://doi.org/10.1109/EUROCON52738.2021.9535582 (2021).
35. Dosovitskiy, A. CARLA: an open urban driving simulator. *Proc. 1st Annu. Conf. Robot Learn.* **78**, 1–16 (2017).
36. ISO 34501: 2022 Road vehicles — Test scenarios for automated driving systems — Vocabulary (2022).
37. ISO-21448: Road vehicles - Safety of the intendedfunctionality (2019).
38. ISO/DIS 34502. DRAFT INTERNATIONAL STANDARD (2022).

Acknowledgements

This research was funded by VR tool to visualize energy-saving management for ZEB considering work style, Project/Area Number: 23K04156.

Author contributions

L.H.， X.T and Z.S designed the study and conducted the experiments. X. T. and L.H. analyzed the data, wrote the main manuscript text, and prepared all figures and tables. Z.S. and W.L. contributed to manuscript revisions. All authors reviewed and approved the final manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-14827-x.

Correspondence and requests for materials should be addressed to Z.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.