Driving with Regulation: Trustworthy and Interpretable Decision-Making for Autonomous Driving with Retrieval-Augmented Reasoning

Tianhui Cai*, Yifan Liu*, Zewei Zhou, Haoxuan Ma, Seth Z. Zhao, Zhiwen Wu, Xu Han, Zhiyu Huang†, Jiaqi Ma

University of California, Los Angeles
{tianhui, bmmliu, zhiyuh, jiaqima}@ucla.edu

Abstract

Understanding and adhering to traffic regulations is essential for autonomous vehicles to ensure safety and trustworthiness. However, traffic regulations are complex, context-dependent, and differ between regions, posing a major challenge to conventional rule-based decision-making approaches. We present an interpretable, regulation-aware decision-making framework, **DriveReg**, which enables autonomous vehicles to understand and adhere to region-specific traffic laws and safety guidelines. The framework integrates a Retrieval Augmented Generation (RAG)-based Traffic Regulation Retrieval Agent, which retrieves relevant rules from regulatory documents based on the current situation, and a Large Language Model (LLM)-powered Reasoning Agent that evaluates actions for legal compliance and safety. Our design emphasizes interpretability to enhance transparency and trustworthiness. To support systematic evaluation, we introduce **DriveReg Scenarios Dataset**, a comprehensive dataset of driving scenarios across Boston, Singapore, and Los Angeles, with both hypothesized text-based cases and real-world driving data, specifically constructed and annotated to evaluate models' capacity for regulation understanding and reasoning. We validate our framework on the DriveReg Scenarios Dataset and real-world deployment, demonstrating strong performance and robustness across diverse environments.

Code — https://github.com/Vickycth/driving-with-regulation
Extended version — https://arxiv.org/abs/2410.04759

1 Introduction

Autonomous driving technologies have advanced significantly in recent years, showing potential to improve safety and efficiency (Zhao et al. 2023; Han et al. 2024; Huang et al. 2025; Huang, Liu, and Lv 2023; Li et al. 2023c, 2022, 2023a). However, the societal deployment of autonomous vehicles (AVs) depends not only on technical progress in perception or control, but also on their ability to operate legally, transparently, and in accordance with human expectations (Kubica 2022; Administration et al. 2016). Understanding and complying with traffic regulations is vital for safety and public trust (Mehdipour et al. 2023). Therefore, beyond technical robustness, AVs must ensure legally compliant and interpretable decision-making to enable trustworthy, socially-aligned deployment.

Despite its importance, understanding traffic rules remains a major challenge for AVs. Regulations are complex, context-dependent, and vary across regions, and small contextual changes can lead to different rule interpretations. While prior work encodes traffic rules via hand-crafted logic (Xiao et al. 2021; Manas, Zwicklbauer, and Paschke 2022; Sun et al. 2022), such approaches lack scalability and adaptability. Moreover, traffic rules span strict legal codes and flexible safety guidelines, demanding diverse reasoning strategies. This semantic richness exceeds the capabilities of conventional models, but recent advances in Large Language Models (LLMs) offer a promising solution through their ability to interpret natural language and perform context-aware reasoning.

In addition to system development challenges, there is no comprehensive benchmark for evaluating traffic rules understanding in AV decision-making. Datasets like Argoverse 2 (Wilson et al. 2023) and nuPlan (Karnchanachari et al. 2024) focus on perception or planning, without annotations linking actions to specific rules. Existing traffic rule-aware datasets either lack real-world scenarios (Li et al. 2024b; Deng et al. 2025), or cover only a limited subset of regulations (Sun et al. 2022; Deng et al. 2025; Li et al. 2024b).

To address these limitations, we introduce **DriveReg**, an interpretable, traffic regulation-aware decision-making framework for AVs, along with the **DriveReg Scenarios Dataset**, a benchmark of 500 scenarios spanning Boston, Singapore, and Los Angeles. DriveReg integrates a Traffic Regulation Retrieval (TRR) Agent built upon Retrieval-Augmented Generation (RAG), with an LLM-based Reasoning Agent. Given a driving scenario, the system retrieves relevant traffic rules from an extensive collection of regulatory documents applicable to the city and evaluates each candidate action on two levels: (1) ***Compliance***, whether it satisfies mandatory legal constraints; and (2) ***Safety***, whether it also adheres to non-mandatory guidelines. To ensure transparency, DriveReg outputs reasoning steps with references to the specific rules used in the decision. The **DriveReg Scenarios Dataset** includes 360 hypothesized cases (120 per city), designed to cover all categories defined in each region's regulatory documents, and 140 real-world scenarios

*These authors contributed equally.
†Corresponding author.

sourced from the nuScenes dataset and our in-house dataset. Each scenario is annotated with the relevant traffic rules and labeled at the action level for compliance and safety, enabling fine-grained evaluation across different cities. The main contributions of our paper are summarized as follows:

- We propose **DriveReg**, an interpretable, LLM-driven decision-making framework for autonomous driving that integrates a Traffic Regulation Retrieval Agent and a Reasoning Agent to assess action-level compliance and safety, enabling regulation-adherent and region-aware decision-making.
- We introduce the **DriveReg Scenarios Dataset**, and a benchmark of 500 driving scenarios (hypothesized and real-world) from Boston, Singapore, and Los Angeles, annotated with retrieved traffic rules and compliance/safety labels for action-level evaluation.
- We demonstrate the effectiveness, robustness, and regional generalization of DriveReg through extensive experiments on the DriveReg Scenarios Dataset and validate the DriveReg framework in real-world deployment to assess its practical performance.

2 Related Work

2.1 Traffic Regulation in Autonomous Driving

To integrate traffic regulations into autonomous driving systems, early approaches included rule-based systems (Li et al. 2023b) and finite state machines (Bae et al. 2020), which encoded traffic laws through explicit if-then rules or state transitions. To handle complex scenarios, more sophisticated methods emerged, such as using behavior trees (Tadewos, Shamgah, and Karimoddini 2019), and formal methods (Maierhofer et al. 2020). However, these methods often struggled with the ambiguity and regional variations of real-world traffic rules.

Recently, Large Language Models (LLMs) have demonstrated remarkable capabilities in understanding natural language and interpreting complex scenarios (Wen et al. 2024; Zhang et al. 2024; Sima et al. 2023; Malla et al. 2023; Wei et al. 2024). LLMs can process traffic rules in a more flexible and context-aware manner. For example, LLaDA (Li et al. 2024a) utilizes LLMs to interpret traffic rules from local handbooks, while Agent-Driver (Mao et al. 2024) incorporates traffic rules into an LLM-based cognitive framework. However, ensuring LLMs accurately apply relevant traffic rules without hallucinations remains a key challenge.

2.2 Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) (Lewis et al. 2021) addresses LLM hallucinations by combining neural retrieval with sequence-to-sequence generation based on relevant documents. Recent studies (Borgeaud et al. 2022; Nakano et al. 2022) have demonstrated significant improvements in LLM accuracy and factual correctness across domains. For autonomous driving systems, RAG's dynamic retrieval capability (Lewis et al. 2021) enables real-time access to region-specific traffic rules, while its enhanced factual grounding (Borgeaud et al. 2022) reduces the risk of fabricating or misapplying regulations. Additionally, RAG's ability to handle complex contextual (Nakano et al. 2022) is well-suited for interpreting nuanced traffic regulations with multiple conditions, and its inherent transparency improves decision-making interpretability, crucial for regulatory compliance and public trust.

2.3 Decision-Making of Autonomous Driving

Decision-making methods for autonomous driving have evolved from rule-based (Wang et al. 2021) to learning-based methods (Kiran et al. 2021), which offer greater adaptability in dynamic environments. Typical learning-based approaches include imitation learning (Bansal, Krizhevsky, and Ogale 2018; Tang et al. 2023) and reinforcement learning (Yuan et al. 2024). More recently, GPT-Driver (Mao et al. 2023) has reformulated motion planning as a language modeling problem. However, the integration of diverse semantic traffic rules into decision-making using a unified model remains underexplored.

3 Method

3.1 Overview

Our proposed method, as shown in Fig. 1, comprises two main components: a **Traffic Rules Retrieval (TRR) Agent** that retrieves relevant traffic rules from regulation documents using a retrieval query, and a **Reasoning Agent** that assesses the traffic rule adherence of each action in the proposal set based on environment information, ego vehicle's state, and retrieved traffic rules.

To support traffic rule retrieval, we perform an environment analysis using a Vision-Language Model (VLM), denoted as $\mathcal{V}$. Given multi-view images I from the ego vehicle and navigation intent $g \in \{\text{Left, Right, Forward}\}$, $\mathcal{V}$ follows a structured Chain-of-Thought (CoT) reasoning process: it first provides a high-level overview of the driving scene, including road layout and general traffic structure. It then performs a more focused analysis of critical elements, such as traffic signs, road users, and lane markings, particularly those relevant to the intended maneuver g. Finally, the VLM generates a concise natural language query that summarizes the current scenario for retrieving relevant traffic rules. We denote this reasoning process as $\mathcal{V}(I, g) \rightarrow (c, q)$, where c is the structured description of the scenario and q is the retrieval query provided to the TRR Agent. An example of this environment analysis output is shown in Fig. 3.

For decision-making, we extract an action proposal set containing possible actions from the Action Space $\mathcal{A}$ based on the Global Planning output g. For simplicity and to maintain the focus of this work on traffic rule adherence, the Action Space $\mathcal{A}$ only consists of a predefined set of actions: turning right, turning left, going forward (with current speed, acceleration, or deceleration), changing lane to the left, and changing lane to the right. The candidate set $\mathcal{A}_{\text{cand}}$ is obtained by selecting actions from $\mathcal{A}$ that align with the current global intent g. For instance, if $g = \text{Left}$, then $\mathcal{A}_{\text{cand}} = \{\text{turning left with current speed, turning left with an acceleration, turning left with a deceleration}\}$.

Given the retrieval query q, the Traffic Rules Retrieval Agent (TRR) fetches a set of relevant traffic rules $\mathcal{R}_q$ from

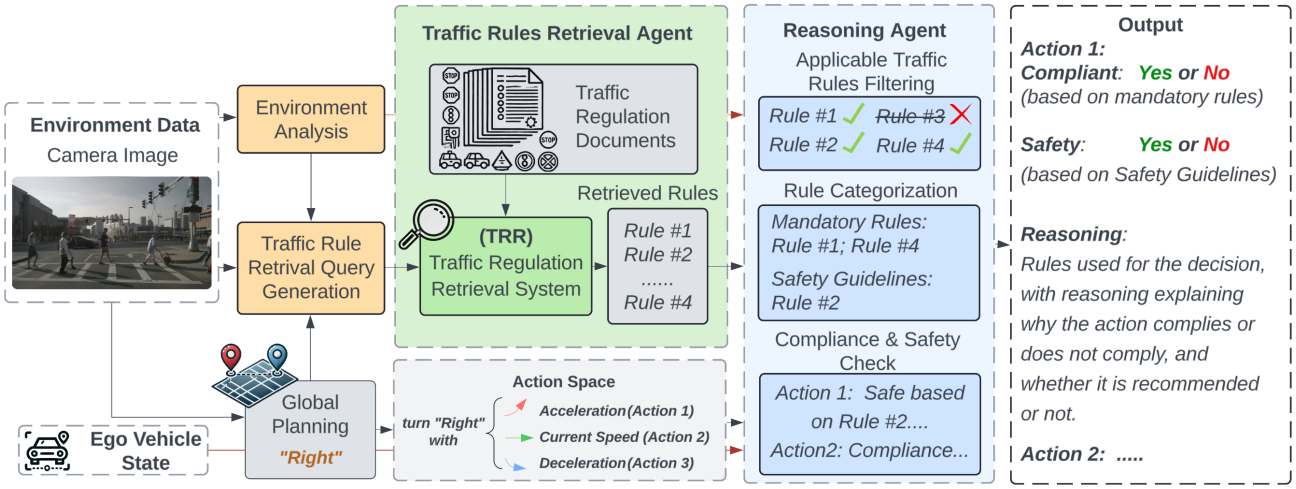

Figure 1: Overview of Driving with Regulation (DriveReg) Framework. The framework consists of two main components: the Traffic Rules Retrieval Agent and the Reasoning Agent. The Traffic Rules Retrieval Agent retrieves relevant rules from traffic regulation documents based on the generated traffic rule retrieval query. The Reasoning Agent then identifies the applicable rules from the retrieved set and performs compliance and safety checks based on those applicable rules.

a collection of traffic regulatory documents $\mathcal{D}$. The Traffic Reasoning Agent (TRA) evaluates each candidate action $a \in \mathcal{A}_{\text{cand}}$ for both compliance and safety, leveraging the structured scene description c and the set of retrieved traffic rules $\mathcal{R}_q$, obtained by querying the Traffic Rules Retrieval Agent with q. Formally, the agents perform:

$$\text{TRR}(q; \mathcal{D}) \to \mathcal{R}_q,\ \text{TRA}(a, c, \mathcal{R}_q) \to (l_c, l_s), \quad (1)$$

where $l_c, l_s \in \{0, 1\}$ are binary outputs for compliance and safety, respectively, with 1 indicating compliant or safe, and 0 otherwise.

The final decision of the system is the set of actions that are both compliant and safe:

$$\mathcal{A}^* = \{a \in \mathcal{A}_{\text{cand}} \mid l_c(a) = 1 \ \wedge \ l_s(a) = 1\}. \quad (2)$$

3.2 Traffic Rules Retrieval Agent

To enhance the model's understanding of local traffic rules and norms, and to fully consider all the related rules from available sources, we developed the Traffic Regulation Retrieval Agent, which employs a two-level retrieval strategy as illustrated in Fig. 2.

Since different regions have varying sources of traffic rules, we use the United States as an example to illustrate the comprehensive collection of regulatory documents $\mathcal{D}$ used by the TRR Agent, including state-level traffic laws, official state driving manuals (which also include safety guidelines), city-level regulations, relevant court cases that establish legal precedent, and widely accepted traffic norms.

During the retrieval process, we first encode both the traffic rule retrieval query q and the regulatory documents $\mathcal{D}$ into dense vector representations using a sentence embedding model $\mathcal{E}(\cdot)$. Each paragraph $p_i \in \mathcal{D}$ is embedded as $\mathbf{v}_{p_i} = \mathcal{E}(p_i)$, and the query is embedded as $\mathbf{v}_q = \mathcal{E}(q)$. We utilize FAISS (Johnson, Douze, and Jégou 2017) to efficiently retrieve a top-k candidate set of paragraphs based on vector similarity:

$$\mathcal{P}_q^{\text{cand}} = \text{TopK}_{\text{FAISS}}\left(\mathbf{v}_q, \{\mathbf{v}_{p_i}\}_{i=1}^{|\mathcal{D}|}\right). \quad (3)$$

We retain only paragraphs with similarity scores exceeding a threshold τ_p:

$$\mathcal{P}_q = \{p_i \in \mathcal{P}_q^{\text{cand}} \mid \text{sim}(\mathbf{v}_q, \mathbf{v}_{p_i}) \geq \tau_p\}. \quad (4)$$

To address the sparsity and granularity issues that arise from long-form regulatory text, we further refine the retrieval by segmenting each paragraph in $\mathcal{P}_q$ into individual sentences and performing sentence-level retrieval. Specifically, each paragraph $p \in \mathcal{P}q$ is segmented into a set of sentences s_j, each encoded as a sentence-level embedding $\mathbf{v}_{s_j} = \mathcal{E}'(s_j)$, where $\mathcal{E}'$ is a lighter-weight sentence encoder. A second FAISS-based similarity search is then conducted to identify the most relevant sentences to $\mathbf{v}'_q = \mathcal{E}'(q)$:

$$\mathcal{R}_q = \text{TopK}_{\text{FAISS}}(\mathbf{v}'_q, \{\mathbf{v}_{s_j}\}_{j=1}^{|\mathcal{S}|}), \quad (5)$$

where $\mathcal{S} = \bigcup_{p \in \mathcal{P}_q} \text{Sentences}(p)$.

Our two-level retrieval strategy combines paragraph-level retrieval for broad coverage with sentence-level refinement for precision, yielding concise, context-aligned rule snippets while improving efficiency and semantic relevance. By anchoring the Reasoning Agent in retrieved regulations, the TRR Agent helps mitigate hallucinations of LLMs and enhance decision reliability and transparency.

3.3 Reasoning Agent

The Reasoning Agent is responsible for determining whether each action in the proposal set complies with traffic rules, leveraging an LLM (e.g., GPT-4o) with CoT prompting and few-shot learning. The Reasoning Agent receives three key inputs: (1) the current environment information

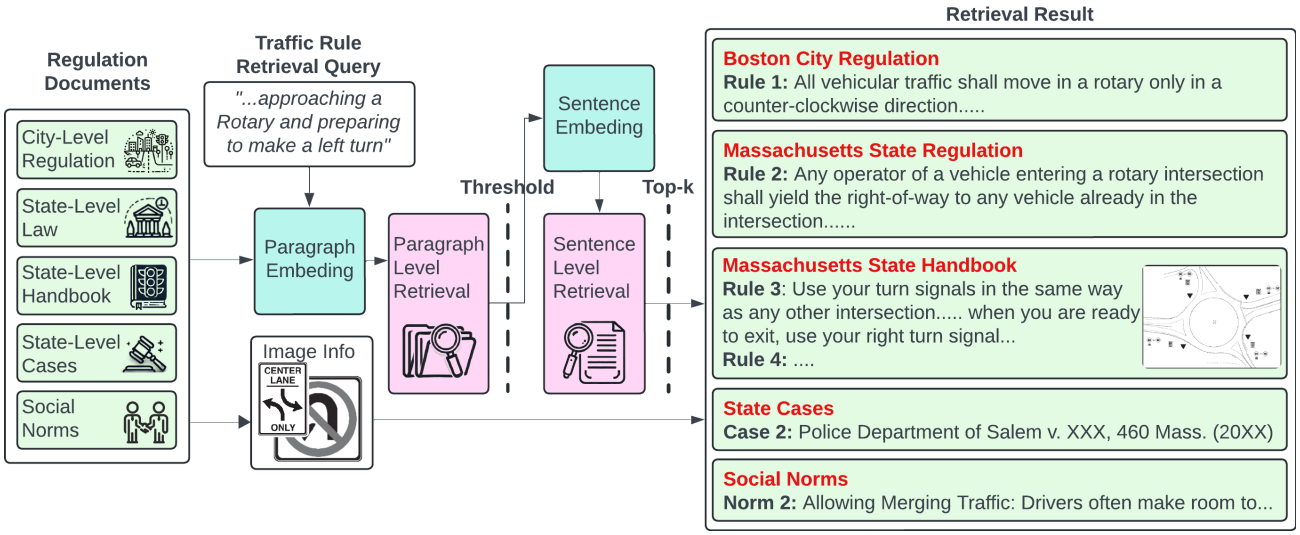

Figure 2: Illustration of the proposed Traffic Regulation Retrieval (TRR) Agent.The retrieval results are obtained through the similarity score between scene description and well-curated regulation documents with a pre-defined relevance metric.

from the environment analysis c, (2) the ego vehicle's action proposal set $\mathcal{A}_{\text{cand}}$, and (3) a set of retrieved traffic rules from TRR Agent $\mathcal{R}_q$.

In the reasoning pipeline, the agent first filters $\mathcal{R}_q$ to identify rules relevant to the current scene and the ego vehicle's intended maneuver. The filtered rules are categorized into two types: **mandatory rules**, which must be satisfied to ensure legal compliance, and **safety guidelines**, which represent best practices not legally binding but recommended for safe operation.

For each action $a \in \mathcal{A}_{\text{cand}}$, the Reasoning Agent first outputs a compliance indicator $l_c = 1$ if the action does not violate any relevant mandatory rule in $\mathcal{R}_q$, and $l_c = 0$ otherwise. It then produces a safety indicator $l_s = 1$ if the action satisfies both the mandatory rules and any retrieved safety guidelines, and $l_s = 0$ if any of these are violated. Alongside each binary indicator, the Reasoning Agent provides a concise explanation referencing the specific rules involved, clarifying why the action is considered compliant or non-compliant, safe or unsafe. This interpretability is critical for transparency in decision-making. The framework then selects the actions that are marked as both compliant and safe as the final output for decision-making. An example output of the Reasoning Agent is shown in Fig. 3.

3.4 DriveReg Scenarios Dataset

We introduce the **DriveReg Scenarios Dataset**, the first dataset designed to evaluate models on understanding and reasoning over traffic regulations in diverse driving scenarios. It contains 360 hypothesized text-based scenarios and 140 real-world cases with front-camera images drawn from the nuScenes dataset and data collected during our deployment testing. Hypothesized scenarios offer greater diversity, while real-world data evaluates the framework's practical performance in real driving conditions. For each city,

Dataset	# Rules	# Scenarios	Real-World	Multi-Region
VioHawk (Li et al. 2024b)	27	42	×	×
TARGET (Deng et al. 2025)	54	284	×	×
LawBreaker (Sun et al. 2022)	24	173	×	×
DriveReg	**562**	**500**	✓	✓

Table 1: Comparison of existing traffic rule-aware driving datasets. DriveReg provides the most comprehensive coverage with real-world data, multi-region scenarios, and fine-grained compliance and safety labels.

we preprocess comprehensive traffic regulation sources, and each scenario is annotated with the relevant traffic rules and the two-level decision labels indicating whether an action is (1) legally compliant and (2) safe, based on the applicable regulations. An example is shown in Fig. 4, with additional cases provided in the extended version, and comparison with existing traffic rule-aware datasets is shown in Table 1.

Traffic Regulation Sources. We use extensive traffic regulation sources for the three cities included in the DriveReg Scenarios Dataset, covering city ordinances, state or national laws, official driver manuals, and selected legal cases or norms. For the Boston region, this includes the Boston City Traffic Rules and Regulations, the Massachusetts General Laws, the Massachusetts Driver's Manual, and twelve selected state court cases on traffic violations. We also include ten representative U.S. driving norms that reflect lawful behavior without explicit violations. More details and source descriptions for Singapore and Los Angeles are provided in the extended version.

Scenario Collection and Annotation Process. *1) Hypothesized Scenarios.* We construct 120 hypothesized scenarios per city, including 100 **normal** and 20 **hard** cases. **Normal scenarios** are designed to cover all regulatory categories and

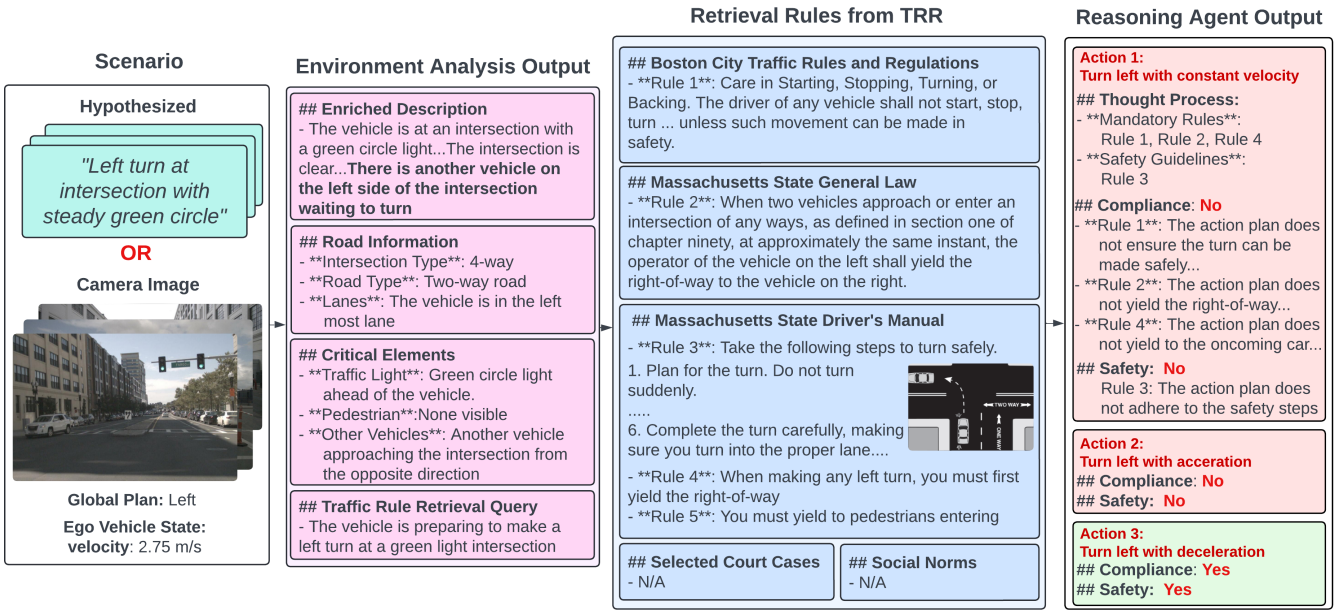

Figure 3: Pipeline of processing the selected scenario. The correct action is labeled in **green** background.

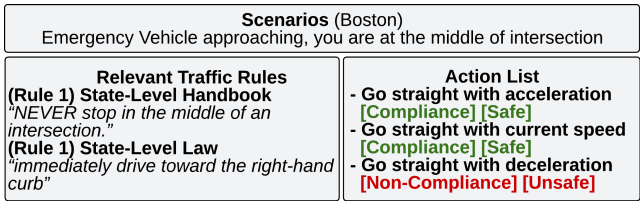

Figure 4: Example from the DriveReg Scenarios Dataset showing scenario descriptions, relevant traffic rules, and action-level compliance/safety annotations.

Category	Count	Category	Count
Right-of-Way Rules	20	Special Driving Situations	11
Intersections	15	Rules for Passing	8
Pedestrians	15	Traffic Signs & Signals	8
Pavement Markings	10	Roadway Construction	5
Service Vehicles	8	Hard Cases	20

Table 2: Distribution of Hypothesized Scenarios in Boston.

document sections, reflecting common traffic situations such as right-of-way decisions and pedestrian interactions. Each involves 1-2 straightforward rules to evaluate basic regulation understanding. **Hard scenarios** include cases with region-specific regulations and complex conditions involving three or more applicable rules, such as school-bus handling at intersections with malfunctioning signals.

The distribution of hypothesized scenarios for Boston across regulatory categories is shown in Table 2, demonstrating comprehensive coverage across key regulatory domains. Similar distributions were constructed for Singapore and Los Angeles and are provided in the extended version.

2) Real-World Scenarios. To evaluate models' performance in real-world driving contexts, we selected and annotated 140 scenarios based on two sources: (1) nuScenes dataset (Caesar et al. 2020), which captures urban driving scenes in Boston and Singapore, and (2) an in-house dataset collected in Los Angeles. Although nuScenes is not originally designed for traffic regulation analysis and lacks rule-level annotations, we manually reviewed front-camera images to identify samples where vehicle behavior is meaningfully influenced by traffic rules. Similarly, we sampled from our deployment logs in Los Angeles under comparable criteria.

4 Experiments

4.1 Experimental Setup

Metrics. We evaluate each model's decision-making by assessing its ability to determine the compliance and safety of candidate actions in each scenario. For each action $a \in \mathcal{A}_{\text{cand}}$, the model outputs a predicted compliance indicator $l_c \in \{0, 1\}$ and safety indicator $l_s \in \{0, 1\}$. *Compliance accuracy (Compl.)* and *safety accuracy (Safety)* are computed as the percentage of actions for which the predicted compliance and safety indicators match the ground truth labels.

Baseline Models. We compare DriveReg with several state-of-the-art LLMs and VLMs (Liu et al. 2024; Guo et al. 2025; Dubey et al. 2024; Bai et al. 2023; Hurst et al. 2024). These baselines do not incorporate external traffic regulations but may reflect some rule understanding from pretraining. Their performance without explicit regulation input is denoted as *Compl.*$_{(-R)}$ and *Safety*$_{(-R)}$. To assess the impact of our TRR module, we also evaluate these models with retrieved traffic

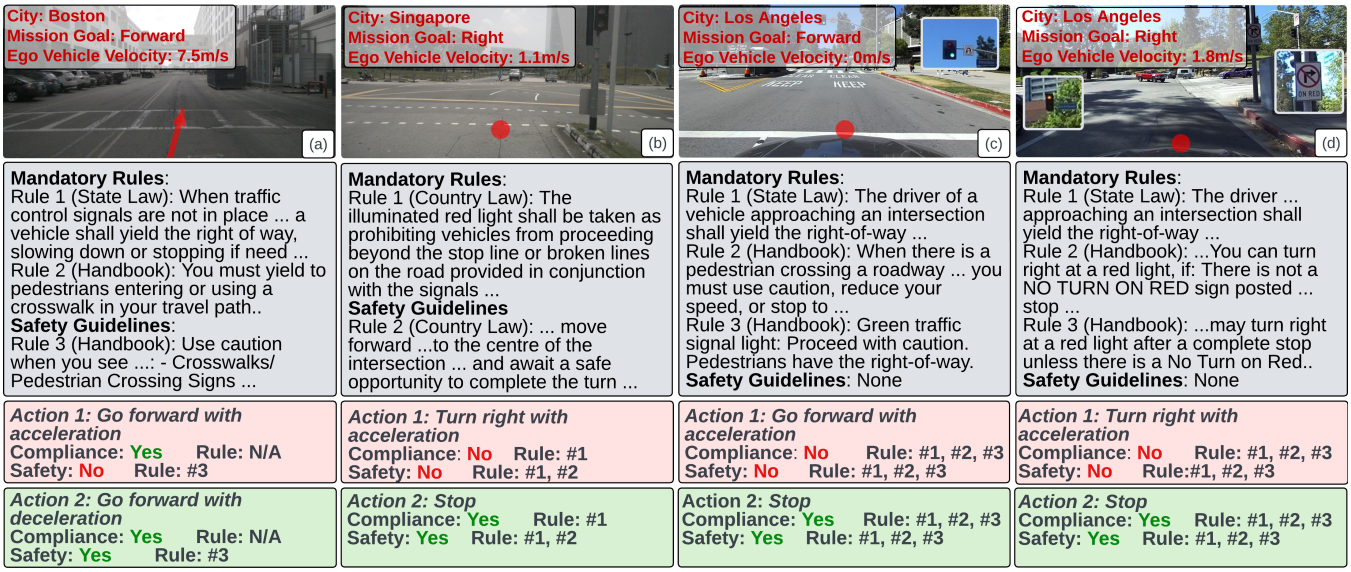

Figure 5: Inference results from the DriveReg (Real-World) Scenarios Dataset. Actions in green boxes represent the final decision-making outputs, which are both compliant and safe. Results demonstrate that our framework successfully retrieves relevant traffic rules and interprets and assesses actions based on those rules. The correct action is labeled in a green background. (Due to space constraints, some actions and reasoning details are omitted.)

rules of TRR, reported as *Compl.*$_{(+R)}$ and *Safety*$_{(+R)}$.

Implementation Details. We adopt "text-embedding-ada-002" as $\mathcal{E}$ and "paraphrase-MiniLM-L6-v2" as $\mathcal{E}'$. τ_p is set to 0.28 and $K = 5$ for FAISS similarity search. Experiments are conducted over 5 random seeds. Additional settings are included in the extended version.

4.2 Main Results

Hypothesized Scenarios Results. We evaluated our framework's performance on DriveReg hypothesized scenarios, with results presented in Table 3. Regulations retrieved via TRR improve performance across most models, particularly in hard cases, underscoring the effectiveness of our TRR agent. We observe that the improvement is notably evident in scenarios involving region-specific traffic laws, such as Boston's requirement that drivers must remain in the right lane unless passing or preparing for a left turn, which are often overlooked by pretrained models. These results reveal the limitations of relying solely on general world knowledge encoded in foundation models, highlighting the need for explicit rule grounding in decision-making systems.

The DriveReg framework, incorporating the TRR Agent and rule-aware Reasoning Agent, achieves the highest overall performance. With access to regulatory input, DriveReg reaches 98% compliance and 98% safety accuracy in normal scenarios (F1: 97.6%, p-value: 4.2e-03), and 99% compliance and 94% safety accuracy in hard scenarios (F1: 94.6%, p-value: 3.5e-03). In particular, for hard cases, DriveReg improves compliance accuracy from 86%, as obtained by the underlying GPT-4o model, to 99% when enhanced with TRR and TRA. These results demonstrate that TRA, utilizing CoT reasoning, effectively interprets and applies retrieved traffic regulations to guide decision-making in scenarios requiring precise legal understanding.

Real-World Scenarios Results. Table 4 compares the performance of our DriveReg framework to the baseline GPT-4o on real-world scenarios across the three cities. Our model consistently outperforms the baseline in both compliance and safety metrics. Notably, GPT-4o achieves higher safety than compliance scores, suggesting VLMs capture general safety patterns (e.g., avoiding pedestrians or maintaining distance) better than precise legal constraints. In contrast, compliance requires accurate interpretation of traffic laws, which our framework addresses through retrieval-augmented reasoning. We also observe that some failure cases come from perception issues, such as misidentifying solid lane markings as broken, highlighting the importance of fine-grained visual understanding in urban environments.

Examples of output from our framework are shown in Fig. 5. In (a), we present a scenario in Boston where the vehicle approaches a crosswalk. Here, the framework correctly identifies that accelerating forward is "compliant but not safe", which aligns with the common-sense guideline to use caution when approaching crosswalks, even when they are clear. In (b), we demonstrate the framework's ability to handle region-specific regulations: The ego vehicle attempts to turn right at a red light, an action that is illegal in Singapore but often permitted in the U.S. Our model correctly outputs "non-compliant", aligning with local traffic regulations. These examples highlight the framework's capacity to reason about both safety and region-specific laws. Additional results are included in the extended version.

Retrieval Method Comparison. We compare our TRR approach against traditional information retrieval methods on

Method	Hypothesized Scenarios - Normal				Hypothesized Scenarios - Hard			
	Compl.(-R)	Compl.(+R)	Safety(-R)	Safety(+R)	Compl.(-R)	Compl.(-R)	Safety(-R)	Safety(+R)
DeepSeek-V3 (Liu et al. 2024)	0.87 ± 0.02	0.92 ± 0.03	0.90 ± 0.02	0.92 ± 0.03	0.82 ± 0.02	0.92 ± 0.04	0.84 ± 0.03	0.84 ± 0.03
DeepSeek-R1 (Guo et al. 2025)	0.92 ± 0.02	0.93 ± 0.02	0.93 ± 0.04	0.95 ± 0.01	0.85 ± 0.01	0.96 ± 0.03	0.86 ± 0.05	0.90 ± 0.03
Llama3.3-70b (Dubey et al. 2024)	0.90 ± 0.04	0.91 ± 0.02	0.93 ± 0.02	0.94 ± 0.04	0.81 ± 0.03	0.93 ± 0.02	0.86 ± 0.03	0.90 ± 0.03
Qwen2.5-72b (Bai et al. 2023)	0.93 ± 0.02	0.94 ± 0.01	0.92 ± 0.04	0.95 ± 0.02	0.88 ± 0.02	0.95 ± 0.04	0.85 ± 0.04	0.86 ± 0.03
GPT-4o (Hurst et al. 2024)	0.91 ± 0.02	0.93 ± 0.03	0.94 ± 0.01	0.97 ± 0.02	0.86 ± 0.03	0.95 ± 0.01	0.88 ± 0.02	0.90 ± 0.02
DriveReg (Ours)	–	$\mathbf{0.98 \pm 0.01}$	–	$\mathbf{0.98 \pm 0.01}$	–	$\mathbf{0.99 \pm 0.02}$	–	$\mathbf{0.94 \pm 0.02}$

Table 3: Average compliance and safety accuracy across three cities (Boston, Los Angeles, and Singapore) under DriveReg-Hypothesized driving scenarios. (-R) indicates baseline models relying solely on pretrained knowledge, while (+R) indicates that models provided with relevant traffic regulations retrieved by the TRR agent.

Method	Boston		Los Angeles		Singapore	
	Comp.	Safety	Comp.	Safety	Comp.	Safety
GPT-4o	0.86	0.91	0.91	**0.93**	0.84	0.89
DriveReg (Ours)	**0.94**	**0.92**	**0.95**	**0.93**	**0.91**	**0.93**

Table 4: Comparison with GPT-4o on compliance and safety accuracy for real-world scenarios.

Method	Best Match 25	QLM	LLaDa TRE†	**TRR (Ours)**
Accuracy	0.60	0.50	0.55	**1.00**

Table 5: Traffic regulation retrieval accuracy comparison on 20 scenarios using top-5 retrieval. †: LLaDA TRE is reproduced following the original paper (Li et al. 2024a).

20 Boston-Hard scenarios using top-5 retrieval: Best Match 25 (BM25) (Robertson, Zaragoza et al. 2009), Query Likelihood Model (QLM) (Ponte and Croft 2017), and the Traffic Rule Extractor (TRE) in LLaDa (Li et al. 2024a). Accuracy is measured as the successful extraction of all ground truth regulations. As shown in Table 5, our semantic similarity-based TRR Agent achieves superior retrieval accuracy, significantly outperforming keyword-based search methods.

Comparison with Agent-Driver. We further evaluate DriveReg against Agent-Driver (Mao et al. 2024). Since Agent-Driver outputs trajectories without explicit compliance or safety indicators, we assess its reasoning performance by comparing its high-level plan outputs against a randomly selected action from DriveReg's compliant-safe action set $\mathcal{A}^*$. As shown in Table 6, DriveReg outperforms Agent-Driver in both compliance and safety, indicating a stronger adherence to traffic rules.

4.3 Real-World Testing

To validate our framework in a real-world setting, we conducted deployment testing on an autonomous vehicle, as shown in Fig. 6 (a), in Los Angeles. For safety reasons, a human driver operated the vehicle while our system ran in parallel, fully functional but without actual control over the vehicle. This setup allowed us to evaluate the system's decision-making capabilities in real-time without compromising safety. We tested the system in urban environments

Method	High-Level Comp.	High-Level Safety
Agent-Driver (Mao et al. 2024)	0.91	0.86
DriveReg (Ours)	**0.98**	**0.97**

Table 6: High-level decision-making evaluation on the Boston real-world data split.

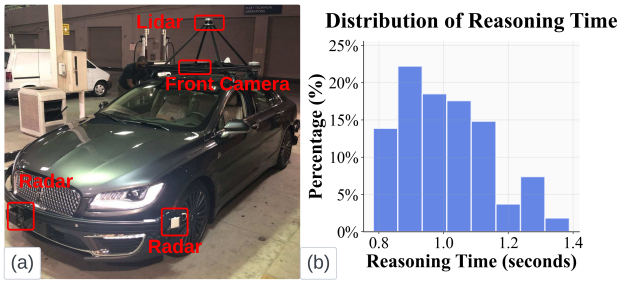

Figure 6: (a) The vehicle used for deployment. (b) Inference time distribution of the Reasoning Agent when additional reasoning steps are omitted for real-time efficiency.

involving dynamic traffic signals, intersections, and interactions with vulnerable road users. The framework reliably retrieved relevant rules and produced safe, compliant decisions. Fig. 5 shows two examples: (c) a pedestrian crossing with a green light, and (d) a "No Turn on Red" intersection, both handled correctly by the system.

In addition to decision-making accuracy, we assess the inference time of DriveReg to evaluate real-time feasibility as shown in Fig. 6 (b). Using GPT-4o, detailed reasoning outputs average 2 seconds per decision, while shorter label-only outputs reduce latency to around 1 second, which is suitable for high-level decision-making in most driving scenarios.

5 Conclusion

We introduce an LLM-driven, traffic regulation-aware decision-making framework that enhances the interpretability and trustworthiness of AV systems. We also introduce a traffic rules scenario dataset to evaluate models on their ability to reason over traffic regulations. Experiments demonstrate the strong performance and adaptability of our approach. Future work will focus on refining the framework for broader real-world applicability.

Acknowledgments

This work was supported by the U.S. Department of Transportation Federal Highway Administration (FHWA) Center of Excellence on New Mobility and Automated Vehicles.

References

Administration, N. H. T. S.; et al. 2016. *Federal automated vehicles policy: Accelerating the next revolution in roadway safety*. US Department of Transportation.

Bae, S.; Joo, S.; Pyo, J.; Yoon, J.-S.; Lee, K.; and Kuc, T.-Y. 2020. Finite State Machine based Vehicle System for Autonomous Driving in Urban Environments. *2020 20th International Conference on Control, Automation and Systems (ICCAS)*, 1181–1186.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Bansal, M.; Krizhevsky, A.; and Ogale, A. 2018. Chauffeurnet: Learning to drive by imitating the best and synthesizing the worst. *arXiv preprint arXiv:1812.03079*.

Borgeaud, S.; Mensch, A.; Hoffmann, J.; Cai, T.; Rutherford, E.; Millican, K.; Van Den Driessche, G. B.; Lespiau, J.-B.; Damoc, B.; Clark, A.; De Las Casas, D.; Guy, A.; Menick, J.; Ring, R.; Hennigan, T.; Huang, S.; Maggiore, L.; Jones, C.; Cassirer, A.; Brock, A.; Paganini, M.; Irving, G.; Vinyals, O.; Osindero, S.; Simonyan, K.; Rae, J.; Elsen, E.; and Sifre, L. 2022. Improving Language Models by Retrieving from Trillions of Tokens. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvari, C.; Niu, G.; and Sabato, S., eds., *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 2206–2240. PMLR.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11621–11631.

Deng, Y.; Tu, Z.; Yao, J.; Zhang, M.; et al. 2025. TARGET: LLM-Guided Generation of Traffic-Rule Test Scenarios for Autonomous Vehicles. *IEEE Transactions on Software Engineering*. Early Access.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv e-prints*, arXiv–2407.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Han, X.; Meng, Z.; Xia, X.; Liao, X.; He, Y.; Zheng, Z.; Wang, Y.; Xiang, H.; Zhou, Z.; Gao, L.; et al. 2024. Foundation intelligence for smart infrastructure services in transportation 5.0. *IEEE Transactions on Intelligent Vehicles*.

Huang, Z.; Liu, H.; and Lv, C. 2023. Gameformer: Game-theoretic modeling and learning of transformer-based interactive prediction and planning for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3903–3913.

Huang, Z.; Weng, X.; Igl, M.; Chen, Y.; Cao, Y.; Ivanovic, B.; Pavone, M.; and Lv, C. 2025. Gen-drive: Enhancing diffusion generative driving policies with reward modeling and reinforcement learning fine-tuning. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 3445–3451. IEEE.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Johnson, J.; Douze, M.; and Jégou, H. 2017. Billion-scale similarity search with GPUs. *arXiv preprint arXiv:1702.08734*.

Karnchanachari, N.; Geromichalos, D.; Tan, K. S.; Li, N.; Eriksen, C.; Yaghoubi, S.; Mehdipour, N.; Bernasconi, G.; Fong, W. K.; Guo, Y.; et al. 2024. Towards learning-based planning: The nuPlan benchmark for real-world autonomous driving. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 629–636. IEEE.

Kiran, B. R.; Sobh, I.; Talpaert, V.; Mannion, P.; Al Sallab, A. A.; Yogamani, S.; and Pérez, P. 2021. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(6): 4909–4926.

Kubica, M. L. 2022. Autonomous Vehicles and Liability Law. *The American Journal of Comparative Law*, 70: i39–i69.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; tau Yih, W.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv:2005.11401.

Li, B.; Wang, Y.; Mao, J.; Ivanovic, B.; Veer, S.; Leung, K.; and Pavone, M. 2024a. Driving Everywhere with Large Language Model Policy Adaptation. arXiv:2402.05932.

Li, Q.; Peng, Z.; Feng, L.; Liu, Z.; Duan, C.; Mo, W.; and Zhou, B. 2023a. ScenarioNet: Open-Source Platform for Large-Scale Traffic Scenario Simulation and Modeling. *Advances in Neural Information Processing Systems*.

Li, Q.; Peng, Z.; Feng, L.; Zhang, Q.; Xue, Z.; and Zhou, B. 2022. Metadrive: Composing diverse driving scenarios for generalizable reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Li, X.; Bai, Y.; Cai, P.; Wen, L.; Fu, D.; Zhang, B.; Yang, X.; Cai, X.; Ma, T.; Guo, J.; Gao, X.; Dou, M.; Li, Y.; Shi, B.; Liu, Y.; He, L.; and Qiao, Y. 2023b. Towards Knowledge-driven Autonomous Driving. arXiv:2312.04316.

Li, Y.; Zhao, S. Z.; Xu, C.; Tang, C.; Li, C.; Ding, M.; Tomizuka, M.; and Zhan, W. 2023c. Pre-training on Synthetic Driving Data for Trajectory Prediction. arXiv:2309.10121.

Li, Z.; Dai, J.; Huang, Z.; You, N.; Zhang, Y.; and Yang, M. 2024b. VioHawk: Detecting Traffic Violations of Autonomous Driving Systems through Criticality-Guided Sim-

ulation Testing. In *Proceedings of the ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA)*.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Maierhofer, S.; Rettinger, A.-K.; Mayer, E. C.; and Althoff, M. 2020. Formalization of Interstate Traffic Rules in Temporal Logic. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, 752–759.

Malla, S.; Choi, C.; Dwivedi, I.; Choi, J. H.; and 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.; Zwicklbauer, S.; and Paschke, A. 2022. Robust Traffic Rules and Knowledge Representation for Conflict Resolution in Autonomous Driving. In *RuleML+ RR (Companion)*.

Mao, J.; Qian, Y.; Zhao, H.; and Wang, Y. 2023. Gpt-driver: Learning to drive with gpt. *arXiv preprint arXiv:2310.01415*.

Mao, J.; Ye, J.; Qian, Y.; Pavone, M.; and Wang, Y. 2024. A Language Agent for Autonomous Driving. arXiv:2311.10813.

Mehdipour, N.; Althoff, M.; Tebbens, R. D.; and Belta, C. 2023. Formal methods to comply with rules of the road in autonomous driving: State of the art and grand challenges. *Automatica*, 152: 110692.

Nakano, R.; Hilton, J.; Balaji, S.; Wu, J.; Ouyang, L.; Kim, C.; Hesse, C.; Jain, S.; Kosaraju, V.; Saunders, W.; Jiang, X.; Cobbe, K.; Eloundou, T.; Krueger, G.; Button, K.; Knight, M.; Chess, B.; and Schulman, J. 2022. WebGPT: Browser-assisted question-answering with human feedback. arXiv:2112.09332.

Ponte, J. M.; and Croft, W. B. 2017. A language modeling approach to information retrieval. In *ACM SIGIR Forum*, volume 51, 202–208. ACM New York, NY, USA.

Robertson, S.; Zaragoza, H.; et al. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4): 333–389.

Sima, C.; Renz, K.; Chitta, K.; Chen, L.; Zhang, H.; Xie, C.; Luo, P.; Geiger, A.; and Li, H. 2023. DriveLM: Driving with Graph Visual Question Answering. *arXiv preprint arXiv:2312.14150*.

Sun, Y.; Poskitt, C. M.; Sun, J.; Chen, Y.; and Yang, Z. 2022. LawBreaker: An approach for specifying traffic laws and fuzzing autonomous vehicles. In *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering*, 1–12.

Tadewos, T. G.; Shamgah, L.; and Karimoddini, A. 2019. Automatic Safe Behaviour Tree Synthesis for Autonomous Agents. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, 2776–2781.

Tang, X.; Yuan, K.; Li, S.; Yang, S.; Zhou, Z.; and Huang, Y. 2023. Personalized Decision-Making and Control for Automated Vehicles Based on Generative Adversarial Imitation Learning. In *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, 4806–4812. IEEE.

Wang, X.; Qi, X.; Wang, P.; and Yang, J. 2021. Decision making framework for autonomous vehicles driving behavior in complex scenarios via hierarchical state machine. *Autonomous Intelligent Systems*, 1: 1–12.

Wei, Y.; Wang, Z.; Lu, Y.; Xu, C.; Liu, C.; Zhao, H.; Chen, S.; and Wang, Y. 2024. Editable Scene Simulation for Autonomous Driving via Collaborative LLM-Agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wen, C.; Liu, Y.; Bethala, G. C. R.; Peng, Z.; Lin, H.; Liu, Y.-S.; and Fang, Y. 2024. Enhancing Socially-Aware Robot Navigation through Bidirectional Natural Language Conversation. arXiv:2409.04965.

Wilson, B.; Qi, W.; Agarwal, T.; Lambert, J.; Singh, J.; Khandelwal, S.; Pan, B.; Kumar, R.; Hartnett, A.; Pontes, J. K.; et al. 2023. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*.

Xiao, W.; Mehdipour, N.; Collin, A.; Bin-Nun, A. Y.; Frazzoli, E.; Tebbens, R. D.; and Belta, C. 2021. Rule-based optimal control for autonomous driving. In *Proceedings of the ACM/IEEE 12th International Conference on Cyber-Physical Systems*, 143–154.

Yuan, K.; Huang, Y.; Yang, S.; Zhou, Z.; Wang, Y.; Cao, D.; and Chen, H. 2024. Evolutionary decision-making and planning for autonomous driving based on safe and rational exploration and exploitation. *Engineering*, 33: 108–120.

Zhang, M.; Jin, D.; Gu, C.; Hong, F.; Cai, Z.; Huang, J.; Zhang, C.; Guo, X.; Yang, L.; He, Y.; et al. 2024. Large motion model for unified multi-modal motion generation. *arXiv preprint arXiv:2404.01284*.

Zhao, J.; Zhao, W.; Deng, B.; Wang, Z.; Zhang, F.; Zheng, W.; Cao, W.; Nan, J.; Lian, Y.; and Burke, A. F. 2023. Autonomous driving system: A comprehensive survey. *Expert Systems with Applications*, 122836.