Talk2Traffic: Interactive and Editable Traffic Scenario Generation for Autonomous Driving with Multimodal Large Language Model

Zihao Sheng[1] Zilin Huang[1] Yansong Qu[2] Yue Leng[3] Sikai Chen[1*]
[1]University of Wisconsin-Madison [2]Purdue University [3]Google
https://zihaosheng.github.io/Talk2Traffic/

Abstract

Deploying autonomous vehicles (AVs) requires testing in diverse and challenging scenarios to ensure safety and reliability, yet collecting real-world data remains prohibitively expensive. While simulation-based approaches offer cost-effective alternatives, most existing methods lack sufficient support for intuitive, interactive editing of generated scenarios. This paper presents Talk2Traffic, a novel framework that leverages multimodal large language models (MLLMs) to enable interactive and editable traffic scenario generation. Talk2Traffic allows human users to generate various traffic scenarios through multimodal inputs (text, speech, and sketches). Our approach first employs an MLLM-based interpreter to extract structured representations from these inputs. These representations are then translated into executable Scenic code using a retrieval-augmented generation mechanism to reduce hallucinations and ensure syntactic correctness. Furthermore, a human feedback guidance module enables iterative refinement and editing of scenarios through natural language instructions. Experiments demonstrate that Talk2Traffic outperforms state-of-the-art methods in generating challenging scenarios. Qualitative evaluations further illustrate the framework can handle diverse input modalities and support scenario editing.

1. Introduction

Recent years have witnessed remarkable progress in autonomous vehicles (AVs), yet safety concerns continue to pose significant challenges for their widespread deployment on public roads [17, 19, 34]. Ensuring the reliability of AVs requires extensive training and validation across diverse and complex traffic scenarios that encompass the full spectrum of real-world driving conditions [22, 28, 32, 33]. While existing large-scale autonomous driving datasets, including Waymo Open Dataset [15], nuScenes [4], and Argoverse [5], have provided valuable traffic data, the process of collecting and annotating real-world scenarios remains prohibitively expensive and time-consuming. Furthermore, these datasets inherently capture a limited subset of possible traffic interactions, leaving AVs potentially vulnerable to unforeseen scenarios. This gap motivates the development of more cost-effective and flexible approaches through simulation-based traffic scenario generation.

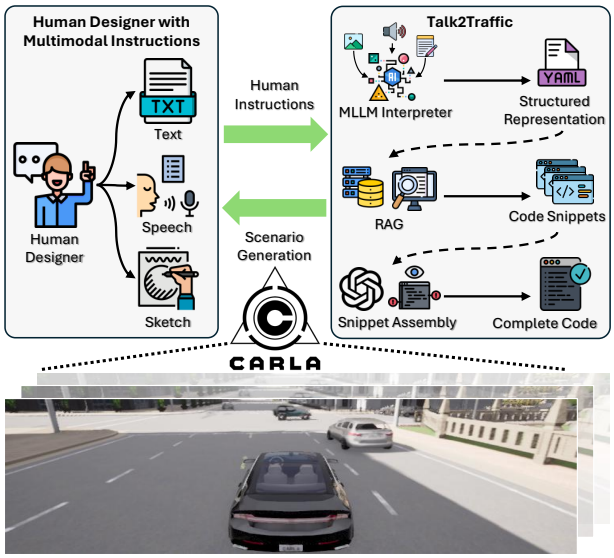

Figure 1. Talk2Traffic enables human designers to create traffic scenarios through multimodal inputs (text descriptions, speech commands, and freehand sketches).

Existing approaches for traffic scenario generation in simulation environments have evolved along several directions [10]. Traditional rule-based methods, such as CARLA ScenarioRunner [7], enable structured scenario configuration through parameterized templates, offering repeatable test cases under predefined conditions. However, their reliance on hand-crafted rules restricts scenario diversity and requires extensive domain expertise to model complex agent interactions. Data-driven approaches, particularly those leveraging diffusion models [21, 27], have demon-

*Corresponding author: sikai.chen@wisc.edu

strated promising results in generating diverse and realistic traffic scenarios by learning from real-world data distributions. These methods can capture complex interaction patterns that are difficult to specify through explicit rules. Nevertheless, a critical limitation of these approaches lies in their inability to adequately support instruction-guided interactive editing. Consequently, users cannot dynamically adjust agent behaviors or modify scene layouts through feedback during generation. This limitation hampers the efficient creation and iterative editing of traffic scenarios to meet customized testing needs.

The emergence of multimodal large language models (MLLMs) presents opportunities to address this limitation. Trained on Internet-scale multimodal data encompassing text, images, and videos, state-of-the-art MLLMs such as GPT-4 [1], Llama [13], and Qwen [2] have demonstrated unprecedented capabilities in unified cross-modal understanding and reasoning [8]. Building upon this foundation, these models could understand complex spatial relationships and temporal dynamics that are essential for traffic scenario generation. Moreover, MLLMs' inherent strength in following interactive instructions naturally aligns with the iterative workflow of human designers who need to progressively refine or modify traffic scenarios [23, 35]. However, despite these promising characteristics, existing approaches to traffic scenario generation have not fully explored the potential of MLLMs to enable interactive editing and semantic control of traffic scenarios.

In this paper, we present **Talk2Traffic**, a novel framework that leverages the capabilities of MLLMs for interactive and editable traffic scenario generation, as illustrated in Fig. 1. Our framework processes multimodal inputs (text, speech, and sketches) through a specialized interpreter that extracts structured representations of traffic scenarios. These representations are then translated into executable Scenic code using a retrieval-augmented generation (RAG) approach that grounds MLLMs' outputs in verified code snippets, effectively reducing hallucinations and ensuring syntactic correctness. The generated scenarios can be iteratively refined or edited through natural language feedback in a human-in-the-loop pipeline.

The main contributions of our work include:

- We introduce a multimodal instruction interface that enables users to specify traffic scenarios through text descriptions, speech commands, and sketch-based images, enhancing the intuitiveness and expressiveness of scenario creation.
- We develop a retrieval-augmented code generation mechanism with a curated database of Scenic snippets that reduces hallucinations and ensures valid simulator specifications.
- We implement a human feedback guidance module that enables refinement and editing of generated scenarios, supporting various adjustments like entity behaviors, environmental conditions, and road layouts.

2. Related work

2.1. Traffic scenario generation

Traffic scenario generation for AVs has attracted widespread attention from the academic community. Researchers have developed various approaches to create diverse and realistic driving scenarios that can thoroughly evaluate AV performance. For instance, SceneGen [36] introduces a neural autoregressive model that sequentially generates the initial states of traffic agents to compose a static scene. TrafficGen [16] further decomposes the generation process into two stages: vehicle placement and trajectory generation. More recently, UniGen [24] proposes a unified model that generates agent positions, initial states, and trajectories from a shared scenario embedding, ensuring consistency between agent placement and motion. While these approaches demonstrated progress in generating diverse and realistic traffic scenarios, they typically suffer from limited controllability.

Recent advances in diffusion models present promising solutions to this challenge. CTG [46] leverages Signal Temporal Logic to guide diffusion models, effectively translating traffic rules into differentiable objectives that control the trajectory generation process. Safe-Sim [6] takes a different approach by combining adversarial guidance with partial diffusion, allowing fine-grained control over safety-critical scenarios and specific collision types. DragTraffic [40] further enhances controllability by guiding the diffusion process through cross-attention with user-specified context to refine initial trajectories into realistic traffic behavior.

Despite these advances, existing methods primarily require users to express their intentions through low-dimensional parametric inputs rather than high-level semantic descriptions, which limits users' ability to specify complex scenario requirements. In contrast, Talk2Traffic leverages MLLMs to enable users to express their design intentions through natural language, sketches, and voice commands, enhancing the intuitiveness of scenario generation.

2.2. MLLMs for autonomous driving

Recent breakthroughs in MLLMs have fueled significant advancements across various domains of autonomous driving research. In scene perception and understanding, models such as ELM [47], BEV-InMLLM [11], and TABot [48] have demonstrated impressive capabilities in interpreting driving environments. For planning and decision-making tasks, frameworks like DriveVLM [38], VLP [26], and LMDrive [31] leverage multimodal inputs to generate robust driving strategies. In reinforcement learning, VLM-RL [20] utilizes contrasting language goals as rewards while

CurricuVLM [35] enables personalized curriculum learning based on agent behavior analysis.

Language models have also shown promising potential in traffic scenario generation [37, 41]. CTG++ [45] is one of the pioneers in this direction, which transforms textual queries into differentiable loss functions that guide diffusion models to generate scenarios aligned with descriptions. TTSG [30] analyzes user text inputs for road retrieval and agent planning, supported by pre-constructed databases of road configurations and agent actions. ChatScene [43] parses safety-critical scenario descriptions and retrieves relevant Scenic code snippets from predefined libraries to compose executable scenarios. ScenicNL [14] combines LLMs with techniques like Tree-of-Thought to directly generate executable Scenic code [18] from crash reports.

While these approaches demonstrate significant progress, most existing works support only text-based inputs, overlooking the multimodal instructions that human designers naturally use, such as sketch-based images and voice commands. These additional modalities could enhance the intuitiveness and efficiency of scenario generation. Furthermore, compared to these recent approaches, our work introduces human feedback guidance as an essential module of the generation process. This not only mitigates potential LLM hallucination issues but also aligns with the dynamic adjustment of traffic scenario design.

3. Method

In this section, we present our framework Talk2Traffic, which leverages MLLMs to enable interactive and editable traffic scenario generation for AV testing. Fig. 2 illustrates the overall architecture, which consists of three main modules: (1) a multimodal instruction interpreter that processes diverse user inputs including text, speech, and sketches; (2) a retrieval-augmented code generation module that translates structured representations into executable Scenic code; and (3) a human feedback guidance mechanism that enables iterative editing of the generated scenarios.

3.1. Multimodal instruction interpreter

The multimodal instruction interpreter unifies and transforms intuitive human expressions across diverse input modalities into a standardized structured representation. This intermediate representation then serves as the foundation for subsequent code generation processes.

Multimodal input processing. Talk2Traffic accepts a variety of input modalities to accommodate different user preferences and scenario description needs. Users can provide natural language instructions through speech commands or text descriptions $l \in \mathcal{L}^{\leq k}$, where $\mathcal{L}$ represents the language vocabulary and k is the maximum length. For speech inputs, we employ an open-source speech-to-text model Whisper [29] to convert audio signals into textual format before processing. Additionally, users can supplement descriptions with sketch images or freehand drawings $s \in \mathcal{I}$, where $\mathcal{I}$ represents the 2D RGB image space.

This multimodal approach significantly enhances the expressiveness and intuitiveness of scenario specifications. While textual or speech descriptions effectively communicate behavioral attributes and temporal dynamics (e.g., "a vehicle suddenly brakes"), sketches and drawings excel at conveying spatial relationships and scene layouts (e.g., the relative positions of vehicles at an intersection). By combining these complementary modalities, users can more naturally express their intended scenarios.

Structured representation extraction. The core function of the interpreter is to transform multimodal inputs into a structured representation $\mathbf{z}$ that can be processed by subsequent modules of the framework:

$$\mathbf{z} = \text{MLLM}(p, l, s), \tag{1}$$

where $p \in \mathcal{L}^{\leq k}$ is the task description that guides the MLLM's interpretation process. It is worth noting that users could provide either a single modality input or combine multiple modalities based on their preferences and the complexity of the scenario they wish to describe.

The structured representation $\mathbf{z}$ consists of multiple components:

$$\mathbf{z} = [z^{\text{m}}, z^{\text{w}}, z^{\text{e}}_1, \ldots, z^{\text{e}}_n], \tag{2}$$

where $z^{\text{m}} \in \mathcal{L}^{\leq k}$ describes the map configuration, including the number of lanes, intersection configurations, traffic light presence, and other relevant infrastructure elements. The component $z^{\text{w}} \in \mathcal{L}^{\leq k}$ captures weather and temporal conditions (e.g., sunny, cloudy, rainy, daytime, night). The elements $z^{\text{e}}_i \in \mathcal{L}^{\leq k}$ for $i \in \{1, \ldots, n\}$ represent each entity in the scene.

Scene entities are categorized into two primary types: dynamic participants and static elements. Dynamic participants include vehicles, pedestrians, cyclists, and other traffic agents, which are characterized by both position and behavior attributes. Static elements comprise traffic cones, warning signs, and other stationary objects, which are primarily defined by their positions and physical properties.

To ensure consistency and facilitate downstream processing, we leverage GPT-4o to translate the user's multimodal instructions into a standardized YAML format. To enhance the quality and reliability of the structured output, we implement in-context learning techniques [3] when prompting GPT-4o. This approach provides the model with exemplars of high-quality YAML representations for several scenario types, effectively guiding it to produce structured outputs aligned with the user's desired scenarios.

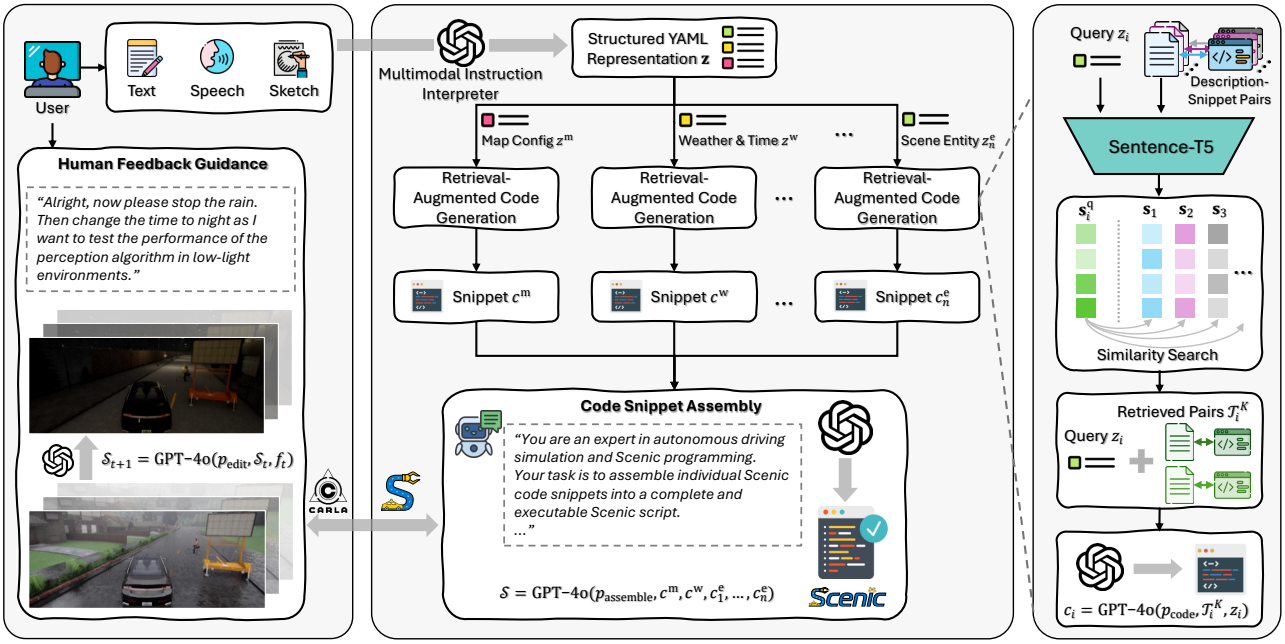

Figure 2. Overview of the Talk2Traffic framework. Our approach leverages MLLMs to enable interactive and editable traffic scenario generation through a unified pipeline of multimodal instruction interpretation, retrieval-augmented code generation, and human feedback.

3.2. Retrieval-augmented code generation with scenic snippets

Building upon the structured representation extracted by the multimodal instruction interpreter, we now focus on translating this representation into an executable simulation scenario. We leverage Scenic [18], a domain-specific probabilistic programming language designed for scenario specification in simulators like CARLA [12]. Scenic provides a natural syntax for describing traffic scenes, with built-in abstractions for spatial relationships and agent behaviors.

Despite MLLMs like GPT-4o demonstrate remarkable capabilities across various tasks, directly prompting them to generate Scenic code poses significant challenges. These models tend to produce hallucinations, resulting in syntax errors or references to non-existent APIs [43]. To address these limitations, we implement a retrieval-augmented generation (RAG) approach that grounds the code generation process in verified Scenic code snippets. Unlike ChatScene [43] which directly retrieves and applies code snippets from predefined libraries, our approach provides the MLLM with relevant examples for generating syntactically correct code. This distinction significantly enhances flexibility and adaptability, allowing our Talk2Traffic framework to handle more diverse and complex traffic scenarios.

Scenic code snippets database and retrieval. We construct a comprehensive database of Scenic code snippets to support our RAG approach. This database is compiled by adapting and extending examples from the official Scenic repository and code libraries provided by ChatScene [43]. These code snippets are carefully curated to align with the components of our structured representation $\mathbf{z}$.

For each code snippet in our database, we create a corresponding natural language description to form description-snippet pairs. These descriptions are crafted to capture the semantic meaning of each snippet in plain language, making them suitable for semantic matching with user instructions. Our database $\mathcal{D}$ can be formally represented as:

$$\mathcal{D} = \{(d_j, c_j) | j \in \{1, \ldots, m\}\}, \tag{3}$$

where $d_j \in \mathcal{L}^{\leq k}$ denotes the j-th natural language description and c_j represents the corresponding code snippet.

To retrieve the most relevant code snippets for a given query from $\mathbf{z}$, we encode each description using the Sentence-T5 model [25] to generate embedding vectors:

$$\mathbf{s}_j = \text{Sentence-T5}(d_j), \tag{4}$$

where $\mathbf{s}_j \in \mathbb{R}^{768}$ is the embedding vector. Similarly, we encode each component of the structured representation $\mathbf{z}$ to create query vectors:

$$\mathbf{s}_i^{\text{q}} = \text{Sentence-T5}(z_i), \tag{5}$$

where z_i represents one component of $\mathbf{z}$. We then compute the cosine similarity between each query vector and all description vectors in our database:

$$\text{sim}(\mathbf{s}_i^{\text{q}}, \mathbf{s}_j) = \frac{\mathbf{s}_i^{\text{q}} \cdot \mathbf{s}_j}{||\mathbf{s}_i^{\text{q}}|| \cdot ||\mathbf{s}_j||}, \tag{6}$$

where $||\cdot||$ denotes the Euclidean norm. For each component z_i, we select the top-K most similar description-snippet pairs $\mathcal{T}_i^K$ based on the computed similarity scores.

Code snippet generation and assembly. These selected pairs $\mathcal{T}_i^K$ serve as exemplars for in-context learning to guide the MLLM in generating appropriate Scenic code for each component in $\mathbf{z}$. We prompt GPT-4o with these examples and the target component description:

$$c_i = \text{GPT-4o}(p_{\text{code}}, \mathcal{T}_i^K, z_i), \tag{7}$$

where p_{code} is a prompt template instructing the model to generate Scenic code, and c_i is the generated code snippet.

Finally, we again leverage GPT-4o to assemble the individually generated code snippets into a complete Scenic script:

$$\mathcal{S} = \text{GPT-4o}(p_{\text{assemble}}, c^{\text{m}}, c^{\text{w}}, c_1^{\text{e}}, \ldots, c_n^{\text{e}}), \tag{8}$$

where c^{m}, c^{w}, and c_i^{e} are the generated code snippets for the map, weather, and agent components, respectively, and $\mathcal{S}$ represents the complete Scenic script. This assembly process, guided by a specialized prompt p_{assemble}, ensures consistent variable declarations and references, as well as overall syntactic correctness.

3.3. Human feedback guidance

Despite recent advances in traffic scenario generation, most existing approaches fundamentally lack interactive editing capabilities, limiting their utility in AV testing workflows. AV testing is inherently iterative, requiring human designers to continuously refine scenarios with carefully calibrated difficulty levels and specific edge cases that challenge system limitations [35]. Additionally, when a generated scenario does not perfectly encapsulate the intended test case, designers need a way to provide feedback on specific elements requiring adjustment rather than regenerating entire scenarios from scratch.

To address these challenges, we implement a human feedback guidance mechanism that leverages the multi-turn dialogue capabilities of MLLMs. This interactive approach allows users to edit and refine scenarios through natural language feedback while maintaining the context of previous interactions. The feedback process can be generally formulated as:

$$\mathcal{S}_{t+1} = \text{GPT-4o}(p_{\text{edit}}, \mathcal{S}_t, f_t), \tag{9}$$

where $\mathcal{S}_t$ represents the scenario script at iteration t, f_t denotes the user feedback at iteration t, and p_{edit} is a prompt template guiding the editing and refinement process.

The human feedback mechanism supports various types of scenario adjustments. For instance, since Scenic is a probabilistic programming language, users can modify parameter distributions to introduce controlled variability. When a user provides feedback such as "vehicles should maintain greater following distances but with less variability," our Talk2Traffic adjusts the relevant probability distributions in the Scenic code to reflect this requirement. Similarly, users can refine scene entity behaviors (e.g., "make the pedestrian cross more hesitantly"), environmental conditions (e.g., "increase rainfall intensity to test perception performance"), or road layouts (e.g., "change the three-lane road to two lanes").

4. Experiments

4.1. Setup

We evaluate Talk2Traffic using the CARLA simulator [12], which provides a high-fidelity driving environment with realistic rendering and diverse road layouts. To support our RAG approach, we construct a comprehensive database of description-snippet pairs, which contains 283 pairs categorized into three main components:

- **Map Configuration**: 63 pairs covering various road layouts, intersection types, and infrastructure elements like traffic lights.
- **Weather and Time**: 32 pairs describing different weather conditions (e.g., clear, rainy, cloudy, foggy), and time of day settings that affect visibility and driving conditions.
- **Entity Specification**: 188 pairs detailing various traffic participants and their behaviors, including vehicles (passenger cars, trucks, motorcycles), pedestrians, cyclists, and static objects (traffic cones, barriers).

During the retrieval process, we select the top-$K = 3$ most similar description-snippet pairs to guide the code generation process.

4.2. Generation quality evaluation

To quantitatively assess the quality of traffic scenarios generated by Talk2Traffic, we follow the evaluation protocols established in ChatScene [43] and SafeBench [42] for safety-critical scenario generation. This allows us to compare our framework against existing methods while maintaining consistency with established benchmarks.

Evaluation metrics. We employ a comprehensive set of metrics to evaluate the generated scenarios across three key dimensions: safety, functionality, and etiquette. **Safety metrics** include collision rate (CR), red light running frequency (RR), stop sign violation frequency (SS), and out-of-road average distance (OR). Higher values indicate more challenging scenarios. **Functionality metrics** consist of route following stability (RF), route completion (Comp), and time spent (TS). Lower values for RF and Comp and higher values for TS indicate scenarios that better challenge the AV's ability. **Etiquette metrics** cover average acceleration (ACC), average yaw velocity (YV), and lane invasion

Metric	Model	Scenario Type								Avg.
		Straight Obstacle	Turning Obstacle	Lane Change	Vehicle Passing	Red Light Running	Unprotected Left Turn	Rigth Turn	Crossing Negotiation	
CR ↑	LC	0.223	0.088	0.710	0.807	0.317	0.403	0.350	0.273	0.396
	AS	0.470	0.350	0.703	0.833	0.497	0.647	0.637	0.607	0.593
	AT	0.343	0.217	0.667	0.830	0.623	0.533	0.357	0.393	0.495
	CS	0.893	0.697	**0.950**	0.927	0.787	0.753	0.777	0.863	0.831
	T2T	**0.913**	**0.780**	0.893	**0.947**	**0.900**	**0.833**	**0.860**	**0.893**	**0.877**
OS ↓	LC	0.809	0.849	0.566	0.508	0.804	0.753	0.681	0.748	0.715
	AS	0.697	0.708	0.563	0.504	0.711	0.628	0.540	0.575	0.616
	AT	0.747	0.776	0.583	0.507	0.643	0.684	0.680	0.684	0.663
	CS	**0.470**	0.522	**0.434**	0.440	0.537	0.560	0.474	**0.421**	0.482
	T2T	0.475	**0.481**	0.461	**0.437**	**0.496**	**0.539**	**0.449**	0.422	**0.471**

Table 1. Performance comparison of Talk2Traffic (T2T) against baseline methods in safety-critical scenario generation. Best results are highlighted in **bold**.

frequency (LI). Higher values indicate scenarios that better test an AV's ability to maintain driving comfort and adherence to traffic norms. Additionally, we compute an overall score (OS) that aggregates performance across all metrics, with lower values indicating more challenging scenarios.

Baselines. For our evaluation, we selected four representative baseline methods: (1) **Learning-to-Collide (LC)** [9]: a reinforcement learning-based approach that trains adversarial agents to create collision scenarios. (2) **AdvSim (AS)** [39]: a trajectory perturbation-based method that employs black-box search algorithms to generate adversarial scenarios. (3) **Adversarial Trajectory Optimization (AT)** [44]: an approach that uses trajectory optimization to identify failure cases. (4) **ChatScene (CS)** [43]: a state-of-the-art LLM-based approach for safety-critical scenario generation.

Experimental protocol. Our experiments span eight common yet challenging scenario types: Straight Obstacle, Turning Obstacle, Lane Change, Vehicle Passing, Red Light Running, Unprotected Left Turn, Right Turn, and Crossing Negotiation. Each scenario type contains approximately 100 challenging configurations. Detailed specifications of these scenario types can be referred to ChatScene [43].

For evaluation, we first employ a surrogate ego vehicle trained by SafeBench using SAC. Then, different scenario generation methods create challenging scenarios that test the limits of this surrogate. Next, these generated scenarios are used to evaluate three different ego vehicles trained by SAC, PPO, and TD3. We report their average performance for all metrics.

For Talk2Traffic, we use the same prompts provided by ChatScene to generate scenarios for each category. However, unlike ChatScene, which relies on sampling 50 different parameter configurations to identify challenging scenarios, our approach leverages human feedback guidance to iteratively refine scenarios based on testing objectives.

Results and analysis. Tab. 1 presents the comparison of our Talk2Traffic framework (T2T) against baseline methods across different scenario types, focusing on the collision rate (CR) and overall score (OS). The results demonstrate that Talk2Traffic outperforms all baseline methods across most scenario types.

In terms of collision rate, Talk2Traffic achieves the highest average CR, surpassing the second best method (ChatScene) by 4.6%. This indicates that our framework generates scenarios that more effectively challenge AVs' collision avoidance capabilities. Notably, Talk2Traffic shows substantial improvements in complex scenarios such as Red Light Running (CR of 0.900 vs. ChatScene's 0.787) and Unprotected Left Turn (CR of 0.833 vs. ChatScene's 0.753). For overall score, Talk2Traffic achieves the lowest (best) average score, slightly outperforming ChatScene. This suggests that our framework generates scenarios with comprehensive complexity that challenge AVs across multiple dimensions simultaneously.

Tab. 2 provides a more detailed breakdown of performance across all evaluation metrics. The results reveal several interesting insights. Talk2Traffic excels in safety-related challenges, achieving the highest CR among all methods. In terms of functionality, Talk2Traffic achieves the lowest route completion and highest time spent driving, indicating that our generated scenarios effectively challenge the vehicle's ability to maintain functionality while navigating difficult conditions. For etiquette metrics, ChatScene generally outperforms Talk2Traffic. This difference highlights a trade-off in scenario generation approaches: while

Model	Safety Level				Functionality Level			Etiquette Level			OS ↓
	CR ↑	RR ↑	SS ↑	OR ↑	RF ↓	Comp ↓	TS ↑	ACC ↑	YV ↑	LI ↑	
LC	0.396	0.316	**0.150**	0.045	0.883	0.809	0.255	0.228	0.228	0.090	0.715
AS	0.593	0.301	0.148	0.041	0.884	0.742	0.255	0.242	0.226	0.094	0.616
AT	0.495	**0.315**	**0.150**	**0.052**	0.884	0.769	0.238	0.249	0.233	0.103	0.663
CS	0.831	0.179	0.143	0.035	**0.833**	0.544	0.223	**0.705**	**0.532**	**0.243**	0.482
T2T	**0.877**	0.226	0.148	0.013	0.895	**0.519**	**0.284**	0.322	0.262	0.060	**0.471**

Table 2. Comprehensive evaluation of Talk2Traffic (T2T) against baseline methods in safety-critical scenario generation, considering safety, functionality, and etiquette dimensions. Best results are highlighted in **bold**.

Talk2Traffic prioritizes safety challenges and functional difficulties, ChatScene's scenarios tend to induce more aggressive driving patterns. Both approaches effectively test AVs but emphasize different aspects of driving performance. Overall, these results demonstrate the effectiveness of the human feedback guidance module in refining scenarios to target specific challenge aspects.

4.3. Ablation study

To systematically assess the contribution of each component within the retrieval-augmented code generation module of our Talk2Traffic framework, we conduct a comprehensive ablation study focused on the code execution success rate. This metric is defined as the percentage of generated Scenic scripts that could be executed without errors.

We construct four variants of our framework to isolate the effect of each component:

- **Baseline**: Direct prompting of GPT-4o to generate Scenic code from user instructions without any structured representation, RAG, or code assembly.
- **SR**: Employing structured representation extraction but using direct prompting for code snippet generation, followed by simple concatenation of generated snippets.
- **SR+RAG**: Using structured representation and retrieval-augmented code generation, with generated code snippets directly concatenated without intelligent assembly.
- **SR+RAG+CA**: Our complete Talk2Traffic framework where GPT-4o intelligently assembles the code snippets, ensuring consistency between variable declarations and references, and overall syntactic correctness.

We provide 100 diverse natural language descriptions to each variant to generate corresponding traffic scenarios. The scenario descriptions cover various traffic conditions, including different road types, weather conditions, and multi-agent interactions.

The results in Tab. 3 reveal a clear progression in code execution success rates as we incorporate each component. The baseline configuration, which directly prompts GPT-4o without any supporting mechanisms, only achieves a 15% success rate, highlighting the inherent challenges of generating domain-specific code from language descriptions.

Structured Representation	RAG	Code Assembly	Code Execution Success Rate
✗	✗	✗	0.15
✓	✗	✗	0.41
✓	✓	✗	0.80
✓	✓	✓	**0.89**

Table 3. Ablation study results evaluating the contribution of components in the Talk2Traffic framework.

When we introduce structured representation extraction, the success rate increases to 41%, demonstrating how breaking down complex scenarios into sub-components enhances the generation process. Further improvement comes with the integration of the RAG approach, which raises the success rate to 80% by grounding code generation in verified Scenic snippets. Despite this substantial gain, both the SR and SR+RAG suffer from similar issues when code snippets are simply concatenated, resulting in naming conflicts and import inconsistencies. Our complete framework achieves an 89% success rate, reflecting the importance of the intelligent code assembly process. The 11% of scenarios that still fail even with our complete framework primarily involve complex multi-agent interactions or occasional hallucinations, which highlights directions for future improvement.

4.4. Scenario visualization

In this subsection, we present qualitative results that demonstrate Talk2Traffic's ability to interpret multimodal inputs and enable interactive scenario editing.

Multimodal input. Fig. 3 demonstrates how Talk2Traffic effectively processes different input modalities to generate realistic traffic scenarios in CARLA.

Fig. 3(a) illustrates how a simple text description specifying an intersection with traffic lights where cars move in an orderly manner according to changing signals is interpreted by our framework. After analyzing this description, Talk2Traffic successfully generates multiple vehicles

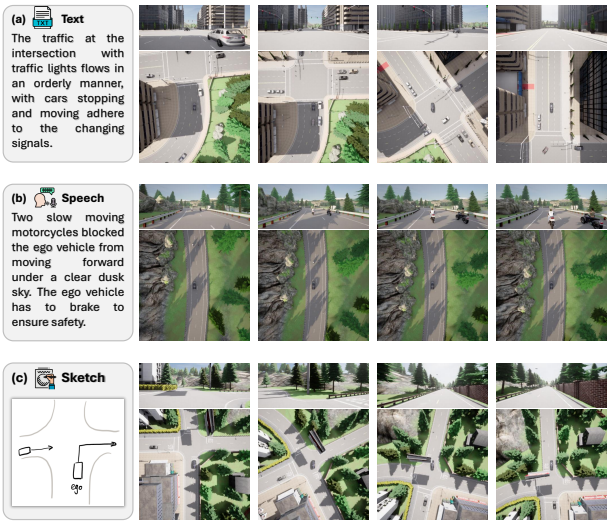

Figure 3. Qualitative examples. Each scenario is downsampled to four frames for visualisation.

at the intersection and configures them to operate in autopilot mode, correctly responding to the traffic signals.

Fig. 3(b) showcases a scenario generated from a speech command describing two slow-moving motorcycles blocking the ego vehicle from moving forward under a clear dusk sky, requiring the ego vehicle to brake for safety. The simulation sequence clearly aligns with this requirement. Notably, the framework correctly modifies CARLA's time setting to sunset, accurately reflecting the environmental condition specified in the speech input.

Fig. 3(c) illustrates how a simple hand-drawn sketch can be transformed into a realistic traffic scenario. Despite the sketch's simplicity, the framework successfully interprets the spatial layout and the relative positioning and behavior of vehicles, producing a scenario where the ego vehicle approaches the intersection to make a right turn while another vehicle travels straight.

Human feedback for scenario editing. Fig. 4 demonstrates Talk2Traffic's capability to edit scenarios through natural language feedback, showcasing the framework's adaptability to evolving test requirements without requiring users to modify code directly.

Fig. 4(a) illustrates an editing of a previously generated scenario from Fig. 3(b). Upon receiving the instruction, the framework reconfigures the original two-lane road into a three-lane one. It also reprograms the ego's behavioral logic to execute lane changes and overtaking maneuvers when encountering slower agents.

Fig. 4(b) exhibits the framework's ability to convert regular driving situations from Fig. 3(c) into safety-critical test cases. When provided with the feedback "Increase

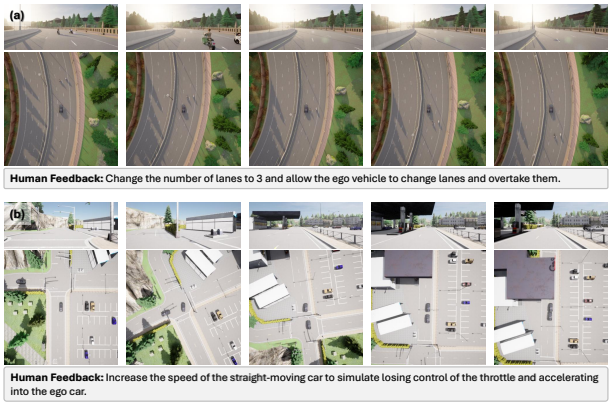

Figure 4. Editing result under human feedback commands. Each scenario is downsampled to five frames for visualisation.

the speed of the straight-moving car to simulate losing control of the throttle and accelerating into the ego car," Talk2Traffic correctly interprets this request, modifying the approaching vehicle's velocity profile to collide with the ego vehicle.

5. Conclusions

This paper introduces Talk2Traffic, a novel framework leveraging multimodal large language models to enable interactive and editable traffic scenario generation for autonomous driving. By integrating a multimodal instruction interface, retrieval-augmented Scenic code generation, and human feedback guidance, our approach bridges the gap between intuitive human expressions and executable simulation code. Experimental results demonstrate Talk2Traffic's superior performance over existing methods. The framework's ability to accurately interpret multimodal inputs and implement human-guided modifications facilitates the efficient creation of diverse test scenarios without requiring specialized programming knowledge. Future work will focus on expanding simulator compatibility and enhancing semantic reasoning capabilities.

Acknowledgment

This work was supported by the University of Wisconsin-Madison's Center for Connected and Automated Transportation (CCAT), a part of the larger CCAT consortium, a USDOT Region 5 University Transportation Center funded by the U.S. Department of Transportation, Award #69A3552348305. The contents of this paper reflect the views of the authors, who are responsible for the facts and the accuracy of the data presented herein, and do not necessarily reflect the official views or policies of the sponsoring organization.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2

[3] Tom Brown, Benjamin Mann, Nick Ryder, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020. 3

[4] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11621–11631, 2020. 1

[5] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, et al. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8748–8757, 2019. 1

[6] Wei-Jer Chang, Francesco Pittaluga, Masayoshi Tomizuka, Wei Zhan, and Manmohan Chandraker. Safe-sim: Safety-critical closed-loop traffic simulation with diffusion-controllable adversaries. In *European Conference on Computer Vision (ECCV)*, pages 242–258, 2024. 2

[7] Scenario Runner Contributors. Carla ScenarioRunner. https://github.com/carla-simulator/scenario_runner, 2019. 1

[8] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, et al. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 958–979, 2024. 2

[9] Wenhao Ding, Baiming Chen, Minjun Xu, and Ding Zhao. Learning to collide: An adaptive safety-critical scenarios generating method. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2243–2250, 2020. 6

[10] Wenhao Ding, Chejian Xu, Mansur Arief, Haohong Lin, Bo Li, and Ding Zhao. A survey on safety-critical driving scenario generation—a methodological perspective. *IEEE Transactions on Intelligent Transportation Systems*, 24(7): 6971–6988, 2023. 1

[11] Xinpeng Ding, Jianhua Han, Hang Xu, Xiaodan Liang, Wei Zhang, and Xiaomeng Li. Holistic autonomous driving understanding by bird's-eye-view injected multi-modal large models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13668–13677, 2024. 2

[12] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on Robot Learning (CoRL)*, pages 1–16, 2017. 4, 5

[13] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 2

[14] Karim Elmaaroufi, Devan Shanker, Ana Cismaru, Marcell Vazquez-Chanlatte, Alberto Sangiovanni-Vincentelli, Matei Zaharia, and Sanjit A. Seshia. ScenicNL: Generating probabilistic scenario programs from natural language. In *First Conference on Language Modeling (COLM)*, 2024. 3

[15] Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, et al. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9710–9719, 2021. 1

[16] Lan Feng, Quanyi Li, Zhenghao Peng, Shuhan Tan, and Bolei Zhou. Trafficgen: Learning to generate diverse and realistic traffic scenarios. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 3567–3575, 2023. 2

[17] Shuo Feng, Haowei Sun, Xintao Yan, Haojie Zhu, Zhengxia Zou, Shengyin Shen, and Henry X Liu. Dense reinforcement learning for safety validation of autonomous vehicles. *Nature*, 615(7953):620–627, 2023. 1

[18] Daniel J Fremont, Edward Kim, Tommaso Dreossi, Shromona Ghosh, Xiangyu Yue, Alberto L Sangiovanni-Vincentelli, and Sanjit A Seshia. Scenic: A language for scenario specification and data generation. *Machine Learning*, 112(10):3805–3849, 2023. 3, 4

[19] Zilin Huang, Zihao Sheng, Chengyuan Ma, and Sikai Chen. Human as ai mentor: Enhanced human-in-the-loop reinforcement learning for safe and efficient autonomous driving. *Communications in Transportation Research*, 2024. 1

[20] Zilin Huang, Zihao Sheng, Yansong Qu, Junwei You, and Sikai Chen. Vlm-rl: A unified vision language models and reinforcement learning framework for safe autonomous driving. *arXiv preprint arXiv:2412.15544*, 2024. 2

[21] Max Jiang, Yijing Bai, Andre Cornman, Christopher Davis, et al. Scenediffuser: Efficient and controllable driving simulation initialization and rollout. *Advances in Neural Information Processing Systems*, 37:55729–55760, 2024. 1

[22] Keke Long, Zihao Sheng, Haotian Shi, Xiaopeng Li, Sikai Chen, and Sue Ahn. A physics enhanced residual learning (perl) framework for vehicle trajectory prediction. *arXiv preprint arXiv:2309.15284*, 2024. 1

[23] Renze Lou, Kai Zhang, and Wenpeng Yin. Large language model instruction following: A survey of progresses and challenges. *Computational Linguistics*, 50(3):1053–1095, 2024. 2

[24] Reza Mahjourian, Rongbing Mu, Valerii Likhosherstov, Paul Mougin, Xiukun Huang, Joao Messias, and Shimon Whiteson. Unigen: Unified modeling of initial agent states and trajectories for generating autonomous driving scenarios. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 16367–16373, 2024. 2

[25] Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith B Hall, Daniel Cer, and Yinfei Yang. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1864–1874, 2022. 4

[26] Chenbin Pan, Burhaneddin Yaman, Tommaso Nesti, Abhirup Mallik, Alessandro G Allievi, Senem Velipasalar, and Liu

Ren. Vlp: Vision language planning for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14760–14769, 2024. 2

[27] Ethan Pronovost, Meghana Reddy Ganesina, Noureldin Hendy, Zeyu Wang, Andres Morales, Kai Wang, and Nick Roy. Scenario diffusion: Controllable driving scenario generation with diffusion. *Advances in Neural Information Processing Systems*, 36:68873–68894, 2023. 1

[28] Yansong Qu, Zixuan Xu, Zilin Huang, Zihao Sheng, Tiantian Chen, and Sikai Chen. Metassc: Enhancing 3d semantic scene completion for autonomous driving through meta-learning and long-sequence modeling. *arXiv preprint arXiv:2411.03672*, 2025. 1

[29] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning (ICML)*, pages 28492–28518, 2023. 3

[30] Bo-Kai Ruan, Hao-Tang Tsui, Yung-Hui Li, and Hong-Han Shuai. Traffic scene generation from natural language description for autonomous vehicles with large language model. *arXiv preprint arXiv:2409.09575*, 2024. 3

[31] Hao Shao, Yuxuan Hu, Letian Wang, Guanglu Song, Steven L Waslander, Yu Liu, and Hongsheng Li. Lmdrive: Closed-loop end-to-end driving with large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15120–15130, 2024. 2

[32] Zihao Sheng, Zilin Huang, and Sikai Chen. Ego-planning-guided multi-graph convolutional network for heterogeneous agent trajectory prediction. *Computer-Aided Civil and Infrastructure Engineering*, 39(22):3357–3374, 2024. 1

[33] Zihao Sheng, Zilin Huang, and Sikai Chen. Kinematics-aware multigraph attention network with residual learning for heterogeneous trajectory prediction. *Journal of Intelligent and Connected Vehicles*, 7(2):138–150, 2024. 1

[34] Zihao Sheng, Zilin Huang, and Sikai Chen. Traffic expertise meets residual rl: Knowledge-informed model-based residual reinforcement learning for cav trajectory control. *Communications in Transportation Research*, 4:100142, 2024. 1

[35] Zihao Sheng, Zilin Huang, Yansong Qu, Yue Leng, Sruthi Bhavanam, and Sikai Chen. Curricuvlm: Towards safe autonomous driving via personalized safety-critical curriculum learning with vision-language models. *arXiv preprint arXiv:2502.15119*, 2025. 2, 3, 5

[36] Shuhan Tan, Kelvin Wong, Shenlong Wang, Sivabalan Manivasagam, Mengye Ren, and Raquel Urtasun. Scenegen: Learning to generate realistic traffic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 892–901, 2021. 2

[37] Shuhan Tan, Boris Ivanovic, Xinshuo Weng, Marco Pavone, and Philipp Kraehenbuehl. Language conditioned traffic generation. In *Conference on Robot Learning (CoRL)*, 2023. 3

[38] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. In *Conference on Robot Learning (CoRL)*, 2024. 2

[39] Jingkang Wang, Ava Pun, James Tu, Sivabalan Manivasagam, Abbas Sadat, Sergio Casas, Mengye Ren, and Raquel Urtasun. Advsim: Generating safety-critical scenarios for self-driving vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9909–9918, 2021. 6

[40] Sheng Wang, Ge Sun, Fulong Ma, Tianshuai Hu, Qiang Qin, Yongkang Song, Lei Zhu, and Junwei Liang. Dragtraffic: Interactive and controllable traffic scene generation for autonomous driving. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 14241–14247, 2024. 2

[41] Yuxi Wei, Zi Wang, Yifan Lu, Chenxin Xu, Changxing Liu, Hao Zhao, Siheng Chen, and Yanfeng Wang. Editable scene simulation for autonomous driving via collaborative llm-agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15077–15087, 2024. 3

[42] Chejian Xu, Wenhao Ding, Weijie Lyu, Zuxin Liu, Shuai Wang, Yihan He, Hanjiang Hu, Ding Zhao, and Bo Li. Safebench: A benchmarking platform for safety evaluation of autonomous vehicles. *Advances in Neural Information Processing Systems*, 35:25667–25682, 2022. 5

[43] Jiawei Zhang, Chejian Xu, and Bo Li. Chatscene: Knowledge-enabled safety-critical scenario generation for autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15459–15469, 2024. 3, 4, 5, 6

[44] Qingzhao Zhang, Shengtuo Hu, Jiachen Sun, Qi Alfred Chen, and Z Morley Mao. On adversarial robustness of trajectory prediction for autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15159–15168, 2022. 6

[45] Ziyuan Zhong, Davis Rempe, Yuxiao Chen, Boris Ivanovic, Yulong Cao, Danfei Xu, Marco Pavone, and Baishakhi Ray. Language-guided traffic simulation via scene-level diffusion. In *Conference on Robot Learning (CoRL)*, pages 144–177, 2023. 3

[46] Ziyuan Zhong, Davis Rempe, Danfei Xu, Yuxiao Chen, Sushant Veer, Tong Che, Baishakhi Ray, and Marco Pavone. Guided conditional diffusion for controllable traffic simulation. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 3560–3566, 2023. 2

[47] Yunsong Zhou, Linyan Huang, Qingwen Bu, Jia Zeng, Tianyu Li, Hang Qiu, Hongzi Zhu, Minyi Guo, Yu Qiao, and Hongyang Li. Embodied understanding of driving scenarios. In *European Conference on Computer Vision (ECCV)*, pages 129–148, 2024. 2

[48] Yixuan Zhou, Long Bai, Sijia Cai, Bing Deng, Xing Xu, and Heng Tao Shen. TAU-106k: A new dataset for comprehensive understanding of traffic accident. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. 2