Any2Critical: Safety-Critical Scenario Generation from Arbitrary Real-World Driving Contexts

Yao Huang[1,2,3*], **Yubo Chen**[1*], **Ruochen Zhang**[1], **Yitong Sun**[1], **Shouwei Ruan**[1], **Zhenyu Wu**[1], **Yinpeng Dong**[2,3], **Xingxing Wei**[1†]

[1]Institute of Artificial Intelligence, Beihang University, Beijing 100191, China
[2]College of AI, Tsinghua University, Beijing 100084, China
[3]Shanghai Qi Zhi Institute
y_huang@buaa.edu.cn, chenyubo@buaa.edu.cn, xxwei@buaa.edu.cn

Abstract

Autonomous driving systems have achieved remarkable capabilities in real-world deployment, yet ensuring safety under corner cases remains a significant challenge due to the scarcity and constrained diversity of safety-critical scenarios. Existing generation methods may either lead to irrational vehicle behaviors or be limited by fixed collision patterns, while both heavily rely on existing map datasets, restricting the diversity. To address these fundamental limitations, we introduce **Any2Critical**, the first framework that can encode arbitrary real-world scenarios and generate contextually relevant safety-critical scenarios with realistic driving behaviors. Specifically, Any2Critical addresses two key challenges: (1) developing comprehensive, diverse map data by successfully leveraging everyday traffic situations as the most abundant source of real-world driving contexts, and (2) proposing an RAG-based Safety-Critical Scenario Generation Strategy based on our curated NHTSA-5K database for achieving an optimal balance between scenario diversity and behavioral rationality. Through comprehensive evaluation, we demonstrate that Any2Critical consistently achieves collision rates with an average of 89.69% across diverse scenarios and autonomous driving systems, significantly outperforming current state-of-the-art generation methods.

Code — https://github.com/Steven-iai/Any2Critical

1 Introduction

The rapid advancement of Autonomous Driving (AD) has brought remarkable capabilities for real-world deployment, such as navigating complex environments (Almalioglu et al. 2022; Hasanujjaman, Chowdhury, and Jang 2023), seamless multi-vehicle coordination (Pei et al. 2019; Cui et al. 2022), and precise lane-keeping under adverse conditions (Lee et al. 2022; Dong et al. 2023). However, ensuring the AD safety under corner cases (Sun et al. 2021; Feng et al. 2021, 2023) remains a significant challenge. For example, an autonomous vehicle (AV) may struggle to react to a vehicle abruptly merging from a blind alley, risking a collision. To improve safety under these scenarios, it is essential to collect and test on diverse datasets that capture such corner cases, thereby enhancing robust decision-making for AD deployment. To achieve this, conventional real-world testing (Kalra and Paddock 2016), though effective, is costly and time-consuming, typically demanding billions of miles driven to gather enough safety-critical scenarios. In contrast, the generation of simulated scenarios (Zhang, Xu, and Li 2024; Gao et al. 2025) offers a cost-effective and scalable alternative.

To effectively broaden the scope of simulated scenarios for AD safety testing, several methods called *safety-critical scenario generation techniques* have been proposed. For instance, some methods leveraging reinforcement learning or optimization-based perturbations (Feng et al. 2021; Rempe et al. 2022; Feng et al. 2023; Zhang et al. 2023) adversarially generate rare, high-risk scenarios to target long-tail collision cases. However, these approaches often produce irrationable vehicle behaviors, such as implausible collision trajectories. In contrast, techniques using predefined templates (Cai et al. 2020; Klischat et al. 2020) ensure logically consistent scenarios compliant with traffic regulations but are limited by fixed collision patterns, reducing scenario variety. Moreover, both approaches heavily rely on existing map datasets, which restricts diversity across diverse geographical and traffic contexts.

To further advance the AD safety testing, this paper seeks to fundamentally address the scarcity and constrained diversity of safety-critical scenarios while ensuring realistic driving behaviors by tackling two key challenges: (1) developing comprehensive, diverse map data that captures varied real-world driving conditions, overcoming the constraints of existing map datasets; and (2) balancing scenario diversity with behavioral rationality to prevent unrealistic outcomes from adversarial optimization methods, while surpassing the fixed patterns of template-based approaches.

For the challenge of developing comprehensive, diverse map data, we draw inspiration from the rich tapestry of real-world traffic scenarios that surround us daily: everyday traffic situations actually represent the most abundant and diverse source of map data. To harness this potential, we propose a framework called **Any2Critical**, which can encode arbitrary real-world scenarios and generate contextually rel-

*These authors contributed equally.
†Corresponding Author.

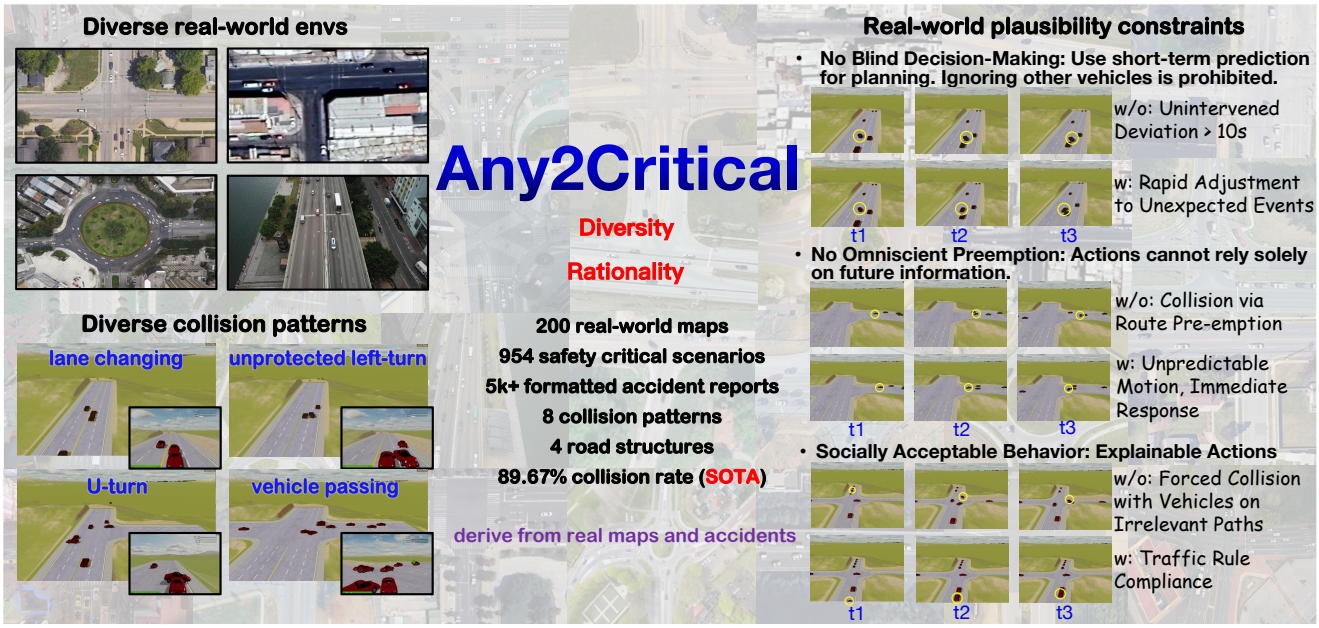

Figure 1: Our Any2Critical emphasizes two key principles, **Diversity** and **Rationality**, to ensure that the generated scenarios are both safety-critical and abundant, while being grounded in real-world conditions with plausible behaviors. To the best of our knowledge, this is the first work to generate scenarios from arbitrary real-world road maps with complex traffic conditions.

evant safety-critical scenarios. Specifically, Any2Critical first employs a multi-modal perception pipeline that extracts semantic map features via Vision Language Models (VLMs) and captures real-world vehicle positions through Grounding DINO (Liu et al. 2024) and SAM (Kirillov et al. 2023). A LLM planner then intelligently integrates these features with the MetaDrive (Li et al. 2022) simulator to generate realistic map configurations that could be used to perform safety-critical scenario generation.

Then, to address the challenge of balancing scenario diversity with behavioral rationality, we design a novel RAG-based Safety-Critical Scenario Generation Strategy for identifying appropriate safety-critical patterns. Specifically, we first manually curate **NHTSA-5K**, a meticulously structured accident database derived from National Highway Traffic Safety Administration records, which captures a rich diversity of real-world collision scenarios spanning multiple accident types, vehicle configurations, and road conditions, serving as our comprehensive knowledge base of collision patterns. Our RAG-based module then retrieves the most relevant historical accident reports from this database to align with pre-encoded driving scenes, using a frequency-aware matching strategy to capture both common and critical rare patterns. Subsequently, LLMs synthesize these retrieved historical patterns with scene-specific contextual details to generate realistic safety-critical scenarios. To ensure behavioral rationality throughout this process, we incorporate a rule-based validation mechanism that verifies the generated scenarios stem from plausible risk patterns rather than unrealistic fabrications, thereby maintaining the authenticity and credibility of the synthesized scenarios.

Overall, the contributions of this paper are as follows:

- We propose Any2Critical, the first framework capable of encoding arbitrary real-world driving contexts into diverse safety-critical simulations, thereby eliminating the limitations imposed by fixed dataset constraints.
- We design a novel RAG-based Safety-Critical Scenario Generation Strategy that could ensure diversity through frequency-aware retrieval of both common and critically rare accident patterns from our curated NHTSA-5K database, while maintaining behavioral rationality via novel rule-based validation, achieving an optimal balance between scenario diversity and driving plausibility.
- Through comprehensive evaluation, we demonstrate that Any2Critical consistently achieves collision rates with an average of 89.69% across diverse scenarios and autonomous driving systems, significantly outperforming existing state-of-the-art methods.

2 Related Work

2.1 Safety-Critical Scenario Generation

Safety-critical scenario generation techniques have been extensively studied to address the limitations of real-world testing in autonomous driving validation (Kalra and Paddock 2016). Current approaches primarily fall into two categories: adversarial generation methods and template-based generation methods. Adversarial approaches (Feng et al. 2021; Rempe et al. 2022; Feng et al. 2023; Zhang et al. 2023) utilize reinforcement learning or optimization techniques to discover rare collision scenarios by perturbing vehicle behaviors or environmental conditions, effectively targeting long-tail safety events. However, these methods often

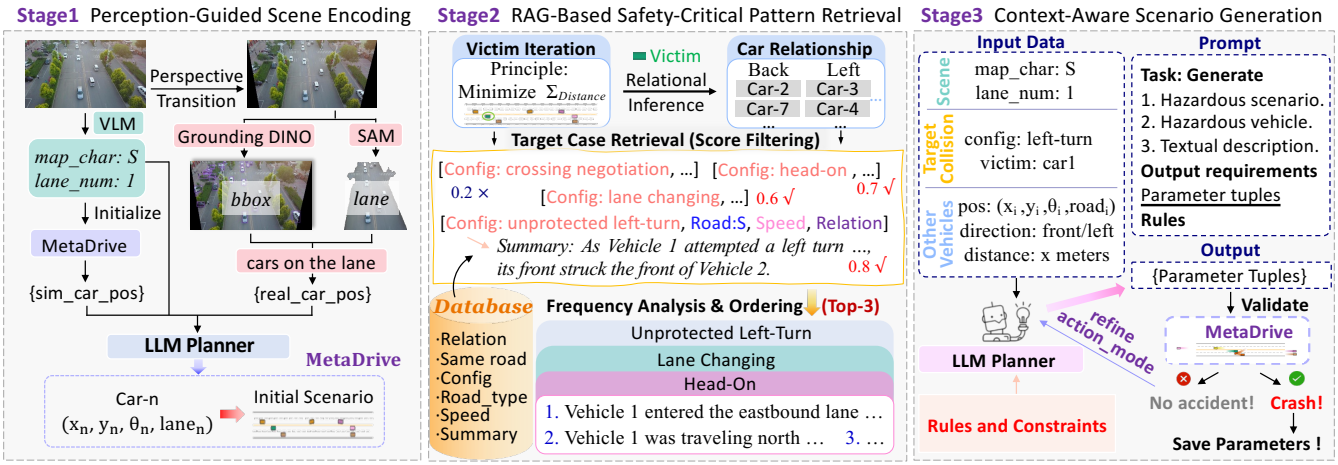

Figure 2: An overview of Any2Critical framework. Stage 1: Semantic and visual information are extracted to determine road and vehicle initialization parameters via an LLM planner. Stage 2: Accident cases are then retrieved and filtered by similarity score; the top three configurations are selected, with the three most similar cases per config retained. Stage 3: Based on scene information, target collision cases and vehicle orientations, the LLM planner selects the hazardous vehicle and sets parameters for all involved vehicles to generate a realistic collision scenario during refinement with the promise of plausibility constraints.

struggle with generating realistic driving behaviors, leading to scenarios with implausible vehicle trajectories. In contrast, template-based methods (Cai et al. 2020; Klischat et al. 2020) rely on predefined scenario structures to ensure behavioral consistency and traffic rule compliance, but suffer from limited diversity due to their dependence on fixed patterns. Recent works have also begun exploring language model-based generation (Zhang, Xu, and Li 2024; Gao et al. 2025) to improve collision diversity, yet existing approaches remain constrained by their reliance on specific map datasets and geographical contexts, motivating the need for more flexible generation frameworks. To meet this, we propose the Any2Critical framework, which could encode arbitrary real-world driving contexts into diverse safety-critical simulations, greatly enriching the scope of testing scenarios.

2.2 Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG)(Lewis et al. 2020) has rapidly gained traction due to its ability to combine external knowledge retrieval with the generative power of LLMs, thereby significantly improving the factual accuracy and domain relevance of generated outputs. Building upon traditional vector-based retrieval (Guu et al. 2020; Gao et al. 2023; Zhao et al. 2024), recent advances have evolved toward structured knowledge representation through graph-based approaches(Han et al. 2024; Edge et al. 2024), which enable more precise pattern matching and relationship reasoning. To further enhance retrieval effectiveness, hybrid strategies have emerged that combine semantic embeddings with exact pattern matching (Karpukhin et al. 2020; Xu, Shi, and Choi 2024). Motivated by these advances, we adapt RAG principles to safety-critical scenario generation by constructing a structured knowledge base of real-world accident patterns from NHTSA databases, employing frequency-aware retrieval strategies to systematically discover both prevalent and rare collision patterns, and utilizing structured pattern matching to ensure contextual relevance and behavioral rationality in generated scenarios.

3 Any2Critical Framework

3.1 Framework Overview

The Any2Critical framework generates diverse, realistic safety-critical scenarios for autonomous driving (AD) safety testing from arbitrary real-world scenes. As shown in Figure 2, it integrates a Perception-Guided Scene Encoding module to encode scenes into MetaDrive configurations, and a RAG-Based Safety-Critical Scenario Generation module, which retrieves patterns from the NHTSA-5K database and applies rule-based validation to ensure plausible scenarios.

3.2 Perception-Guided Scene Encoding

To encode real-world scenarios into simulation-compatible configurations, we develop a perception-guided scene encoding method that systematically extracts and reconstructs key environmental elements. This hybrid approach strategically combines coarse-grained perception from a VLM with precise extraction from specialized visual models including Grounding DINO and SAM, effectively mitigating potential hallucinations inherent in VLMs and ultimately achieving seamless mapping from real-world scenes to the simulator.

Specifically, given an input image or video frame I, our pipeline begins by employing the VLM to extract coarse-grained semantic configurations, $\mathbf{e}_S = \{s_1, s_2, \ldots, s_m\}$, where each s_i represents essential attributes such as road structures, lane counts, and maximum speed limits. This semantic extraction can be formulated as:

$$\mathbf{e}_S = f_{\text{VLM}}(I, P_{\text{VLM}}), \tag{1}$$

where $f_{\text{VLM}}(\cdot)$ denotes the VLM's extraction function, and P_{VLM} represents the corresponding prompt. These coarse se-

mantic parameters $\mathbf{e}_S$ then serve to initialize the MetaDrive simulator with approximate scene configurations.

Concurrently, we leverage specialized visual models to perform precise geometric extraction. Prior to perception processing, a perspective transformation $T \in \mathbb{R}^{3\times3}$ is applied to transform the input I from its original perspective view to a perception-friendly bird's-eye view. Subsequently, Grounding DINO performs vehicle detection, yielding bounding boxes $B = \{(x_j, y_j, w_j, h_j)\}_{j=1}^{n}$ for n detected vehicles. Meanwhile, SAM segments the lane structures, generating segmentation masks $M = \{m_j\}_{j=1}^{n}$. The precise positional information is then aggregated as:

$$\mathbf{e}_P = \sum_{j=1}^{n} \psi(b_j, m_j, T), \tag{2}$$

where $\psi(\cdot,\cdot,\cdot)$ represents a fusion function that combines bounding box coordinates, lane masks, and transformation parameters to yield calibrated vehicle and lane layout.

Finally, both the semantic configuration $\mathbf{e}_S$ and positional information $\mathbf{e}_P$ are integrated through an LLM Planner with designated prompt P_{LLM} for comprehensive alignment:

$$\Theta = f_{\text{LLM}}(\mathbf{e}_S, \mathbf{e}_P; P_{\text{LLM}}), \tag{3}$$

where $f_{\text{LLM}}(\cdot,\cdot;P_{\text{LLM}})$ intelligently reconciles the initialized MetaDrive configuration with the more precise geometric data, performing fine-grained adjustments to parameters such as lane numbers and vehicle positions to ensure high-fidelity scene representation.

This systematic alignment process creates a coherent and accurate simulation environment by leveraging the LLM's reasoning capabilities to reconcile coarse VLM configurations with precise geometric observations, ultimately yielding a well-calibrated MetaDrive map Θ for the following safety-critical scenario generation.

3.3 RAG-Based Safety-Critical Scenario Generation

This module generates safety-critical scenarios by adapting and retrieving real-world accident patterns from a manually curated database, balancing diversity and rationality through frequency-aware retrieval and rule-based validation.

Construction of NHTSA-5K We first construct NHTSA-5K, a structured database derived from National Highway Traffic Safety Administration accident reports. The selection of NHTSA database is strategically motivated by its authoritative data reliability and comprehensive coverage across diverse geographical regions and traffic conditions. More importantly, the rich textual descriptions in NHTSA reports contain sufficient detail for extracting complex accident patterns essential for autonomous vehicle testing.

To systematically capture the essence of safety-critical events, our workflow employs specialized LLM prompting strategies to transform unstructured NHTSA narratives T_k into structured tuples $D_k = (L_k, C_k, V_k, P_k, B_k)$. This multidimensional framework begins with **Lane Structure (L_k)**, which defines the spatial foundation and road geometry that fundamentally constrain vehicle interactions. Building on this context, **Collision Type (C_k)** identifies the specific vehicle-to-vehicle interaction mechanisms and their associated risk patterns. The dynamic nature of the accident is captured by **Maximum Speed (V_k)**, establishing a kinetic framework where velocity dictates both avoidance feasibility and impact severity. This is complemented by **Relative Position (P_k)**, which records the precise geometric configuration at the critical moment to enable accurate scenario reconstruction. Finally, **Behavioral Abstraction (B_k)** encodes the causal sequence of vehicle actions, linking environmental constraints to collision outcomes by revealing the underlying decision-making failures that autonomous systems must navigate and avoid.

This systematic extraction process finally produces over 5k structured entries that preserve the statistical distribution of real-world accident patterns while enabling computational analysis. The resulting frequency-aware knowledge base also maintains authentic collision probabilities, ensuring that generated scenarios reflect genuine traffic safety challenges while providing sufficient diversity for comprehensive autonomous vehicle evaluation.

RAG-Based Safety-Critical Pattern Retrieval To retrieve relevant safety-critical patterns from the NHTSA-5K database, we propose a two-stage frequency-aware matching mechanism that balances pattern relevance with collision type diversity. This approach addresses two key challenges: ensuring retrieved patterns are contextually similar to the target scenario while maintaining representation of both common and critically rare collision types.

Our method begins with the encoded scene parameters $\Theta = (\Theta^L, \Theta^V, \Theta^P)$, representing lane structure, maximum speed, and victim vehicle orientations respectively. The vehicle with the minimal total distance to others is prioritized as the victim. For each database entry D_k with corresponding $(\Theta_k^L, \Theta_k^V, \Theta_k^P)$, we compute a weighted similarity score:

$$\begin{aligned}\text{sim}(D_k, \Theta) = w_1 \cdot \delta(\Theta_k^L, \Theta^L) + w_2 \cdot \delta(\Theta_k^V, \Theta^V) \\ + w_3 \cdot \delta(\Theta_k^P, \Theta^P),\end{aligned} \tag{4}$$

where $\delta(\cdot,\cdot)$ represents a normalized similarity metric. For embedded representations, we employ cosine similarity:

$$\delta(a, b) = \frac{\mathbf{v}_a \cdot \mathbf{v}_b}{\|\mathbf{v}_a\|\|\mathbf{v}_b\|}, \tag{5}$$

with $\mathbf{v}_a$ and $\mathbf{v}_b$ denoting the vector embeddings of features a and b. The weights w_i are empirically tuned to reflect the relative importance of different scene characteristics.

The two-stage retrieval process operates as follows:

Stage 1: Similarity-based Filtering and Frequency Analysis. We first filter database entries using a similarity threshold τ, retaining only those with $\text{sim}(D_k, \Theta) > \tau$ and aligned victim vehicle orientations. The filtered entries are then grouped by collision type C_k, and we compute the frequency of each collision type c as:

$$f_c = \sum_{k:C_k=c} \mathbb{I}(\text{sim}(D_k, \Theta) > \tau), \tag{6}$$

where $\mathbb{I}(\cdot)$ is the indicator function. Collision types are subsequently sorted in descending order of frequency, ensuring that common collision patterns receive priority while

preserving representation of rare but critical scenarios. The top three configurations based on this frequency ranking are then considered for advancement to the second stage.

Stage 2: Top-k Selection with Context Augmentation. For each retained collision type c identified in Stage 1, we retrieve the top-3 entries with highest similarity scores, forming the subset $R_c = \{D_{k_1}, D_{k_2}, D_{k_3}\}$. These selected patterns are then augmented with comprehensive scene context information Θ to provide the LLM with sufficient environmental details for accurate scenario generation.

Rule-Constrained Scenario Generation The LLM synthesizes retrieved collision patterns R_c with aligned scene features Θ to generate safety-critical vehicle control sequences, i.e., parameters of all vehicles like positions, speed and action mode. Such parameters determine acceleration profiles and trajectory modifications, which are directly compatible with the MetaDrive simulation environment.

During this phase, we design a unique rule-based validation mechanism using a predefined rule library $\mathcal{R}$ to maintain scenario rationality and prevent the generation of physically impossible or legally invalid behaviors. Our rule library, detailed in Table 1, encodes essential traffic regulations and physical constraints that govern realistic driving scenarios. These rules span multiple domains including speed compliance, physical feasibility, spatial constraints, safety protocols, and regulatory adherence.

The validation process operates through a two-tier scoring mechanism. For a generated scenario G, we first compute a basic validity score that evaluates compliance across rules:

$$v(G) = \sum_{r \in \mathcal{R}} \mathbb{I}(r(G)), \tag{7}$$

where $\mathbb{I}(\cdot)$ represents the indicator function that returns 1 if rule r is satisfied by scenario G, and 0 otherwise.

Recognizing that certain rules carry greater importance for scenario realism and safety, we extend this with a weighted validation scheme:

$$v_w(G) = \sum_{r \in \mathcal{R}} \alpha_r \mathbb{I}(r(G)), \tag{8}$$

where α_r denotes the weight assigned to rule r, reflecting its criticality for scenario validity. Rules related to physical constraints and safety protocols typically receive higher weights than those governing minor regulatory details.

Scenarios failing to meet the validation threshold, defined as $v_w(G) < \eta$, undergo adjustments during the iterative refinement. This process specifically addresses violated rules while preserving the safety-critical nature, ultimately ensuring the generation of plausible risk transformations while eliminating artificially contrived collision sequences.

In all, iterative refinements continue until a successful collision occurs under the latest settings in MetaDrive. This process involves feeding the trajectories of the victim and hazardous vehicles from the previous iteration back to the LLM, prompting it to adjust the hazardous vehicle's action mode to increase collision likelihood. Concurrently, trajectory analysis aids in identifying rule violations, enabling the LLM to generate more plausible risk transformations.

Rule	Description
Driving Behavior Rationality Constraints	
R1.1	Prohibition of random behaviors without clear causal chains or objectives
R1.2	Detection and prevention of abnormal non-task-oriented repetitive actions
Hazardous Behavior & Collision Modeling Rationality	
R2.1	Rational target object selection based on perception and trajectory intersection
R2.2	Validation of collision causal chains (intent $\rightarrow$ conflict $\rightarrow$ intersection $\rightarrow$ timing)
R2.3	Limitation of threat modeling to reachable areas without obstacle penetration
Perception & Temporal Rationality	
R3.1	Prohibition of omniscient prediction based on future known information
R3.2	Prevention of blind decision-making by utilizing short-term prediction capabilities
R3.3	Enforcement of temporal smoothness in velocity, direction, and acceleration changes
Traffic Rules & Social Interaction	
R4.1	Compliance with basic traffic regulations (signals, speed limits, right-of-way)
R4.2	Maintenance of the minimum safety distance with dynamic thresholds
R4.3	Socially acceptable behavior modeling considering human driver expectations

Table 1: Rule Library for Scenario Validation

4 Experiments and Results

In this section, we conduct a series of qualitative and quantitative experiments to evaluate the superior performance of our method in generating safety-critical scenarios. First, we examine the authenticity and rationality of the generated scenarios. Second, we assess the collision rate and diversity of the generated scenarios. Finally, we perform ablation experiments to evaluate the effectiveness of each component.

4.1 Experimental Setup

Input Data. The road shapes consist of straight sections, T-junctions, crossroads, and roundabouts. In this work, we collect 200 images of different traffic conditions, covering these four common basic road types, as inputs to the algorithm. These 200 images include 26 images of O (roundabouts), 44 images of X (crossroads), 63 images of T (T-junctions), and 67 images of S (straight sections), which can effectively represent daily traffic scenarios.

Collision Patterns. In this paper, we define eight common collision patterns from real-world statistics: Rear-end, Head-on, Lane Changing, Right-Turn, Unprotected Left-turn, U-turn, Crossing Negotiation, and Vehicle Passing. Our NHTSA-5K database is classified accordingly, and our method can generate multiple safety-critical scenarios with different collision patterns from a single original scenario.

Methods	IDM		PPO		TD3	
	Collision Rate↑	Overall Score↓	Collision Rate↑	Overall Score↓	Collision Rate↑	Overall Score↓
Clean	0.02	0.91	0.39	0.58	0.34	0.58
ChatScene (Zhang, Xu, and Li 2024)	0.76	0.27	0.70	0.31	0.89	0.13
FWC (Gao et al. 2025)	0.63	0.39	0.68	0.33	0.91	0.11
Any2Critical	**1.00**	**0.07**	**0.77**	**0.25**	**0.92**	**0.10**

Table 2: Performance Comparison of Driving Algorithms Across Multiple Scenarios

Metric	Road Shape	# Lanes	# Vehicles	Velocity Plaus.	Behavior Plaus.
Accuracy	0.78	0.98	0.80	0.99	0.95

Table 3: Accuracy Evaluation Across Scene Attributes

Metric	Straight	T-Junction	Crossroads	Roundabout
Number of Cases	350	215	311	78
Collision Rate	0.86	0.89	0.93	0.97
Overall Score	0.17	0.14	0.11	0.08

Table 4: Performance Metrics for Different Road Shapes

Baselines. We compare Any2Critical with two latest safety-critical generation techniques ChatScene (Zhang, Xu, and Li 2024) and FWC (Gao et al. 2025). For a fair comparsion, we migrate these two algorithms to the MetaDrive simulator.

AD Models. Here, we test three representative AD models: the rule-based Intelligent Driver Model (IDM) (Treiber, Hennecke, and Helbing 2000), on-policy Proximal Policy Optimization (PPO) (Schulman et al. 2017), and off-policy Twin Delayed Deep Deterministic Policy Gradient (TD3) (Fujimoto, van Hoof, and Meger 2018).

Evaluation Metrics. For evaluation, we adopt Collision Rate (CR), the proportion of collisions in total simulations, and Overall Score (OS), a metric covering Safety, Functionality, and Etiquette levels (Xu et al. 2022; Zhang, Xu, and Li 2024) (fine-grained details in *Appendix*). For truthfulness and rationality, we compute encoding accuracy against ground truth and obtain plausibility scores via LLM-based rule validation to measure behavioral rationality.

▷ For more specific *Implementation Details* of Any2Critical and the evaluation templates, please refer to the *Appendix*.

4.2 Main Results

Safety-Critical Scenario Generation Performance. As shown in Table 2, our Any2Critical significantly outperforms existing methods in generating safety-critical scenarios. Compared to ChatScene and FWC, Any2Critical consistently achieves the highest CRs and lowest ORs across all three autonomous driving algorithms tested. Notably, our method achieves a perfect 1.00 collision rate with IDM, substantially exceeding ChatScene's 0.76 and FWC's 0.62. This superior performance demonstrates not only the enhanced criticality of our generated scenarios but also their excellent transferability across different driving algorithms. As shown in Figure 5 and Table 4, Any2Critical maintains consistently high CRs across eight collision patterns and four road structures, with performance ranging from 0.78 to 0.97. This comprehensive effectiveness across diverse conditions validates both the robustness of our method and its superior capability in generating challenging safety-critical scenarios compared to existing techniques.

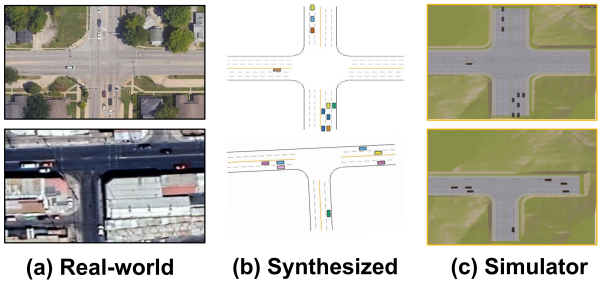

Figure 3: Real2Sim Scenario Mapping.

Accuracy of Encoded Scenarios. Table 3 evaluates the accuracy of our perception-guided scenario encoding against ground truth traffic conditions. The results demonstrate that our method achieves consistently high accuracy across all scene attributes, with performance ranging from 0.78 to 0.99. This validates our method's capability to accurately perceive traffic conditions and generate realistic safety-critical scenarios. The precise road geometry and vehicle mapping illustrated in Figure 3 further confirm the effectiveness of our perception-guided encoding approach.

Performance under Varied Collisions Figure 5 demonstrates our method's exceptional capability to generate diverse yet consistently challenging safety-critical scenarios across nine collision patterns. Any2Critical achieves remarkably high CRs ranging from 0.78 to 0.92, with crossing negotiation scenarios reaching the best performance of 0.92 CR and 0.12 OS. This comprehensive coverage validates our method's superior ability to explore the full spectrum of potential safety-critical situations, ensuring thorough testing for autonomous driving systems.

Performance under Varied Road Shapes Table 4 validates Any2Critical's strong generalization capability across different road geometries. Our method consistently maintains high CRs exceeding 0.85 across all road types, with

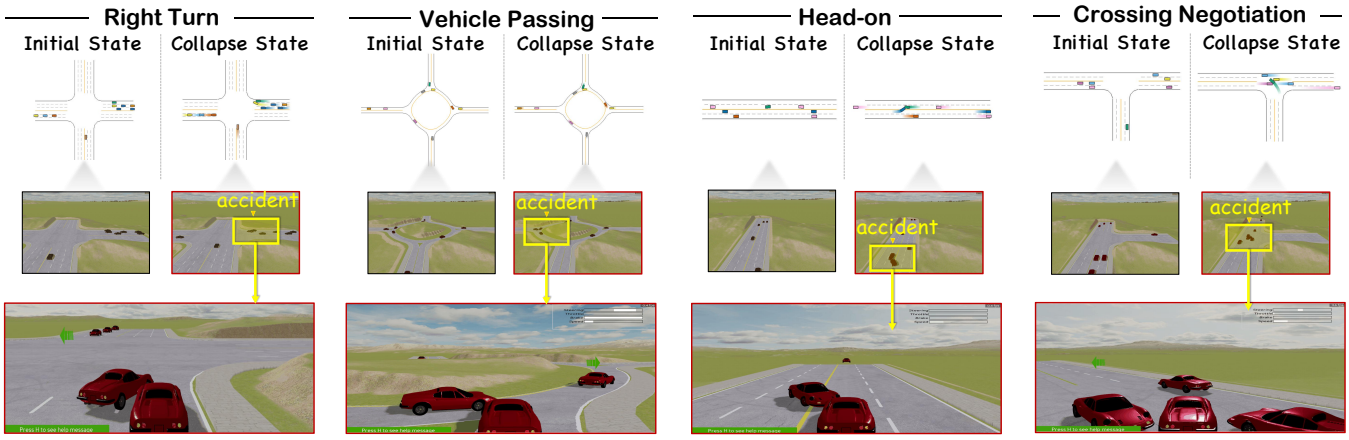

Figure 4: Examples of Any2Critical's generation results for various safety-critical scenarios. We include different road structures and vehicle manipulations for specific collision types.

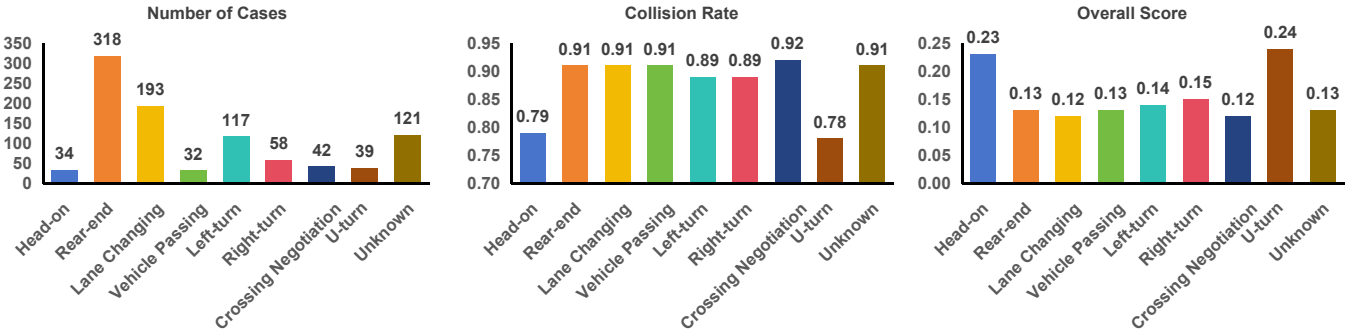

Figure 5: Safety-Critical Scenario Generation Analysis across Various Collisions: Quantity, Collision Rate, and Overall Score.

roundabouts achieving the highest challenge level (0.97 CR). The progressive performance scaling from straight roads (0.86 CR) to complex roundabouts (0.97 CR) demonstrates our method's intelligent adaptation to geometric complexity, ensuring effective safety-critical scenario generation regardless of road configuration.

4.3 Ablation Study

To validate each component's contribution, we individually remove the RAG module and plausibility rule constraints. Table 5 shows our full setting achieves optimal performance. Removing the RAG module causes significant degradation, demonstrating its critical role in enhancing scenario criticality and authenticity. Similarly, eliminating plausibility constraints reduces performance, because excessively abnormal vehicle behaviors are easily perceived, confirming that rules are necessary to maintain the rationality of the generation.

5 Limitations

For concerns about scenario realism, we conduct a double-blind human evaluation showing 0.98 of scenarios receive high plausibility ratings from drivers and AD engineers, accompanied by detailed traffic rule compliance analysis in the *Appendix*. However, two limitations remain: (1) the simulation-to-reality gap requires validation as real-world deployment introduces environmental variations not fully captured in simulation; (2) our NHTSA-5K necessitates geographically diverse subsets for broader applicability.

Setting	Collision Rate	Overall Score	Plausibility
Full	0.90	0.14	0.95
w/o RAG DB	0.79	0.23	0.87
w/o Plausibility Rule	0.81	0.23	0.88

Table 5: Ablation Study Results on CR, OS, and Plausibility

6 Conclusion

This paper presents Any2Critical, the first framework that generates safety-critical scenarios from arbitrary real-world driving contexts. By combining perception-guided scene encoding with a RAG-based generation strategy using our curated NHTSA-5K database, Any2Critical achieves an optimal balance between scenario diversity and behavioral rationality. The framework significantly outperforms existing methods across diverse road configurations and collision patterns, demonstrating superior effectiveness in generating challenging safety-critical scenarios for comprehensive autonomous vehicle safety testing.

Acknowledgments

This work was supported by NSFC Projects (62576020, 62276149, U2341228) and was also supported by the Fundamental Research Funds for the Central Universities.

References

Almalioglu, Y.; Turan, M.; Trigoni, N.; and Markham, A. 2022. Deep learning-based robust positioning for all-weather autonomous driving. *Nature machine intelligence*, 4(9): 749–760.

Cai, P.; Lee, Y.; Luo, Y.; and Hsu, D. 2020. Summit: A simulator for urban driving in massive mixed traffic. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 4023–4029. IEEE.

Cui, J.; Qiu, H.; Chen, D.; Stone, P.; and Zhu, Y. 2022. Coopernaut: End-to-end driving with cooperative perception for networked vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17252–17262.

Dong, Y.; Kang, C.; Zhang, J.; Zhu, Z.; Wang, Y.; Yang, X.; Su, H.; Wei, X.; and Zhu, J. 2023. Benchmarking robustness of 3d object detection to common corruptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1022–1032.

Edge, D.; Trinh, H.; Cheng, N.; et al. 2024. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. *arXiv preprint arXiv:2404.16130*.

Feng, S.; Sun, H.; Yan, X.; Zhu, H.; Zou, Z.; Shen, S.; and Liu, H. X. 2023. Dense reinforcement learning for safety validation of autonomous vehicles. *Nature*, 615(7953): 620–627.

Feng, S.; Yan, X.; Sun, H.; Feng, Y.; and Liu, H. X. 2021. Intelligent driving intelligence test for autonomous vehicles with naturalistic and adversarial environment. *Nature communications*, 12(1): 748.

Fujimoto, S.; van Hoof, H.; and Meger, D. 2018. Addressing function approximation error in actor-critic methods. *arXiv preprint arXiv:1802.09477*.

Gao, Y.; Piccinini, M.; Moller, K.; Alanwar, A.; and Betz, J. 2025. From Words to Collisions: LLM-Guided Evaluation and Adversarial Generation of Safety-Critical Driving Scenarios. *arXiv preprint arXiv:2502.02145*.

Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Wang, H.; and Wang, H. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1).

Guu, K.; Lee, K.; Tung, Z.; Pasupat, P.; and Chang, M. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, 3929–3938. PMLR.

Han, H.; Wang, Y.; Shomer, H.; Guo, K.; Ding, J.; Lei, Y.; Halappanavar, M.; Rossi, R. A.; Mukherjee, S.; Tang, X.; et al. 2024. Retrieval-augmented generation with graphs (graphrag). *arXiv preprint arXiv:2501.00309*.

Hasanujjaman, M.; Chowdhury, M. Z.; and Jang, Y. M. 2023. Sensor fusion in autonomous vehicle with traffic surveillance camera system: detection, localization, and AI networking. *Sensors*, 23(6): 3335.

Kalra, N.; and Paddock, S. M. 2016. Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability? *Transportation research part A: policy and practice*, 94: 182–193.

Karpukhin, V.; Oğuz, B.; Min, S.; et al. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6769–6781.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.

Klischat, M.; Liu, E. I.; Holtke, F.; and Althoff, M. 2020. Scenario factory: Creating safety-critical traffic scenarios for automated vehicles. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, 1–7. IEEE.

Lee, Y.; Ko, Y.; Kim, Y.; and Jeon, M. 2022. Perception-friendly video enhancement for autonomous driving under adverse weather conditions. In *2022 International Conference on Robotics and Automation (ICRA)*, 7760–7767. IEEE.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, 9459–9474. Curran Associates, Inc.

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*, 45(3): 3461–3475.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. 2024. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, 38–55. Springer.

Pei, H.; Feng, S.; Zhang, Y.; and Yao, D. 2019. A cooperative driving strategy for merging at on-ramps based on dynamic programming. *IEEE Transactions on Vehicular Technology*, 68(12): 11646–11656.

Rempe, D.; Philion, J.; Guibas, L. J.; Fidler, S.; and Litany, O. 2022. Generating useful accident-prone driving scenarios via a learned traffic prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17305–17315.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Sun, H.; Feng, S.; Yan, X.; and Liu, H. X. 2021. Corner case generation and analysis for safety assessment of autonomous

vehicles. *Transportation research record*, 2675(11): 587–600.

Treiber, M.; Hennecke, A.; and Helbing, D. 2000. Congested traffic states in empirical observations and microscopic simulations. *Physical Review E*, 62(2): 1805.

Xu, C.; Ding, W.; Lyu, W.; Liu, Z.; Wang, S.; He, Y.; Hu, H.; Zhao, D.; and Li, B. 2022. SafeBench: A benchmarking platform for safety evaluation of autonomous vehicles. *Advances in Neural Information Processing Systems*, 35: 25494–25509.

Xu, F.; Shi, W.; and Choi, E. 2024. RECOMP: Improving retrieval-augmented LMs with context compression and selective augmentation. In *The Twelfth International Conference on Learning Representations*.

Zhang, J.; Xu, C.; and 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, L.; Peng, Z.; Li, Q.; and Zhou, B. 2023. Cat: Closed-loop adversarial training for safe end-to-end driving. In *Conference on Robot Learning*, 2357–2372. PMLR.

Zhao, P.; Zhang, H.; Yu, Q.; Wang, Z.; Geng, Y.; Fu, F.; Yang, L.; Zhang, W.; Jiang, J.; and Cui, B. 2024. Retrieval-augmented generation for ai-generated content: A survey. *arXiv preprint arXiv:2402.19473*.