Language-Guided Generation of Specific Driving Scenarios Using Diffusion Models

Jiongqi Li

{ jl2136@hw.ac.uk }

Department of Electrical, Electronic and Computer Engineering, Heriot-Watt University, Edinburgh, United Kingdom

Abstract. Scenario-based simulation has now become essential for validating autonomous driving vehicles, and real world trials alone cannot enable the validation of rare but critical-to-safety situations. However, simulation-based validation is effective when the simulated scenarios that are generated are realistic, diverse and controllable. Simulated scenarios from scripted methods are very controllable but have poor scalability; fully data-driven methods fail to be semantic in nature. Recent advances that combine natural language interfaces with diffusion-based generative models give a promising and complementary approach. The paper reviews representative methods at this intersection, discusses their system designs and evaluation schemes, and also the open challenges pertaining to controllability, executability and reproducibility and particularly focuses on scene-level conditional diffusion models with language. This review summarizes the current research trends and provides guidelines for achieving more controllable, executable and repeatable scenario-generation pipelines in the future autonomous driving validation framework.

Keywords: Autonomous driving, Scenario generation, Diffusion models, Large language models, Simulation-based validation

1 Introduction

As autonomous driving systems begin to enter mass deployment, systems and repeatable validation under corner case conditions is becoming critical. Despite the necessity of on-road testing, it is also well accepted that the real world driving alone is not feasible enough to ensure that the long tail of rare yet safety-critical events can be uncovered within a reasonable time and cost. For this reason, simulation-based scenario testing has emerged as a cornerstone of autonomous driving validation pipelines [1]. By enabling controlled manipulation of actors, environments, and behaviors, scenario-based simulation allows developers to stress-test perception, planning, and control modules under conditions that may only occur infrequently in the real world [2].

Nevertheless, the practical value of scenario testing then is critically dependent on properties of the scenarios that are generated. Effective scenarios should have realistic agent behaviour,

ICIAAI 2026, May 22-24, Singapore, Singapore

DOI 10.4108/eai.22-5-2026.2365115

sufficient diversity so as to capture a wide spectrum of operating states and conditions, and high controllability such that engineers have fine grained control of testing intent. Traditional scenario generation tools work largely based on scripted, hand-made scripts, expert rules, or parameterized templates and are expressed in standard description languages such as ASAM OpenSCENARIO [3]. Although such methods are transparent and directly executable in simulators, they entail considerable manual effort and do not scale with scenario complexity and richer interactivity [4].

There are also data-driven generation methods which learn driving behavior from large-scale naturalistic driving datasets such as Waymo open motion dataset and nuScenes [5,6], which provide enhanced realism. Early work in multi-agent modelling describes the interactions between different traffic participants using probabilistic or graph-based neural networks and obtaining socially meaningful distributions over trajectories [7]. More recently, diffusion probabilistic models have been proposed for multi-agent motion generation and prediction, showing great performance on modelling highly dimensional, multimodal and temporally coherent behaviours, respectively [8,9]. These models are particularly attractive for the purpose of scenario generation because they are naturally suited to stochastic sampling and can be biased by given auxiliary conditions during inference.

Large language models have also recently received attention as a way of expressing scenario intent in a natural and flexible form. The natural language allows the tester to freely describe complex situations, such as conflicts, priorities or unplanned behaviour, without writing low-level scenario code directly. Several of the latest systems use LLMs to convert free-form text descriptions into scenarios in a structured way or into simulator specifications that can be run [10], whereas other systems use LLMs to obtain scenario seeds from the text sources such as in accident reports or data set annotations [11].

Combining language-based intent specification and diffusion-based trajectory synthesis generated a new class of hybrid scenario generation pipelines. The language model runs at the semantic level and infers intent or constraints, while the diffusion model generates physically and socially plausible multi-agent scenes that satisfy such constraints. Scene-level conditional diffusion models illustrate such coupling between language model and sampling process by projecting language-derived conditions into guidance signals for the probabilistic sampler, resulting in flexible control yet diversity preservation [1].

Overall, this paper reviews and summarizes recent work in language-guided and diffusion-based driving scenario generation. The review makes three important contributions. Firstly, the paper places language-guided diffusion methods in the landscape of scenario generation approaches. Secondly, the paper analyses representative systems and their design choices, shedding light on common architectural patterns. Thirdly, the paper analyses evaluation practices and identifies gaps that stifle reproducibility and fair comparisons, offering recommendations for future work.

2 Methodological Categories

Approaches to scenario generation for autonomous driving currently fall into broad categories on the basis of how scenarios are specified and synthesised.

2.1 Rule-Based and Scripted Scenario Generation

Rule-based and scripted methods are still very popular for industry testing workflows, for the interpretability, ease of use, and conformity to standard formats such as the ASAM OpenSCENARIO [3]. These approaches enable precise control over actor behaviors and event timing, making them suitable for regression testing and safety case construction. These approaches provide precise control over the actors' behaviour and event timing, and are suited to regression testing and construction of safety cases. However, handcrafted scripts and parameterized templates require painstaking and manual effort and scale poorly as scenario complexity and richness of interaction grow [4].

2.2 Search- and Optimization-Based Methods

Search-based and optimization-based techniques formalize scenario generation as an adversarial or exploratory search process. By setting criticality metrics or objectives for failure, they use evolutionary algorithms or heuristic search to search for scenarios exposing a potential weakness of a particular autonomous driving system [2]. They are effective in uncovering system-level failures but depend heavily on the objective design and on computational budget and are often strongly coupled to a particular ADS implementation.

2.3 Data-Driven Generative Models

Data-driven approaches aim to learn distributions of traffic behavior directly from large-scale naturalistic driving datasets such as the Waymo Open Motion Dataset and nuScenes [5, 6]. Early multi-agent models, such as graph based probabilistic models, including Trajectron [7], capture patterns of interactions amongst traffic participants and provide socially intuitive trajectory distributions. More recently, diffusion probabilistic models have successfully modeled high-dimensional, multimodal, and temporally coherent multi-agent motion [8,9]. These models are especially useful for generating scenarios because they support both stochastic sampling and conditional guidance naturally.

2.4 Language-Guided and Hybrid Pipelines

Large language models have recently been introduced as flexible APIs for expressing scenario intent in natural language. Systems such as ChatScene that use LLMs to convert free-form descriptions to structured or simulator-executable scenarios [10], and extraction-based methods that reconstruct a scenario from accident reports or textual annotations [11].

The integration of language-based intent specification and diffusion-based trajectory synthesis has given rise to a new class of hybrid scenario generation pipelines, in which language models operate at the semantic level and diffusion models generate physically and socially plausible multi-agent scenes that satisfy inferred constraints [1]. These constraints are then encoded as guidance terms that influence the denoising of a diffusion model, biasing samples towards a story that satisfies the given intent while maintaining diversity [1].

3 Representative Approaches

Scene-level language-guided diffusion models provide a clear illustration of the hybrid paradigm. In these frameworks, a user's textual query is first parsed by a large language model to identify semantic constraints or objectives. These constraints are then encoded as guidance terms that influence the denoising process of a diffusion model, biasing samples toward scenarios that satisfy the specified intent while preserving diversity [1]. Together, this approach exploits the differentiability of diffusion guidance, to provide adaptable control, without collapsing the generative distribution.

Another line of work stresses guaranteed conversion from language to executable scenarios. Systems like ChatScene aim to produce safety critical scenario descriptions by assimilating knowledge and templates based on LLM reasoning, end up being simulator-ready descriptions [10]. Chat2Scenario tackles a related problem. It extracts structured scenarios from dataset text and models them in a suitable format for OpenSCENARIO. These methods have a high executability and integration with popular simulation platforms. These methods prioritize executability and integration with existing simulation platforms.

Extraction based approaches complement generation by rooting scenarios in real-world events. SoVAR showed how accident reports can be parsed and reconstructed into parameterised situations that can be passed from maps and simulators, making realistic and re-usable test cases [11]. By anchoring generation to known failures, such systems provide a useful seed for additional stochastic augmentation.

Across these representative systems, there is a design principle that is common. Semantic intent is separated from motion realization, with language models operating on expressive but ambiguous inputs and diffusion models producing concrete trajectories that satisfy physical and social constraints. Results of some of our empirical studies reported above show that such hybrid systems are able to produce a higher percentage of semantically related and safety-critical scenarios than those derived from templates or data sampling alone [1,10].

4 Evaluation and Reproducibility

Amid evaluation practices in the literature are so varied that it is very difficult to compare directly different methods. Some approaches to realism evaluation use distributional similarity measures computed on the trajectory features, for example Wasserstein distance or KL divergence [5-7]. These metrics provide how closely generated trajectories fit the statistical properties of real data, thus measuring the plausibility globally. But they can miss important safety events or interactions that are rare but still relevant.

Other studies focus on conditional compliance rates, measured by rule-based validators that check whether generated scenarios satisfy predefined constraints, such as lane adherence, speed limits, or traffic rules [3,4]. It directly judges for the semantic correctness and safety-critical compliance, however, it is too rigid and rule specification sensitive, and may penalize the correct but unconventional behaviors.

Another set evaluates downstream utility by having generated scenarios run against an autonomous driving system and reporting on performance measures such as failure rate, near miss counts, or number of unsafe events [2]. This evaluation gives practical information on system robustness, but is strongly based on the choice of simulator, map, and vehicle dynamics model and hence not reproducible on other platforms.

Executability also differentiates evaluation pipelines. Taking generated trajectories and translating them into formats needed by simulators requires handling timing, agent priorities and dynamic feasibility. Methods that achieve a common encoding standard such as OpenSCENARIO lead to repeatability and industry adoption [3,4], but many research prototypes end up with a workflow, but not a fully executable one.

Clearly, a full assessment should jointly assess realism, semantic soundness, scenario importance and diversity with common datasets, common canonical maps, and common encodings [1,5]. Each type of metric has trade-offs: distributional metrics measure general realism but miss the rare events; rule-based validators measure semantic validity but lack flexibility; ADS-in-the-loop testing means of practical impact, but simulator-compatibility issues may arise. Recognizing these strengths and limitations can inform robust and interpretable evaluation of scenario generation approaches.

5 Discussion

5.1 Strengths and Limitations

Advantages of language-guided diffusion-based scenario generation include: natural language allows effortless specification of complex testing intent, reducing the barrier of scenario authoring. Diffusion models output temporally coherent and multimodal joint behaviours that better capture interactions than many handcrafted scripts [1,5,6]. Together, these elements support generation of diverse but targeted scenarios that are usable for stress-testing.

However, there are also important challenges ahead. Ambiguities of language and LLM hallucination can give rise to unintended interpretations of scenarios, even when retrieval or schema constraints are considered [10]. Conditioning diffusion models to respect high-level constraints also leads to lower quality samples or higher computational costs [9]. Fragmented benchmarks and often counter-productive evaluation measures inhibit reproducibility, and the absence of explicit map and dynamics constraints can produce unrealistic trajectories [5,6]. It should be mentioned finally that the computational burden associated with combining large language models with high-fidelity diffusion samplers remains a barrier to large-scale industrial deployment [1,8].

5.2 Future Directions

Future advances in this area will benefit from standardized benchmarks including canonical maps, language intent corpora, executable encodings and consistent metrics of evaluation [3,4]. More research should compare diffusion conditioning strategies in controlled experiments and hybrid combinations of symbolic constraints and differentiable guidance [1,9]. Also, improving language-to-schema reliability with retrieval augmentation and constrained decoding is essential for robust executability [10]. Introducing map aware priors and vehicle dynamics into diffusion models will also bridge the gap between plausibility and deployability [5,6]. Human-in-the-loop tools and closed-loop evaluation with ADS feedback are promising avenues towards practical deployment [2].

6 Conclusion

This paper reviews recent advances in language-guided and diffusion-based scenario generation for autonomous driving. Survey representative methods such as rule-based, optimization-based, data-driven and hybrid language-guided diffusion methods, analyze their system models, synthetic method of scenario and assessment practice.

The review concludes that hybrid pipelines that combine large language models with diffusion-based multi-agent trajectory generation strike a balance between high-level semantic control and low-level realistic motion, generating diverse and safety-relevant situations. Language-guided methods are more accessible and interpretable, while diffusion methods provide temporal consistency and social plausibility of interactions between multiple agents.

Despite this progress, there are still some challenges ahead. Ambiguity of natural language, LLM hallucinations, lack of comprehensive evaluation metrics, and insufficient consideration of vehicle dynamics and map limitations mean reproducibility and deployability are limited.

And looking ahead, overcoming these issues will require standardized assessments, increased language-to-schema reliability, and better incorporation of map- and dynamics-aware priors into scenario generation pipelines.

References

[1] Zhong, Z., Rempe, D., Xu, D. et al.: Language-Guided Traffic Simulation via Scene-Level Diffusion, in Proceedings of The 7th Conference on Robot Learning (CoRL), PMLR (2023)

[2] Birte, N., Eike, M.: Scenario-based Verification and Validation of Automated Transportation Systems, published by INCOSE (2022)

[3] ASAM e. V.: ASAM OpenSCENARIO: User Guide, version 1.2.0 (2022)

[4] Feng, S., Feng, Y., Yu, C. et al.: Testing Scenario Library Generation for Connected and Automated Vehicles, Part I: Methodology, published in IEEE Transactions on Intelligent Transportation Systems (2021)

[5] Waymo-Research: Waymo Open Dataset, Online dataset and benchmark (2021)

[6] Caesar, H., Bankiti, V. et al.: nuScenes: A Multimodal Dataset for Autonomous Driving, in Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)

[7] Ivanovic, B. & Pavone, M.: The Trajectron: Probabilistic Multi-Agent Trajectory Modeling With Dynamic Spatiotemporal Graphs, in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019)

[8] Ho, J., Jain, A. & Abbeel, P.: Denoising Diffusion Probabilistic Models, in Advances in Neural Information Processing Systems (NeurIPS) (2020)

[9] Jiang, C., Cornman, A., Park, C. et al.: MotionDiffuser: Controllable Multi-Agent Motion Prediction using Diffusion, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023)

[10] Zhang, J., Xu, C. et al.: ChatScene: Knowledge-Enabled Safety-Critical Scenario Generation for Autonomous Vehicles, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)

[11] Liu, Y., Guo, A., Zhou, Y. et al.: SoVAR: Building Generalizable Scenarios from Accident Reports for Autonomous Driving Testing, 39th IEEE/ACM International Conference on Automated Software Engineering (ASE '24) (2024)