Commun Transp Res

Review

Enhancing Autonomous Vehicle Safety with Knowledge Graphs and Large Language Models: A Comprehensive Review

Jiaxin Liu[1,4,†], Liang Peng[1,†], Xiangyu Yan[2], Lingjun Zhang[1], Chenye Yang[1], Yueming Tao[3], Ashton Yu Xuan Tan[1], Tianli Xu[1], Sai Guo[4], Hong Wang[1,✉], Jun Li[1]

Cite this article: https://doi.org/10.26599/COMMTR.2026.9640023

ABSTRACT: As autonomous vehicle technology continues to evolve, ensuring its safety in complex and dynamic environments has become a critical challenge. Knowledge Graphs (KGs) and Large Language Models (LLMs), as two cutting-edge approaches representing the forefront of knowledge-driven and connectionist paradigms in modern artificial intelligence, are emerging as powerful tools for enhancing autonomous vehicle safety. In this paper, we provide a comprehensive review of the applications of KGs and LLMs in enhancing the safety of autonomous driving systems. Building upon a brief introduction to the fundamental concepts and underlying technologies of KGs and LLMs, we then present their respective applications in autonomous vehicle safety from complementary perspectives. We further compare the advantages and limitations of KGs and LLMs in terms of knowledge representation, inference capability, scalability, and real-time performance. To leverage the complementary strengths of structured knowledge and language-based reasoning, we review existing research efforts that integrate KGs and LLMs in the context of AV safety enhancement. Based on this analysis, we propose a hybrid safety-enhancement framework that combines explicit knowledge structures of KGs with the flexible reasoning capabilities of LLMs, offering a promising direction toward more robust, interpretable, and adaptive autonomous driving systems.

KEYWORDS: Knowledge Graph (KG), Large Language Model (LLM), Bidirectional reasoning, Autonomous vehicle safety, Risk cognition

1 Introduction

1.1. Motivation

Autonomous vehicles (AVs) show great promise for improving traffic safety and efficiency (D. Yang et al., 2018). However, complex real-world scenarios—with diverse participants and unpredictable interactions—remain a major challenge to current AV systems (Peng et al., 2023). Despite progressive advances in sensors and learning-based models, current AVs still suffer in understanding scene context, interpreting interactions, and handling rare or uncertain situations (McCarthy, 2022). This highlights the need for stronger reasoning, better knowledge representation, and more robust decision-making.

To address these issues, Knowledge Graphs (KGs) and Large Language Models (LLMs) have attracted attention for their potential to inject structured knowledge and advanced reasoning abilities into AV systems. KGs represent structured, symbolic knowledge through entities and their interrelations, enabling human-understandable explicit reasoning. By contrast, LLMs are grounded in deep neural architectures and trained on massive text corpus to acquire statistical patterns of language or multimodal data. From the perspective of artificial intelligence (AI) development, KGs and LLMs represent the latest advancements of two typical paradigms of AI, i.e., symbolism and connections, and both have been widely used in numerous areas.

[1] School of Vehicle and Mobility, Tsinghua University, Beijing 100084, China. [2] School of Mechanical Engineering, Beijing Institute of Technology, Beijing 100081, China.

[3] School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu 611731. [4]Xiaomi EV, Beijing 100085, China.

* Jiaxin Liu and Liang Peng contributed equally to this work.

✉ Corresponding author. E-mail: hong_wang@tsinghua.edu.cn

Received: October 28, 2025; Revised: February 17, 2026; Accepted: April 14, 2026

Nomenclature (or others such as Abbreviation, if applicable; nomenclature entries should have the units identified)

A-Box	Assertion Box
ADS	Autonomous Driving System
AV	Autonomous Vehicle
CoT	Chain-of-Thought
DKG	Dynamic Knowledge Graph
DSL	Domain Specific Language
EKG	Event Knowledge Graph
E2E	End to End
GNN	Graph Neural Network
KG	Knowledge Graph
KGE	Knowledge Graph Embedding
LLM	Large Language Model
MLLM	Multimodal Large Language Model
RAG	Retrieval-Augmented Generation
SKG	Static Knowledge Graph
T-Box	Terminological Box
TKG	Temporal Knowledge Graph
TTC	Time to Collision
VLM	Vision Language Model

In the context of AV safety, KGs' strengths in structured representation, causal reasoning, and knowledge traceability contribute to more reliable scene interpretation and risk avoidance. By modeling human driving knowledge, such as traffic rules, object relations, driving experience, and commonsense, KGs help AV systems maintain semantic consistency and ensure explainable behavior. However, KGs often require manual construction or expert curation and lack scalability in open-world environments. Besides, the limited reasoning capability makes KG struggle with handling complex and highly semantic driving situations. In contrast, LLMs excel at capturing semantic information and generalizing across diverse contexts, enabling AV systems to interpret unstructured inputs, adapt to rare scenarios, and interact naturally through language. Yet, they may suffer from hallucinations, lack grounded knowledge, and provide limited interpretability in safety-critical decisions.

To this end, a growing body of promising research has begun to explore the integration of KGs and LLMs to enhance AV safety. These efforts aim to combine the explicit, domain-grounded reasoning offered by KGs with the adaptive, context-aware understanding enabled by LLMs. Such complementary synergy has already demonstrated potential in tasks such as prediction, decision-making, and risk recognition, especially in scenarios that are rare, ambiguous, and dynamic. However, as LLM remains a technology that emerged in recent years, current research remains fragmented, in an early stage, and facing challenges in knowledge alignment, reasoning efficiency, and real-time applicability.

1.2. Contribution

To this end, we conduct a comprehensive review to synthesize recent advances, identify research gaps, and provide insights into the integration of KGs and LLMs for enhancing AV safety.

To the best of our knowledge, existing surveys cover relevant pieces from different angles, such as traffic ontologies, LLM applications in intelligent transportation, and neuro-symbolic methods for AV. However, they typically focus on a single methodology or a single stage of the safety workflow. In contrast, as summarized in Table.1, this review provides an **explicit, lifecycle-oriented comparison** across the development, V&V, and operation phases, and systematically discusses both stand-alone and synergistic roles of **KGs and LLMs** in **enhancing autonomous vehicle safety**. Therefore, this review aims to fill this gap by:

- Systematically summarizing the current roles of KGs and LLMs in AV safety;
- Analyzing their respective strengths and limitations in safety-related applications;
- Reviewing recent research efforts that explore their integration;
- And proposing a unified perspective and a hybrid framework for future development.

This work aims to serve as both a reference for researchers and a roadmap for advancing safety-enhancing AI in AV. The structure of the remainder of this paper is shown in Fig.1, organized as follows: in Sec.2, we briefly introduce the foundational concepts and core technologies of KGs and LLMs. Methods for enhancing AV safety using KG and LLM are respectively reviewed in Sec.3 and Sec.4. Furthermore, we compare their pros and cons in Sec.5, with the research on the complementary integration of the two summarized in Sec.6. In Sec.6.3.3, we propose a potential framework to integrate them for AV safety. We discussed the current challenges and promising researches in Sec.6.3.3. The paper is finally concluded in Sec.8.

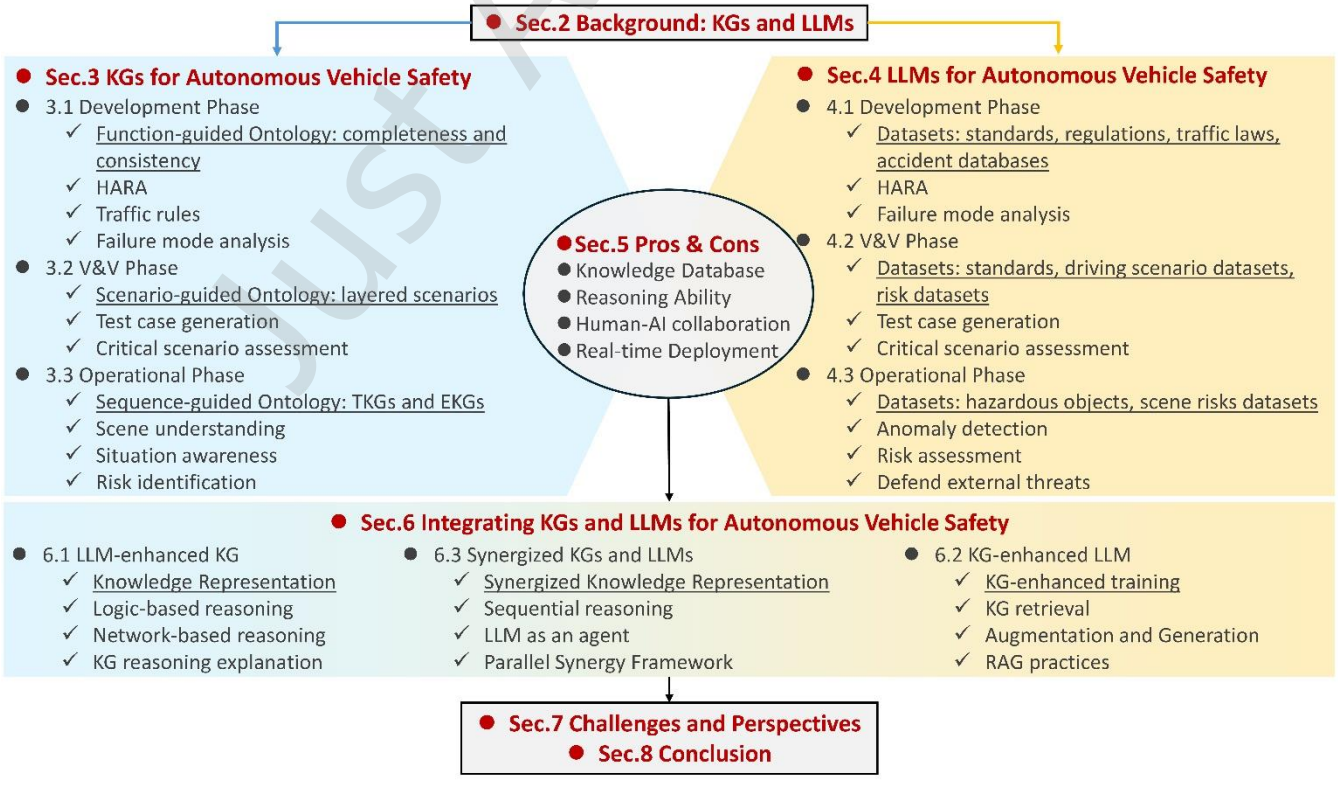


Fig. 1. The framework and structure of this review

Table 1. A Comparative Overview of Representative Surveys and Reviews on Knowledge Graphs and Large Language Models for AV Safety

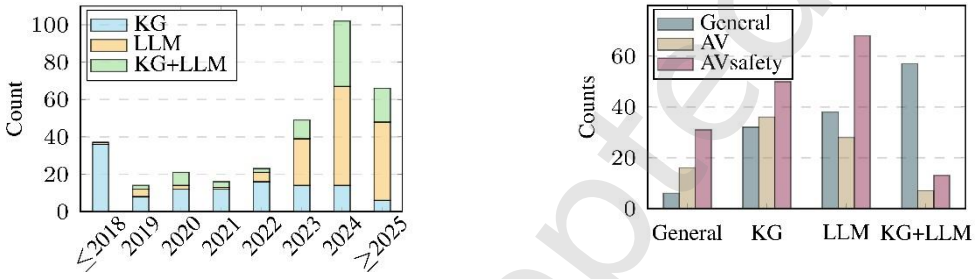
References	Domain Scope	LifeCycle Coverage			Methodological Coverage							Highlights
					KG		LLM		Synergizing KG & LLM			
		Dev	V & V	Ope	ontology & extraction	reasoning application	data & adaptation	reasoning application	KG-enhanced LLM	LLM-enhanced KG	Hybrid	
Hogan et al. (2021); X. Jiang et al. (2023)	KG General				✓	✓						Concepts, tasks and evolution of KGs
S. Xu et al. (2023); Guan et al. (2022)	KG General				✓	✓						Definition, taxonomy and trends of EKGs
L. Zhong et al. (2023)	KG General				✓							KG construction pipeline
Bugueño et al. (2024); J. Zhang et al. (2021); Delong et al. (2023); X. Chen et al. (2020)	KG General					✓						Graph-based explainable reasoning
Qin et al. (2025)	LLM General						✓	✓				Resources and trends of MLLMs
Han et al. (2024); Xing et al. (2024); Sahoo et al. (2024)	LLM General						✓					Efficient fine-tuning and prompt engineering
L. Huang et al. (2025)	LLM General							✓				Challenges and causes of hallucination
P. Zhao et al. (2026); Ni et al. (2025)	LLM General							✓	✓			Framework of trustworthy RAG
S. Pan et al. (2024); D. Li and Xu (2024)	KG-LLM				✓	✓	✓	✓	✓	✓	✓	KG-LLM synergy paradigms and prospects
J. Z. Pan et al. (2023)	KG-LLM				✓	✓		✓	✓	✓	✓	Opportunities and challenges for synthesis
R. Chen (2025); Z. Zhu et al. (2025)	KG-LLM				✓	✓		✓	✓			KG as structured evidence for RAG
Y. Zhu et al. (2024); Ibrahim et al. (2024)	KG-LLM				✓	✓	✓	✓		✓	✓	Methods & metrics for LLMs-enhanced KG
Katsumi and Fox (2018)	AV General	✓	✓	✓	✓							Taxonomy of ontologies in transportation
Luettin et al. (2022)	AV General	✓		✓	✓	✓						KG-based methods in autonomous driving
Htun and Fukuda (2024)	AV General	✓		✓	✓	✓						Integrating KGs into AV technology stack
Manas and Paschke (2025)	AV General	✓		✓	✓	✓				✓		System-level knowledge integration
X. Zhou et al. (2023); Z. Yang et al. (2023)	AV General	✓	✓	✓			✓	✓				VLM/MLLM applications in AD & ITS
Ashqar et al. (2025)	AV General			✓			✓	✓				VLM/MLLM-based object detection in ITS
Y. Gao et al. (2026)	AV General		✓	✓			✓	✓				Foundation Models in Scenario Generation & Analysis
Zipfl et al. (2023)	AV Safety		✓		✓							Ontologies for scenario-based testing
Riedmaier et al. (2020)	AV Safety		✓		✓	✓						Scenario-based safety assessment
D. Xiao et al. (2023)	AV Safety			✓		✓						Graph-based hazardous event detection
Q. Wei (2025)	AV Safety			✓				✓				LLM-based AV self-protection mechanisms
Tami et al. (2025b); D. Zhang et al. (2024)	AV Safety			✓			✓	✓				VLM/MLLM applications in traffic safety
H. Xu et al. (2024)	AV Safety	✓		✓				✓	✓			RAG for ITS and traffic safety
Ours	AV Safety	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	Enhancing AV Safety with KGs and LLMs

1.3. Literature Coverage

To ensure comprehensive coverage in this survey, we collected and filtered the literature from multiple public sources, including major publisher databases such as IEEE and Elsevier, preprint repositories such as arXiv, and official project websites when relevant. In total, we reviewed and cited nearly 400 papers in this review. Within these sources, beyond the core terms of KG, LLM, and AV safety, we performed combinational searches using KG-related keywords such as scene graph, ontology, and OWL, LLM and synergy-related keywords such as VLM, language, fine-tune, and RAG, and AV safety-related keywords such as edge case, long tail, and critical scenario, and then selected representative works from

the retrieved results. Considering the fast-evolving nature of LLM research, where influential advances are often published first as preprints and are later consolidated through peer-reviewed publication, we did not restrict our coverage to peer-reviewed papers only. Instead, we also extensively reviewed preprints (e.g., from arXiv) and conducted additional quality checks to ensure that the included preprint studies are technically sound and relevant to the scope of this survey.

Fig. 2(a) summarizes the reviewed papers by publication year. Early studies (up to 2018) are dominated by KG-related research. With the emergence of the transformer architecture and representative models such as BERT, a small number of works on LLMs and KG-LLM integration began to appear. Since 2023, when GPT made LLMs widely recognized by the community, researchers have increasingly explored the potential of LLM-based and KG+LLM technical routes for AV safety, and most of the reviewed studies fall into this recent period. Fig. 2(b) further shows the application areas of the reviewed works, where "General" refers to foundational studies or research outside AV. The statistics suggest that KG–LLM synergy is an emerging direction and remains underexplored in AV and AV safety settings, leaving substantial room to explore its potential. Accordingly, this review aims to summarize existing efforts of KGs and LLMs toward AV safety and to introduce mainstream KG-LLM synergy directions, providing a structured reference for researchers in this area.



(a) Publication years of the reviewed papers. (b) Research scope of the reviewed papers

Fig. 2. The temporal and domain distributions of the reviewed paper.

2 Backgrounds: KGs and LLM

In this section, we will briefly introduce the basic concepts, development, and key technologies of KGs and LLMs, as well as their applications in AVs. The development timeline of these two technologies with their integrations is shown in Fig.3.

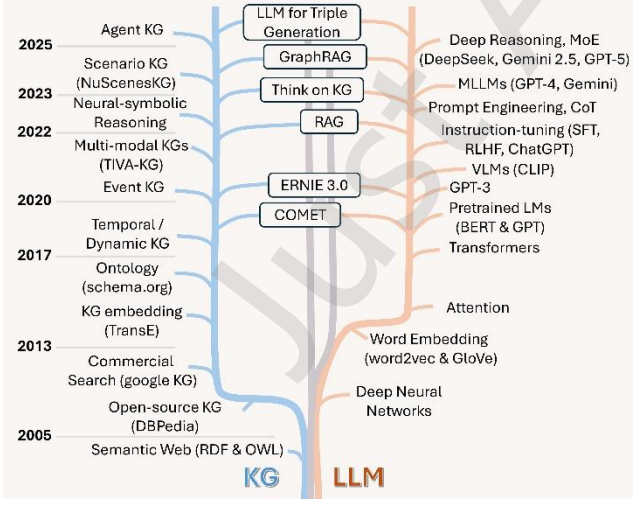


Fig. 3. The development of KGs, LLMs and their integrations.

2.1. Knowledge Graphs in AV

2.1.1 Knowledge Graphs

A KG organizes information in the form of a graph, where nodes represent entities or concepts and edges denote the relationships between them (Hogan et al., 2021). By organizing vast amounts of fragmented data into highly structured information, KGs enable the modeling of complex interentity relations and support reasoning over these associations. The structured reasoning paradigm ensures transparency and reliability of KG-based reasoning systems, thereby providing a foundation for responsible AI research (Bugueño et al., 2024).

KGs have evolved significantly to meet the increasing complexity and dynamism of real-world applications (X. Jiang et al., 2023). Initially, static knowledge graphs (SKG) stored knowledge that did not change over time, and dynamic knowledge graphs (DKG) then introduced update mechanisms to reflect evolving environments. Temporal knowledge graphs (TKG) incorporated a time dimension to represent the temporal dependency of knowledge, and event knowledge graphs (EKG) further integrated events along with their evolution over time (S. Xu et al., 2023). EKGs can summarize changes in entity relationships with high-dimensional semantic features. Their sequential and causal relationships also support the prediction of event sequences, making them particularly suitable for applications such as traffic accident prediction (Guan et al., 2022). Future developments are expected to include multi-modal KGs, enriching their semantic representation capabilities.

Constructing a domain-specific KG involves three main steps: knowledge extraction, fusion, and completion (L. Zhong et al., 2023). In this process, ontologies are usually introduced to define concepts, entities, and relationships within a specific domain, thereby forming a structured knowledge hierarchy. The ontology defines domain-specific concepts, relationships, and rules, corresponding to the terminological box (T-Box) in description logic. The associated data are structured based on the ontology, corresponding to the Assertion box (ABox) in description logic, enabling knowledge reasoning and integration. Extraction derives entities and relationships from

unstructured data, while fusion aligns and integrates multiple sources. Completion addresses the inherent incompleteness of real-world KGs by inferring missing triples through reasoning methods (J. Zhang et al., 2021), e.g., link prediction.

To leverage the reasoning capability of neural networks, knowledge graph embedding (KGE) techniques represent entities and relations as vectors (Delong et al., 2023). These embeddings support efficient downstream tasks such as link prediction and semantic retrieval. Methods have evolved from early translation-based models (e.g., TransE) to deep learning–based approaches using graph neural networks (GNNs), improving the ability to handle noisy, multimodal, and dynamic driving scenarios (Bordes et al., 2013; B. Yang et al., 2014; Dettmers et al., 2018; B. Jiang et al., 2019).

2.1.2 KGs in Autonomous Vehicles

KGs have been widely applied across various aspects of AV, involving traffic flow prediction and management, perception enhancement and path planning, scene description and test case generation, as well as the prediction of hazardous events in dynamic scenarios (Zipfl et al., 2023; Luettin et al., 2022; Riedmaier et al., 2020; D. Xiao et al., 2023). The existing research on vehicle-related ontologies spans from high-level concepts of intelligent transportation system management to specific scene descriptions for ADSs, encompassing aspects such as roads, traffic signals, and vehicle behaviors (Katsumi and Fox, 2018). These ontologies provide semantic support for traffic flow prediction, route planning, and scene generation but remain insufficient in risk identification and causal analysis with a safety-oriented focus (Htun and Fukuda, 2024), and thus lack dynamic reasoning ability to complex scenarios and unforeseen events.

2.2 Large Language Models in AV

2.2.1. Large Language Models

Large Language Models mainly refer to transformer-based models for Natural Language Processing tasks (NLP), which contain millions or even billions of parameters and are trained on a massive corpus. Benefiting from their wide coverage of training data, LLMs exhibit impressive language understanding, generation, and zero-shot transfer abilities, which are not present in previous small models. The GPT series is the work that made LLM widely known to the public (Achiam et al., 2023; Brown, 2020; Radford et al., 2019; Radford, 2018). Other popular LLMs include BERT (Devlin et al., 2019), T5 (Raffel et al., 2020), Qwen (Bai et al., 2023a), LLaMa (Touvron et al., 2023), DeepSeek (D. Guo et al., 2025) etc. The processing ability can be further extended to multi-modal, such as vision and point cloud, by corresponding modality encoders, projectors, and training on multi-modal datasets, i.e., Multi-Modal Large Language Models (MLLMs) (Qin et al., 2025; J. Li et al., 2022; Bai et al., 2023b;Y. Shen et al., 2023). This multi-modal capability enables LLMs to handle a broader range of tasks beyond language, particularly in domains such as AV that involve multi-modal perceptual inputs and complex reasoning.

Although pre-trained MLLMs possess extensive knowledge from large corpora, they may struggle with complex tasks or generating strictly formatted outputs in specialized domains. A common solution is to **Fine-Tune** pre-trained LLMs on domain-specific datasets, enhancing their capabilities in task-corresponding areas like AV (Han et al., 2024). Compared to training from scratch, fine-tuning requires less data and computational resources while leveraging the base model's language understanding, reasoning, and internal knowledge (Xing et al., 2024).

Another key technique to enhance MLLMs' performance is **Prompt Engineering**, which involves designing, optimizing, and managing prompts to ensure accurate and efficient execution of user instructions (Sahoo et al., 2024). Early methods provided few input-output examples to induce task understanding, known as one-shot or few-shot prompting (Lee et al., 2024; Yong et al., 2023). Recent advances focus on chain-of-thought (CoT) prompting (J. Wei et al., 2022; Diao et al., 2023), decomposing problems into ordered reasoning steps to guide reasoning paradigm and model outputs. Variants include automatic CoT (Z. Zhang et al., 2022), Logical CoT (X. Zhao et al., 2024), Tree of Thought (Long, 2023), and Graph of Thought (Besta et al., 2024).

A major challenge in deploying MLLMs on AVs is the **Hallucination** problem, that models generate incorrect or harmful outputs inconsistent with human values owing to their large but unfiltered training data (L. Huang et al., 2025). To mitigate this, LLM outputs are usually verified against external retrievals (Min et al., 2023), internal consistency (Dhuliawala et al., 2024), or uncertainty measures (Varshney et al., 2023). Among them, **Retrieval-Augmented Generation (RAG)** (Lewis et al., 2020) improves MLLMs by incorporating relevant external knowledge, enhancing accuracy, enabling real-time information access, and reducing hallucinations (Ni et al., 2025). The detailed RAG procedure and applications will be discussed in Sec.6.2.2.

2.2.2 MLLMs in Autonomous Vehicles

MLLMs' ability in wide-world knowledge understanding, reasoning, and multi-modal processing provides promising application potential for solving complicated AV tasks, e.g., perception, scene understanding, decision-making, etc. For example, MLLMs trained on large-scale image–text data enable zero-shot recognition and broad category generalization, which remain highly challenging for traditional perception systems (X. Zhou et al., 2023; Parisot et al., 2023), as well as improving Bird's Eye View (BEV) perception by modeling global interactions (Q. Xu et al., 2025). For scene understanding, their contextual understanding ability is applied in the Visual Question Answering (VQA) task to infer object behaviors and intentions from visual inputs (Qian et al., 2024; Sima et al., 2024), and in the Captioning task to produce structured, scene-level descriptions that support downstream AV tasks (Malla et al., 2023; J. Fang et al., 2021). In the decision-making field, LLMs have been used to encode motion planning in natural language (Mao et al., 2023b), such as representing trajectories via way-point lists for better generalization (Mao et al., 2023a), or converting human commands into objectives and constraints for personalized planning (W. Huang et al., 2023).

With the rapid development of end-to-end (E2E) driving systems, some researchers use MLLM as the main backbone to handle the whole autonomous driving process with promising onboard deployment, e.g., (R. Zhao et al., 2024; Z. Guo et al., 2024; X. Tian et al., 2024). Through MLLM, ADSs can provide

more explainable decisions, deal with many zero-shot long-tail scenarios, and revolutionize the human-vehicle interactions (Z. Xu et al., 2024b).

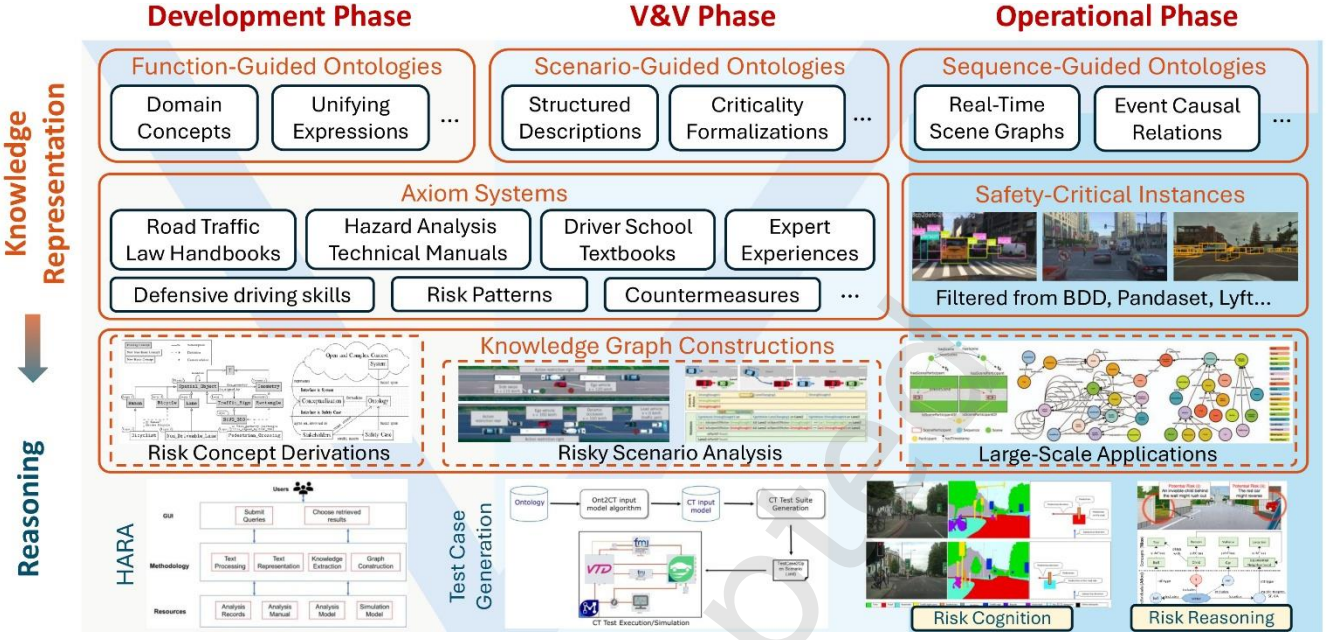


Fig. 4. Research on KGs for AV Safety structured by the applications

3 KGs for Autonomous Vehicle Safety

Research on applying KGs to AV safety is still in its early stages, limited by scarce annotated data and challenges in real-time processing. Despite this, the field shows great promise, especially for dynamic scene understanding and risk prediction in complex scenarios with high safety demands. The use of KGs in AV safety usually begins with constructing ontologies that embed safety semantics and define domain concepts, relationships, and constraints in a consistent manner. Building on this foundation, KG-based techniques apply these structures to support semantic integration, safety reasoning, and dynamic risk analysis across different phases of the system lifecycle.

This chapter aligns with the system safety lifecycle framework (International Organization for Standardization (ISO), 2018, 2022a, 2022b), as shown in Fig. 4, reviewing the application of KGs across three phases: functionality implementation and enhancement during the development phase, functionality testing and optimization during the verification and validation (V&V) phase, and functionality monitoring and maintenance during the operational phase of ADSs. KG-based research directions for AV safety are summarized in Table.2, which organizes representative tasks across safety lifecycle, and reports for each task a set of representative works together with the underlying data modalities, typical evaluation settings, and the major practical limitations and challenges observed in current studies.

Table 2. Summary table of KG-supported safety tasks across the lifecycle.

Phase	Task	Representative works	Data modality	Evaluation Setting	Main Challenges / Limitations
Dev.	Unify and complete data annotations	Halilaj et al. (2022, 2023); Nag Chowdhury et al. (2021)	Multiple datasets with structural annotations	Case study, downstream model training	Costly manual alignment
	HARA, rule formalization	L. Zhao et al. (2015b); M. Zhao et al. (2025)	Textual standards and documents, textual rulebooks	Case study, rule satisfaction	Ambiguous knowledge and rules
	Accident analysis	L. Zhang et al. (2022); Rakhmawati et al. (2022)	Multi-modal accident data, structural accident reports	Case study	High-dimensional, heterogeneous data
V&V	Structured scenario description	Schuldt (2017); Scholtes et al. (2021); S. Wu et al. (2021)	Scenario DSL	Case study	Standard compatibility and scenario complexity
	Test case generation	Feilhauer (2018); Panagiotaki et al. (2025); J. Zhao et al. (2025)	Scenario DSL, constraints	Scenario feasibility, criticality and rarity	Combinatorial explosion, physical feasibility, unknown generation
	Critical scenario assessment	S. Y. Yu et al. (2021); Westhofen et al. (2022, 2023)	Scene graphs	Assess accuracy	High false-positive rate, limited to pre-defined risk patterns
Ope.	Real-time scene graph construction	Suryawanshi et al. (2019); Qiu et al. (2021)	Perception and localization outputs	Missed part, response time	Real-time EKG construction for events and causality cognition
	Situation awareness and risk identification	Buechel et al. (2017); Takahashi et al. (2016)	Real-time scene graphs, risk pattern rules or models	Case study, risk detection accuracy	Limited risk coverage, weak generalization

3.1 KGs for AV Safety During Development Phase

3.1.1. Function-guided Ontology

In the development phase of AV, many functionalities rely on models trained on datasets. However, manual annotations are often incomplete, lacking small but safety-critical objects or temporal, spatial, and causal relationships. Besides, the inconsistency across datasets due to divergent labeling schemes can lead to the omission or misinterpretation of safety-critical scenarios. Ontology design guided by functional requirements and safety requirements enables the construction of task-relevant KGs that embed safety semantics and explicitly capture spatial, temporal, functional, and causal relationships to address these issues. Representative designs include multi-layer KG frameworks integrating heterogeneous datasets for perception and prediction tasks (Halilaj et al., 2022, 2023), core ontologies supporting semantic traffic rule reasoning with SWRL (L. Zhao et al., 2015a), and ontology-aware annotation tools that generate RDF triples during labeling (Warren et al., 2021).

To mitigate incompleteness, well-designed ontologies enable KGs to infer missing entities and relations using commonsense and domain knowledge, such as predicting a "cyclist" when a "person rides bicycle" relation exists (Gouidis et al., 2019; Nag Chowdhury et al., 2021), or identifying a "child" from the presence of a rolling ball (Wickramarachchi et al., 2021, 2022). For inconsistency, ontologies integrate heterogeneous datasets into a unified semantic framework, enabling consistent labeling and improving interoperability across multiple sources (Tran et al., 2022; Henson et al., 2019; Schmid et al., 2019).

3.1.2. Key Techniques

Hazard analysis is essential in the development phase to identify potential hazards, assess associated risks, and ensure that safety requirements are satisfied before deployment. The structural characteristics of KGs support knowledge management in established processes such as STPA, enabling hazards and reasoning procedures to be formally represented and systematically analyzed (Roh and Lee, 2015; X. Liu et al., 2021). Additional KG structures, such as operational scenario ontologies, can be introduced to extend traditional hazard analysis, enhancing analysis completeness by capturing scenario-specific hazards and control relationships (M. Zhao et al., 2025). Furthermore, ontology-based methods can integrate hazards and mitigation strategies into a unified representation, even across different domains (A. Xiao et al., 2024), which improves hazard identification and ensures vocabulary consistency among experts and stakeholders (Ali et al., 2024).

During the development phase, it is essential to incorporate road traffic rules to ensure that ADS can interpret and follow legal requirements under diverse operational conditions. The finite number of rules and their feature to be decomposed into atomic components (W. Yu et al., 2024) make them inherently suitable for knowledge-driven processing using KGs, where ontologies are employed to formalize traffic regulations. Early studies (Hülsen et al., 2011; L. Zhao et al., 2015b) adopted simplified ontology structures to represent specific rules for different ADS levels, e.g., AEB, while some approaches (Regele, 2008) employed abstract traffic models incorporating relations such as "Conflicting" and "Neighboring" to support compliance decisions in more complex scenarios. Subsequent research enhanced these methods by incorporating location- and domain-specific factors to dynamically construct sub-KGs for different scenarios, enabling real-time reasoning with greater efficiency (L. Zhao et al., 2015c, 2016; Hovi and Ichise, 2020). Other work addressed exceptional conditions through ontology-based softening rules, allowing context-dependent exceptions such as crossing solid lines to bypass obstacles (Morignot and Nashashibi, 2012).

Accident analysis in the development phase provides essential insights into the causes, patterns, and contributing factors of road traffic incidents, serving as a foundation for safety improvement and risk mitigation in ADSs. The structured nature of KGs supports the integration and reasoning over heterogeneous accident datasets, enabling experts to interpret large-scale and complex information (D. Yu et al., 2024; L. Zhang et al., 2022; Yue et al., 2009). Incompleteness in accident records, such as missing spatial attributes or causal links, can be addressed through ontology-based frameworks that enrich datasets with inferred relationships, as demonstrated by ontology-driven risk mapping approaches using spatial clustering to retrieve location-specific risk data (J. Wang and X. Wang, 2011; C. Liu and Yang, 2022). In addition, KGs help address inconsistencies in accident data from different sources through unified ontological representations, such as integrating official reports, sensor data, and social media information (Rakhmawati et al., 2022; Barrachina et al., 2012), and can be further enhanced with advanced KG embeddings for high-dimensional, multi-source, heterogeneous, and spatio-temporal datasets (X. Liu et al., 2025).

3.2 KGs for AV Safety During V&V Phase

3.2.1. Scenario-guided Ontology

The V&V phase of ADS relies on scenario-based methods to assess the scenario criticality, enabling the identification of high-priority scenarios and the subsequent test case generation. By focusing on such high-criticality scenarios, it reduces the scope and cost of testing while improving the detection of system defects (Feng et al., 2023). Effective V&V requires scenario-guided ontology design, in which functional requirements are decomposed into specific scenario components through a top-down process.

Table 3. Evolution of scenario ontology for autonomous vehicle V&V Phase.

References	Ontology Layers	Compatibility / Standards
Schuldt et al. (2013)	*(1) Road geometry and topology, (2) Situation-dependent road adaptation, (3) Traffic situation, (4) Environment*	/
Geyer et al. (2014)	*(1) Ego vehicle, (2) Scenery, (3) Scene, (4) Situation, (5) Scenario, (6) Driving mission, (7) Route*	/
Ulbrich et al. (2015)	*(1) Scene (environment snapshot), (2) Situation (goal- and value-based interpretation), (3) Scenario (temporal sequence of scenes)*	ISO 26262 V-model
Schuldt (2017)	*(1) Base road network, (2) Situation-specific adaptation of the road network, (3) Description and control of actors, (4) Environmental conditions*	ISO 26262 functional safety

Bagschik et al. (2018)	*(1) Road level (L1), (2) Traffic infrastructure (L2), (3) Temporary manipulation of L1 and L2 (L3), (4) Objects (L4), (5) Environment (L5)*	ISO 26262 functional safety
Sauerbier et al. (2019)	*(1) Road geometry, (2) Road furniture and rules, (3) Temporal modifications and events, (4) Moving objects, (5) Environmental conditions, (6) Digital information*	OpenDRIVE / OpenSCENARIO
Scholtes et al. (2021)	*(1) Road network and traffic guidance objects, (2) Roadside structures, (3) Temporary modifications of L1 and L2, (4) Dynamic objects, (5) Environmental conditions, (6) Digital information*	OpenDRIVE / OpenSCENARIO
S. Wu et al. (2021)	*(1) Road structure, (2) Infrastructure, (3) Temporary manipulation, (4) Traffic participants, (5) Nature environment, (6) Intelligent connected, (7) Vehicle inside information*	ISO 21448

The key to constructing scenario ontologies is to determine the layers from which a scenario should be comprehensively described. Its evolution is shown in Table.3. In early research, the definitions and boundaries of different layers were relatively vague, but they were generally organized around several aspects, such as static scenes, dynamic information, and temporal interactions (Schuldt et al., 2013; Geyer et al., 2014; Ulbrich et al., 2015). Schuldt refined these models and developed a four-layer scenario ontology framework, consisting of the base road network, its adaptations, actor control, and environmental conditions (Schuldt, 2017). In later developments, the base road network was further differentiated into permanent and temporary road elements (Bagschik et al., 2018), and digital information was incorporated with the development of V2X and HDmap (Sauerbier et al., 2019). Scholtes et al. (2021) adapted this framework for urban traffic, explicitly distinguishing spatial (Layers 1–3) and temporal (Layers 4–6) components, while ensuring compatibility with OpenDRIVE and OpenSCENARIO standards. Building on this, S. Wu et al. (2021) introduced a seventh layer representing the ego-vehicle state, addressing algorithm failures and user misuse in SOTIF-related research. The six- and seven-layer architectures represent the current mainstream approach to scenario description, and the content of each layer can be further built, refined, and enriched through dataset-driven methods (C. Zhang et al., 2024; Kaleeswaran et al., 2019).

3.2.2 Key Techniques

Ontology and relevant KGs provide precise scenario descriptions for test case generation. By applying constraints such as traffic regulations and physical limitations, ontology-based methods can produce a large number of test scenarios (Geyer et al., 2014; W. Chen and Kloul, 2018), while the increasing size of the ontology can make exhaustive generation impractical. To address this, the combinatorial testing method can be applied by transforming concepts, attributes, and relationships from the ontology into input models (Wotawa and Li, 2018; Klück et al., 2018), which effectively covers diverse parameter combinations and interactions while retaining the ability to satisfy the specified constraints (Y. Li et al., 2020). These methods can directly generate edge test cases by assigning extreme parameter values (Feilhauer, 2018), but research on generating complex corner scenarios remains relatively limited, especially where the combinations of non-extreme individual parameters lead to criticality (Zipfl et al., 2023). Recent works have introduced accident causality into scene graphs, thereby generating scenes that are more representative of real-world accidents (J. Zhao et al., 2025).

KG-based scenario generation typically follows a paradigm of element composition under explicit constraints. With standardized DSL (domain-specific language) representations such as OpenSCENARIO, KG-generated scenarios inherently align with simulator-executable code, e.g., importing OpenSCENARIO descriptions into CARLA via its OpenSCENARIO interface. In contrast, the physical feasibility of KG-generated scenarios critically depends on the constraint design during generation, including hard physical constraints on speed, acceleration, curvature, timing, and collision-related conditions.

However, fewer studies directly report KGs' effectiveness in generating unknown, corner or safety-critical scenarios. For instance, Bogdoll et al. (2022) used an ontology to generate corner cases and reported 27 of 94 individuals were newly created in a new scenario, but did not further analyze the safety impact of them. GraphSCENE (Panagiotaki et al., 2025) reported that only 40–50% of the generated near-collision scenarios satisfy the intended criticality requirement across different driving agents. Instead, CRADLE (J. Zhao et al., 2025) reported a more outcome-oriented result: training with KG-generated scenarios reduced the model's collision rate by 74%.

Leveraging KGs for critical scenario assessment aids in identifying high-impact scenarios related to safety and functional validation, particularly in extreme or high-risk situations. These studies typically encode raw observations with spatial and temporal relations into a KG, and then apply GNNs to learn relational representations for classifying critical scenarios and their subtypes. For instance, S. Y. Yu et al. (2021) construct a KG capturing spatial interactions among traffic participants and leverage it to predict lane-change risk, reporting 87% accuracy on the 571-Honda naturalistic driving dataset. Similarly, Halilaj et al. (2021) builds structural driving data along with temporal relations such as occursAfter into KG, and uses MRGCN to classify different types of critical or normal scenarios, achieving an accuracy of 95%.

However, when scenario criticality is inferred solely from KG-defined relations, these approaches often suffer from a high false-positive rate, which similarly arises in enumeration-like KG-driven scenario generation methods (Zipfl et al., 2023). For example, Westhofen et al. (2022) report that KG-based filtering yields more than 1.74 potential conflicts per scene per participant, indicating substantial over-triggering. Consequently, KG-based scenario generation or critical-scenario mining is commonly paired with an additional rule-based filtering stage to suppress spurious positives and retain only physically plausible and safety-relevant cases. Early research uses ontologies to identify critical scenarios from split frames with predefined safety thresholds, such as emergency braking deceleration or Time-to-Collision (TTC) in cut-in scenarios (Westhofen et al., 2022, 2023). Subsequent developments expanded this by considering dynamic interactions and event causality analysis (Neurohr et al., 2021). The ontology is enriched in temporal by formally defining events like braking and activities like overtaking (De Gelder et

al., 2022). Methods like Metric Temporal Conjunctive Queries (MTCQ) enabled reasoning over dynamic, time-dependent scenarios, improving the critical scenario assessment (Westhofen et al., 2024).

3.3 KGs for AV Safety During the Operational Phase

3.3.1. Sequence-guided Ontology

During the operational phase, the absence of ground-truth labels and complete scene replays necessitates real-time scene graph construction from perception and localization outputs. In this process, KGs and ontologies are leveraged to integrate prior driving experience and commonsense knowledge, dynamically compensating for or correcting deficiencies caused by incomplete or imperfect onboard perception and localization. A representative example is the work by BMW (Suryawanshi et al., 2019), who introduced a hierarchical ontology to process map data from heterogeneous sources, enabling semantic abstraction of map elements and spatial reasoning (Qiu et al., 2020a,b). To ensure data reliability, they further proposed the Map Quality Violation Ontology for detecting and correcting errors in map data (Qiu et al., 2021).

The aforementioned methods facilitate the real-time construction of DKGs but inevitably degrade into SKGs at each definite moment, lacking mechanisms to preserve environmental context and analyze risk causality. Spatiotemporal scene graphs address this limitation by enabling TKG construction through the capture of key frame information and retention of temporal dimensions via sliding time windows. However, research on the real-time construction of EKGs that further incorporate events and causality remains under development. Existing studies primarily focus on macro-level transportation applications such as traffic management (Xiong et al., 2018; W. Luo et al., 2020) and accident response (Muppalla et al., 2017; Y. Li and Liu, 2022), with limited attention to fine-grained, real-time causal analysis in driving scenarios for ADS.

3.3.2 Key Techniques

During the operational phase, KGs can enhance the scene understanding and situation awareness of ADSs by providing a structured representation of the road network (Hummel et al., 2008; L. Huang et al., 2019; Kunze et al., 2018), object attributes (Wickramarachchi et al., 2020), spatial and temporal relations (Wickramarachchi et al., 2020; Dianov et al., 2015; Oltramari et al., 2020), interactions (C. Li et al., 2020), and events (P. Xiao et al., 2021; Matheus et al., 2003). Such enriched knowledge enables ADSs to perceive their operational situation more comprehensively and to reason about potential future hazardous states, such as chained braking events (Armand et al., 2014) or potential conflicts (Buechel et al., 2017), thereby supporting proactive risk mitigation, safer decision-making and spatial-temporal resilience (Dui et al., 2025).

By incorporating risk representations and safe-driving knowledge into general ontologies and TKGs, KGs can support onboard risk identification during the operational phase. Early studies relied on predefined risk patterns to identify risks, mapping each pattern to corresponding risk levels (W. Fang et al., 2020; B. Zhong et al., 2020). Representative examples include the analysis of pedestrian rule violations (Mohammad et al., 2015) and the development of over 200 risk rules by Wolf et al. (2019) for complex risk assessment. With risk patterns formalized like SWRL, ontologies can further support abductive reasoning for risk inference (Inoue et al., 2015), while typical safety metrics such as TTC can be integrated in this procedure to more accurately prioritize imminent threats (Takahashi et al., 2016). However, manually designed risk patterns are inherently difficult to generalize and enumerate in complex, open-ended driving scenarios. To overcome these limitations, data-driven and embedding-based approaches have been introduced, leveraging scene graphs and KGE models to identify risk nodes and infer potential hazard patterns, demonstrating strong potential for handling endless long-tail scenarios (S. Y. Yu et al., 2021; Malawade et al., 2022; C. Li et al., 2023; Suman et al., 2023).

Table 4. Representative industrial applications of ontologies and KGs for autonomous vehicle safety. Items without specified scales are not publicly available.

Phase	Enterprise	Knowledge Source	Ontology	Knowledge Graph	Primary Contribution
Dev.	Bosch Germany (Hülsen et al., 2011)	Expert experience	Intersection ontology	/	Right-of-way determination
	Toyota Japan (L. Zhao et al., 2015b,c; 2016)	Specific road segment	Map & car & control ontologies (14 rules)	Real-world cases (37566 triples)	Fast decision making
	Geely China (X. Liu et al., 2021)	Hazard analysis technical manuals	Risk & cause & countermeasure ontologies	/	Export safety requirements
	Bosch USA (Wickramarachchi et al., 2021, 2022) & Bosch Germany (Halilaj et al., 2023; Młodzian et al., 2023)	nuScenes & BDD & Pandaset	Core concepts (8 classes & 22 properties)	Global (~43m + ~27m + ~3.3m triples)	Unify semantic expression
V&V	ZF AG Germany (Scholtes et al., 2021)	Expert experience	6-layer scenario model	/	Structured description
	AVL List GmbH Austria (Klück et al., 2018; Y. Li et al., 2020)	Simulation data	AEB function ontology (523 parameters)	/	Efficient combination testing
	Bosch Germany (Neurohr et al., 2021; Westhofen et al., 2022, 2023)	inD dataset	Automotive urban traffic ontology (6405 axioms)	Identified risks per frame	Formalize and recognize criticality phenomena
Ope.	Renault France (Armand et al., 2014)	Expert experience	Core ontology (14 logical rules)	/	Chain reaction reasoning
	Denso Japan (Inoue et al., 2015; Takahashi et al., 2016)	Driver school textbook	Scenario risk (32 causal relations & 11 risk patterns)	Simulation case study	Risk abductive reasoning
	Continental AG Germany (Buechel et al., 2017)	Multinational traffic law handbooks	Traffic regulations ontology	/	General definition of rules

BMW Car IT Germany (Suryawanshi et al., 2019; Qiu et al., 2020a,b; 2021)	Self-collected data	Layered map ontology (879 axioms) & Map quality violation ontology (79 rules)	Real map tiles (~430k + ~146k triples)	Sliding time window to update the local map
Bosch USA (Oltramari et al., 2020; Wickramarachchi et al., 2020)	nuScenes & Lyft	Scene graph ontology	KG embedding (~5.95m + ~3.94m triples)	Infer missing information

Despite recent progress, operational KG-based AV safety research still largely centers on SKGs and DKGs, while onboard TKG-based safety reasoning remains scarce. A practical workflow starts with online TKG construction that aligns objects and relations across frames via tracking and association pipelines to form temporally linked scene graphs. It then performs temporal-aware reasoning on the constructed TKG, for example by applying temporal GNNs for feature learning and risk inference. A representative work is nuScenesKG (Mlodzian et al., 2023) which introduces temporal scene connectivity and releases a trajectory prediction dataset, supporting further TKG-based trajectory prediction model training (Z. Sun et al., 2024; Z. Wang et al., 2024). Extending this construction to reasoning pipeline to operational safety tasks such as risk recognition and scene understanding is promising because these tasks depend on temporal consistency and interaction evolution to detect early risk cues and prevent hazardous behaviors.

EKG studies remain scarcer in AV domain, whereas related efforts in traffic management and unmanned robotics provide a practical paradigm. In general, an EKG reasoning workflow can be organized into three stages: offline definition of an event ontology, online event recognition, and online reasoning over the EKG. For event ontology construction, a feasible way is to explicitly link unintended behaviors to hazard events, for example, associating abnormal deceleration or abrupt lane changes with blind-spot related risks, thereby enabling structured mapping from observable behaviors to safety-relevant hazards (T. Zhang et al., 2026). Online event recognition often relies on predefined rules, including simple kinematic thresholds or richer constraints in MTL or SWRL format (Urbieta et al., 2023). Reasoning over complex EKGs usually relies on probabilistic or neural models, such as using Bayesian networks to infer the impact of uncertain events on unmanned robot task (H. Yao et al., 2022), or heterogeneous graph transformers to predict subsequent events after an initial incident in traffic management (Y. Li and Liu, 2022).

3.4 Industrial Progress of KGs for AV Safety

Beyond academic advances, industrial efforts have also substantially shaped KG-based safety engineering, as KGs naturally support traceability, auditability, and controlled updates required by practical safety processes. A summary of representative industrial applications is shown in Table.4.

Industrial adoption of KGs is more evidently driven by the safety development lifecycle. In the development phase, KGs are used to organize safety knowledge and support structured hazard analysis, e.g., Geely extracts multiple ontologies from technical manuals to facilitate hazard analysis and export safety requirements (X. Liu et al., 2021). In the V&V phase, ontologies are commonly employed to formalize scenario elements and enable combinatorial construction of complex test cases, as exemplified by AVL's industrial-grade AEB testing where an AEB function ontology organizes 5233 parameters for efficient combination testing and generation of critical cases (Klück et al., 2018; Y. Li et al., 2020). In the operational stage, ontologies built upon traffic regulations and driving experience are frequently used for rule-based compliance and right-of-way checking. However, constructing reliable, real-time scene topology remains challenging, and BMW addresses this by dynamically extracting spatiotemporal map fragments from layered maps and further introducing map quality violation ontologies to support consistent updates (Suryawanshi et al., 2019; Qiu et al., 2020b,a, 2021).

In practice, industry often treats KGs as a governed backbone that links regulations, requirements, hazard analyses, and test evidence to enable E2E traceability, while supporting scenario engineering tasks such as compositional synthesis, coverage analysis, and risk-oriented filtering. Current trends emphasize lifecycle-aware governance and interoperability across scenario domain-specific languages, simulators, and safety toolchains. Hybrid workflows are also emerging, where KGs provide auditable structure and LLMs serve as an auxiliary layer for retrieval and human-facing interaction, with retrieval grounding and verification used to preserve safety-critical reliability.

4 LLMs for Autonomous Vehicle Safety

Unlike KGs, LLMs provide powerful capabilities for processing unstructured data, capturing implicit knowledge, and enabling flexible reasoning across diverse modalities, making them effective in addressing long-tail scenarios and complex interactions. Large-scale pre-trained MLLMs already demonstrate strong general reasoning and perception abilities, with encouraging results on certain AV safety tasks. However, in this specialized domain, fine-tuned models are generally more reliable, as task-specific datasets provide essential supervision to align LLMs with driving semantics, safety rules, and domain knowledge. In addition, effective task decomposition and tailored prompt design further enhance performance by simplifying individual queries and facilitating the use of domain-specific information. Yet, the task decomposition, the logic design of prompts for each subtask, and the formulation of inputs at the word or sentence level remain highly task-specific, lacking a unified design framework. Accordingly, this section reviews the application of LLMs for AV safety across the development, V&V, and operational phases, as in Fig. 5. LLM-based research directions for AV safety are summarized in Table.5, which organizes representative tasks across safety lifecycle, and reports for each task a set of representative works together with the underlying data modalities, typical evaluation settings, and the major practical limitations and challenges observed in current studies.

Table 5. Summary table of LLM/MLLM-supported safety tasks across the lifecycle.

Phase	Task	Representative works	Data modality	Evaluation Setting	Main Challenges / Limitations
Dev.	HARA, rule formalization	Zheng et al. (2023); Nouri et al. (2024a); Qi et al. (2025)	textual hazard documents, human feedback, textual rulebooks	Case study, rule satisfaction, QA accuracy	heavy workflow engineering
	accident analysis	Jaradat et al. (2024); Y. Lu et al. (2024)	multi-modal accident datasets, unstructured online content	Accident factor accuracy, case study	complex accident causal analysis
V&V	critical scenario assessment	Y. Yang et al. (2024); Tami et al. (2024)	heterogeneous driving data, semantic expert requirements	Assessment accuracy	higher latency, hallucinated judgements
	DSL-based test scenario generation	J. Zhang et al. (2024); X. Cai et al. (2026)	semantic descriptions and requirements, DSL format constraint	scenario accuracy, scenario feasibility, ADS performance	constraint satisfaction to expert requirements and format
	visual test scenario generation	S. Gao et al. (2024); G. Zhao et al. (2025a); Y. Zhou et al. (2026)	expert-language prompts, initial images or action-trajectory specifications	physical feasibility, trajectory-prompt alignment, video consistency	physical and expert constraint satisfaction, unreal videos, high cost
Ope.	anomaly detection	Shafiq et al. (2024); Sinha et al. (2024)	visual inputs, trajectories	Detection accuracy	real-time cost, poor adaptability in long-tail
	hazard object identification	G. Wang et al. (2024); H. Zhang et al. (2024)	visual inputs	Detection accuracy	real-time cost, imprecise spatial localization
	risk recognition	W. Zhong et al. (2025); Meng et al. (2025)	multi-modal perceptual inputs, kinematic information, task-specific prompts, etc.	safety-related task accuracy, safety-related driving metrics, closed-loop metrics	real-time cost, complex scenario understanding paradigm
	Adversarial attack defense	Song et al. (2024); Aldeen et al. (2024)	visual inputs, perception outputs, kinematic information, V2X data	Attack detection accuracy	real-time cost

4.1 LLMs for AV Safety During Development Phase

4.1.1. Datasets

Regulatory and standard documents inject authoritative institutional knowledge into LLMs. Standards such as the United Nations Economic Commission for Europe (2021) Regulations R157, ISO 26262, ISO 21448, and ISO 34502 define structured requirements for AV safety and verification, serving as key corpora for fine-tuning LLMs in safety engineering and supporting hazard analysis tasks such as STPA. Meanwhile, road traffic laws and driver manuals provide behavioral norms for real-world driving. In addition to raw legal texts, curated QA datasets such as TrafficSafety2K (Zheng et al., 2023) offer structured supervision that helps align LLMs with compliance reasoning, supporting functional definition and violation scenario modeling during the development phase. Besides, technical documents generated during the system development process can constitute another key data source for understanding the structure and functionality of specific ADS, while such materials are often for internal use and rarely publicly available.

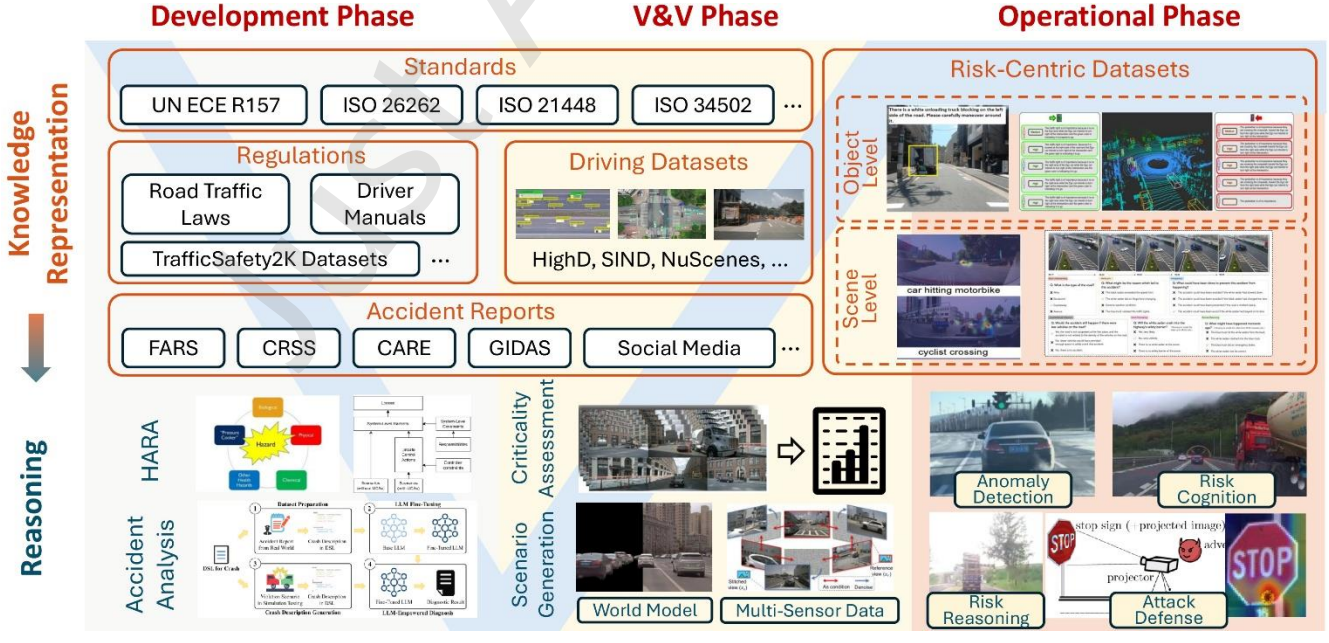


Fig. 5. Research on LLMs for AV Safety structured by the applications

Textual accident databases and reports provide LLMs with accident instances and contextual information, enabling models to learn accident structures and identify human or system errors, thereby offering semantic support for further failure mode analysis and module development. For example, the Crash Report Sampling System (CRSS) (National Highway

Traffic Safety Administration (NHTSA), 2024a) and the Fatality Analysis Reporting System (FARS) (NHTSA, 2024b) cover accidents of varying severity with detailed contextual descriptions, as well as CARE (a Community Road Accident Database) (European Commission, 2024), GIDAS (German InDepth Accident Study) (German in-depth accident study cooperation, 2024), etc. Besides, social media accident records include a large amount of information absent from conventional reports. While their multimodal and semantically rich characteristics pose challenges for traditional methods, they are particularly suitable for MLLMs (Z. Liu et al., 2023; Jaradat et al., 2024).

4.1.2 Key Techniques

Integrating LLMs into the development phase helps reduce the labor-intensive expert effort typically required for complex hazard analysis, especially under the current rapid iteration of onboard hardware and software. Early research applied LLMs to automate and manage HARA procedures such as STPA within existing analysis frameworks, aiming to substitute specific expert-driven steps with LLMs through well-designed prompts for improved efficiency (Raeisdanaei et al., 2025; Nouri et al., 2024b; Diemert and Weber, 2023). Later studies enhanced applicability by decomposing tasks into modular subtasks, designing subtask-specific prompts, and transforming system models to improve compatibility with LLM reasoning, thereby constructing structured and LLM-adapted workflows for tasks such as safety requirements specification (Nouri et al., 2024a; Geissler et al., 2024). To improve the reliability and practicality, subsequent studies explored various expert–LLM collaboration strategies, including injecting prior knowledge, assessment proposals, and illustrative examples before inference (Hillen et al., 2024; Odu et al.), conducting human review and refinement of LLM outputs (Sivakumar et al., 2024), and incorporating feedback through recurring duplex collaboration loops (Qi et al., 2025).

In the development phase, MLLMs can assist experts in analyzing historical accidents by extracting key features such as fatality, severity, and crash type, thereby supporting system hazard analysis (Mumtarin et al., 2023). The diversity of accident data sources, including textual reports, images, and online content, can be effectively handled through MLLMs, enabling the integration and summarization of critical accident characteristics for system improvement (Jaradat et al., 2024; K. Wu et al., 2024; R. Zhang et al., 2025). To enhance analytical performance, some researchers have proposed unified data structures that standardize multimodal accident descriptions, allowing consistent reasoning about causes and prediction of detailed outcomes such as accident type, severity level, and injury count (Kung et al., 2024; Fan et al., 2024). A representative effort is DIAVIO (Y. Lu et al., 2024), which introduces a domain-specific language (Crash DSL) to describe crash events using five components: Road, Environment, Obstacles, Vehicles, and Diagnosis.

4.2 LLMs for AV Safety During V&V Phase

4.2.1. Datasets

During the V&V phase of ADS, LLMs have been widely adopted to generate critical scenarios according to experts' requirements and constraints, as well as to construct these test scenarios with simulators or as videos. Existing V&V efforts typically leverage pre-trained LLMs or only fine-tune them with a limited amount of custom-built data, relying primarily on prompt design and sub-task decomposition to activate the LLMs' inherent capabilities in scene understanding and task execution.

In general, the datasets used for fine-tuning MLLMs during the V&V stage can be grouped into several categories. The first consists of standards and traffic regulations (ISO, 2018, 2022a, 2022b), which can guide MLLMs in producing standardized test scenarios, as well as evaluating risks of a scene, e.g., breaking the traffic rule. The second category includes conventional driving datasets, such as HighD, SIND and NuScenes (Qian et al., 2024; Y. Xu et al., 2022; Krajewski et al., 2018), which provide multi-view and multi-sensor information for training scenario generation models, thereby improving the realism of generated scenes. Finally, risk scenario datasets (Pham et al., 2020; Bao et al., 2020; Y. Yao et al., 2022) can enhance the quality of risk-oriented scene generation and also serve as a base for retrieving risk data that meet specific requirements, as well as accident reports (NHTSA, 2024a,b).

4.2.2. Key Techniques

On the critical scenario assessment task, LLMs can facilitate discovering safety-critical cases from large-scale driving data, which can be directly used for testing or serve as anchors for subsequent scenario generalization. Compared with rule- or ontology-based approaches that are deterministic and computationally efficient yet inherently bounded by pre-defined metrics, events, and relation templates, MLLMs offer a more flexible interface for long-tail risk discovery and assessment. By directly processing heterogeneous inputs such as replay videos and structured scenario descriptions, MLLMs can evaluate risk in a more holistic and context-aware manner, rather than being restricted to handcrafted indicators (e.g., TTC, acceleration, or simple behavior thresholds), predefined scenario layers, or ontology-specified event or causal chains (Y. Yang et al., 2024; Lu et al., 2024a; Tami et al., 2024). This broader understanding further enables MLLMs to identify key actors, interaction dynamics, and contextual factors, and to identify more realistic and comprehensive risk logic chains, thereby supporting targeted scenario generalization and uncovering long-tail hazards beyond pre-enumerated expert patterns in rule- or ontology-based approaches. (Tami et al., 2024; H. Tian et al., 2024).

In addition, MLLMs can translate hard-to-quantify expert requirements into explicit, customizable evaluation criteria, such as normalizing "smoothness" into reasonable acceleration bounds for scenario selection (S. Yao et al., 2025). Nevertheless, simple MLLM-based critical scenario assessment may introduce higher latency and hallucinated judgments. Thus, rule- or ontology-based checks are usually employed as lightweight initial filters and constraints to prune candidates and support downstream MLLM critical scenario assessment (Y. Gao et al., 2025; H. Tian et al., 2024).

Besides, LLMs can generate test scenarios directly based on expert descriptions or testing requirements. Early studies often produced textual scene descriptions in a DSL format. In addition to widely adopted standards such as OpenSCENARIO (X. Cai et al., 2026), some research proposed

task-specific DSLs that remove task-irrelevant layers and define the details of retained layers to improve efficiency (H. Tian et al., 2024; Fernandez et al., 2025). These DSL-formatted outputs can be directly parsed by simulators such as CARLA or SUMO, enabling automated construction of executable scenarios from LLM-generated descriptions (Petrovic et al., 2024; W. Xu et al., 2024; S. Li et al., 2024). More recently, with the advancement of video generation models and world model (Y. Wang et al., 2024a), several approaches utilize expert commands, trajectory-level inputs, or edits to real-world image sequences to generate multi-view images for validation with high reality (G. Zhao et al., 2025b; Y. Wei et al., 2024).

Compared with conventional test generation methods, LLM-based methods present better potential for exploring unknown corner cases and exposing latent failure modes of the ADS. In ChatScene (J. Zhang et al., 2024), the authors synthesize safety-critical scenarios for evaluating RL driving policies and report that their generated tests lead to a 15% higher collision rate than knowledge- and adversary-based baselines, while TrafficComposer (Tu et al., 2025) reports 33%–124% improvements in bug-finding. Training or fine-tuning the policy with these newly discovered scenarios can further improve robustness, where ChatScene reported a collision rate drop by 9% (J. Zhang et al., 2024). Beyond testing, the diversity of LLM-generated scenarios is also helpful for downstream perception module training (G. Zhao et al., 2025b). However, LLMs' instability and hallucinations can produce invalid or non-executable scenarios, especially when generating DSLs or simulator code. Text2Scenario reports an executable success rate of 87.31% when generating OpenSCENARIO-format DSL (with GPT4) (X. Cai et al., 2026), whereas H. Yang et al. (2025) report only around 50% running success rate when directly generating MATLAB and Simulink code. Finally, even when executable, LLM-generated scenarios may be semantically inconsistent with the prompt intent. For example, TrafficComposer reports 97.0% accuracy on its scenario construction evaluation, implying about 3% inaccurate scenarios (Tu et al., 2025).

In addition, for test-scenario video generation models conditioned on expert-language prompts, initial images or action–trajectory specifications, whether the synthesized video satisfies physical feasibility directly determines the validity of downstream testing. In general, physical feasibility can be assess along the following aspects: whether the generated trajectory is plausible with kinematic limits and reasonable with common patterns, whether the video evolution aligns with the provided action or trajectory condition, and whether foreground agents and background structure remain coherent across frames and across diverse scenes. For kinematic constraints, existing studies often integrate explicit kinematic models into the generation pipeline so that the model jointly accounts for motion states such as velocity, acceleration, and steering (Hanselmann et al., 2022; J. Kim et al., 2022). Some works additionally introduce motion-related loss signals to improve kinematic plausibility (S. Gao et al., 2024). Beyond kinematic constraints, motion signals can further be encoded in the 4D spatiotemporal space by constructing 4D occupancy grids or more general trajectory-injected latent representations as intermediate features to guide video generation (Z. Yang et al., 2025; J. Zhu et al., 2026). These features can act as bases to instruct the video generation, therefore enhancing trajectory-video consistency. Moreover, generating multi-frame and multi-view visuals from a shared 4D feature evolution helps strengthen cross-frame and cross-view consistency in safety-critical testing scenarios (G. Zhao et al., 2025a). To better evaluate the physical feasibility, DrivingGen (Y. Zhou et al., 2026) introduces a comprehensive benchmark for driving video generation evaluation. It proposes Trajectory Quality (TQ) to quantify motion plausibility under vehicle dynamics via signals such as jerk, lateral acceleration, and yaw rate. It further proposes ADE (Average Displacement Error) and DTW (Dynamic Time Warping) as complementary trajectory-alignment metrics, capturing the agreement with and without temporal alignment, respectively. DrivingGen also evaluates temporal consistency and agent-level consistency by checking their existence and features across frames. Together with data-distributional and visual quality indicators, it forms a comprehensive benchmark for assessing driving video generators under testing-oriented conditions. Related efforts such as DriveDreamer4D (G. Zhao et al., 2025a) provides multi-view consistency measures like Novel Trajectory Agent IoU (NTA-IoU) and Novel Trajectory Lane IoU (NTL-IoU), to quantify whether generated foreground objects and roadway semantics remain physically and geometrically consistent. Importantly, these metrics are most meaningful for within-benchmark, same-dataset comparisons. For example, under DrivingGen's reported results, Vista (S. Gao et al., 2024) achieves markedly stronger trajectory alignment (lower ADE/DTW), whereas VaViM (Bartoccioni et al., 2025) is reported to be more competitive on spatiotemporal coherence.

Given the complexity of test scenarios and the large number of involved elements, decomposing scenario generation into sub-tasks or applying chain-of-thought prompting, rather than generating the entire scene in one step, contributes to improving the quality of the overall V&V process. Although layered scenario representations naturally define sub-tasks as in Sec.3.2.1, such fine-grained decompositions are often overly detailed for practical applications. In most cases, road structure and traffic participants constitute the fundamental elements of test scenarios, while some studies further incorporate additional layers such as road infrastructure, behavior hierarchies, or temporal information (Q. Lu et al., 2024; J. Zhang et al., 2024; Chang et al., 2024)

4.3 LLMs for AV Safety During Operational Phase

4.3.1. Datasets

At the object level, a major subset of traditional risk-centric datasets targets open-world object detection, capturing numerous objects that are hard for conventional perception yet pose safety threats (Pinggera et al., 2016; K. Li et al., 2022; Lis et al., 2019; S. Shao et al., 2018). However, these datasets typically contain only bounding box and category label annotations, offering limited semantic information. With the development of LLMs, a growing number of language-based object-level risk understanding datasets are emerging. These datasets often construct natural language understanding frameworks for hazardous objects containing basic attributes, spatiotemporal information, and semantic relations, e.g., DRAMA, Rank2Tell, CODA-LM (Malla et al., 2023; Sachdeva et al., 2024; K. Chen et al., 2025). Some datasets further provide QA pair formats,

allowing direct use in supervised fine-tuning, e.g., HazardQA (Tami et al., 2025a). By learning from such datasets, LLMs can identify risk-relevant targets during operation, understanding the risk events with the targets, and supporting ADS in risk avoidance.

Beyond object-level risk cognition, scene-level risk assessment commonly relies on accident datasets. Traditional datasets record accidents through video segments with annotations focused on coarse visual labels, such as bounding boxes, object classifications, collision frame markers, and collision types, as in A3D (Pham et al., 2020). Some datasets, like DADA-2000 (J. Fang et al., 2021), CCD (Bao et al., 2020), and DoTA (Y. Yao et al., 2022), incorporate natural language descriptions of accident causes, but these are typically limited to a small set of predefined templates, lacking sufficient semantic information. With the advancement of MLLMs, natural language descriptions in datasets have become more comprehensive, capturing complex causal structures and fine-grained details. Organized in sub-task or QA formats, datasets such as MM-AU (J. Fang et al., 2024), SUTD-TrafficQA (L. Xu et al., 2021) and RoadSocial (Parikh et al., 2025) provide essential resources for MLLMs to perform road risk event understanding, causal reasoning, and risk anticipation. A summary of these datasets is shown in Table.6.

Table 6. Risk datasets used for LLM training in AV safety. CCTV: Closed-Circuit Television, referring to Fixed Roadside Monitor.

Dataset	Primary Task	Language Annotation	Viewpoint	Sensor Modality	Additional Annotations	Scale	Risk Level
DRAMA (Malla et al., 2023)	Joint risk localization	Caption + Q/A	dashcam	Video clips	Bounding boxes	~18k clips	Object-level
Rank2Tell (Sachdeva et al., 2024)	Object importance ranking	Caption + Q/A	dashcam	Multi-view video clips + LiDAR	Bounding boxes + Importance rank	116 clips	Object-level
CODA-LM (K. Chen et al., 2025)	Corner-case detection and evaluation	Caption	dashcam	Images	Bounding boxes	~10k images	Object-level
HazardQA (Tami et al., 2025a)	Hazard recognition	Q/A	dashcam	Images	/	~17k images	Object-level
CCD (Bao et al., 2020)	Traffic Accident Analysis	Pre-defined caption templates	dashcam	Video clips	Bounding boxes + scenario info + accident temporal info	~1.5k clips	Scene-level
DoTA (Y. Yao et al., 2022)	Traffic anomaly detection	Pre-defined caption templates	dashcam	Video clips	Bounding boxes + scenario info + accident temporal info	~4.7k clips	Scene-level
DADA-2000 (J. Fang et al., 2021)	Driver attention prediction in accidents	Pre-defined caption templates	dashcam	Video clips	Attention maps	~2k clips	Scene-level
MM-AU (J. Fang et al., 2024)	Accident understanding	Q/A	dashcam	Video clips	Bounding boxes + accident temporal info	~11k clips	Scene-level
SUTD-TrafficQA (L. Xu et al., 2021)	Traffic event QA	Q/A	dashcam + CCTV	Video clips	/	~10k QA pairs	Scene-level
PotentialRisk QA (J. Liu et al., 2025)	Potential risk detection	Caption + Q/A	dashcam	Video clips	Accident temporal info	~500 clips	Scene-level
DriveSOTIF (S. Huang et al., 2025)	Perception SOTIF challenges	Caption + Q/A	dashcam	Images	Bounding boxes	~1.1k images	Scene-level
RoadSocial (Parikh et al., 2025)	Road event understanding	Caption + Q/A	dashcam + CCTV + handheld	Video clips	/	13.2k clips	Scene-level

4.3.2 Key Techniques

During the operational phase, anomaly detection aims to identify unusual traffic participants or behaviors, thereby supporting risk awareness and safety decision-making. While traditional anomaly detection methods focus on small or out-of-distribution objects, LLM-enhanced approaches enable the recognition of complex interactions and abnormal behaviors. The basic idea is to design prompts that instruct LLMs to directly process video or image inputs and detect anomalies, such as unexpected vehicle maneuvers or abnormal traffic infrastructure (Cao et al., 2023; Shafiq et al., 2024; Elhafsi et al., 2023). Beyond visual data, LLMs can also incorporate additional modalities like trajectories to improve detection accuracy (Sinha et al., 2024). Detected anomalies can be encoded into structured representations for downstream safety decisions and emergency responses (H. Wang et al., 2024). As these methods leverage LLMs' internal understanding of normal patterns, they suffer from poor adaptability, long-tail distribution, and limited reasoning capacity. Recent research adopts a two-stage approach in which the induction stage learns normals from few-shot normal samples, and the deduction stage applies these patterns to identify long-tail anomalies (Y. Yang et al., 2025).

The broad world knowledge of MLLMs makes them particularly effective in identifying long-tail hazard objects which are often missed by traditional perception modules. Empirically, both close-world detectors and earlier open-world detectors that rely on latent-space clustering can struggle on corner-case object benchmarks. Taking CODA (K. Li et al., 2022) as an example, RetinaNet (T. Y. Lin et al., 2017) reports only 10.8% average recall (AR), and ORE (Joseph et al., 2021) achieves 8.3% AR on CODA corner-object detection, which is one representative open-world method at the time. In contrast, VLM-based approaches are not restricted to a fixed label

vocabulary and can leverage open-world semantics to better recognize rare hazardous objects, such as LLaVA-Grounding (H. Zhang et al., 2024) improving AR to 18.4%. Moreover, purely VLM-based grounding often suffers from imprecise spatial localization. To address this limitation, recent work has explored hybrid designs that couple VLM knowledge with stronger localization backbones. For example, MR-GroundingDINO (G. Wang et al., 2024) distills open-world knowledge from an MLLM into a detector and reports 56.1% AR on CODA, while Z. Lin et al. (2024) integrate a VLM with SAM by using VLM attention as prompts for SAM-based segmentation and localization, boosting CODA performance to 40.1% AR.

For the risk cognition task, the multimodal reasoning ability of LLMs enables direct processing of perception information to recognize basic risk patterns (Luu et al., 2024; Y. Wang et al., 2023). Traditional risk modeling techniques, such as risk fields, can be incorporated into the LLM's input to enhance its contextual understanding of safety-critical situations (Z. Zhou et al., 2026). Memory modules are also commonly introduced, allowing the LLM to retrieve and reason with similar events by leveraging stored collision or hazard data (Y. Wang et al., 2023; Z. Zhou et al., 2026; W. Zhong et al., 2025). Fine-tuning on task-specific datasets further enhances the ability of MLLMs to identify particular risk types, such as driver distraction or category-level inconsistencies in perception (K. Zhang et al., 2024; H. Pan et al., 2024). In addition, fine-tuning on general accident datasets enables LLMs to capture a broader pattern of risks at both the object and scene levels (Charoenpitaks et al., 2024).

During risk assessment, structured risk descriptions assist LLMs in understanding risk information, clarifying reasoning steps, and constraining output, functioning similarly to CoT. The description of risk is typically developed from six aspects or a subset thereof, including the risk time (when), risk objects (what), their attributes (which), their location (where), risk analysis (why), and avoidance strategies (how). For example, Tami et al. (2024) used a "what-which-where" structure to formalize risks, while H. Liao et al. (2024) involved when, what, and where, and the 5W framework proposed in Malla et al. (2023) lacks temporal information due to its use of static images.

To quantitatively evaluate VLM-based models for risk recognition in AV, researchers have developed a broad range of benchmarks. In open-loop evaluation, CODALM (K. Chen et al., 2025b) decomposes the tasks into general perception, regional perception, and driving suggestion in corner cases, and measures model outputs along multiple reliability-related dimensions, including accuracy, hallucination penalty, consistency, and rationality. The results indicate that fine-tuned VLMs can achieve strong performance in corner-case scenarios. DSBench (Meng et al., 2025) covers diverse in-cabin and out-of-cabin risk scenarios and adopts a GPT-4o-based score for assessment, whose results suggest that current VLMs remain insufficient in understanding driver behaviors and recognizing risks related to dynamic obstacles. SafeDriveQA (Yamazaki et al., 2025) treats traffic-signal and rule understanding as core test items, reporting that VLMs exhibit a violation probability above 60% under yellow-light conditions and over 30% of violations in car-following distance, which highlights clear limitations in compliant driving. For future-risk anticipation, RiskCueBench (S. Luo et al., 2026) formulates "whether a risk will occur" and "when it will occur" as measurable tasks, using metrics such as risk detection F1, reporting that popular VLMs achieve a baseline F1 of roughly 0.6–0.7 for risk anticipation.

Beyond task accuracy, some open-loop studies specifically examine the safety-related reliability and robustness of VLMs. DriveBench (S. Xie et al., 2025) investigates reliability and consistency of model outputs with visual evidence, showing that some VLMs overly rely on textual prompts and neglect visual inputs, leading to unreliable conclusions. RoboDriveBench (D. Liao et al., 2025) evaluates robustness to sensor corruption and prompt corruption, and uses metrics such as collision rate and L2 error under corrupted conditions to demonstrate that current VLMs can be highly sensitive to such perturbations For instance, under dark-condition sensor corruption, DriveVLM (X. Tian et al., 2024) can reach a collision rate of 0.5. While open-loop metrics may not directly translate to safety outcomes in interactive driving, closed-loop evaluations typically focus on task-level indicators such as route success rate (SR) and driving score (DS), where safety and compliance are usually the key components of these aggregate scores. Bench2Drive (Jia et al., 2024) adopts a pseudo closed-loop mechanism to better approximate closed-loop evaluation and reports DS and SR. For example, SimLingo (Renz et al., 2025), as a VLM-based model, improves DS from 64.22 to 85.07 and SR from 33.08% to 67.27% compared to representative E2E driving models. In CARLA Leaderboard 2.0 (CARLA Team, 2025), the official metrics include DS, Route Completion (RC), and Infraction Score (IS), where violations and collisions are reflected via IS and DS. On this benchmark, SimLingo achieves slightly higher DS and SR than the E2E baseline.

Compared to such simulation-based benchmarks, NuPlan (Caesar et al., 2021) provides a planning benchmark built upon real logs and maps closer to real-world driving, including safety metrics like collision-related measures, off-road, traffic-rule violations, and TTC. On nuPlan Close-Hard20, under the nonreactive setting, VLMPlanner (Z. Tang et al., 2025) improves TTC from 78.26 to 81.27 compared to the original baseline. Under the more challenging reactive setting, VLMPlanner improves the collision-related score by about 4 points over an advancing game-theory-based planner. In real-vehicle deployments, the safety impact of LLM is commonly quantified with metrics such as takeover rate, TTC, and aggregate driving scores (Cui et al., 2023, 2024a). However, the set of safety-specific and cross-system comparable metrics reports remains limited, while most of them predominantly emphasize human-in-the-loop instruction understanding or time latency rather than safety outcome.

In addition to the inherent risk within the scene, LLMs can also defend against external threats such as perception and V2X attacks. Traditional perception modules are vulnerable to adversarial attacks, including the injection of noise through invisible light (Köhler et al., 2021) or electromagnetic waves (Köhler et al., 2022), as well as adversarial patches or projected noise on traffic signs (Lovisotto et al., 2021), while LLM shows greater robustness. Thus, LLMs are used to verify the consistency between visual inputs and perception outputs, or identify decision misalignments, thereby mitigating potential

attack patterns (Song et al., 2024). A representative form of attack is Natural Denoising Diffusion (NDD) attacks, where attackers use diffusion models to create information that is harmless to humans but easily misclassified by ADS, for example, placing a sticker on the rear of a vehicle that causes the perception system to falsely detect a stop sign. MLLMs have shown stronger robustness to these attacks and can provide reliability judgements by verifying image-perception consistency or assessing the plausibility of motion data provided by V2X (Hu et al., 2025b; Aldeen et al., 2024).

5 Pros and Cons of KGs and LLMs for Autonomous Vehicle Safety

Having explored the wide contribution of KGs and MLLMs, this section will discuss their respective strengths and limitations in enhancing AV safety. KGs provide structured, reliable, and scalable knowledge representation, whereas LLMs demonstrate superior capabilities in complex reasoning, semantic understanding, and adaptive decision-making. A comparative figure is illustrated in Fig. 6, highlights their complementary roles in improving AV safety.

5.1 Knowledge Database

5.1.1. Knowledge Representation

KGs store knowledge explicitly by triples, offering a stable, adaptable, and precise foundation for well-defined and exhaustive databases such as road traffic rules and road elements, while also enabling the modeling of more complex spatio-temporal relations and causal chains. Their explicit storage of knowledge provides advantages of structure, traceability, and efficient querying. However, the fixed structural form constrains their capacity to represent knowledge involving multiple entities and inevitably leads to omissions of relations or entities. In contrast, LLMs encode knowledge implicitly through large-scale parameter training, allowing them to cover a broader semantic space and capture highly complex relational patterns. But this implicit representation makes it difficult to precisely locate knowledge sources or guarantee the accuracy of specific rules.

5.1.2. Knowledge Update and Maintenance

The explicit and structural representation of KGs supports incremental updates and modifications, enabling new traffic rules, scenario elements, or risk knowledge to be incorporated manually at low cost and with high efficiency. In contrast, updating and maintaining LLMs requires retraining or fine-tuning on new datasets, which makes it difficult and costly to keep pace with the rapid iteration of software and hardware, or follow external knowledge changes such as traffic law modifications. Consequently, KGs are better suited for quickly adapting to dynamic knowledge changes, whereas LLMs are more effective in stable contexts involving large-scale knowledge.

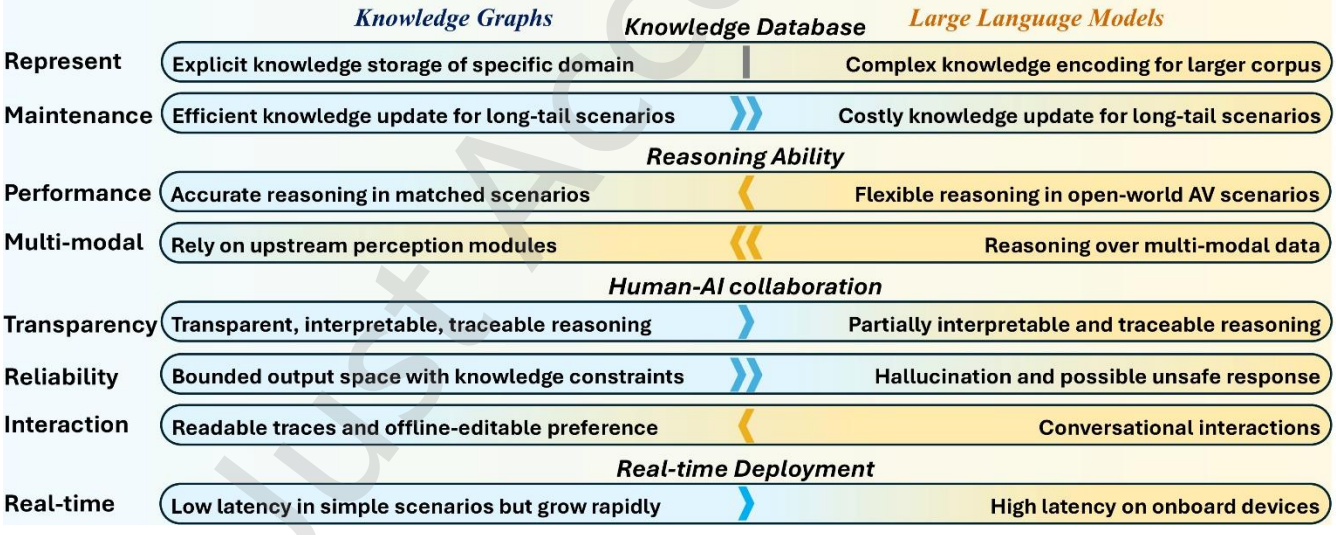


Fig. 6. The comparison of KGs and LLMs in autonomous vehicle safety.

5.2. Reasoning Ability

5.2.1. Reasoning Performance

KGs leverage explicitly defined domain knowledge to deliver accurate and reliable reasoning when the target knowledge can be well captured by entities, relations, and constraints. Consequently, they are particularly effective for tasks with scoped and well-modelled knowledge, such as ASAM OpenSCENARIO-style ontology-grounded scenario representations in V&V toolchains, Waymo's road-graph based map semantics (P. Sun et al., 2020), and Bosch's road-sign KG for traffic element semantics (Monka et al., 2021).

However, in complex scenarios, numerous interfering factors affect the matching, applicability, and validity of predefined knowledge. Meanwhile, the limited semantic reasoning capability of KGs further constrains their effectiveness in handling cases characterized by long reasoning chains, probabilistic uncertainty, and ambiguity. In contrast, the statistical nature of MLLMs enables reasoning under ambiguity and uncertainty by exploiting patterns learned from diverse data, supporting multi-hypothesis interpretation, fuzzy reasoning and proactive responses in evolving scenarios. Accordingly, LLMs are often positioned as holistic driving agents that go beyond isolated safety subtasks, for example NVIDIA's Alpamayo-R1 (Y. Wang et al., 2025), which integrates explicit reasoning for safer decision making in long tail and rare scenarios.

5.2.2. Multi-modal Reasoning

Usually, KGs cannot directly process multimodal inputs and typically rely on existing perception modules, which may introduce information loss and grounding errors. In contrast, MLLMs can directly interpret and reason over multimodal sensory inputs, bridging the gap between high-level reasoning and real-world sensor observations, which reduces reliance on complex encoding processes and minimizes information loss while preserving high-level semantics for reasoning. This direction has already emerged in industrial AV stacks that integrate multimodal understanding with action prediction, such as Wayve's LINGO-2 closed-loop vision language action driving model tested on public roads (Wayve, 2024), and NVIDIA's Alpamayo-R1 of reasoning-based vision language action models that couple multimodal interpretation with driving decisions and evaluation toolchains (Y. Wang et al., 2025).

5.3. Human-AI Collaboration

5.3.1. Transparency, Interpretability and Traceability

KGs offer inherent transparency and interpretability because domain knowledge is explicitly represented and the reasoning process can be auditably traced. This property is well aligned with safety engineering tasks such as safety and compliance checking, since the outputs are usually predictable, stable, and controllable when a rule reasoner (e.g., Pellet (Sirin et al., 2007)) is used and the task knowledge is well covered. Such traceability also supports accident analysis and improvement, as engineers can locate which knowledge gaps or rule conflicts cause failures and update them in a controlled manner. Moreover, auditable traces enable safety governance and accountability practices, which can improve social trust in AV safety as well.

In contrast, MLLMs often lack stable, verifiable provenance for their outputs, and their generation process is difficult to trace. This hinders systematic management and optimization of failure causes in complex driving scenarios and makes it hard to completely ensure rule-consistent validity even in routine scenarios. MLLMs' limited interpretability typically relies on self-generated explanations (e.g., CoT), but such explanations can be unfaithful and mismatch the true generation process of the models' generation. This mismatch may miscalibrate user trust and model capability boundaries, leading to possible risks in safety-critical deployment (Cui et al., 2024a).

5.3.2. Reliability and Hallucination

KGs perform inference over explicitly defined entities, relations, and rules, and thus tend to produce stable and consistent outputs under identical or similar inputs. With predefined knowledge constraints, KG reasoning effectively bounds the output space, reducing hallucination errors and enforcing adherence to specified rules, especially when using rule reasoners such as Pellet within adequate knowledge coverage (Sirin et al., 2007), which supports reliability in covered scenarios.

In contrast, MLLMs can be sensitive to minor input perturbations and are therefore more exposed to adversarial attacks (Bhagwatkar et al., 2024). Moreover, data-driven LLMs may learn spurious correlations or biases and exhibit a heightened risk of hallucination, particularly under out-of-distribution conditions not covered in training data. In hazard analysis, such hallucinations can yield incorrect failure-mode hypotheses or omit critical ones, undermining the completeness and validity of the assessment (Dokas, 2026). Similarly, in LLMdriven test scenario generation, hallucinated constraints or biased patterns can produce invalid scenarios or over-emphasize low-frequency cases, misaligning safeguards with core risks (J. Zhou et al., 2025). When deployed in complex operational safety-critical situations, these hallucination failures may propagate into unsafe responses of the ADS, posing serious safety risks.

5.3.3. Human-AI interaction

KGs support human–AI interaction by exposing explicit, human-readable, and updatable knowledge, together with auditable reasoning traces. User-specific preferences can be incorporated at design time by encoding them as nodes, relations or rule conditions in the KG (Halilaj et al., 2021), and the resulting inference trace can further provide structured explanations of the ADS decision procedure. However, such interaction remains rule-bound and schema-constrained, which limits flexibility and makes it difficult to satisfy real-time, open-ended user demands beyond the predefined representation and rule coverage.

In contrast, MLLMs provide direct conversational interaction with users or regulators via natural language understanding and generation. This allows the ADS to communicate its scene understanding, highlight safety-relevant risks, and justify safe driving behaviors in a human-readable language-form (Z. Xu et al., 2024b). Moreover, MLLMs can accept driver preferences or high-level driving guidance in real time, and can further proactively query users when uncertainty is high or additional constraints are needed, supporting a more adaptive human-in-the-loop interaction paradigm (Deruyttere et al., 2019; H. Shao et al., 2024).

5.4. Real-time Deployment

In practical AV safety domain, an update rate of 10Hz is generally considered acceptable for onboard use, i.e., 100 ms latency (Ulbrich et al., 2014; Zipfl and Zöllner, 2022). Owing to KGs' structured and lightweight nature, query-style KG reasoning, such as SPARQL or compact SWRL rule sets over modest graph sizes, are reported to meet this requirement typically with a latency of tens of milliseconds or around 100 ms in some worse cases (Ulbrich et al., 2014; L. Zhao et al., 2017). However, when the reasoning requires long or intertwined rule chains and when it must operate over large TBox content with complex axioms under OWL-DL semantics, it often relies on tableau-based reasoners such as Pellet or HermiT (Sirin et al., 2007; Shearer et al., 2008), and the latency can degrade to around 1 s and even exceed 10 s in extreme cases (L. Zhao et al., 2015c, 2016; Grabowski and Wang, 2023). Moreover, the computational cost can grow rapidly with graph scale and the number of rules, which makes pruning, modular pre-selection, and knowledge pre-compilation essential for maintaining predictable onboard latency in complex driving scenarios.

VLMs impose substantial challenges for their onboard AV safety applications due to their large parameter counts and compute resource requirements. In practice, their latency is

highly sensitive to model scale and architecture, input and output length, quantization precision, inference engine, and the target hardware platform. Recent onboard deployments of VLMs commonly rely on low-precision quantization (e.g., 4-bit AWQ (J. Lin et al., 2024)) together with optimized inference engine such as LMDeploy. For example, Cui et al. (2024b) reaches a latency about 1.6-2 s with a 9B model and the previous deployment setting, while LLM4AD reports end-to-end latencies of 1.2-1.8 s including an onboard 8B VLM (Cui et al., 2024a). These latencies are only tolerable for high-level, nonurgent adjustments instead of safety-critical, time-sensitive maneuvers. Instead, NVIDIA's Alpamayo-R1 (AR1) substantially optimizes both input and output token counts by preprocessing the input images and performing backend trajectory decoding rather than directly handling these through VLM, and achieves an end-to-end 10 Hz update rate with a 10B-scale model (Y. Wang et al., 2025).

Another practical design is to decouple the VLM as a low-frequency "slow" module rather than synchronizing it with a 10 Hz primary driving loop. This is examplified by DriveVLM-Dual (X. Tian et al., 2024), which runs a 1.5B VLM reasoning on an OrinX platform asynchronously and reports an average speed of 410 ms. In addition, LSRE encodes low-frequency semantic understanding from a VLM into a latent space and performs real-time risk discrimination with a lightweight margin classifier, enabling 10 Hz continuous risk identification despite the VLM's roughly 3 s output latency (Cheng et al., 2025). However, in emergent safety-critical events, VLMs may not be able to respond timely and the fast module may fail to recognize such rare or complex hazards, potentially leading to risks.

6 Integrating KGs and LLMs for Autonomous Vehicle Safety

Meeting AV safety requirements for real-time performance, reliability, and traceability remains difficult with a single technical solution, either KGs or LLMs. Given their complementary strengths and limitations, the integrated use of both offers a promising path to improving AV safety. To systematically explore this potential, we categorize current and emerging integration approaches into three paradigms based on the enhancement direction, as in Fig.7, i.e., LLM-enhanced KG for AV Safety, KG-enhanced LLM for AV Safety, and Synergized KGs and LLMs for AV Safety. For each paradigm, we mainly discuss it through two aspects: knowledge building and reasoning, where the former serves as a prerequisite for the latter.

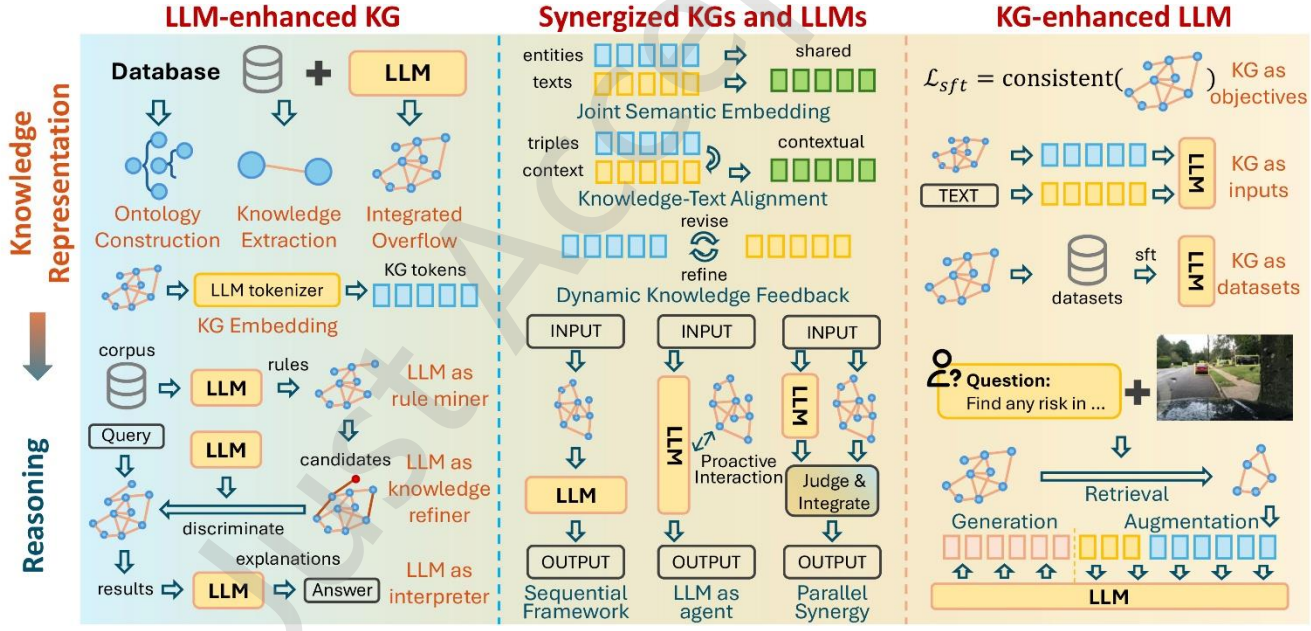


Fig. 7. Categorization of research on integrating KGs and LLMs for AV Safety.

However, current studies of integrating KGs and LLMs remain scarce in the AV safety field. Therefore, in this section, we primarily introduce the paradigms of integrating KGs and LLMs, and then present several practices relevant to AV and AV safety.

6.1. LLM-enhanced KG for Autonomous Vehicle Safety

6.1.1. LLM-enhanced KG Knowledge Representation

From a process perspective, traditional domain-specific KG construction workflows can be divided into two main stages: ontology construction and knowledge extraction. Ontology construction depends heavily on experts, while knowledge extraction requires handling large-scale unstructured data, leading to high time and labor costs. Furthermore, KG embedding maps the KG into an embedding space, enabling reasoning via vector-based approaches such as deep learning. Leveraging LLMs offers a promising alternative in reducing reliance on experts and improving the efficiency and cost-effectiveness of KG construction and embedding (Manas and Paschke, 2025). As knowledge representation typically occurs offline, the real-time ability of LLMs is not required.

Ontology Construction: Traditionally, ontology construction begins with large-scale knowledge extraction based on semantic structures, followed by analysis and abstraction. However, LLMs trained on diverse corpora enable alternative approaches by distilling implicit knowledge directly from the

models. Petroni et al. (2019) first demonstrated the potential of LLMs as static knowledge bases using the LAMA probing framework, which evaluates knowledge recall via cloze-style queries. Subsequently, Bosselut et al. (2019) introduced COMET, a generative Transformer-based model that translates LLM-encoded linguistic knowledge into explicit commonsense knowledge. More recently, Y. Tang et al. (2023) proposed the first LLM-driven domain knowledge distillation framework for AV, using a three-stage prompting and iteration strategy to guide ChatGPT in automatically expanding traffic ontologies from minimal seed inputs.

Despite their advantages, LLM-generated knowledge may lack consistency and precision, especially in specialized domains. To address this, incorporating domain-specific corpora and expert feedback from traditional workflows remains necessary. Hofer et al. (2024) proposed a two-stage generate-and-refine approach for automatic RDF Mapping Language (RML) file generation using LLMs. Their method combines LLMs' semantic understanding with expert-driven ontology refinement and the scalability of traditional mapping tools. Similarly, Tupayachi et al. (2024) developed an AI-based ontology engineering framework for next-generation Urban Decision Support Systems, integrating ChatGPT-4, NLP modules, and the classical Methontology in a prompt-based pipeline for automated concept extraction and ontology construction from scientific and technical texts.

Knowledge Extraction: LLMs significantly enhance knowledge extraction by improving accuracy and efficiency, particularly for complex syntax and ambiguous information. Kumar et al. (2020) increased relation classification accuracy by adding entity position markers to BERT inputs. Melnyk et al. (2021) proposed Grapher, a multi-stage KG construction framework that generates nodes with LLMs and constructs edges based on node features, effectively addressing edge sparsity. In the AV domain, Hou et al. (2025) developed a semi-automatic framework for constructing a traffic accident KG using ChatGLM4, where experts participate only in ontology design and triple verification. H. Dong et al. (2024) introduced the Iterative-LLM Prompting Pipeline, which applies zero-shot iterative prompting with GPT-4-Turbo for large-scale triple extraction from in-cabin textual data, incorporating graph alignment and fuzzy matching for semantic disambiguation without relying on elaborate ontologies. They further proposed GLM-TripleGen (H. Dong et al., 2025), a domain-specific model for intelligent cockpits that reveals implicit links between driving behaviors and vehicle states, enabling efficient, high-quality knowledge extraction and KG construction through lightweight fine-tuning.

The generative ability of LLMs allows the production of large-scale data from few-shot or zero-shot examples, effectively enriching the extracted knowledge, especially under the sparsity of corner and safety-critical data. For example, Chang et al. (2024) proposed the LLMScenario framework, which establishes a closed-loop process of prompt engineering, LLM generation, dual-dimension evaluation (realism and rarity), and expert feedback. J. Zhang et al. (2024) proposed a safety-critical scenario generation agent that integrates LLMs with Scenic, which constructs a large-scale knowledge base of Scenic code snippets and employs a retrieval-based LLM-to-Sim pipeline to assemble high-risk driving scenarios, effectively avoiding hallucinations typically associated with direct code generation.

Integrated Workflow: Applying LLMs to the entire workflow of integrated KG construction is beneficial for maintaining semantic consistency across different stages. However, few studies have so far explored the simultaneous enhancement of ontology construction and knowledge extraction using LLMs (Kommineni et al., 2024). Moreover, human intervention remains necessary in existing frameworks to ensure KG quality and reliability.

KG Embedding: LLMs naturally possess word embedding ability from a large corpus, facilitating high-quality KG embedding for vector-based reasoning. For instance, ERNIE (Y. Sun et al., 2019) employs a dual-encoder structure that integrates tokenized KG triples with sentence-level textual representations. Similarly, KEPLER unifies knowledge embedding and pretrained language models by encoding entity descriptions into entity vectors, achieving deep alignment between linguistic features and KG facts (X. Wang et al., 2021). However, either LLM-enhanced KG construction or embedding methods have scarcely been developed specifically for the AV domain based on our investigation.

6.1.2. LLM-enhanced KG reasoning

Existing KG reasoning methods are mainly divided into logic-based and neural network-based paradigms. The former emphasizes explainability and deductive reasoning, while the latter is well-suited for efficient multi-hop reasoning and knowledge completion (X. Chen et al., 2020). LLMs can enhance both paradigms with their strong reasoning ability, while their high computational demands and latency limit their suitability for real-time reasoning tasks such as AV safety.

For symbolic reasoning, LLMs can generate semantic rules while maintaining the explainability. L. Luo et al. (2025) introduced ChatRule, a logic rule mining framework that integrates KG structural information into LLM prompts through reasoning paths, and proposed metrics to filter erroneous rules from hallucinations. LLMs further enable translating rules into structured representations applicable to KG reasoning. Manas et al. (2024) combined stepwise reasoning with logical templates, showing that GPT-4 can accurately translate complex traffic rules into Metric Temporal Logic (MTL) formulas.

For the KG completion task, traditional learning-based methods focus on validating and verifying the plausibility of predicted triples, but with limitations like insufficient modeling of relational semantics and difficulty distinguishing correct answers among lexically similar candidates. LLMs can be introduced to facilitate these limitations by their wide knowledge and reasoning ability (B. Kim et al., 2020). Traditional KGE methods can be further integrated with CoT-driven LLM to reduce omissions and hallucinations in LLM-enhanced KG completion (Ji et al., 2024). In addition to traditional discriminative task, LLM's generative ability enables the triple generation for KG completion with a "prompt–cloze" paradigm. For example, X. Xie et al. (2022) proposed GenKGC, the first method to formulate KG completion as a sequence generation task using a generative

Transformer, addressing the limitations of traditional discriminative approaches dependent on exhaustive candidate enumeration. B. Zhang et al. (2023) introduced LLMKE, a two-stage framework combining knowledge probing and entity disambiguation, where the LLM-generated answers are refined through external corpus alignment to ensure consistency with the KG. In the AV domain, KG completion research is still nascent and limited in scope.

Beyond improving reasoning performance, LLMs can also transform KG reasoning outputs into human-readable natural language explanations, thereby facilitating system debugging and enhancing interpretability. Manzour et al. (2024) constructed task-specific PedFeatKG and DriverKG for pedestrian crossing intent and vehicle lane-change intent prediction, respectively, developing corresponding KGEs and applied Bayesian reasoning on the HighD dataset for road user behavior prediction. Furthermore, Hussien et al. (2025) proposed integrating structured reasoning outputs into GPT-4, combining KG traceability paths, fuzzy rule semantics, and LLM-generated explanations to reveal probabilistic reasoning chains and providing human-interpretable justifications.

6.2. KG-enhanced LLM for Autonomous Vehicle Safety

In the context of LLMs, knowledge representation involves training or fine-tuning LLMs on general or domain-specific datasets, embedding the required knowledge into model parameters. In KG-enhanced LLM reasoning, KG RAG is a central and widely adopted approach for improving LLM reasoning capabilities. Thus, we will focus on the application of KG RAG for AV safety in the final part of this subsection.

6.2.1. KG-enhanced training for LLM

The structured domain knowledge contained in KGs can serve as supervision for LLM fine-tuning, enabling better capture of the intrinsic relationships among knowledge entities and relevant information. Following S. Pan et al. (2024), we divide the methods of enhancing LLM training with KGs into three categories: incorporating KGs into the training objectives, integrating KGs into the input, and using KGs to fine-tune the LLM.

Incorporating KGs into training objectives focuses on designing KG-aware objectives. A common approach is assigning higher masking probabilities to KG entities during training, enhancing LLM attention to these entities (T. Shen et al., 2020; D. Zhang et al., 2020). Alternatively, some methods (Rosset et al., 2020; S. Li et al., 2022) directly use KG-derived knowledge as training objectives, like using KGs to generate high-quality explanations for QA pairs (H. Chen et al., 2025). Integrating KGs into the input involves incorporating transformed subgraphs during training. A basic method is converting knowledge triples into token sequences appended to input texts (Y. Sun et al., 2021). Subsequent designs reduce noise from knowledge integration by constraining interactions between language tokens and KG tokens (W. Liu et al., 2020). Using KGs to fine-tune LLMs leverages their factual and structural information to construct instruction-tuning datasets. By transforming structured triples into textual form, LLMs can directly acquire KG knowledge, thereby enhancing their ability to extract and apply factual and structural relations for improved reasoning (H. Zhu et al., 2023; J. Wang et al., 2022; L. Luo et al., 2023).

In the AV domain, Qasemi et al. (2023) introduced TRIVIA, a VQA framework for the traffic domain. The framework constructed a KG from collected and coarsely annotated traffic videos, and further leveraged this KG to automatically generate subtitles for LLM fine-tuning. H. Zhou et al. (2025) proposed a KG-based symbolic MLLM training method by incorporating the Nuscenes knowledge graph, i.e., NuscenesKG (Mlodzian et al., 2023), which encodes road topology, traffic rules, and complex interactions. Based on this KG, a symbolic BEV (bird's-eye view) representation was built to capture spatiotemporal information, and this representation was used to fine-tune the LLM, enhancing its understanding of element co-occurrence and improving sequential scene prediction.

6.2.2. KG-enhanced LLM reasoning

Enhancing LLMs with KGs during training lacks the ability to dynamically update knowledge or target specific information. In domains with rapidly evolving knowledge or requiring explicit grounding, integrating KGs at the reasoning stage enables better performance and cooperation with updating knowledge S. Pan et al. (2024). Knowledge-graph-based Retrieved Augmented Generation (RAG) is the most commonly used technique for KG-enhanced LLM reasoning and has been widely adopted in AV. RAG is a paradigm that enhances generative models like MLLMs by integrating external knowledge sources into the generation process, shown in Fig,8. There are three critical components of a RAG framework, i.e., an information retrieval base, a knowledge augmentation part, and a context generator.

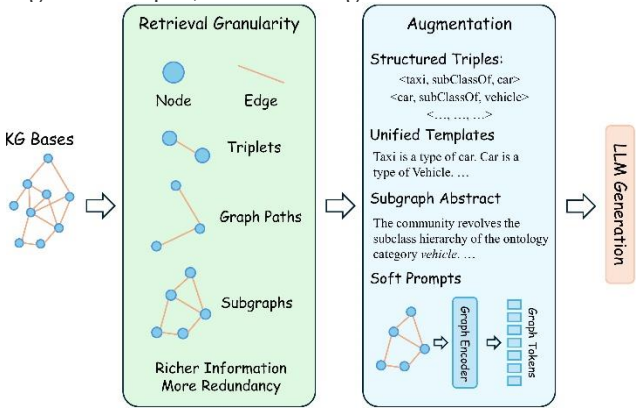


Fig. 8. An illustration of different retrieval granularity and augmentation methods in the RAG framework

Retrieval: In KG-based RAG, KGs function as structured retrieval sources, providing more organized information than plain text in typical RAG systems. Retrieval methods can be classified by the granularity of their results into nodes or edges, triplets, graph paths, and subgraphs (Chen, 2025). The finest nodes and edges level provides the most specific descriptions of entity attributes (Y. Wang et al., 2024b; Munikoti et al., 2023), while the triples capture direct relationships between entities (S. Li et al., 2023). Furthermore, graph paths represent more distant connections (Mavromatis and Karypis, 2024) and subgraph-level retrieval offers richer contextual information (He et al., 2024). Some studies further consider the combination

of different retrieval granularities. For example, Edge et al. (2024) adopts fine-grained nodes and edges for highly relevant knowledge and subgraphs for less relevant information.

In terms of retrieval method, a common approach is to extract entities from the query and retrieve related nodes, edges, or triplets from the KG to obtain relevant knowledge (S. Li et al., 2023; Gutiérrez et al., 2024). Further iterative querying (He et al., 2024; Edge et al., 2024) can be applied based on these elements with traditional graph algorithms for multi-hop information (Mavromatis and Karypis, 2024; Gutiérrez et al., 2024; Delile et al., 2024). To avoid redundant information during iteration, the similarity between retrieved results and the query is assessed as a filter criterion (He et al., 2024; Sanmartin, 2024). Learning-based approaches, such as graph convolutional networks (Kang et al., 2022; G. Liu et al., 2024) and attention mechanisms (H. Sun et al., 2023; J. Dong et al., 2023), can better encode historical interactions and context, highlight critical relations, and improve retrieval efficiency by filtering irrelevant information (J. Zhu et al., 2026).

To quantitatively measure the retrieval quality, prior studies commonly report retrieval accuracy and recall. When groundtruth relevant evidence is available, retrieval performance can be typically measured by Precision@K and Recall@K, which reflect whether the retrieved set covers the relevant information and how much irrelevant information is retrieved. Rank-sensitive metrics such as Mean Reciprocal Rank further consider the positions of correct items in the top-K list (H. Yu et al., 2024). As tasks become more complex and retrieval corpora scale up, annotating ground-truth evidence becomes costly, and many recent evaluations therefore rely on answer-conditioned relevance signals to proxy retrieval precision and recall (Strich et al., 2025). In KG-based RAG, retrieval evaluation is often extended to graph-structured evidence by measuring whether the retrieved triples or subgraphs contain the target entities, or whether they recover key multi-hop reasoning structures such as shortest paths between query entities and answer nodes, i.e., shortest path triple recall (Y. Xu et al., 2025; M. Li et al., 2024).

Augmentation and Generation: The augmentation process feeds the retrieved knowledge into the LLM with a suitable format. A straightforward method is to convert structured triples or graph paths into plain text (Mavromatis and Karypis, 2024; J. Sun et al., 2023), joined with commas or simple connecting words, which retains rich structural information and functioning similarly to CoT (Chen, 2025). Alternatively, unified templates can be used to transform structured knowledge into formats more consistent with natural language (Wen et al., 2024; Sen et al., 2023). When the retrieved information exceeds the LLM's context window or contains excessive irrelevant content, directly converting it to plain text may reduce the model's ability to capture key relations. To address this, some studies extract essential information or filter out irrelevant parts before augmentation. For instance, Edge et al. (2024) measured the relevance between retrieved subgraphs and the query, replacing less relevant segments with concise subgraph abstracts. J. Wu et al. (2024) employed an additional LLM to summarize and refine the retrieved knowledge. The resulting plain texts are then integrated into prompts to support LLM generation.

Besides converting retrieved knowledge into plain text, some approaches use graph encoders to transform the graph structure into soft prompts for the LLM (He et al., 2024; S. Wang et al., 2025). This preserves semantic and structural information more effectively, improving reasoning performance. Some works (Hu et al., 2025a; Ao et al., 2025) extended this by integrating both textual descriptions and latent graph embeddings, providing the LLM with complementary text-view and graph-view representations.

6.2.3 RAG Practices in AV safety

RAG can facilitate efficient safety analysis and design during the development phase. Balu et al. (2025) proposed an agent-based RAG approach for automatically deriving safety requirements, introducing agents to enhance retrieval by extracting more relevant information, enabling more efficient and complex querying. Kieu and Bergstrand (2024) introduced a RAG-based model to improve the explainability and usability of AUTOSAR specifications, demonstrating improved efficiency. Y. Liu et al. (2025) proposed a RAG-based computer numerical control fault diagnosis system that incorporates fault knowledge and real-time engineering data into the decision-making process, an idea that can be borrowed for autonomous driving safety.

In the operational phase, RAG is increasingly applied to enhance the transparency and explainability of ADSs. Hussien et al. (2025) extracted features from public datasets to construct KGs centered on road user behavior and context, which were embedded and retrieved via RAG to enable explainable predictions. Similarly, Manzour et al. (2025) built a KG for near-crash scenarios during lane changes, integrating Bayesian inference with RAG to provide interpretable natural language explanations. In decision-making, J. Huang et al. (2024) proposed DriveRP, a RAG-integrated framework combining real-time and historical data for trajectory planning and motion control. The working memory captures dynamic inputs, while long-term memory stores environmental knowledge. Furthermore, Hernandez-Salinas et al. (2024) introduced IDAS, a RAG-based system that enhances vehicle decision transparency by querying automotive manuals through LLMs in human–machine interactions.

A database of historical driving experiences also supports user-preference–driven AV, enhancing the trustworthiness and reliability of LLM-based AV systems. Dai et al. (2024) proposed VistaRAG, which integrates driving experience, traffic rules, and user preferences to enable reliable and explainable decision-making. Ma et al. (2024) introduced a method that stores positively evaluated LLM outputs in a database, allowing a RAG-based approach to reference user preferences during generation, thereby improving safety and personalization. Z. Xu et al. (2024a) developed a RAG-based warning system that retrieves user profiles to identify driver characteristics, enabling personalized interaction and tailored warnings.

In the AV safety domain, a typical KG-RAG workflow for scenario risk cognition can be decomposed into three stages: offline KG construction, online KG retrieval and augmentation, and online LLM reasoning. In offline KG construction, safety-related knowledge can be encoded into KGs from road traffic laws and defensive-driving experience, predefined risk

patterns, or large-scale historical driving logs and accident records. In the onboard stage, the system first queries the KG for knowledge relevant to the current scene. We explain this workflow by an intersection scenario, where a neighboring vehicle brakes and a pedestrian suddenly emerges in front of it from a blind area to cross. Retrieval can be triggered by key observed behaviors such as surrounding-vehicle braking, or by a scene-graph similarity query. The system then fetches regulation or experience-derived statements in triple form (e.g., "no overtaking in intersections") or similar-scenario subgraphs together with their associated risk sources. The retrieved evidence is then provided to the LLM, typically as textual context or graph-derived embeddings, enabling grounded reasoning to recognize the evolving risk and support downstream avoidance-oriented decisions.

For road traffic laws retrieval, T. Cai et al. (2024) proposed a RAG framework that retrieves regulatory documents and employs an LLM to evaluate whether real-time decisions comply with safety requirements. In terms of driving experience, RAG-Driver (Yuan et al., 2024) retrieves expert demonstrations of similar scenarios as in-context examples, enabling safer and more interpretable driving behaviors in new scenarios. Y. Wang et al. (2026) introduced RAC3, a corner case understanding framework built on a CODA-based image-text database. Through a cross-modal embedding, retrieval, and generation pipeline, RAC3 improves the interpretation of rare scenarios and reduces hallucination in VLMs by similar corner scenario experiences.

6.3. Synergized KGs and LLMs for Autonomous Vehicle Safety

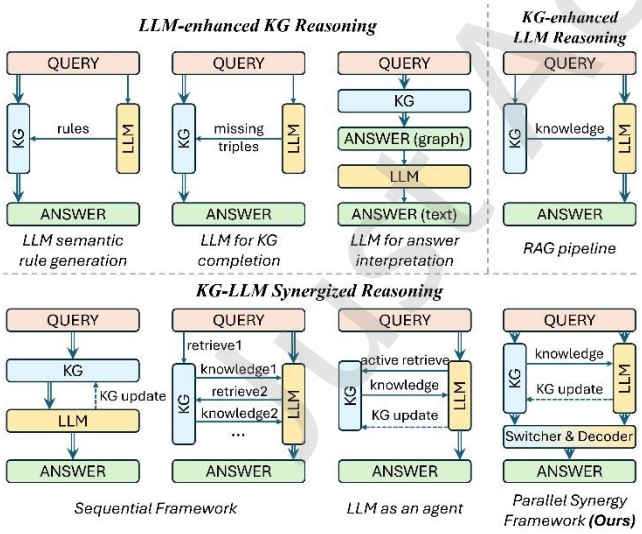


Fig. 9. The information flow of typical KG-LLM integrating reasoning paradigms. In this figure, double-line arrows indicate the primary reasoning f low, solid single arrows denote information flow between the KG and the LLM, and dashed single arrows represent closed-loop knowledge updates.

Currently, the integration of LLMs and KGs is shifting from unidirectional enhancement to deep bidirectional reasoning synergy. In this work, **Bidirectional Reasoning** refers to synergistic reasoning enabled by explicit information exchange between the two components, where both the KG and the LLM act as reasoning modules and the overall reasoning procedure fully exploits the exchanged information. This allows the agent to better capitalize on their respective strengths, selecting appropriate reasoning mechanisms across different task types, while using mutual information to mitigate typical failure modes, such as LLM hallucination and the limited generalization capability of KGs. It naturally distinguishes Bidirectional Reasoning from unidirectional enhancement paradigms, where a single component serves as the primary reasoning module and the information flow is predominantly one-way, such as in RAG-style pipelines. The information flow of typical KG-LLM integrating reasoning paradigms is shown in Fig.9.

This section primarily discusses key paradigms and representative studies of the synergy of KGs and LLMs in two dimensions: Synergized Knowledge Representation and Retrieval, and Synergized Reasoning Interaction and Feedback. Building on the notion of Bidirectional Reasoning, we characterize typical reasoning paradigms by analyzing their information flow patterns, highlighting whether and how the coupling between KGs and LLMs. Finally, we propose a promising Parallel Synergy Framework towards bidirectional-reasoning-enhanced real-time AV safety.

6.3.1. Synergized Knowledge Representation and Retrieval

The unidirectional enhanced knowledge representation and retrieval methods fall short for interactive reasoning in complex tasks due to the lack of dynamic feedback. Beyond this, recent studies emphasize bidirectional complementarity of KGs and LLMs in such tasks.

At the knowledge representation level, the synergy between KGs and LLMs is primarily on the integration of structured knowledge into the embedding space of LLM, as well as the enhanced completion of KG by LLM. A representative work is the series of ERNIE models, progressively achieving vertical refinement of knowledge granularity and horizontal extension of model paradigms. In the early stage, ERNIE works integrates KGs in embedding phase, with multi-task injection to support more precise generation of LLMs (Y. Sun et al., 2019, 2020). Their recent work further constructing a dual-tower framework for both understanding and generation of natural languages, wherein the generation tower supports reverse extraction and completion of KGs (Y. Sun et al., 2021). In AV safety, synergistic knowledge representation via KG alignment facilitates the injection of domain-specific knowledge like traffic rules and safety protocols into AV-oriented LLMs. Conversely, LLMs can analyze driving data, extract safety strategies from long-tail scenarios, and iteratively enrich the knowledge base.

Building upon the foundation of the unidirectional enhancement of KG retrieval for LLM in typical RAG procedure, recent developments have moved toward a more synergistic paradigm, where the LLM's outputs further update or refine the KG, further enabling structural symbolic reasoning from KGs and implicit reasoning from LLMs to complement each other. Microsoft's GraphRAG (Edge et al., 2024) exemplifies this approach by using LLMs to extract knowledge from raw text to build KGs, which are then processed via community detection and integrated into the RAG pipeline for both localized retrieval and global

summarization. However, generating domain-specific KGs from scratch can be costly and incompatible with existing KGs. To mitigate this, Ojima et al. (2024) proposed IR-based GraphRAG, which connects expert-curated failure KGs in AV with the GraphRAG framework through a retrieval–subgraph–filtering pipeline. This allows the KG to provide logically coherent reasoning chains, demonstrating the strengths of KG–LLM collaboration in complex fault diagnosis.

6.3.2. Synergized Reasoning Interaction and Feedback

Sequential Framework: The simplest synergy paradigm assigns distinct reasoning tasks to KGs and LLMs within a multistep workflow. Qi et al. (2024) proposed a safety control framework for service robots integrating LLMs, Embodied Robotic Control Prompts (ERCP), and an Embodied KG. Here, the LLM parses natural language instructions into API-level sequences, while the KG validates the resulting Cypher queries for safety. Choudhary and Reddy (2023) introduced the LARK framework, which decomposes complex logical queries into a reasoning chain with multiple RAG steps. In detail, compared to the typical RAG framework, it performs retrieval and generation at each step of the logical chain, and the output of each LLM step dynamically influences subsequent queries.

In the AV domain, J. Wang et al. (2025) proposed the HybridDriving framework, integrating dynamic risk reasoning and static rule filtering into LLM prompts via a Scene Evolution KG. In this framework, KG handles risk reasoning for vehicle actions, while the LLM generates safe and interpretable driving decisions based on KG outputs. H. Zhou et al. (2025) introduced the FM4SU framework, which fuses KG-driven BEV symbolic representations with LLMs for scene understanding and prediction. In turn, the LLM supports KG completion by inferring occluded elements and predicting future actions, enabling real-time updates for downstream reasoning or decision-making.

LLM as an agent: Another emerging KG–LLM reasoning paradigm treats the LLM as an agent for actively controlling KG retrieval and reasoning. leveraging LLM's capabilities in language understanding and inference. J. Sun et al. (2023) introduced Think-on-Graph (ToG), where the LLM iteratively executes a search–prune–reason loop via Beam Search, overcoming the retrieval bottlenecks of traditional pipelines like LLM-SPARQL-KG. Extending this idea, T. Zhou et al. (2024) proposed CogMG, a collaborative framework addressing challenges such as incomplete KG coverage and asynchronous knowledge updates. While ToG focuses on guiding LLMs to reason over existing KGs, CogMG allows the LLM to decompose queries, infer missing triples using context, and update the KG in real time for subsequent reuse. J. Jiang et al. (2025) proposed the KG-Agent framework, empowering an LLM to autonomously make decisions regarding its interaction with the KG. This design enables the LLM to independently decide how to interact with the KG, including tool selection and relation retrieval, and to iteratively update its knowledge state for complex reasoning tasks. Nevertheless, applications of such agent-based paradigms in the AV domain remain underexplored.

6.3.3. Parallel Synergy Framework for Real-Time AV Safety

Despite the progress in synergizing KGs and LLMs for safety-related reasoning, most existing systems still collaborate in a sequential manner, where information is passed step-by-step between the two components and the knowledge source remains essentially static during operation. In contrast, bidirectional reasoning, characterized by explicit two-way interaction between the LLM and the KG together with collaborative reasoning, has rarely been realized in practice. This capability is particularly important for AV safety, as two-way coupling enables more effective joint utilization of structured knowledge and model-based inference under complex, long-tail scenarios.

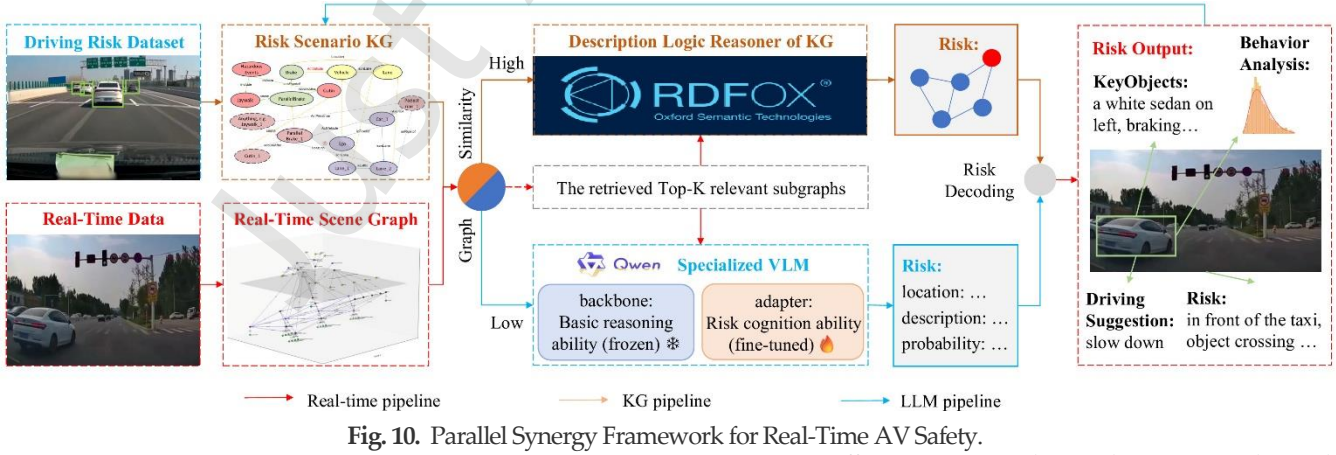


Fig. 10. Parallel Synergy Framework for Real-Time AV Safety.

To meet the real-time and computational demands of AV, we propose a parallel synergy paradigm for KG–LLM collaboration, featuring a dual-system architecture with KG-driven fast and LLM-driven slow channels, illustrated in Fig.10. In detail, on the offline side, a risk scenario KG is built via ontology specification and knowledge extraction from a driving risk dataset. During the operational phase, the observed real-time data is built as a real-time scene graph comprising traffic-participant nodes, road-structure nodes, and typed relational edges. This scene graph is continuously compared against the risk scenario KG database to retrieve the top-K similar subgraphs. These subgraphs will be fed into the KG reasoner to directly infer risk when the scene-subgraph similarity is high, thereby enabling low-latency recognition of pre-defined risk modes.

Conversely, when no sufficiently similar subgraph exists, the

scene is treated as potentially containing an unseen risk pattern, and the retrieved Top-K subgraphs are used as RAG evidence for a parallel VLM-based reasoner. This VLM employs a popular backbone (e.g., Qwen) to provide general multimodal reasoning capability, while a lightweight adapter is finetuned on risk datasets to specialize the model for risk cognition. The outputs from the parallel reasoners are subsequently unified through a shared decoding interface for the subsequent decision-making or E2E models. Finally, for low-similarity cases, the outcome of scene progression, together with the VLM reasoning results, is validated either by human feedback or VLM-assisted verification and then incorporated back into the risk scenario KG, thereby enabling continuous improvement in risk recognition.

This design balances and combines responsiveness and strategic depth, supports redundancy and fallback at the task level, and improves reasoning stability in safety-critical contexts. Furthermore, in risky scenarios, the system activates data logging alongside risk reasoning and safety response, enabling later human–machine collaboration to feed experiential knowledge back into the KG, evolving the architecture into a closed-loop paradigm. The key features of the parallel proposed synergy reasoning framework lie in following aspects:

Parallel reasoning with adaptive switching: Unlike sequential pipelines, this framework treats both the KG reasoner and the VLM as first-class reasoning engines that run in parallel and mutually reinforce each other via bidirectional information support. With a scenario-aware switcher, it can adaptively route the more suitable engine, thereby exploiting the KG's strength in high-similarity, rule-consistent cases and the LLM's strength in complex, uncertain, or long-tail situations at the same time.

Robustness via redundancy and fallback: By enabling the KG reasoner and VLM to jointly infer risk, the framework introduces task redundancy. While one side is unreliable or mismatches the scenario, e.g., by insufficient KG coverage or unstable VLM response, the system can fall back to the other reasoning engine to maintain a valid risk assessment. This redundancy provides a practical robustness guarantee across routine, long-tail, and uncertain scenarios for the overall risk identification pipeline, without collapsing into a single-point failure.

Real-time inference ability: With the fast–slow dual-channel design, the KG-driven channel delivers high-frequency, low-latency risk identification for predefined risk patterns, while the VLM-driven channel runs at a lower frequency to conduct a more holistic risk assessment by better scene understanding. The two risk evaluations are mutually used as reference inputs, enabling consistent cross-conditioning rather than isolated outputs. As a result, the framework preserves real-time responsiveness without sacrificing the ability to handle complex long-tail situations.

Continuous improvement by closed-loop feedback: By leveraging complementary signals from both reason engines and the observed evolution or outcomes of scenarios, the framework enables automatic closed-loop knowledge updates. On the KG side, it can incrementally enrich missing risk-scenario knowledge, strengthening future avoidance capability in similar cases. On the VLM side, it can automatically collect and curate failure or uncertainty cases for continual refinement, driving sustained robustness gains over time.

7 Perspectives and Challenges

KGs and LLMs have both been widely adopted in AV-safety tasks, and recent studies have increasingly demonstrated the potential of their bidirectional enhancement. In this section, we summarize current challenges and future perspectives on enhancing AV safety with the synergy of KGs and LLMs.

7.1. KG Perspectives and Open Challenges

Multi-modal KGs: Most existing KGs are built on textual ontologies and symbolic relations, making them inherently misaligned with the multi-modal sensor inputs of AV systems. Moreover, cross-modal alignment is costly and often incurs semantic, temporal, and spatial information loss. Nevertheless, multi-modal KGs are increasingly viewed as structured memory that bridges perception and reasoning, providing richer context for safety-critical tasks and enabling tighter integration with MLLMs. Representative efforts, such as TRIVIA (Qasemi et al., 2023) and the NuScenes KG-based symbolic representation (H. Zhou et al., 2025), highlight both the promise and the practical bottlenecks of this direction, motivating unified multi-modal knowledge representation and reasoning frameworks that preserve spatiotemporal fidelity with manageable computational cost.

Event and Causal KGs: Current AV KGs are mainly built on static and long-lived entity relations, making them less suitable for dynamic traffic participants and event-level interactions during operation. EKGs and causal augmentation provide richer temporal and causal support for accident mechanism modeling and counterfactual reasoning, and can thus complement MLLMs in reasoning about invisible or potential risks. However, constructing fine-grained event-centric causal graphs remains difficult, as it requires reliable event definitions, robust event detection from heterogeneous data, and trustworthy causal structure extraction. Existing studies already expose these bottlenecks, such as limited transferability in Y. Li and Liu (2022) and assumption-driven causal modeling in Jaimini et al. (2024).

Real-time Deployment: KG-based reasoning often faces a trade-off between completeness and timeliness, as longer temporal context expands the candidate space and runtime in dynamic traffic. This motivates attention-like subgraph selection and incremental reasoning over only affected graph regions to achieve faster risk recognition within a smaller effective scope. In ontology-based rule reasoning, the main bottleneck is rule matching over large spatiotemporal candidate sets. L. Zhao et al. (2016) reduced reasoning time from 470 ms to 53 ms by restricting inference to a sub-knowledge-base near the ego vehicle and to single-frame entities. At larger scales, this real-time restriction becomes more severe and is further limited by representation alignment quality. Ye et al. (2025) reported that SafeDriveRAG, built on a 1.32m-token multi-modal KG, requires 884.10s for retrieval, making online deployment infeasible.

7.2. LLM Perspectives and Open Challenges

Closed-loop Data Collection: Fine-tuning pipelines for LLMs are become increasingly standardized, and multi-agent workflows can be used to generate, cross-validate, and refine instruction data for SFT. However, risk data is inherently vast and long-tail, and high-quality samples are still largely dependent on manual heuristics or rule-based triggers. Looking ahead, scalable AV safety requires automated closed-loop mechanisms for continuously mining and reusing long-tail cases. Existing studies, such as TrafficSafetyGPT and RAC3 (Zheng et al., 2023; Y. Wang et al., 2026), have injected safety-critical experience into LLMs, but still lack a reliable framework for identifying, generalizing, and validating new safety knowledge at scale. Meanwhile, current models remain inaccurate under low iteration counts and prone to false alarms, making them more suitable for coarse pre-screening than direct deployment (Sinha et al., 2024).

Safe and Constraint-aligned Outputs: A major limitation of LLMs in AV is hallucination-induced unsafe or non-compliant outputs. Improving safety therefore requires more accurate intermediate reasoning and tighter integration with reliable external knowledge constraints. Yet this remains difficult, as confidence estimation and uncertainty quantification increase computational burden, and models may still violate constraints even when scene-specific knowledge and structured prompts are provided (An et al., 2025). This shows that directly adding knowledge does not guarantee compliance. Regulation-grounded retrieval-augmented reasoning is a promising direction (T. Cai et al., 2024), but reliable adherence under open-ended generation remains challenging and requires verification-aware decoding and in-loop constraint checking.

Real-time Deployment: Current MLLMs still face high computational demands, limiting real-time deployment in safety-critical driving scenarios. Looking forward, lightweight or distilled models for onboard hardware, more efficient architectures, and accelerated inference pipelines have shown initial progress toward practical AV deployment. Nevertheless, although such attempts can reduce part of the model footprint but still face a persistent difficulty in retaining sufficiently strong capability under a strict parameter budget, as evidenced in Y. Tang et al. (2023). Meanwhile, multi-route reasoning frameworks and adaptive thinking strategies (Q. Yang et al., 2025; Y. Luo et al., 2025) can accelerate responses in routine situations, yet they still fall short when urgent and complex events require deeper reasoning and more stable decision making. Consistently, DriveVLM (X. Tian et al., 2024) reports about 600 ms for image compression and 410 ms for inference even on an NVIDIA OrinX platform, showing that latency remains fundamentally constrained by onboard computing and that robust reasoning under time pressure is still difficult.

7.3. Synergistic Perspectives and Open Challenges

Accurate Knowledge Retrieval: Knowledge retrieval is essential for reliable risk reasoning, yet incomplete coverage and poor retrieval strategies may provide limited or even misleading evidence for downstream reasoning. Looking forward, retrieval should support multi-granularity evidence and go beyond naive top-K filtering to retain scenario-relevant context. This remains challenging because AV knowledge spans documents and passages, nodes, triples, subgraphs, frames, and event chains, while accurate cross-level retrieval often conflicts with efficiency. Recent methods point to possible directions, such as GraphRAG, which replaces less relevant segments with concise subgraph abstracts (Edge et al., 2024), and GNN-RAG, which uses graph encoding and neural retrieval to improve evidence selection (Mavromatis and Karypis, 2024). However, the tension between retrieval accuracy and real-time readiness remains.

Robust Knowledge Utilization: Even when a retrieval module returns top K evidence, correctness and applicability are not guaranteed, and weakly related, partially correct, or conflicting items may mislead generation if naively trusted. Looking forward, robust utilization should reduce exposure to such evidence by structuring retrieved content before generation, combining rank with absolute similarity for gating, and applying validation-based reranking when conflicts or low relevance are detected. Existing studies have explored these directions through evidence refinement, where an additional LLM summarizes retrieved knowledge before prompting (J. Wu et al., 2024), and value-guided utilization, such as KnowVal (Xia et al., 2025), which uses a Value Model to assess knowledge utility and guide trajectory decisions. However, robust utilization remains difficult because scenario-dependent calibration, heterogeneous and partial observations, and aggressive filtering may remove rare but safety-critical cues under real-time constraints.

Closed-loop Knowledge Update: KGs are more extensible and their update cycle can be shorter than LLM weight updates, while injecting new knowledge into a KG does not induce catastrophic forgetting. From a future perspective, a practical closed-loop strategy is to update KGs more frequently with new entities, relations, and risk patterns from operational data and expert supervision, while refining LLMs less often through continual training, enabling KG and LLM components to jointly cope with endless operational risks. Nevertheless, building an automated framework that can reliably identify novel knowledge, incorporate it with minimal human intervention, and maintain consistency across KG and LLM components remains challenging. building an automated framework that can reliably identify novel knowledge, incorporate it with minimal human intervention, and maintain KG-LLM consistency remains challenging. Agent-based paradigms provide early attempts, such as Think-on-Graph (J. Sun et al., 2023) and CogMG (T. Zhou et al., 2024). Yet their gap to safety-critical AV deployment still calls for stronger consistency maintenance, provenance tracking, and update validation under shifting operational distributions.

7.4. Benchmarking and Evaluation Infrastructure

For KGs, a practical roadmap calls for open and unified ontology frameworks that jointly cover static scene descriptions, traffic participant behavior semantics, and risk concepts, so as to reduce the current schema fragmentation and repeated engineering overhead across studies. Beyond defining schemas, the community also needs standardized construction and evaluation protocols for converting heterogeneous multimodal data sources into graph-structured knowledge. Such pipelines should be assessed not only by construction accuracy but also by cross-source consistency, schema compliance, and the degree of required human intervention, since these factors directly determine whether

multi-modal KGs can be built and maintained at scale. Moreover, dedicated driving-risk knowledge-base benchmarks are needed. The associated deployment-oriented metrics should capture ontology and graph scale, knowledge freshness and update cost, and the correctness and stability of downstream risk reasoning under bounded runtime budgets. The benchmarks should also define reproducible splits across cities, weather, and time to avoid distribution leakage and to quantify generalization.

For LLMs, the roadmap should prioritize AV-safety-specific data curation and annotation workflows that support long-tail risk discovery and iterative reuse. When stepwise supervision is adopted, standardized reasoning traces are also needed, enabling datasets to systematically cover rare but safety-critical scenarios rather than relying on manual heuristics or rule triggers. Evaluation should be anchored by protocols that report not only end-task accuracy but also model size and runtime, stepwise reasoning correctness and consistency, and auditable safety and compliance outcomes such as constraint violation rates, because safety failures often arise from intermediate reasoning errors and cannot be diagnosed by a single final-score metric. To support closed-loop improvement, benchmarks should further include incremental settings that simulate low-iteration data loops. New risk cases should arrive over rounds, and the model should be evaluated for how reliably it absorbs new knowledge, how stable its behavior remains across updates, and how well uncertainty or confidence calibration reflects risk-critical mistakes.

For KG-LLM synergy, benchmarks should pair AV-safety datasets with matched knowledge bases. This enables controlled and reproducible evaluation of retrieval, utilization, and system-level behavior, instead of treating the KG as an unobserved implementation detail. In addition to retrieval accuracy and query relevance, protocols should explicitly test robustness to imperfect evidence. This includes low-similarity yet highly ranked items, conflicting retrieved facts, and partial knowledge coverage, which are common failure modes that can mislead generation. The evaluation should jointly measure the correctness of generated risk conclusions or actions, the degree of evidence attribution and verifiability, and the full-stack latency broken down into retrieval, verification, and generation under realistic onboard budgets. Finally, closed-loop update benchmarks are required to quantify how newly observed risks trigger knowledge updates, how consistency is maintained across KG and LLM components over time, and how performance evolves over incremental updates without sacrificing safety-critical reliability.

8 Conclusion

Ensuring safety remains a central challenge in the development of autonomous driving systems. This review systematically explored how Knowledge Graphs (KGs) and Large Language Models (LLMs), as representatives of knowledge-driven and connectionist AI paradigms, can individually and jointly contribute to enhancing autonomous vehicle safety. By analyzing their respective strengths and limitations, as well as reviewing recent efforts toward their integration, we highlighted the potential of combining explicit knowledge reasoning with flexible semantic understanding to address safety-critical challenges in complex driving environments, and provided a feasible framework for real-time AV safety. Finally, we discussed the potential and challenges in this field. Despite the progress, the field remains in an early stage, and we believe this review will provide insight for readers. Looking forward, we believe that advancing the synergy between KGs and LLMs will play a pivotal role in building the next generation of robust, interpretable, and trustworthy autonomous driving systems.

Author contributions

Jiaxin Liu: Investigation, Methodology, Validation, Visualization, Writing - original draft, Writing - review & editing. **Liang Peng:** Investigation, Methodology, Validation, Visualization, Writing - original draft, Writing - review & editing. **Xiangyu Yan:** Data curation, Formal analysis, Investigation. **Lingjun Zhang**: Data curation, Formal analysis, Investigation. **Chenye Yang**: Data curation, Investigation. **Yueming Tao**: Data curation, Formal analysis. **Ashton Yu Xuan Tan**: Data curation, Investigation, Writing - review & editing. **Tianli Xu**: Data curation, Writing - review & editing. **Sai Guo**: Funding acquisition, Resources, Supervision. **Hong Wang:** Funding acquisition, Supervision, Writing - review & editing. **Jun Li**: Funding acquisition, Project administration, Supervision

Replication and data sharing

This study is based solely on a review of literature and does not involve the collection or generation of new datasets. All sources referenced in this paper are publicly available and have been appropriately cited. No additional data were created or analyzed beyond those presented in the referenced works.

Acknowledgements

The authors would like to appreciate the financial support by Xiaomi Cooperation under Project: 20252002178, the National Science Foundation of China Project: 52072215, National Natural Science Foundation of China grant 52221005, National Natural Science Foundation of China grant: 52075213, Beijing Natural Science Foundation L243025, and State Key Laboratory of Intelligent Green Vehicle and Mobility.

Declaration of competing interest

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

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author(s) used ChatGPT for polishing and paper organization. After utilizing this tool/service, the author(s) thoroughly reviewed and edited the content as necessary and take full responsibility for the final content of the published article.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., et al., 2023. Gpt-4 technical report. https://doi.org/10.48550/arXiv.2303.08774.

Aldeen, M., MohajerAnsari, P., Ma, J., Chowdhury, M., Cheng, L., Pesé, M.D., 2024. Wip: A first look at employing large multimodal models against autonomous vehicle attacks. In: ISOC Symposium on Vehicle Security and Privacy (VehicleSec'24). 1–7.

Ali, N., Lundqvist, K., Hänninen, K., 2024. Mitigation ontology for analysis of safety-critical systems. In: the 34-th European Safety and Reliability Conference (ESREL). Polish Safety and Reliability Association. 9–17.

An, B., Zhang, S., Dredze, M., 2025. Rag llms are not safer: A safety analysis of retrieval-augmented generation for large language models. In: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). 5444–5474.

Ao, T., Yu, Y., Wang, Y., Deng, Y., Guo, Z., Pang, L., et al., 2025. Lightprof: A lightweight reasoning framework for large language model on knowledge graph. In: Proceedings of the AAAI Conference on Artificial Intelligence. 23424–23432.

Armand, A., Filliat, D., Ibañez-Guzman, J., 2014. Ontology-based context awareness for driving assistance systems. In: 2014 IEEE intelligent vehicles symposium proceedings. 227–233.

Ashqar, H.I., Jaber, A., Alhadidi, T.I., Elhenawy, M., 2025. Advancing object detection in transportation with multimodal large language models (mllms): A comprehensive review and empirical testing. Computation 13, 133.

Bagschik, G., Menzel, T., Maurer, M., 2018. Ontology based scene creation for the development of automated vehicles. In: 2018 IEEE Intelligent Vehicles Symposium (IV). 1813–1820.

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., et al., 2023a. Qwen technical report. https://doi.org/10.48550/arXiv.2309.16609.

Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., et al., 2023b. Qwen-vl: A versatile vision-language model for understanding, localization. Text Reading, and Beyond 2, 1.

Balu, B.V., Geissler, F., Carella, F., Zacchi, J.V., Jiru, J., Mata, N., et al., 2025. Towards automated safety requirements derivation using agent-based rag. In: Proceedings of the AAAI Symposium Series. 299–307.

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. 2682–2690.

Barrachina, J., Garrido, P., Fogue, M., Martinez, F.J., Cano, J.C., Calafate, C.T., et al., 2012. Veacon: A vehicular accident ontology designed to improve safety on the roads. Journal of Network and Computer Applications 35, 1891–1900.

Bartoccioni, F., Ramzi, E., Besnier, V., Venkataramanan, S., Vu, T.H., Xu, Y., et al., 2025. Vavim and vavam: Autonomous driving through video generative modeling. https://doi.org/10.48550/arXiv.2502.15672.

Besta, M., Blach, N., Kubicek, A., Gerstenberger, R., Podstawski, M., Gianinazzi, L., et al., 2024. Graph of thoughts: Solving elaborate problems with large language models. In: Proceedings of the AAAI Conference on Artificial Intelligence. 17682–17690.

Bhagwatkar, R., Nayak, S., Bashivan, P., Rish, I., 2024. Improving adversarial robustness in vision-language models with architecture and prompt design. In: Findings of the Association for Computational Linguistics: EMNLP 2024. 17003–17020.

Bogdoll, D., Guneshka, S., Zöllner, J.M., 2022. One ontology to rule them all: Corner case scenarios for autonomous driving. In: European Conference on Computer Vision. 409–425.

Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O., 2013. Translating embeddings for modeling multi-relational data. Advances in neural information processing systems 26.

Bosselut, A., Rashkin, H., Sap, M., Malaviya, C., Celikyilmaz, A., Choi, Y., 2019. Comet: Commonsense transformers for automatic knowledge graph construction. In: Proceedings of the 57th annual meeting of the association for computational linguistics. 4762–4779.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., et al., 2020. Language models are few-shot learners. Advances in neural information processing systems 33, 1877–1901.

Buechel, M., Hinz, G., Ruehl, F., Schroth, H., Gyoeri, C., Knoll, A., 2017. Ontology-based traffic scene modeling, traffic regulations dependent situational awareness and decision-making for automated vehicles. In: 2017 IEEE intelligent vehicles symposium (IV). 1471–1476.

Bugueño, M., Biswas, R., de Melo, G., 2026. Explainable ai for graph-based learning: A survey beyond graph neural networks. https://hal.science/hal-04660442v2.

Caesar, H., Kabzan, J., Tan, K.S., Fong, W.K., Wolff, E., Lang, A., et al., 2021. nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. https://doi.org/10.48550/arXiv.2106.11810.

Cai, T., Liu, Y., Zhou, Z., Ma, H., Zhao, S.Z., Wu, Z., et al., 2024. Driving with regulation: Interpretable decision-making for autonomous vehicles with retrieval-augmented reasoning via llm. https://doi.org/10.48550/arXiv.2410.04759.

Cai, X., Bai, X., Cui, Z., Xie, D., Fu, D., Yu, H., et al., 2026. Text2scenario: Text-driven scenario generation for autonomous driving test. Automotive Innovation , 1–26.

Cao, Y., Xu, X., Sun, C., Huang, X., Shen, W., 2023. Towards generic anomaly detection and understanding: Large-scale visual-linguistic model (gpt-4v) takes the lead. https://doi.org/10.48550/arXiv.2311.02782.

CARLA Team, 2025. Carla autonomous driving leaderboard (v2.x, includes leaderboard 2.0). https://leaderboard.carla.org/.

Chang, C., Wang, S., Zhang, J., Ge, J., Li, L., 2024. Llmscenario: Large language model driven scenario generation. IEEE Transactions on Systems, Man, and Cybernetics: Systems .

Charoenpitaks, K., Nguyen, V.Q., Suganuma, M., Takahashi, M., Niihara, R., Okatani, T., 2024. Exploring the potential of multi-modal ai for driving hazard prediction. IEEE Transactions on Intelligent Vehicles .

Chen, H., Shen, X., Wang, J., Wang, Z., Lv, Q., He, J., et al., 2025. Knowledge graph finetuning enhances knowledge manipulation in large language models. In: The Thirteenth International Conference on Learning Representations. 1–15.

Chen, K., Li, Y., Zhang, W., Liu, Y., Li, P., Gao, R., et al., 2025. Automated evaluation of large vision-language models on self-driving corner cases. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). 7817–7826.

Chen, R., 2025. Retrieval-augmented generation with knowledge graphs: A survey. In: Computer Science Undergradaute Conference 2025@ XJTU. 1–8.

Chen, W., Kloul, L., 2018. An ontology-based approach to generate the advanced driver assistance use cases of highway traffic. In: 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management.. 1–11.

Chen, X., Jia, S., Xiang, Y., 2020. A review: Knowledge reasoning over knowledge graph. Expert systems with applications 141, 112948.

Cheng, Q., Zhou, W., Jing, C., Deng, N., Wen, J., Liu, Z., et al., 2025. Lsre: Latent semantic rule encoding for real-time semantic risk detection in autonomous driving. https://doi.org/10.48550/arXiv.2512.24712.

Choudhary, N., Reddy, C.K., 2023. Complex logical reasoning over knowledge graphs using large language models. https://doi.org/10.48550/arXiv.2305.01157.

Cui, C., Ma, Y., Yang, Z., Zhou, Y., Liu, P., Lu, J., et al., 2024a. Large language models for autonomous driving (llm4ad): Concept, benchmark, experiments, and challenges. https://doi.org/10.48550/arXiv.2410.15281.

Cui, C., Yang, Z., Zhou, Y., Ma, Y., Lu, J., Wang, Z., 2023. Large language models for autonomous driving: Real-world experiments. CoRR .

Cui, C., Yang, Z., Zhou, Y., Peng, J., Park, S.Y., Zhang, C., et al., 2024b. On-board vision-language models for personalized autonomous vehicle motion control: System design and real-world validation. https://doi.org/10.48550/arXiv.2411.11913.

Dai, X., Guo, C., Tang, Y., Li, H., Wang, Y., Huang, J., et al., 2024. Vistarag: Toward safe and trustworthy autonomous driving through retrieval-augmented generation. IEEE Transactions on Intelligent Vehicles 9,

4579–4582.

De Gelder, E., Paardekooper, J.P., Saberi, A.K., Elrofai, H., den Camp, O.O., Kraines, S., et al., 2022. Towards an ontology for scenario definition for the assessment of automated vehicles: An object-oriented framework. IEEE Transactions on Intelligent Vehicles 7, 300–314.

Delile, J., Mukherjee, S., Van Pamel, A., Zhukov, L., 2024. Graph-based retriever captures the long tail of biomedical knowledge. https://doi.org/10.48550/arXiv.2402.12352.

Delong, L.N., Mir, R.F., Whyte, M., Ji, Z., Fleuriot, J.D., 2023. Neurosymbolic ai for reasoning on graph structures: A survey. https://doi.org/10.48550/arXiv.2302.07200.

Deruyttere, T., Vandenhende, S., Grujicic, D., Van Gool, L., Moens, M.F., 2019. Talk2car: Taking control of your self-driving car. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). 2088–2098.

Dettmers, T., Minervini, P., Stenetorp, P., Riedel, S., 2018. Convolutional 2d knowledge graph embeddings. In: Proceedings of the AAAI conference on artificial intelligence. 1811–1818.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). 4171–4186.

Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A., et al., 2024. Chain-of-verification reduces hallucination in large language models. In: Findings of the association for computational linguistics: ACL 2024. 3563–3578.

Dianov, I., Ramirez-Amaro, K., Cheng, G., 2015. Generating compact models for traffic scenarios to estimate driver behavior using semantic reasoning. In: IROS, International Conference on Intelligent Robots and Systems. 69–74.

Diao, S., Wang, P., Lin, Y., Pan, R., Liu, X., Zhang, T., 2023. Active prompting with chain-of-thought for large language models. https://doi.org/10.48550/arXiv.2302.12246.

Diemert, S., Weber, J.H., 2023. Can large language models assist in hazard analysis? In: International Conference on Computer Safety, Reliability, and Security. 410–422.

Dokas, I.M., 2026. From hallucinations to hazards: benchmarking llms for hazard analysis in safety-critical systems. SAFETY SCIENCE 194.

Dong, H., Wang, W., Sun, Z., Kang, Z., Ge, X., 2024. Iterative llm prompting for intelligent cockpit knowledge graph construction. In: 2024 4th International Conference on Computer Systems (ICCS). 136–145.

Dong, H., Wang, W., Sun, Z., Kang, Z., Ge, X., Gao, F., et al., 2025. Knowledge graph construction for intelligent cockpits based on large language models. Scientific Reports 15, 7635.

Dong, J., Zhang, Q., Huang, X., Duan, K., Tan, Q., Jiang, Z., 2023. Hierarchy-aware multi-hop question answering over knowledge graphs. In: Proceedings of the ACM web conference 2023. 2519–2527.

Dui, H., Zhang, H., Wu, S., Xie, M., 2025. Spatiotemporal resilience analysis of iot-enabled unmanned system of systems. Engineering .

Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., et al., 2024. From local to global: A graph rag approach to query-focused summarization. https://doi.org/10.48550/arXiv.2404.16130.

Elhafsi, A., Sinha, R., Agia, C., Schmerling, E., Nesnas, I.A., Pavone, M., 2023. Semantic anomaly detection with large language models. Autonomous Robots 47, 1035–1055.

European Commission, 2024. Community Road Accident Database CARE. https://road-safety.transport.ec.europa.eu/european-road-safety-observatory/statistics-and-analysis-archive/about-care_en.

Fan, Z., Wang, P., Zhao, Y., Zhao, Y., Ivanovic, B., Wang, Z., et al., 2024. Learning traffic crashes as language: Datasets, benchmarks, and what-if causal analyses. https://doi.org/10.48550/arXiv.2406.10789.

Fang, J., Li, L.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. 22030–22040.

Fang, J., Yan, D., Qiao, J., Xue, J., Yu, H., 2021. Dada: Driver attention prediction in driving accident scenarios. IEEE transactions on intelligent transportation systems 23, 4959–4971.

Fang, W., Ma, L., Love, P.E., Luo, H., Ding, L., Zhou, A., 2020. Knowledge graph for identifying hazards on construction sites: Integrating computer vision with ontology. Automation in Construction 119, 103310.

Feilhauer, M., 2018. Simulationsgestützte Absicherung von Fahrerassistenzsystemen. Ph.D. thesis. Stuttgart, DE: Universitätsbibliothek der Universität Stuttgart.

Feng, S., Sun, H., Yan, X., Zhu, H., Zou, Z., Shen, S., et al., 2023. Dense reinforcement learning for safety validation of autonomous vehicles. Nature 615, 620–627.

Fernandez, D., Mohajer Ansari, P., Kokenoz, C., Salarpour, A., Li, B., Pesé, M.D., 2025. Wip: From detection to explanation: Using llms for adversarial scenario analysis in vehicles. mpese .

Gao, S., Yang, J., Chen, L., Chitta, K., Qiu, Y., Geiger, A., et al., 2024. Vista: A generalizable driving world model with high fidelity and versatile controllability. Advances in Neural Information Processing Systems 37, 91560–91596.

Gao, Y., Piccinini, M., Moller, K., Alanwar, A., Betz, J., 2025. From words to collisions: Llm-guided evaluation and adversarial generation of safety-critical driving scenarios. https://doi.org/10.48550/arXiv.2502.02145.

Gao, Y., Piccinini, M., Zhang, Y., Wang, D., Moller, K., Brusnicki, R., et al., 2026. Foundation models in autonomous driving: A survey on scenario generation and scenario analysis. IEEE Open Journal of Intelligent Transportation Systems .

Geissler, F., Roscher, K., Trapp, M., 2024. Concept-guided llm agents for human-ai safety codesign. In: Proceedings of the AAAI Symposium Series. 100–104.

German in-depth accident study cooperation, 2024. German In-Depth Accident Study GIDAS. https://www.gidas.org/start-en.html.

Geyer, S., Baltzer, M., Franz, B., Hakuli, S., Kauer, M., Kienle, M., et al., 2014. Concept and development of a unified ontology for generating test and use-case catalogues for assisted and automated vehicle guidance. IET Intelligent Transport Systems 8, 183–189.

Gouidis, F., Vassiliades, A., Patkos, T., Argyros, A., Bassiliades, N., Plexousakis, D., 2019. A review on intelligent object perception methods combining knowledge-based reasoning and machine learning. https://doi.org/10.48550/arXiv.1912.11861.

Grabowski, M., Wang, Y., 2023. A concept to support ai models by using ontologies-presented on the basis of ger-man technical specifications for lane markings. 27th ESV 2023 .

Guan, S., Cheng, X., Bai, L., Zhang, F., Li, Z., Zeng, Y., et al., 2022. What is event knowledge graph: A survey. IEEE Transactions on Knowledge and Data Engineering 35, 7569–7589.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., et al., 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. https://doi.org/10.48550/arXiv.2501.12948.

Guo, Z., Lykov, A., Yagudin, Z., Konenkov, M., Tsetserukou, D., 2024. Co-driver: Vlm-based autonomous driving assistant with human-like behavior and understanding for complex road scenes. https://doi.org/10.48550/arXiv.2405.05885.

Gutiérrez, B.J., Shu, Y., Gu, Y., Yasunaga, M., Su, Y., 2024. Hipporag: Neurobiologically inspired long-term memory for large language models. Advances in neural information processing systems 37, 59532–59569.

Halilaj, L., Dindorkar, I., Lüttin, J., Rothermel, S., 2021. A knowledge graph-based approach for situation comprehension in driving scenarios. In: The Semantic Web: 18th International Conference, ESWC 2021, Virtual Event, June 6–10, 2021, Proceedings 18. 699–716.

Halilaj, L., Luettin, J., Henson, C., Monka, S., 2022. Knowledge graphs for automated driving. In: 2022 IEEE Fifth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE). 98–105.

Halilaj, L., Luettin, J., Monka, S., Henson, C., Schmid, S., 2023. Knowledge graph-based integration of autonomous driving datasets. International Journal of Semantic Computing 17, 249–271.

Han, Z., Gao, C., Liu, J., Zhang, J., Zhang, S.Q., 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey. https://doi.org/10.48550/arXiv.2403.14608.

Hanselmann, N., Renz, K., Chitta, K., Bhattacharyya, A., Geiger, A., 2022. King: Generating safety-critical driving scenarios for robust imitation via kinematics gradients. In: European Conference on Computer Vision. 335–352.

He, X., Tian, Y., Sun, Y., Chawla, N., Laurent, T., LeCun, Y., et al., 2024. G-retriever: Retrieval-augmented generation for textual graph understanding and question answering. Advances in Neural Information Processing Systems 37, 132876–132907.

Henson, C., Schmid, S., Tran, A.T., Karatzoglou, A., 2019. Using a knowledge graph of scenes to enable search of autonomous driving data. In: ISWC (Satellites). 313–314.

Hernandez-Salinas, B., Terven, J., ChaveZ-Urbiola, E., Cordova-Esparza, D.M., Romero-Gonzalez, J.A., Arguelles, A., et al., 2024. Idas: Intelligent driving assistance system using rag. IEEE Open Journal of Vehicular Technology .

Hillen, D., Helten, C., Reich, J., 2024. Towards llm-augmented situation space analysis for the hazard and risk assessment of automotive systems. In: INFORMATIK 2024. 709–714.

Hofer, M., Frey, J., Rahm, E., 2024. Towards self-configuring knowledge graph construction pipelines using llms-a case study with rml. In: Fifth International Workshop on Knowledge Graph Construction@ ESWC2024. 1–15.

Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., Melo, G.D., Gutierrez, C., et al., 2021. Knowledge graphs. ACM Computing Surveys (Csur) 54, 1–37.

Hou, Y., Shao, Y., Han, Z., Ye, Z., 2025. Construction and Application of Traffic Accident Knowledge Graph Based on LLM. SAE Technical Paper 2025-01-7139.

Hovi, J., Ichise, R., 2020. Feasibility study: Rule generation for ontology-based decision-making systems. In: Semantic Technology: 9th Joint International Conference, JIST 2019, Hangzhou, China, November 25–27, 2019, Revised Selected Papers 9. 88–99.

Htun, S.N.N., Fukuda, K., 2024. Integrating knowledge graphs into autonomous vehicle technologies: A survey of current state and future directions. Information 15, 645.

Hu, Y., Lei, Z., Zhang, Z., Pan, B., Ling, C., Zhao, L., 2025a. Grag: Graph retrieval-augmented generation. In: Findings of the Association for Computational Linguistics: NAACL 2025. 4145–4157.

Hu, Y., Wang, F., Ye, D., Wu, M., Kang, J., Yu, R., 2025b. Llm-based misbehavior detection architecture for enhanced traffic safety in connected autonomous vehicles. IEEE Transactions on Vehicular Technology .

Huang, J., Ma, H., Zhang, T., Lin, F., Ma, S., Wang, X., et al., 2024. Driverp: Rag and prompt engineering embodied parallel driving in cyber-physical-social spaces. In: 2024 IEEE 4th International Conference on Digital Twins and Parallel Intelligence (DTPI). 547–553.

Huang, L., Liang, H., Yu, B., Li, B., Zhu, H., 2019. Ontology-based driving scene modeling, situation assessment and decision making for autonomous vehicles. In: 2019 4th Asia-Pacific Conference on Intelligent Robot Systems (ACIRS). 57–62.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., et al., 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems 43, 1–55.

Huang, S., Shi, F., Sun, C., Zhong, J., Ning, M., Yang, Y., et al., 2025. Drivesotif: Advancing sotif through multimodal large language models. IEEE Transactions on Vehicular Technology .

Huang, W., Wang, C., Zhang, R., Li, Y., Wu, J., Fei-Fei, L.V., 2023. Composable 3d value maps for robotic manipulation with language models. https://doi.org/10.48550/arXiv.2307.05973.

Hülsen, M., Zöllner, J.M., Weiss, C., 2011. Traffic intersection situation description ontology for advanced driver assistance. In: 2011 IEEE intelligent vehicles symposium (IV). 993–999.

Hummel, B., Thiemann, W., Lulcheva, I., 2008. Scene understanding of urban road intersections with description logic. In: Dagstuhl Seminar Proceedings (Volume 8091). 1–16.

Hussien, M.M., Melo, A.N., Ballardini, A.L., Maldonado, C.S., Izquierdo, R., Sotelo, M.A., 2025. Rag-based explainable prediction of road users behaviors for automated driving using knowledge graphs and large language models. Expert Systems with Applications 265, 125914.

Ibrahim, N., Aboulela, S., Ibrahim, A., Kashef, R., 2024. A survey on augmenting knowledge graphs (kgs) with large language models (llms): models, evaluation metrics, benchmarks, and challenges. Discover Artificial Intelligence 4, 76.

Inoue, N., Kuriya, Y., Kobayashi, S., Inui, K., 2015. Recognizing potential traffic risks through logic-based deep scene understanding. In: Proc. 22nd ITS World Congress. 1–12.

International Organization for Standardization (ISO), 2018. Road vehicles – Functional safety, ISO 26262. https://www.iso.org/standard/68383.html.

International Organization for Standardization (ISO), 2022a. Road vehicles — Test scenarios for automated driving systems — Scenario based safety evaluation framework, ISO 34502. https://www.iso.org/standard/78951.html.

International Organization for Standardization (ISO), 2022b. Road vehicles – Safety of the intended functionality, ISO 21448. https://www.iso.org/standard/77490.html.

Jaimini, U., Henson, C., Sheth, A., 2024. Causal knowledge graph for scene understanding in autonomous driving. In: International Semantic Web Conference, Fall 2024.. 1–3.

Jaradat, S., Nayak, R., Paz, A., Ashqar, H.I., Elhenawy, M., 2024. Multitask learning for crash analysis: A fine-tuned llm framework using twitter data. Smart Cities 7, 2422–2465.

Ji, S., Liu, L., Xi, J., Zhang, X., Li, X., 2024. Klr-kgc: Knowledge-guided llm reasoning for knowledge graph completion. Electronics (2079-9292) 13.

Jia, X., Yang, Z., Li, Q., Zhang, Z., Yan, J., 2024. Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. Advances in Neural Information Processing Systems 37, 819–844.

Jiang, B., Zhang, Z., Lin, D., Tang, J., Luo, B., 2019. Semi-supervised learning with graph learning-convolutional networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 11313–11320.

Jiang, J., Zhou, K., Zhao, W.X., Song, Y., Zhu, C., Zhu, H., et al., 2025. Kg-agent: An efficient autonomous agent framework for complex reasoning over knowledge graph. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 9505–9523.

Jiang, X., Xu, C., Shen, Y., Sun, X., Tang, L., Wang, S., et al., 2023. On the evolution of knowledge graphs: A survey and perspective. https://doi.org/10.48550/arXiv.2310.04835.

Joseph, K., Khan, S., Khan, F.S., Balasubramanian, V.N., 2021. Towards open world object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 5830–5840.

Kaleeswaran, A.P., Nordmann, A., ul Mehdi, A., 2019. Towards integrating ontologies into verification for autonomous driving. In: ISWC (Satellites). 319–320.

Kang, M., Kwak, J.M., Baek, J., Hwang, S.J., 2022. Knowledge-consistent dialogue generation with knowledge graphs. In: ICML 2022 Workshop on Knowledge Retrieval and Language Models. 1–23.

Katsumi, M., Fox, M., 2018. Ontologies for transportation research: A survey.

Transportation Research Part C: Emerging Technologies 89, 53–82.

Kieu, K., Bergstrand, O., 2024. Empowering Automotive Software Development with LLM-RAG Integration-A study on leveraging the RAG-framework for AUTOSAR and automotive safety standards and specifications. Ph.D. thesis. Gothenburg, SWE: Gothenburg University.

Kim, B., Hong, T., Ko, Y., Seo, J., 2020. Multi-task learning for knowledge graph completion with pre-trained language models. In: Proceedings of the 28th international conference on computational linguistics. 1737–1743.

Kim, J., Jin, J., Heo, J., Yoon, J., Hwang, S.J., 2022. Realistic world model for autonomous driving: Integrating physical constraints and multi-agent interactions. https://openreview.net/forum?id=r91tAISb88.

Klück, F., Li, Y., Nica, M., Tao, J., Wotawa, F., 2018. Using ontologies for test suites generation for automated and autonomous driving functions. In: 2018 IEEE International symposium on software reliability engineering workshops (ISSREW). 118–123.

Köhler, S., Baker, R., Martinovic, I., 2022. Signal injection attacks against ccd image sensors. In: Proceedings of the 2022 ACM on Asia Conference on Computer and Communications Security. 294–308.

Köhler, S., Lovisotto, G., Birnbach, S., Baker, R., Martinovic, I., 2021. They see me rollin': Inherent vulnerability of the rolling shutter in cmos image sensors. In: Proceedings of the 37th Annual Computer Security Applications Conference. 399–413.

Kommineni, V.K., König-Ries, B., Samuel, S., 2024. Towards the automation of knowledge graph construction using large language models. In: NLP4KGC@ SEMANTiCS. 19–34.

Krajewski, R., Bock, J., Kloeker, L., Eckstein, L., 2018. The highd dataset: A drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems. In: 2018 21st international conference on intelligent transportation systems (ITSC). 2118–2125.

Kumar, A., Pandey, A., Gadia, R., Mishra, M., 2020. Building knowledge graph using pre-trained language model for learning entity-aware relationships. In: 2020 IEEE international conference on computing, power and communication technologies (GUCON). 310–315.

Kung, J.C., Chang, C.H., Huang, H.H., Chien, K.C., 2024. A narrative assistant for traffic accidents based on large language models (llm). In: Legal Knowledge and Information Systems. IOS Press. 84–94.

Kunze, L., Bruls, T., Suleymanov, T., Newman, P., 2018. Reading between the lanes: Road layout reconstruction from partially segmented scenes. In: 2018 21st International Conference on Intelligent Transportation Systems (ITSC). 401–408.

Lee, U., Jung, H., Jeon, Y., Sohn, Y., Hwang, W., Moon, J., et al., 2024. Few-shot is enough: exploring chatgpt prompt engineering method for automatic question generation in english education. Education and Information Technologies 29, 11483–11515.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., et al., 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems 33, 9459–9474.

Li, C., Chan, S.H., Chen, Y.T., 2023. Droid: Driver-centric risk object identification. IEEE transactions on pattern analysis and machine intelligence .

Li, C., Meng, Y., Chan, S.H., Chen, Y.T., 2020. Learning 3d-aware egocentric spatial-temporal interaction via graph convolutional networks. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). 8418–8424.

Li, D., Xu, F., 2024. Synergizing knowledge graphs with large language models: a comprehensive review and future prospects. https://doi.org/10.48550/arXiv.2407.18470.

Li, J., Li, D., Xiong, C., Hoi, S., 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning. 12888–12900.

Li, K., Chen, K., Wang, H., Hong, L., Ye, C., Han, J., et al., 2022. Coda: A real-world road corner case dataset for object detection in autonomous driving. In: European Conference on Computer Vision. 406–423.

Li, M., Miao, S., Li, P., 2024. Simple is effective: The roles of graphs and large language models in knowledge-graph-based retrieval-augmented generation. https://doi.org/10.48550/arXiv.2410.20724.

Li, S., Azfar, T., Ke, R., 2024. Chatsumo: Large language model for automating traffic scenario generation in simulation of urban mobility. IEEE Transactions on Intelligent Vehicles .

Li, S., Gao, Y., Jiang, H., Yin, Q., Li, Z., Yan, X., et al., 2023. Graph reasoning for question answering with triplet retrieval. In: Findings of the Association for Computational Linguistics: ACL 2023. 3366–3375.

Li, S., Li, X., Shang, L., Sun, C.J., Liu, B., Ji, Z., et al., 2022. Pre-training language models with deterministic factual knowledge. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. 11118–11131.

Li, Y., Liu, W., 2022. Sudden event prediction based on event knowledge graph. Applied Sciences 12, 11195.

Li, Y., Tao, J., Wotawa, F., 2020. Ontology-based test generation for automated and autonomous driving functions. Information and software technology 117, 106200.

Liao, D., Qi, M., Shu, P., Zhang, Z., Lin, Y., Liu, L., et al., 2025. Robodrivevlm: A novel benchmark and baseline towards robust vision-language models for autonomous driving. https://doi.org/10.48550/arXiv.2512.01300.

Liao, H., Li, Y., Wang, C., Guan, Y., Tam, K., Tian, C., et al., 2024. When, where, and what? a benchmark for accident anticipation and localization with large language models. In: Proceedings of the 32nd ACM International Conference on Multimedia. 8–17.

Lin, J., Tang, J., Tang, H., Yang, S., Chen, W.M., Wang, W.C., et al., 2024. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. Proceedings of machine learning and systems 6, 87–100.

Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. 2980–2988.

Lin, Z., Wang, Y., Tang, Z., 2024. Training-free open-ended object detection and segmentation via attention as prompts. Advances in Neural Information Processing Systems 37, 69588–69606.

Lis, K., Nakka, K., Fua, P., Salzmann, M., 2019. Detecting the unexpected via image resynthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2152–2161.

Liu, C., Yang, S., 2022. Using text mining to establish knowledge graph from accident/incident reports in risk assessment. Expert Systems with Applications 207, 117991.

Liu, G., Zhang, Y., Li, Y., Yao, Q., 2024. Explore then determine: A gnn-llm synergy framework for reasoning over knowledge graph. https://doi.org/10.48550/arXiv.2406.01145.

Liu, J., Yan, X., Peng, L., Yang, L., Zhang, L., Luo, Y., et al., 2025. Seeing before observable: Potential risk reasoning in autonomous driving via vision language models. https://doi.org/10.48550/arXiv.2511.22928.

Liu, W., Zhou, P., Zhao, Z., Wang, Z., Ju, Q., Deng, H., et al., 2020. K-bert: Enabling language representation with knowledge graph. In: Proceedings of the AAAI conference on artificial intelligence. 2901–2908.

Liu, X., Wu, H., Yu, D., Chen, Y., Wu, H., 2025. A construction and representation learning method for a traffic accident knowledge graph based on the enhanced transd model. Applied Sciences 15, 6031.

Liu, X., Zheng, R., Wang, H., Butala, M.D., Liu, D., Ren, X., et al., 2021. A knowledge management framework for vehicle hazard analysis. In: 2021 IEEE International Conference on e-Business Engineering (ICEBE). 165–169.

Liu, Y., Zhou, Y., Liu, Y., Xu, Z., He, Y., 2025. Intelligent fault diagnosis for cnc through the integration of large language models and domain knowledge graphs. Engineering .

Liu, Z., He, S., Ding, F., Tan, H., Liu, Y., 2023. Exploring the potential of social

media data in interpreting traffic congestion: A case study of jiangsu freeways. In: CICTP 2023. ascelibrary. 1535–1546.

Long, J., 2023. Large language model guided tree-of-thought. https://doi.org/10.48550/arXiv.2305.08291.

Lovisotto, G., Turner, H., Sluganovic, I., Strohmeier, M., Martinovic, I., 2021. {SLAP}: Improving physical adversarial examples with {Short-Lived} adversarial perturbations. In: 30th USENIX Security Symposium (USENIX Security 21). 1865–1882.

Lu, Q., Ma, M., Dai, X., Wang, X., Feng, S., 2024. Realistic corner case generation for autonomous vehicles with multimodal large language model. https://doi.org/10.48550/arXiv.2412.00243.

Lu, Y., Tian, Y., Bi, Y., Chen, B., Peng, X., 2024. Diavio: Llm-empowered diagnosis of safety violations in ads simulation testing. In: Proceedings of the 33rd ACM SIGSOFT International Symposium on Software Testing and Analysis. 376–388.

Luettin, J., Monka, S., Henson, C., Halilaj, L., 2022. A survey on knowledge graph-based methods for automated driving. In: Iberoamerican Knowledge Graphs and Semantic Web Conference. 16–31.

Luo, L., Ju, J., Xiong, B., Li, Y.F., Haffari, G., Pan, S., 2025. Chatrule: Mining logical rules with large language models for knowledge graph reasoning. In: Pacific-asia conference on knowledge discovery and data mining. 314–325.

Luo, L., Li, Y.F., Haffari, G., Pan, S., 2023. Reasoning on graphs: Faithful and interpretable large language model reasoning. https://doi.org/10.48550/arXiv.2310.01061.

Luo, S., Prabhu, Y., Ossowski, T., Chen, K., Hu, J., 2026. Riskcuebench: Benchmarking anticipatory reasoning from early risk cues in video-language models. https://doi.org/10.48550/arXiv.2601.03369.

Luo, W., Zhang, H., Yang, X., Bo, L., Yang, X., Li, Z., et al., 2020. Dynamic heterogeneous graph neural network for real-time event prediction. In: Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining. 3213–3223.

Luo, Y., Li, F., Xu, S., Lai, Z., Yang, L., Chen, Q., et al., 2025. Adathinkdrive: Adaptive thinking via reinforcement learning for autonomous driving. https://doi.org/10.48550/arXiv.2509.13769.

Luu, Q.K., Deng, X., Van Ho, A., Nakahira, Y., 2024. Context-aware llm-based safe control against latent risks. https://doi.org/10.48550/arXiv.2403.11863.

Ma, Y., Cao, X., Ye, W., Cui, C., Mei, K., Wang, Z., 2024. Learning autonomous driving tasks via human feedbacks with large language models. In: Findings of the Association for Computational Linguistics: EMNLP 2024. 4985–4995.

Malawade, A.V., Yu, S.Y., Hsu, B., Muthirayan, D., Khargonekar, P.P., Al Faruque, M.A., 2022. Spatiotemporal scene-graph embedding for autonomous vehicle collision prediction. IEEE Internet of Things Journal 9, 9379–9388.

Malla, S., Choi, C., Dwivedi, I., Choi, J.H., Li, J., 2023. Drama: Joint risk localization and captioning in driving. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. 1043–1052.

Manas, K., Paschke, A., 2025. Knowledge integration strategies in autonomous vehicle prediction and planning: A comprehensive survey. In: 2025 IEEE Intelligent Vehicles Symposium (IV). 694–701.

Manas, K., Zwicklbauer, S., Paschke, A., 2024. Tr2mtl: Llm based framework for metric temporal logic formalization of traffic rules. In: 2024 IEEE Intelligent Vehicles Symposium (IV). 1206–1213.

Manzour, M., Ballardini, A., Izquierdo, R., Sotelo, M., 2024. Vehicle lane change prediction based on knowledge graph embeddings and bayesian inference. In: 2024 IEEE Intelligent Vehicles Symposium (IV). 1893–1900.

Manzour, M., Ballardini, A., Izquierdo, R., Sotelo, M., 2025. Explainable lane change prediction for near-crash scenarios using knowledge graph embeddings and retrieval augmented generation. In: 2025 IEEE Intelligent Vehicles Symposium (IV). 1538–1545.

Mao, J., Qian, Y., Ye, J., Zhao, H., Wang, Y., 2023a. Gpt-driver: Learning to drive with gpt. https://doi.org/10.48550/arXiv.2310.01415.

Mao, J., Ye, J., Qian, Y., Pavone, M., Wang, Y., 2023b. A language agent for autonomous driving. https://doi.org/10.48550/arXiv.2311.10813.

Matheus, C.J., Kokar, M.M., Baclawski, K., 2003. A core ontology for situation awareness. In: Proceedings of the sixth international conference on information fusion. 545–552.

Mavromatis, C., Karypis, G., 2024. Gnn-rag: Graph neural retrieval for large language model reasoning. https://doi.org/10.48550/arXiv.2405.20139.

McCarthy, R.L., 2022. Autonomous vehicle accident data analysis: California ol 316 reports: 2015–2020. ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part B: Mechanical Engineering 8, 034502.

Melnyk, I., Dognin, P., Das, P., 2021. Grapher: Multi-stage knowledge graph construction using pretrained language models. In: NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications. 1–12.

Meng, X., Zhang, Y., Huang, Z., Lu, Z., Ji, Z., Yin, Y., et al., 2025. Is your vlm for autonomous driving safety-ready? a comprehensive benchmark for evaluating external and in-cabin risks. https://doi.org/10.48550/arXiv.2511.14592.

Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.t., Koh, P., et al., 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 12076–12100.

Mlodzian, L., Sun, Z., Berkemeyer, H., Monka, S., Wang, Z., Dietze, S., et al., 2023. nuscenes knowledge graph-a comprehensive semantic representation of traffic scenes for trajectory prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 42–52.

Mohammad, M.A., Kaloskampis, I., Hicks, Y., Setchi, R., 2015. Ontology-based framework for risk assessment in road scenes using videos. Procedia Computer Science 60, 1532–1541.

Monka, S., Halilaj, L., Schmid, S., Rettinger, A., 2021. Learning visual models using a knowledge graph as a trainer. In: The Semantic Web–ISWC 2021: 20th International Semantic Web Conference, ISWC 2021, Virtual Event, October 24–28, 2021, Proceedings 20. 357–373.

Morignot, P., Nashashibi, F., 2012. An ontology-based approach to relax traffic regulation for autonomous vehicle assistance. https://doi.org/10.48550/arXiv.1212.0768.

Mumtarin, M., Chowdhury, M.S., Wood, J., 2023. Large language models in analyzing crash narratives–a comparative study of chatgpt, bard and gpt-4. https://doi.org/10.48550/arXiv.2308.13563.

Munikoti, S., Acharya, A., Wagle, S., Horawalavithana, S., 2023. Atlantic: Structure-aware retrieval-augmented language model for interdisciplinary science. https://doi.org/10.48550/arXiv.2311.12289.

Muppalla, R., Lalithsena, S., Banerjee, T., Sheth, A., 2017. A knowledge graph framework for detecting traffic events using stationary cameras. In: Proceedings of the 2017 ACM on Web Science Conference. 431–436.

Nag Chowdhury, S., Wickramarachchi, R., Gad-Elrab, M.H., Stepanova, D., Henson, C., 2021. Towards leveraging commonsense knowledge for autonomous driving. In: 20th International Semantic Web Conference. 1–5.

National Highway Traffic Safety Administration (NHTSA), 2024a. Crash Report Sampling System (CRSS). https://www.nhtsa.gov/sites/nhtsa.gov/files/morgan-sae2016-datamod-v3.pdf.

National Highway Traffic Safety Administration (NHTSA), 2024b. Fatality Analysis Reporting System (FARS). https://www.nhtsa.gov/research-data/fatality-analysis-reporting-system-fars.

Neurohr, C., Westhofen, L., Butz, M., Bollmann, M.H., Eberle, U., Galbas, R., 2021. Criticality analysis for the verification and validation of automated vehicles. IEEE Access 9, 18016–18041.

Ni, B., Liu, Z., Wang, L., Lei, Y., Zhao, Y., Cheng, X., et al., 2025. Towards trustworthy retrieval augmented generation for large language models: A survey. https://doi.org/10.48550/arXiv.2502.06872.

Nouri, A., Cabrero-Daniel, B., Törner, F., Sivencrona, H., Berger, C., 2024a.

Engineering safety requirements for autonomous driving with large language models. In: 2024 IEEE 32nd International Requirements Engineering Conference (RE). 218–228.

Nouri, A., Cabrero-Daniel, B., Torner, F., Sivencrona, H., Berger, C., 2024b. Welcome your new ai teammate: On safety analysis by leashing large language models. In: Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI. 172–177.

Odu, O., Belle, A.B., Wang, S., 2025. Llm-based safety case generation for baidu apollo: Are we there yet? In: 2025 IEEE/ACM 4th International Conference on AI Engineering–Software Engineering for AI (CAIN). 222–233.

Ojima, Y., Sakaji, H., Nakamura, T., Sakata, H., Seki, K., Teshigawara, Y., et al., 2024. Knowledge management for automobile failure analysis using graph rag. In: 2024 IEEE International Conference on Big Data (BigData). 6624–6631.

Oltramari, A., Francis, J., Henson, C., Ma, K., Wickramarachchi, R., 2020. Neuro-symbolic architectures for context understanding. In: Knowledge Graphs for eXplainable Artificial Intelligence: Foundations, Applications and Challenges. IOS Press. 143–160.

Pan, H., Yi, S., Yang, S., Qi, L., Hu, B., Xu, Y., et al., 2024. The solution for cvpr2024 foundational few-shot object detection challenge. https://doi.org/10.48550/arXiv.2406.12225.

Pan, J.Z., Razniewski, S., Kalo, J.C., Singhania, S., Chen, J., Dietze, S., et al., 2023. Large language models and knowledge graphs: Opportunities and challenges. https://doi.org/10.48550/arXiv.2308.06374.

Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., Wu, X., 2024. Unifying large language models and knowledge graphs: A roadmap. IEEE Transactions on Knowledge and Data Engineering 36, 3580–3599.

Panagiotaki, E., Pramatarov, G., Kunze, L., De Martini, D., 2025. Graphscene: On-demand critical scenario generation for autonomous vehicles in simulation. In: 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). 13801–13808.

Parikh, C., Rawat, D., Ghosh, T., Sarvadevabhatla, R.K., others, 2025. Roadsocial: A diverse videoqa dataset and benchmark for road event understanding from social video narratives. In: Proceedings of the Computer Vision and Pattern Recognition Conference. 19002–19011.

Parisot, S., Yang, Y., McDonagh, S., 2023. Learning to name classes for vision and language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 23477–23486.

Peng, L., Li, B., Yu, W., Yang, K., Shao, W., Wang, H., 2023. Sotif entropy: Online sotif risk quantification and mitigation for autonomous driving. IEEE Transactions on Intelligent Transportation Systems .

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., et al., 2019. Language models as knowledge bases? In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). 2463–2473.

Petrovic, N., Lebioda, K., Zolfaghari, V., Schamschurko, A., Kirchner, S., Purschke, N., et al., 2024. Llm-driven testing for autonomous driving scenarios. In: 2024 2nd International Conference on Foundation and Large Language Models (FLLM). 173–178.

Pham, Q.H., Sevestre, P., Pahwa, R.S., Zhan, H., Pang, C.H., Chen, Y., et al., 2020. A* 3d dataset: Towards autonomous driving in challenging environments. In: 2020 IEEE International conference on Robotics and Automation (ICRA). 2267–2273.

Pinggera, P., Ramos, S., Gehrig, S., Franke, U., Rother, C., Mester, R., 2016. Lost and found: detecting small road hazards for self-driving vehicles. In: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). 1099–1106.

Qasemi, E., Francis, J.M., Oltramari, A., 2023. Traffic-domain video question answering with automatic captioning. https://doi.org/10.48550/arXiv.2307.09636.

Qi, Y., Kyebambo, G., Xie, S., Shen, W., Wang, S., Xie, B., et al., 2024. Safety control of service robots with llms and embodied knowledge graphs. https://doi.org/10.48550/arXiv.2405.17846.

Qi, Y., Zhao, X., Khastgir, S., Huang, X., 2025. safety analysis in the era of large language models: a case study of stpa using chatgpt. Machine Learning with Applications , 100622.

Qian, T., Chen, J., Zhuo, L., Jiao, Y., Jiang, Y.G., 2024. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. In: Proceedings of the AAAI Conference on Artificial Intelligence. 4542–4550.

Qin, L., Chen, Q., Zhou, Y., Chen, Z., Li, Y., Liao, L., et al., 2025. A survey of multilingual large language models. Patterns 6.

Qiu, H., Ayara, A., Glimm, B., 2020a. A knowledge architecture layer for map data in autonomous vehicles. In: 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC). 1–6.

Qiu, H., Ayara, A., Glimm, B., 2020b. Ontology-based processing of dynamic maps in automated driving. In: KEOD. 98–107.

Qiu, H., Ayara, A., Glimm, B., 2021. Ontology-based map data quality assurance. In: The Semantic Web: 18th International Conference, ESWC 2021, Virtual Event, June 6–10, 2021, Proceedings 18. 73–89.

Radford, A., 2018. Improving language understanding by generative pre-training. https://www.mikecaptain.com/resources/pdf/GPT-1.pdf.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al., 2019. Language models are unsupervised multitask learners. OpenAI blog 1, 9.

Raeisdanaei, A., Kim, J., Liao, M., Kochhar, S., 2025. An llm-integrated framework for completion, management, and tracing of stpa. https://doi.org/10.48550/arXiv.2503.12043.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., et al., 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research 21, 1–67.

Rakhmawati, N.A., Awwab, Y., Najib, A.C., Irsyad, A., 2022. Ontology-based traffic accident information extraction on twitter in indonesia. Inteligencia Artificial 25, 1–12.

Regele, R., 2008. Using ontology-based traffic models for more efficient decision making of autonomous vehicles. In: Fourth International Conference on Autonomic and Autonomous Systems (ICAS'08). 94–99.

Renz, K., Chen, L., Arani, E., Sinavski, O., 2025. Simlingo: Vision-only closed-loop autonomous driving with language-action alignment. In: Proceedings of the Computer Vision and Pattern Recognition Conference. 11993–12003.

Riedmaier, S., Ponn, T., Ludwig, D., Schick, B., Diermeyer, F., 2020. Survey on scenario-based safety assessment of automated vehicles. IEEE access 8, 87456–87477.

Roh, K.H., Lee, K.S., 2015. An ontology-based hazard analysis and risk assessment for automotive functional safety. Journal of The Korea Society of Computer and Information 20, 9–17.

Rosset, C., Xiong, C., Phan, M., Song, X., Bennett, P., Tiwary, S., 2020. Knowledge-aware language model pretraining. https://doi.org/10.48550/arXiv.2007.00655.

Sachdeva, E., Agarwal, N., Chundi, S., Roelofs, S., Li, J., Kochenderfer, M., et al., 2024. Rank2tell: A multimodal driving dataset for joint importance ranking and reasoning. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. 7513–7522.

Sahoo, P., Singh, A.K., Saha, S., Jain, V., Mondal, S., Chadha, A., 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. https://doi.org/10.48550/arXiv.2402.07927.

Sanmartin, D., 2024. Kg-rag: Bridging the gap between knowledge and creativity. https://doi.org/10.48550/arXiv.2405.12035.

Sauerbier, J., Bock, J., Weber, H., Eckstein, L., 2019. Definition of scenarios for safety validation of automated driving functions. ATZ worldwide 121, 42–45.

Schmid, S., Henson, C., Tran, T., 2019. Using knowledge graphs to search an enterprise data lake. In: The Semantic Web: ESWC 2019 Satellite Events: ESWC 2019 Satellite Events, Portorož, Slovenia, June 2–6, 2019, Revised Selected Papers 16. 262–266.

Scholtes, M., Westhofen, L., Turner, L.R., Lotto, K., Schuldes, M., Weber, H., et al., 2021. 6-layer model for a structured description and categorization of urban traffic and environment. IEEE Access 9, 59131–59147.

Schuldt, F., 2017. Ein Beitrag für den methodischen Test von automatisierten Fahrfunktionen mit Hilfe von virtuellen Umgebungen, Towards testing of automated driving functions in virtual driving environments. Ph.D. thesis. Braunschweig, DE: Technischen Universitaet Braunschweig.

Schuldt, F., Saust, F., Lichte, B., Maurer, M., Scholz, S., 2013. Effiziente systematische testgenerierung für fahrerassistenzsysteme in virtuellen umgebungen. Automatisierungssysteme, Assistenzsysteme und Eingebettete Systeme Für Transportmittel 4, 1–7.

Sen, P., Mavadia, S., Saffari, A., 2023. Knowledge graph-augmented language models for complex question answering. In: Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (NLRSE). 1–8.

Shafiq, S., Awan, H.M., Khan, A.A., Amin, W., 2024. Driving like humans: Leveraging vision large language models for road anomaly detection. In: 2024 3rd International Conference on Emerging Trends in Electrical, Control, and Telecommunication Engineering (ETECTE). 1–6.

Shao, H., Hu, Y., Wang, L., Song, G., Waslander, S.L., Liu, Y., et al., 2024. Lmdrive: Closed-loop end-to-end driving with large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 15120–15130.

Shao, S., Zhao, Z., Li, B., Xiao, T., Yu, G., Zhang, X., et al., 2018. Crowdhuman: A benchmark for detecting human in a crowd. https://doi.org/10.48550/arXiv.1805.00123.

Shearer, R.D., Motik, B., Horrocks, I., 2008. Hermit: A highly-efficient owl reasoner. In: Owled. 91.

Shen, T., Mao, Y., He, P., Long, G., Trischler, A., Chen, W., 2020. Exploiting structured knowledge in text via graph-guided representation learning. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP). 8980–8994.

Shen, Y., Song, K., Tan, X., Li, D., Lu, W., Zhuang, Y., 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. Advances in Neural Information Processing Systems 36, 38154–38180.

Sima, C., Renz, K., Chitta, K., Chen, L., Zhang, H., Xie, C., et al., 2024. Drivelm: Driving with graph visual question answering. In: European Conference on Computer Vision. 256–274.

Sinha, R., Elhafsi, A., Agia, C., Foutter, M., Schmerling, E., Pavone, M., 2024. Real-time anomaly detection and reactive planning with large language models. https://doi.org/10.48550/arXiv.2407.08735.

Sirin, E., Parsia, B., Grau, B.C., Kalyanpur, A., Katz, Y., 2007. Pellet: A practical owl-dl reasoner. Journal of Web Semantics 5, 51–53.

Sivakumar, M., Belle, A.B., Shan, J., Shahandashti, K.K., 2024. Prompting gpt–4 to support automatic safety case generation. Expert Systems with Applications 255, 124653.

Song, R., Ozmen, M.O., Kim, H., Bianchi, A., Celik, Z.B., 2024. Enhancing llm-based autonomous driving agents to mitigate perception attacks. https://doi.org/10.48550/arXiv.2409.14488.

Strich, J., Scharfenberg, A., Biemann, C., Semmann, M., 2025. Encourage: Evaluating rag local, fast, and reliable. https://doi.org/10.48550/arXiv.2511.04696.

Suman, V., Pham, P., Bera, A., 2023. Raist: Learning risk aware traffic interactions via spatio-temporal graph convolutional networks. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). 5545–5550.

Sun, H., Li, Y., Deng, L., Li, B., Hui, B., Li, B., et al., 2023. History semantic graph enhanced conversational kbqa with temporal information modeling. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 3521–3533.

Sun, J., Xu, C., Tang, L., Wang, S., Lin, C., Gong, Y., et al., 2023. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. https://doi.org/10.48550/arXiv.2307.07697.

Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., et al., 2020. Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2446–2454.

Sun, Y., Wang, S., Feng, S., Ding, S., Pang, C., Shang, J., et al., 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. https://doi.org/10.48550/arXiv.2107.02137.

Sun, Y., Wang, S., Li, Y., Feng, S., Chen, X., Zhang, H., et al., 2019. Ernie: Enhanced representation through knowledge integration. https://doi.org/10.48550/arXiv.1904.09223.

Sun, Y., Wang, S., Li, Y., Feng, S., Tian, H., Wu, H., et al., 2020. Ernie 2.0: A continual pre-training framework for language understanding. In: Proceedings of the AAAI conference on artificial intelligence. 8968–8975.

Sun, Z., Wang, Z., Halilaj, L., Luettin, J., 2024. Semanticformer: Holistic and semantic traffic scene representation for trajectory prediction using knowledge graphs. IEEE Robotics and Automation Letters .

Suryawanshi, Y., Qiu, H., Ayara, A., Glimm, B., 2019. An ontological model for map data in automotive systems. In: 2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE). 140–147.

Takahashi, R., Inoue, N., Kuriya, Y., Kobayashi, S., Inui, K., 2016. Explaining potential risks in traffic scenes by combining logical inference and physical simulation. International Journal of Machine Learning and Computing 6, 248.

Tami, M.A., Ashqar, H.I., Elhenawy, M., Glaser, S., Rakotonirainy, A., 2024. Using multimodal large language models (mllms) for automated detection of traffic safety-critical events. Vehicles 6, 1571–1590.

Tami, M.A., Elhenawy, M., Ashqar, H.I., 2025a. Hazardnet: A small-scale vision language model for real-time traffic safety detection at edge devices. In: 2025 IEEE 4th International Conference on Computing and Machine Intelligence (ICMI). 1–5.

Tami, M.A., Elhenawy, M., Ashqar, H.I., 2025b. Multimodal large language models for enhanced traffic safety: A comprehensive. In: Intelligent Systems, Blockchain, and Communication Technologies: Selected papers from the International Conference on Intelligent Systems, Blockchain, and Communication Technologies (ISBCom25). Volume 1. 132.

Tang, Y., Da Costa, A.A.B., Zhang, X., Patrick, I., Khastgir, S., Jennings, P., 2023. Domain knowledge distillation from large language model: An empirical study in the autonomous driving domain. In: 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC). 3893–3900.

Tang, Z., Zhang, S., Deng, J., Wang, C., You, G., Huang, Y., et al., 2025. Vlmplanner: Integrating visual language models with motion planning. In: Proceedings of the 33rd ACM International Conference on Multimedia. 5040–5049.

Tian, H., Han, X., Zhou, Y., Wu, G., Guo, A., Cheng, M., et al., 2024. Lmm-enhanced safety-critical scenario generation for autonomous driving system testing from non-accident traffic videos. https://doi.org/10.48550/arXiv.2406.10857.

Tian, X., Gu, J., Li, B., Liu, Y., Hu, C., Wang, Y., et al., 2024. Drivevlm: The convergence of autonomous driving and large vision-language models. https://doi.org/10.48550/arXiv.2402.12289.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., et al., 2023. Llama: Open and efficient foundation language models. https://doi.org/10.48550/arXiv.2302.13971.

Tran, T.K., Le-Tuan, A., Nguyen-Duc, M., Yuan, J., Le-Phuoc, D., 2022. Fantastic data and how to query them. https://doi.org/10.48550/arXiv.2201.05026.

Tu, Z., Niu, L., Fan, W., Zhang, T., 2025. Multi-modal traffic scenario generation for autonomous driving system testing. Proceedings of the ACM on Software Engineering 2, 1733–1756.

Tupayachi, J., Xu, H., Omitaomu, O.A., Camur, M.C., Sharmin, A., Li, X., 2024. Towards next-generation urban decision support systems through ai-powered construction of scientific ontology using large language models—a case in optimizing intermodal freight transportation. Smart Cities 7, 2392–2421.

Ulbrich, S., Menzel, T., Reschka, A., Schuldt, F., Maurer, M., 2015. Defining and substantiating the terms scene, situation, and scenario for automated driving. In: 2015 IEEE 18th international conference on intelligent transportation systems. 982–988.

Ulbrich, S., Nothdurft, T., Maurer, M., Hecker, P., 2014. Graph-based context representation, environment modeling and information aggregation for automated driving. In: 2014 IEEE Intelligent Vehicles Symposium Proceedings. 541–547.

United Nations Economic Commission for Europe (UNECE), 2021. Un regulation no. 157 - automated lane keeping systems (alks). https://unece.org/transport/documents/2021/03/standards/un-regulation-no-157-automated-lane-keeping-systems-alks.

Urbieta, I., Lizaso, A., Loyo, E., Nieto, M., Aginako, N., 2023. Introducing mosa: Enhancing traffic event analysis for situational awareness services. Semantic Web , 1–13.

Varshney, N., Yao, W., Zhang, H., Chen, J., Yu, D., 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. https://doi.org/10.48550/arXiv.2307.03987.

Wang, G., Wei, X., Liu, J., Zhang, R., Zhang, Y., Zhang, K., et al., 2024. Mr-mllm: Mutual reinforcement of multimodal comprehension and vision perception. https://doi.org/10.48550/arXiv.2406.15768.

Wang, H., Qin, J., Bastola, A., Chen, X., Suchanek, J., Gong, Z., et al., 2024. Visiongpt: Llm-assisted real-time anomaly detection for safe visual navigation. https://doi.org/10.48550/arXiv.2403.12415.

Wang, J., Huang, W., Qiu, M., Shi, Q., Wang, H., Li, X., et al., 2022. Knowledge prompting in pre-trained language model for natural language understanding. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. 3164–3177.

Wang, J., Wang, X., 2011. An ontology-based traffic accident risk mapping framework. In: International Symposium on Spatial and Temporal Databases. 21–38.

Wang, J., Wu, Z., Dong, Q., Meng, L., Xue, Y., Yang, Y., 2025. Hybrid-driving: An autonomous driving decision framework integrating large language models, knowledge graphs and driving rules. In: Proceedings of the AAAI Conference on Artificial Intelligence. 826–833.

Wang, S., Fan, W., Feng, Y., Shanru, L., Ma, X., Wang, S., et al., 2025. Knowledge graph retrieval-augmented generation for llm-based recommendation. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 27152–27168.

Wang, X., Gao, T., Zhu, Z., Zhang, Z., Liu, Z., Li, J., et al., 2021. Kepler: A unified model for knowledge embedding and pre-trained language representation. Transactions of the Association for Computational Linguistics 9, 176–194.

Wang, Y., He, J., Fan, L., Li, H., Chen, Y., Zhang, Z., 2024a. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 14749–14759.

Wang, Y., Jiao, R., Zhan, S.S., Lang, C., Huang, C., Wang, Z., et al., 2023. Empowering autonomous driving with large language models: A safety perspective. https://doi.org/10.48550/arXiv.2312.00812.

Wang, Y., Lipka, N., Rossi, R.A., Siu, A., Zhang, R., Derr, T., 2024b. Knowledge graph prompting for multi-document question answering. In: Proceedings of the AAAI Conference on Artificial Intelligence. 19206–19214.

Wang, Y., Liu, Q., Fan, J., Hong, J., Chu, H., Tian, M., et al., 2026. Rac3: Retrieval-augmented corner case comprehension for autonomous driving with vision-language models. IEEE Transactions on Multimedia .

Wang, Y., Luo, W., Bai, J., Cao, Y., Che, T., Chen, K., et al., 2025. Alpamayo-r1: Bridging reasoning and action prediction for generalizable autonomous driving in the long tail. https://doi.org/10.48550/arXiv.2511.00088.

Wang, Z., Sun, Z., Luettin, J., Halilaj, L., 2024. Socialformer: Social interaction modeling with edge-enhanced heterogeneous graph transformers for trajectory prediction. https://doi.org/10.48550/arXiv.2405.03809.

Warren, M., Shamma, D.A., Hayes, P.J., 2021. Knowledge engineering with image data in real-world settings. In: AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering. 1–4.

Wayve, 2024. Driving with language: Introducing wayve's multimodal driving model LINGO-2. https://wayve.ai/press/introducing-lingo-2/.

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.

Wei, Q., 2025. Llm-based autonomous vehicle self-protection mechanisms: A comprehensive review. Applied and Computational Engineering 181, 95–102. ACE Vol.181 (ISSN 2755-2721 print; 2755-273X online).

Wei, Y., Wang, Z., Lu, Y., Xu, C., Liu, C., Zhao, H., et al., 2024. Editable scene simulation for autonomous driving via collaborative llm-agents. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 15077–15087.

Wen, Y., Wang, Z., Sun, J., 2024. Mindmap: Knowledge graph prompting sparks graph of thoughts in large language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 10370–10388.

Westhofen, L., Neurohr, C., Butz, M., Scholtes, M., Schuldes, M., 2022. Using ontologies for the formalization and recognition of criticality for automated driving. IEEE Open Journal of Intelligent Transportation Systems 3, 519–538.

Westhofen, L., Neurohr, C., Jung, J.C., Neider, D., 2024. Answering temporal conjunctive queries over description logic ontologies for situation recognition in complex operational domains. In: International Conference on Tools and Algorithms for the Construction and Analysis of Systems. 167–187.

Westhofen, L., Neurohr, C., Koopmann, T., Butz, M., Schütt, B., Utesch, F., et al., 2023. Criticality metrics for automated driving: A review and suitability analysis of the state of the art. Archives of Computational Methods in Engineering 30, 1–35.

Wickramarachchi, R., Henson, C., Sheth, A., 2020. An evaluation of knowledge graph embeddings for autonomous driving data: Experience and practice. https://doi.org/10.48550/arXiv.2003.00344.

Wickramarachchi, R., Henson, C., Sheth, A., 2021. Knowledge-infused learning for entity prediction in driving scenes. Frontiers in big Data 4, 759110.

Wickramarachchi, R., Henson, C., Sheth, A., 2022. Knowledge-based entity prediction for improved machine perception in autonomous systems. IEEE Intelligent Systems 37, 42–49.

Wolf, P., Ropertz, T., Feldmann, P., Berns, K., 2019. Combining onthologies and behavior-based control for aware navigation in challenging off-road environments. In: ICINCO (2). 135–146.

Wotawa, F., Li, Y., 2018. From ontologies to input models for combinatorial testing. In: IFIP International Conference on Testing Software and Systems. 155–170.

Wu, J., Zhu, J., Qi, Y., Chen, J., Xu, M., Menolascina, F., et al., 2024. Medical graph rag: Towards safe medical large language model via graph retrieval-augmented generation. https://doi.org/10.48550/arXiv.2408.04187.

Wu, K., Li, W., Xiao, X., 2024. Accidentgpt: Large multi-modal foundation model for traffic accident analysis. https://doi.org/10.48550/arXiv.2401.03040.

Wu, S., Wang, H., Yu, W., Yang, K., Cao, D., Wang, F., 2021. A new sotif scenario hierarchy and its critical test case generation based on potential risk assessment. In: 2021 IEEE 1st International Conference on Digital Twins and Parallel Intelligence (DTPI). 399–409.

Xia, Z., Chen, W., Wang, Y., Yang, M.H., 2025. Knowval: A knowledge-augmented and value-guided autonomous driving system. https://doi.org/10.48550/arXiv.2512.20299.

Xiao, A., Yan, W., Zhang, X., Liu, Y., Zhang, H., Liu, Q., 2024. Multi-domain fusion for cargo uav fault diagnosis knowledge graph construction. Autonomous Intelligent Systems 4, 10.

Xiao, D., Dianati, M., Geiger, W.G., Woodman, R., 2023. Review of graph-based hazardous event detection methods for autonomous driving systems. IEEE Transactions on Intelligent Transportation Systems 24, 4697–4715.

Xiao, P., Shao, Z., Hao, S., Zhang, Z., Chai, X., Jiao, J., et al., 2021. Pandaset: Advanced sensor suite dataset for autonomous driving. In: 2021 IEEE International Intelligent Transportation Systems Conference (ITSC). 3095–3101.

Xie, S., Kong, L., Dong, Y., Sima, C., Zhang, W., Chen, Q.A., et al., 2025. Are vlms ready for autonomous driving? an empirical study from the reliability, data and metric perspectives. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 6585–6597.

Xie, X., Zhang, N., Li, Z., Deng, S., Chen, H., Xiong, F., et al., 2022. From discrimination to generation: Knowledge graph completion with generative transformer. In: Companion Proceedings of the Web Conference 2022. 162–165.

Xing, J., Liu, J., Wang, J., Sun, L., Chen, X., Gu, X., et al., 2024. A survey of efficient fine-tuning methods for vision-language models—prompt and adapter. Computers & Graphics 119, 103885.

Xiong, H., Vahedian, A., Zhou, X., Li, Y., Luo, J., 2018. Predicting traffic congestion propagation patterns: A propagation graph approach. In: Proceedings of the 11th ACM SIGSPATIAL international workshop on computational transportation science. 60–69.

Xu, H., Yuan, J., Zhou, A., Xu, G., Li, W., Ban, X., et al., 2024. Genai-powered multi-agent paradigm for smart urban mobility: Opportunities and challenges for integrating large language models (llms) and retrieval-augmented generation (rag) with intelligent transportation systems. https://doi.org/10.48550/arXiv.2409.00494.

Xu, L., Huang, H., Liu, J., 2021. Sutd-trafficqa: A question answering benchmark and an efficient network for video reasoning over traffic events. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 9878–9888.

Xu, Q., Chen, S., Chen, G., Wang, Y., Zhang, Y., 2025. Chatbev: A visual language model that understands bev maps. https://doi.org/10.48550/arXiv.2503.13938.

Xu, S., Liu, S., Jing, C., Li, S., 2023. Event knowledge graph: A review based on scientometric analysis. Applied Sciences 13, 12338.

Xu, W., Pei, H., Yang, J., Shi, Y., Zhang, Y., Zhao, Q., 2024. Exploring critical testing scenarios for decision-making policies: An llm approach. https://doi.org/10.48550/arXiv.2412.06684.

Xu, Y., Shao, W., Li, J., Yang, K., Wang, W., Huang, H., et al., 2022. Sind: A drone dataset at signalized intersection in china. In: 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC). 2471–2478.

Xu, Y., Zhang, Y., Fu, N., 2025. Query-aware graph attention for precise subgraph retrieval in knowledge-augmented reasoning. https://openreview.net/forum?id=MGvhGFCFNK.

Xu, Z., Chen, T., Chen, S., 2024a. A llm-based multimodal warning system for driver assistance. In: 2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC). 1527–1532.

Xu, Z., Zhang, Y., Xie, E., Zhao, Z., Guo, Y., Wong, K.Y.K., et al., 2024b. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. IEEE Robotics and Automation Letters 9, 8186–8193.

Yamazaki, S., Inoue, M., Sakuma, M., Kimoto, T., 2025. Safedriveqa: Benchmarking vision language model for safe driving assessment. In: Proceedings of the 6th Workshop on Intelligent Cross-Data Analysis and Retrieval. 27–31.

Yang, B., Yih, W.t., He, X., Gao, J., Deng, L., 2014. Embedding entities and relations for learning and inference in knowledge bases. https://doi.org/10.48550/arXiv.1412.6575.

Yang, D., Jiang, K., Zhao, D., Yu, C., Cao, Z., Xie, S., et al., 2018. Intelligent and connected vehicles: Current status and future perspectives. Science China Technological Sciences 61, 1446–1471.

Yang, H., Kim, H., Kang, J., 2025. A comparative study on self-driving scenario code generation through prompt engineering based on llm-specific characteristics. Applied Sciences 15, 12502.

Yang, Q., Ni, B., Xiang, S., Hu, H., Peng, H., Jiang, J., 2025. R-4b: Incentivizing general-purpose auto-thinking capability in mllms via bi-mode annealing and reinforce learning. https://doi.org/10.48550/arXiv.2508.21113.

Yang, Y., Lee, K., Dariush, B., Cao, Y., Lo, S.Y., 2025. Follow the rules: reasoning for video anomaly detection with large language models. In: European Conference on Computer Vision. 304–322.

Yang, Y., Zhang, Q., Ikemura, K., Batool, N., Folkesson, J., 2024. Hard cases detection in motion prediction by vision-language foundation models. In: 2024 IEEE Intelligent Vehicles Symposium (IV). 2405–2412.

Yang, Z., Jia, X., Li, H., Yan, J., 2023. Llm4drive: A survey of large language models for autonomous driving. https://doi.org/10.48550/arXiv.2311.01043.

Yang, Z., Liu, Z., Lu, Y., Hou, L., Miao, C., Peng, S., et al., 2025. Geniedrive: Towards physics-aware driving world model with 4d occupancy guided video generation. https://doi.org/10.48550/arXiv.2512.12751.

Yao, H., Han, C., Xu, F., 2022. Reasoning methods of unmanned underwater vehicle situation awareness based on ontology and bayesian network. Complexity 2022, 7143974.

Yao, S., Guo, R., Qin, Y., Meng, M., Cao, J., Lin, Y., et al., 2025. Query as test: An intelligent driving test and data storage method for integrated cockpit-vehicle-road scenarios. https://doi.org/10.48550/arXiv.2506.22068.

Yao, Y., Wang, X., Xu, M., Pu, Z., Wang, Y., Atkins, E., et al., 2022. Dota: unsupervised detection of traffic anomaly in driving videos. IEEE transactions on pattern analysis and machine intelligence .

Ye, H., Qi, M., Liu, Z., Liu, L., Ma, H., 2025. Safedriverag: Towards safe autonomous driving with knowledge graph-based retrieval-augmented generation. In: Proceedings of the 33rd ACM International Conference on Multimedia. 11170–11178.

Yong, G., Jeon, K., Gil, D., Lee, G., 2023. Prompt engineering for zero-shot and few-shot defect detection and classification using a visual-language pretrained model. Computer-Aided Civil and Infrastructure Engineering 38, 1536–1554.

Yu, D., Peng, W., Chen, Y., Yang, Y., Chen, H., 2024. Road traffic accident data management and application analysis based on knowledge graph technology. In: Proceedings of the 2024 International Conference on Digital Society and Artificial Intelligence. 297–302.

Yu, H., Gan, A., Zhang, K., Tong, S., Liu, Q., Liu, Z., 2024. Evaluation of retrieval-augmented generation: A survey. In: CCF Conference on Big Data. 102–120.

Yu, S.Y., Malawade, A.V., Muthirayan, D., Khargonekar, P.P., Al Faruque, M.A., 2021. Scene-graph augmented data-driven risk assessment of autonomous vehicle decisions. IEEE Transactions on Intelligent Transportation Systems 23, 7941–7951.

Yu, W., Zhao, C., Wang, H., Liu, J., Ma, X., Yang, Y., et al., 2024. Online legal driving behavior monitoring for self-driving vehicles. Nature communications 15, 408.

Yuan, J., Sun, S., Omeiza, D., Zhao, B., Newman, P., Kunze, L., et al., 2024. Rag-driver: Generalisable driving explanations with retrieval-augmented in-context learning in multi-modal large language model. https://doi.org/10.48550/arXiv.2402.10828.

Yue, D., Wang, S., Zhao, A., 2009. Traffic accidents knowledge management based on ontology. In: 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery. 447–449.

Zhang, B., Reklos, I., Jain, N., Peñuela, A.M., Simperl, E., 2023. Using large

language models for knowledge engineering (llmke): A case study on wikidata. https://doi.org/10.48550/arXiv.2309.08491.

Zhang, C., Hong, L., Wang, D., Liu, X., Yang, J., Lin, Y., 2024. Research on driving scenario knowledge graphs. Applied Sciences 14, 3804.

Zhang, D., Yuan, Z., Liu, Y., Zhuang, F., Chen, H., Xiong, H., 2020. E-bert: A phrase and product knowledge enhanced language model for e-commerce. https://doi.org/10.48550/arXiv.2009.02835.

Zhang, D., Zheng, H., Yue, W., Wang, X., 2024. Advancing its applications with llms: A survey on traffic management, transportation safety, and autonomous driving. In: International Joint Conference on Rough Sets. 295–309.

Zhang, H., Li, H., Li, F., Ren, T., Zou, X., Liu, S., et al., 2024. Llava-grounding: Grounded visual chat with large multimodal models. In: European Conference on Computer Vision. 19–35.

Zhang, J., Chen, B., Zhang, L., Ke, X., Ding, H., 2021. Neural, symbolic and neural-symbolic reasoning on knowledge graphs. AI Open 2, 14–35.

Zhang, J., Xu, C., Li, B., 2024. Chatscene: Knowledge-enabled safety-critical scenario generation for autonomous vehicles. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 15459–15469.

Zhang, K., Wang, S., Jia, N., Zhao, L., Han, C., Li, L., 2024. Integrating visual large language model and reasoning chain for driver behavior analysis and risk assessment. Accident Analysis & Prevention 198, 107497.

Zhang, L., Zhang, M., Tang, J., Ma, J., Duan, X., Sun, J., et al., 2022. Analysis of traffic accident based on knowledge graph. Journal of advanced transportation 2022, 3915467.

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, T., Wang, Z., Wang, H., Li, J., 2026. Intelligent defensive driving for autonomous vehicles: Framework, strategy and verification. Accident Analysis & Prevention 226, 108355.

Zhang, Z., Zhang, A., Li, M., Smola, A., 2022. Automatic chain of thought prompting in large language models. https://doi.org/10.48550/arXiv.2210.03493.

Zhao, G., Ni, C., Wang, X., Zhu, Z., Zhang, X., Wang, Y., et al., 2025a. Drivedreamer4d: World models are effective data machines for 4d driving scene representation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. 12015–12026.

Zhao, G., Wang, X., Zhu, Z., Chen, X., Huang, G., Bao, X., et al., 2025b. Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. 10412–10420.

Zhao, J., Li, W., Zhu, B., Zhang, P., Huang, Y., Tang, R., 2025. Cradle: An accident scenario generation method based on scenario knowledge graph considering accident causation. IEEE Transactions on Software Engineering .

Zhao, L., Ichise, R., Liu, Z., Mita, S., Sasaki, Y., 2017. Ontology-based driving decision making: A feasibility study at uncontrolled intersections. IEICE TRANSACTIONS on Information and Systems 100, 1425–1439.

Zhao, L., Ichise, R., Mita, S., Sasaki, Y., 2015a. Core ontologies for safe autonomous driving. In: ISWC (Posters & Demos). 1–4.

Zhao, L., Ichise, R., Mita, S., Sasaki, Y., 2015b. Ontologies for advanced driver assistance systems. Proceedings of the Second Research Meeting of the Society for Artificial Intelligence 2015, 03.

Zhao, L., Ichise, R., Sasaki, Y., Liu, Z., Yoshikawa, T., 2016. Fast decision making using ontology-based knowledge base. In: 2016 IEEE Intelligent Vehicles Symposium (IV). 173–178.

Zhao, L., Ichise, R., Yoshikawa, T., Naito, T., Kakinami, T., Sasaki, Y., 2015c. Ontology-based decision making on uncontrolled intersections and narrow roads. In: 2015 IEEE intelligent vehicles symposium (IV). 83–88.

Zhao, M., Liang, C., Ci, Y., 2025. Ontology-enhanced stpa method of scenario safety constraint identification for autonomous driving. Journal of Transportation Engineering, Part A: Systems 151, 04025081.

Zhao, P., Zhang, H., Yu, Q., Wang, Z., Geng, Y., Fu, F., et al., 2026. Retrieval-augmented generation for ai-generated content: A survey. Data Science and Engineering , 1–29.

Zhao, R., Yuan, Q., Li, J., Fan, Y., Li, Y., Gao, F., 2024. Drivellava: Human-level behavior decisions via vision language model. Sensors (Basel, Switzerland) 24.

Zhao, X., Li, M., Lu, W., Weber, C., Lee, J.H., Chu, K., et al., 2024. Enhancing zero-shot chain-of-thought reasoning in large language models through logic. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). 6144–6166.

Zheng, O., Abdel-Aty, M., Wang, D., Wang, C., Ding, S., 2023. Trafficsafetygpt: Tuning a pre-trained large language model to a domain-specific expert in transportation safety. https://doi.org/10.48550/arXiv.2307.15311.

Zhong, B., Li, H., Luo, H., Zhou, J., Fang, W., Xing, X., 2020. Ontology-based semantic modeling of knowledge in construction: Classification and identification of hazards implied in images. Journal of construction engineering and management 146, 04020013.

Zhong, L., Wu, J., Li, Q., Peng, H., Wu, X., 2023. A comprehensive survey on automatic knowledge graph construction. ACM Computing Surveys 56, 1–62.

Zhong, W., Huang, J., Wu, M., Luo, W., Yu, R., 2025. Large language model based system with causal inference and chain-of-thoughts reasoning for traffic scene risk assessment. Knowledge-Based Systems , 113630.

Zhou, H., Schimid, S., Li, Y., Halilaj, L., Yao, X., Cao, W., 2025. Predicting the road ahead: A knowledge graph based foundation model for scene understanding in autonomous driving. In: European Semantic Web Conference. 116–132.

Zhou, J., Zhao, Y., Hu, Y., Li, H., Gu, Z., Xu, N., et al., 2025. Can ai generate more comprehensive test scenarios? review on automated driving systems test scenario generation methods. https://doi.org/10.48550/arXiv.2512.15422.

Zhou, T., Chen, Y., Liu, K., Zhao, J., 2024. Cogmg: Collaborative augmentation between large language model and knowledge graph. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations). 365–373.

Zhou, X., Liu, M., Zagar, B.L., Yurtsever, E., Knoll, A.C., 2023. Vision language models in autonomous driving and intelligent transportation systems. https://doi.org/10.48550/arXiv.2310.14414.

Zhou, Y., Shao, H., Wang, L., Zong, Z., Li, H., Waslander, S.L., 2026. Drivinggen: A comprehensive benchmark for generative video world models in autonomous driving. https://doi.org/10.48550/arXiv.2601.01528.

Zhou, Z., Huang, H., Li, B., Zhao, S., Mu, Y., Wang, J., 2026. Safedrive: Knowledge-and data-driven risk-sensitive decision-making for autonomous vehicles with large language models. Accident Analysis & Prevention 224, 108299.

Zhu, H., Peng, H., Lyu, Z., Hou, L., Li, J., Xiao, J., 2023. Pre-training language model incorporating domain-specific heterogeneous knowledge into a unified representation. Expert Systems with Applications 215, 119369.

Zhu, J., Jia, Z., Gao, T., Deng, J., Li, S., Zhang, L., et al., 2026. Other vehicle trajectories are also needed: A driving world model unifies ego-other vehicle trajectories in video latent space. In: Proceedings of the AAAI Conference on Artificial Intelligence. 13934–13942.

Zhu, Y., Wang, X., Chen, J., Qiao, S., Ou, Y., Yao, Y., et al., 2024. Llms for knowledge graph construction and reasoning: Recent capabilities and future opportunities. World Wide Web 27, 58.

Zhu, Z., Huang, T., Wang, K., Ye, J., Chen, X., Luo, S., 2025. Graph-based approaches and functionalities in retrieval-augmented generation: A comprehensive survey. ACM Computing Surveys .

Zipfl, M., Koch, N., Zöllner, J.M., 2023. A comprehensive review on

ontologies for scenario-based testing in the context of autonomous driving. In: 2023 IEEE Intelligent Vehicles Symposium (IV). 1–7.

Zipfl, M., Zöllner, J.M., 2022. Towards traffic scene description: The semantic scene graph. In: 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC). 3748–3755.

Author biography

Jiaxin Liu received the B.Eng. degree in vehicle engineering from the School of Vehicle and Mobility, Tsinghua University, Beijing, China, in 2021. He is currently pursuing the Ph.D. degree in vehicle engineering with Tsinghua University, supervised by Prof. Jun Li and Prof. Hong Wang. His research interests include risk cognition of autonomous driving vehicles and Large Language Models for vehicle safety.

Liang Peng received the B.Eng. degree in Vehicle Engineering from the School of Vehicle and Mobility of Tsinghua University, Beijing, China, in 2020. He is currently working toward the Ph.D. degree in Vehicle Engineering at Tsinghua University, supervised by Prof. Jun Li and Prof. Hong Wang. His research interests include the evaluation and uncertainty analysis of perceptual algorithms, and knowledge graphs for the safety of autonomous driving vehicles.

Xiangyu Yan received the B.E. degree in Intelligent Vehicle Engineering from Harbin Institute of Technology, Weihai, China, in 2023. He is currently working toward the M.E. degree in Mechanical Engineering at Beijing Institute of Technology, Beijing, China. Since 2023, he has been a Visiting Student at the Tsinghua University. His research interests include risk reasoning and decision-making for autonomous driving.

Lingjun Zhang received the B.E. degree in Beijing Institute of Technology, in 2024. He is currently working toward the M.S. degree in School of Vehicle and Mobility in Tsinghua University, Beijing. His research interests include large language models and reinforcement learning for autonomous driving decision-making and safety.

Chenye Yang is a senior undergraduate student in Mechanics and Vehicle Engineering at Xingjian College, Tsinghua University, Beijing, China. She is currently pursuing the B.Eng. degree and is expected to graduate in 2026.

Yueming Tao received the B.Eng. degree in Vehicle Engineering from Wuhan University of Technology, in 2023. He is currently pursuing the M.Eng. degree with the School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu, China. He is currently a visiting student at Tsinghua University, under the supervision of Prof. Jun Li and Hong Wang.

Ashton Tan Yu Xuan received his B.S. degree in automotive engineering from Tsinghua University in 2023. He is now pursuing an M.S. degree in the same field at Tsinghua University. His research interests include risk reasoning and decision-making for autonomous driving.

Tianli Xu received the B.E. degree in Vehicle Engineering from the School of Vehicle and Mobility, Tsinghua University, Beijing, China, in 2019. He is currently pursuing the Ph.D. degree at the same university. His research interests include vehicle and pedestrian trajectory prediction, as well as cooperative safety decision-making for intelligent vehicles.

Sai Guo received his Ph.D. degree in School of Vehicle and Mobility in Tsinghua University. He is currently working in Xiaomi Cooperation. His research interests include simulation-in-the-loop autonomous vehicle testing and test methods and platform development.

Hong Wang received the Ph.D. degree from the Beijing Institute of Technology, China, in 2015. From 2015 to 2019, she was a Research Associate of mechanical and mechatronics engineering with the University of Waterloo. She is currently a Research Associate Professor with Tsinghua University. Her research interests include the safety of the on-board AI algorithm, the safe decision-making for intelligent vehicles, and the test and evaluation of SOTIF.

Jun Li received the Ph.D. degree in vehicle engineering from Jilin University, Changchun, Jilin, China, in 1989. He is currently an Academician with the Chinese Academy of Engineering, a Professor with the School of Vehicle and Mobility, Tsinghua University, the President of the Society of Automotive Engineers of China, and the Director of the Expert Committee of China Industry Innovation Alliance for the Intelligent and Connected Vehicles. His research interests include internal combustion engine, electric drive systems, electric vehicles, intelligent vehicles, and connected vehicles.