

中图法分类号: TP391 文献标识码: A 文章编号: 1006-8961(2025)11-3413-25

论文引用格式: Liu J F, Zhang T Y, Zhong F Z, Yue P, Liu A S and Liu X L. 2025. A survey of safety evaluation data generation techniques for autonomous driving. Journal of Image and Graphics, 30(11):3413-3437(刘江帆, 张天缘, 钟芳桢, 岳鹏, 刘艾杉, 刘祥龙. 2025. 面向自动驾驶的安全评测数据生成技术综述. 中国图象图形学报, 30(11):3413-3437)[DOI:10.11834/jig.250181]

面向自动驾驶的安全评测数据生成技术综述

刘江帆^{1,2}, 张天缘^{1,2}, 钟芳桢^{1,2}, 岳鹏^{1,2}, 刘艾杉^{1,2*}, 刘祥龙^{1,2,3}

1. 北京航空航天大学计算机学院, 北京 100191; 2. 北京航空航天大学复杂关键软件环境全国重点实验室, 北京 100191;
3. 北京中关村实验室, 北京 100194

摘要: 基于深度学习的自动驾驶技术已经在多个场景中广泛应用。然而由于真实世界环境的开放性和复杂性, 自动驾驶系统面临大量潜在的安全性问题。如何生成高质量的安全评测数据, 提升系统在极端场景下的可靠性, 成为当前研究的关键问题之一。基于此, 本文对面向自动驾驶的安全评测数据生成技术进行了系统综述, 旨在把握领域的研究进展与方向, 进一步推动自动驾驶技术的长足发展, 助力其应用安全可靠。首先结合自动驾驶的典型工作流程, 从感知、预测与决策等角度出发, 定义并分析了当前系统面临的安全威胁与技术挑战。随后, 围绕生成模型、优化方法和语义驱动3类关键技术路径, 系统梳理了各类数据生成方法在自动驾驶全流程中的典型应用与技术特点。最后, 展望了融合多模态数据、语义建模与系统反馈机制的重要方向, 提出构建兼具真实性与多样性的生成体系以及依托高精度闭环仿真平台的发展趋势, 为自动驾驶安全验证提供理论与技术参考。

关键词: 数据生成技术; 安全关键数据; 自动驾驶; 生成模型; 优化方法; 语义驱动技术

A survey of safety evaluation data generation techniques for autonomous driving

Liu Jiangfan^{1,2}, Zhang Tianyuan^{1,2}, Zhong Fangzhi^{1,2}, Yue Peng^{1,2}, Liu Aishan^{1,2*}, Liu Xianglong^{1,2,3}

1. School of Computer Science and Engineering, Beihang University, Beijing 100191, China;
2. State Key Laboratory of Complex and Critical Software Environment, Beihang University, Beijing 100191, China;
3. Beijing Zhongguancun Laboratory, Beijing 100194, China

Abstract: The rapid advancement of deep learning has profoundly accelerated the development of autonomous driving (AD) systems, enabling them to operate in increasingly complex, dynamic, and unstructured real-world environments. By leveraging large-scale, high-dimensional data and increasingly powerful neural network architectures, these systems have achieved substantial progress in a wide range of core tasks, including high-precision perception through sensor fusion, accurate prediction of surrounding agent behaviors, and real-time decision-making under complex traffic constraints. This progress has enabled autonomous vehicles (AVs) to navigate structured environments such as urban streets and highways. However, ensuring the safety and reliability of AD remains a fundamental and unresolved challenge, particularly when systems encounter rare, unexpected, or hazardous situations that lie outside the typical distribution of training and test data. Scenarios involving long-tail events, adversarial perturbations, or distribution shifts can severely compromise system perfor-

收稿日期: 2025-04-15; 修回日期: 2025-07-30; 预印本日期: 2025-08-06

* 通信作者: 刘艾杉 liuaishan@buaa.edu.cn

基金项目: 国家自然科学基金(62476018, 62206009); 中央高校基本科研业务专项资金资助; 航空科学基金项目(20230017051001)

Supported by: National Natural Science Foundation of China (62476018, 62206009)

mance, leading to outcomes that may threaten safety despite high overall accuracy under standard conditions. These edge cases are difficult to capture through conventional data collection because of their infrequency and unpredictability, and are often underrepresented in existing benchmarks. Consequently, the research community has increasingly recognized the need to construct evaluation datasets that are not only large in scale but also rich in diversity and semantic relevance to meaningfully assess system robustness under real-world variability. To support such efforts, recent work has begun to explore how to systematically generate challenging and informative scenarios through synthetic data, simulation platforms, and generative modeling techniques. These trends underscore a shift toward more proactive, data-centric approaches to validation, emphasizing the importance of evaluating AD systems not only in nominal cases but also under the full spectrum of environmental complexity and operational risk. To address the pressing need for reliable evaluation under rare and safety-critical scenarios, this survey presents a comprehensive review of data generation techniques aimed at enhancing safety assessment in AD systems. Although such scenarios occur infrequently in real-world operations, they often correspond to critical failure points where system vulnerabilities are most pronounced. Traditional validation pipelines, which rely heavily on precollected datasets or limited real-world testing, are insufficient for uncovering these edge cases given their constrained scenario coverage and the high cost, risk, and unpredictability associated with capturing rare events in physical environments. In light of these limitations, we emphasize artificial data generation as a proactive and scalable approach for constructing diverse, targeted, and informative evaluation scenarios. Rather than concentrating solely on algorithmic refinements within individual modules, we adopt a holistic perspective that investigates how safety vulnerabilities emerge across the entire autonomy stack, encompassing perception, prediction, and planning. While these components are often evaluated in isolation, they operate in a highly interconnected manner in deployed systems. Errors in perception may cascade into inaccurate predictions or unsafe planning decisions, while flaws in predictive reasoning can trigger suboptimal or hazardous behaviors in downstream planning, even when upstream inputs are correct. These safety risks frequently stem not only from the limitations inherent to individual modules but also from complex interdependencies, hidden feedback loops, and susceptibility to environmental uncertainties. Such cross-cutting failure pathways highlight the need for scenario generation frameworks that capture not only isolated errors but also integrated, system-level interactions that more accurately reflect the operational complexities of real-world AD. To support rigorous and targeted evaluation under these conditions, this survey organizes existing data generation approaches into three methodological paradigms: data-driven generative modeling, which employs machine learning models to synthesize diverse and realistic driving data distributions; optimization-based methods, which explicitly seek failure-inducing conditions through adversarial testing, scenario search, or reward-driven exploration; and semantics-guided generation, which leverages high-level constraints, formal specifications, or scene priors to ensure interpretability, control, and relevance in generated scenarios. For each paradigm, we analyze the underlying principles, technical implementations, and typical application domains, and we critically assess their strengths and limitations in capturing failure modes that are essential for robust safety evaluation. Through this analysis, we aim to clarify the landscape of scenario generation techniques, identify gaps in current methodologies, and provide a foundation for future research toward more comprehensive and risk-aware validation of AD systems. Despite the progress made, current data generation techniques still face several critical challenges that constrain their effectiveness in supporting system-level safety validation. A key difficulty lies in balancing the trade-off between data realism and diversity. While high-fidelity samples are necessary to ensure physical plausibility and semantic coherence, excessive emphasis on realism may hinder the generation of rare, safety-critical scenarios that are essential for robustness evaluation. Conversely, approaches that prioritize diversity may produce semantically inconsistent or physically implausible samples, reducing their applicability to real-world systems. These challenges are further complicated by the differing requirements across system modules, where perception tasks emphasize sensory accuracy and geometric consistency, while prediction and planning demand coherent behavioral evolution and decision feasibility. In addition, most existing methods lack mechanisms to incorporate system feedback into the generation process, limiting their ability to adaptively target failure-prone conditions. The absence of an integrated testing pipeline that connects data generation, simulation, and deployment also restricts the overall effectiveness of these methods in realistic safety assessment. Moving forward, future research should focus on developing adaptive and interpretable generation frameworks that can dynamically adjust generation objectives on the basis of

task context and system feedback while building high-precision simulation platforms that support consistent multimodal data modeling and semantically controllable scenario construction. Advancing these capabilities will be essential for transforming data generation into a reliable foundation for the evaluation and improvement of safety-critical AD systems.

Key words: data generation; safety-critical data; autonomous driving; generative model; optimization; semantic-driven technique

0 引言

自动驾驶技术的快速发展正在推动该领域从理论研究逐步迈向实际场景的应用。在商业化层面,民用乘用车广泛配备了不同程度的智能辅助驾驶功能,为消费者带来了显著的驾驶体验提升;与此同时,Uber、Waymo、百度等企业在无人驾驶出租车等领域的深入探索,也显著加速了技术的实践化和市场化进程。然而,伴随着自动驾驶技术的普及,安全性问题愈发引发广泛关注。例如,2018年全球首例自动驾驶致死事故中,一辆搭载自动驾驶系统的Uber沃尔沃测试车在路测过程中撞击行人,导致悲剧发生,此事件引发了社会对自动驾驶技术安全性的广泛质疑。除此之外,多次发生的车祸等事故均暴露出当前自动驾驶技术在复杂交通环境中应对能力的不足,尤其是在面对突发情况时的反应迟缓。这些事件深刻揭示了自动驾驶技术在安全性方面的严重短板,为技术的进一步推广和应用敲响了警钟。

针对自动驾驶系统面临的安全挑战,生成覆盖长尾和极端场景的安全评测数据被认为是提升测试有效性和系统鲁棒性的核心任务之一。尽管这类情况发生概率较低,但一旦发生,往往对系统的安全性提出严峻挑战,甚至可能导致严重的安全后果。传统真实路测在覆盖这类场景时存在显著局限。Kalra和Paddock(2016)的研究表明,为全面验证自动驾驶系统的安全性,可能需要上亿甚至上百亿公里的道路测试,这种规模的测试在资源需求和操作可行性方面均难以实现,且场景复现困难,存在潜在安全风险。因此,如何高效地测试长尾和极端场景,成为当前自动驾驶安全测试的核心需求。

目前,已有多篇综述对自动驾驶安全评测问题进行了探讨。例如,Guo等人(2020)深入分析了影响自动驾驶系统安全性的因素和数据集,并提出了可行的改进方向。然而,如何在缺乏实际数据支持的情况下有效构建和验证新的度量标准,仍然是一

个挑战。Chia等人(2022)全面评估了自动驾驶系统的风险评估方法,强调了其在开发和实时操作中的重要性,但对不同方法的实际应用和局限性讨论较为简略。Lou等人(2022)揭示了自动驾驶系统测试中的现有实践和需求,并通过文献回顾分析了当前测试方法的不足,为未来的研究方向提供了宝贵的见解。然而,以上工作主要集中在测试方法的现状和需求上,尚未深入探讨如何具体实现这些研究方向的技术细节。本文的主要贡献和特色在于以下几个方面:1)提出以“安全评测数据生成”为中心的研究视角,拓展了自动驾驶安全验证研究的技术范式。不同于以往综述主要关注系统层面的测试方法与评估指标,本文强调数据生成机制在构建边界场景与表征系统脆弱性中的基础作用,为理解安全验证的底层支撑提供了新的切入点。2)从生成方法与应用层面双维度出发,理清不同技术路径在自动驾驶系统中的定位与适配关系。本文将主流生成策略划分为3类技术路径,并结合其在感知数据生成与语义行为建模中的实际应用进行系统梳理,为后续研究提供了结构清晰的技术图谱。3)结合当前技术演进趋势,分析生成方法面临的关键问题,并提出值得进一步关注的研究方向。在总结已有成果的基础上,从语义一致性建构、多模态融合策略与闭环验证机制等方面提出若干挑战。

本文聚焦于自动驾驶系统中的安全评测数据生成问题,围绕感知—预测—决策的典型任务链条,梳理了当前深度学习方法在系统运行中的技术结构与作用机制,本文结构如图1所示。在分析自动驾驶系统面临的误差耦合传播、对抗扰动敏感性与长尾场景脆弱性等安全挑战的基础上,指出数据生成技术在构建多样化、高风险、低频测试集中的关键价值。为系统呈现该领域的研究进展,本文从生成建模范式出发,将现有方法划分为3类路径:概率分布学习驱动的生成模型、性能优化导向的策略方法以及以语义结构控制为核心的生成机制。同时,本文结合应用任务的层次特征,将生成实践划分为感知

层输入数据生成与语义层行为/场景建模两个方面,并在第3节和第4节中通过典型研究展开对比与归纳。在已有方法取得初步成果的同时,本文亦指出当前仍存在生成控制能力不足、语义一致性约束薄弱和多模态融合与闭环评测机制尚未完善等关键问题。基于对研究对象、方法体系与应用需求的系统分析,本文构建了面向自动驾驶安全评测的数据生成研究框架,为该领域的后续研究提供了问题导向与方法参考。

为确保综述内容的代表性与前沿性,本文主要依据近年来在计算机视觉与自动驾驶相关领域的国

际主流会议及高水平期刊上发表的研究成果,同时部分参考具有影响力的预印本平台 arXiv 中发布的最新工作。在文献收集过程中,以“autonomous driving data generation”、“safety-critical testing”、“rare scenario synthesis”等关键词为核心,综合检索了2018年—2025年的研究文献,并结合引用影响力与方法多样性进行筛选。该文献收集策略旨在覆盖当前具有代表性的技术路径和方法类别,同时也不可避免地存在对部分最新或未广泛传播方法覆盖不全的局限性。相关说明有助于理解本文综述范围的边界条件,并为读者把握结论的适用性提供参考。

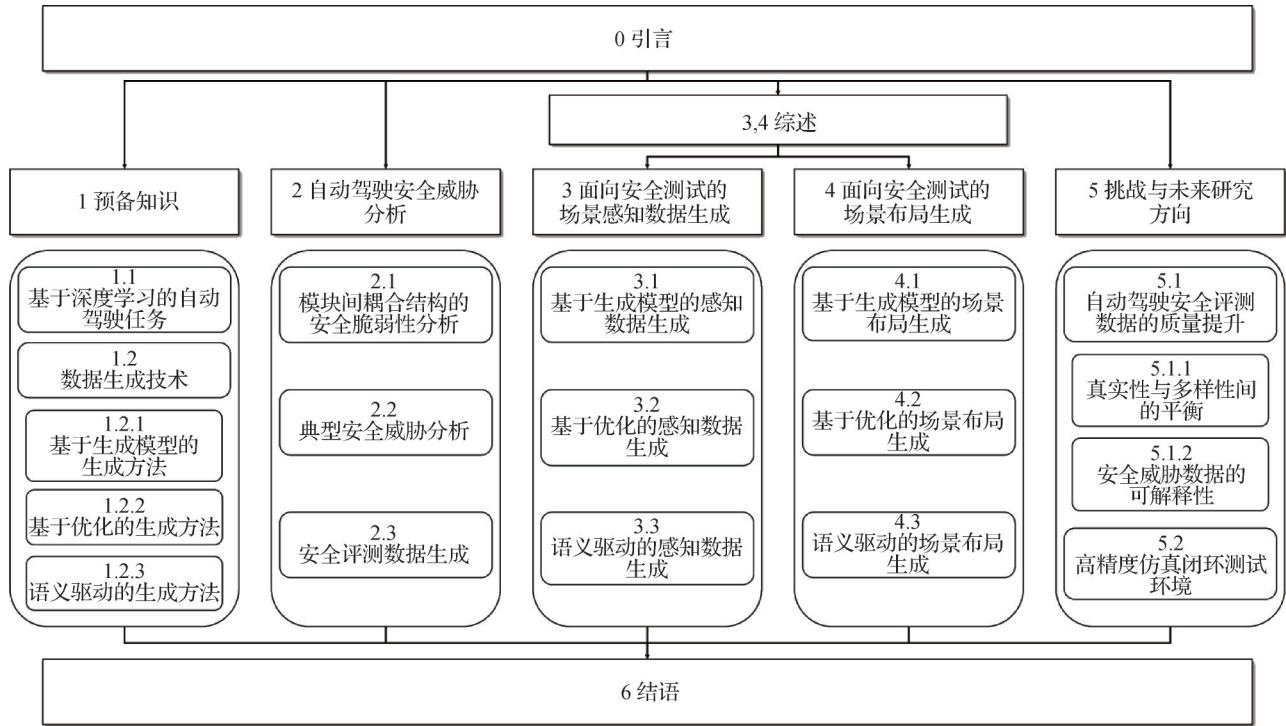


图1 本文的章节结构

Fig. 1 The chapter structure of this article

1 预备知识

1.1 基于深度学习的自动驾驶任务

在自动驾驶技术领域,深度学习的飞速发展推动了系统的革命性变化,从感知到决策的各个环节均得到了显著提升。相较于传统方法,深度学习凭借其强大的特征提取和泛化能力,极大地增强了自动驾驶系统在核心任务——感知(perception)、预测(prediction)和规划(planning)中的表现。这不仅使自动驾驶车辆在复杂多变的真实环境中展现出更高

的智能化水平和可靠性,还在多个方面提供了突破性解决方案,例如提高环境感知的精度和动态适应性、预测交通参与者行为以及优化行驶策略。

在感知任务中,环境理解是自动驾驶系统的基础,主要依赖于多模态传感器数据的处理,例如RGB相机、激光雷达(light detection and ranging, LiDAR)和毫米波雷达(radio detection and ranging, Radar)。这些传感器通过数据融合与映射,综合了视觉相机的语义丰富性与激光雷达的深度精确性,从而显著提升了感知的鲁棒性和准确性(李熙莹等,2023)。其中,目标检测(object detection)是感知

的核心任务之一,它利用深度卷积神经网络从传感器数据中高效提取特征,识别动态与静态目标,例如车辆、行人和交通标志。而目标跟踪(object tracking)则通过分析目标的运动轨迹和动态变化,利用时序信息构建稳定的环境理解模型。在线地图构建(online mapping)模块整合实时几何信息与语义信息,生成高精度动态地图,为车辆精准定位与导航提供支持。

预测任务在自动驾驶系统中扮演着至关重要的角色,其目标是通过推断动态环境中交通参与者的行为,为规划和决策提供前瞻性保障。其中,运动预测(motion prediction)和占用预测(occupancy map prediction)是预测任务的两大核心方向,成为研究领域的热点。运动预测专注于动态目标的轨迹推断,分析交通参与者的历史行为及周围环境关系,推测其未来运动状态。这一任务需要准确捕捉速度、加速度和方向等动态特征,并建模不同交通参与者之间的交互关系,从而生成多模态轨迹分布,为路径规划提供高质量输入。而占用预测不仅关注动态目标轨迹,还扩展到静态障碍物未来空间分布的建模。通过预测占用网格图与占用流,占用预测能够精确描述环境中障碍物的动态变化,提前识别潜在风险区域,为碰撞风险评估提供全面支持。

规划任务是自动驾驶的核心决策模块,其目标是在感知与预测的基础上制定最优行驶策略。规划任务通常分为全局路径规划、行为决策和局部轨迹规划3个层次,三者协同合作以确保车辆能够在动态环境中安全、高效地完成驾驶任务。全局路径规划侧重于静态地图信息,通过搜索算法生成从当前位置到目标点的最优路线;行为决策则专注于动态场景下的驾驶选择,如换道、车道保持和避障操作,强调与其他交通参与者的实时交互;局部轨迹规划结合车辆的实时状态与局部环境信息,在动态约束下优化驾驶轨迹,从而确保行驶的安全性与舒适性。

近年来,针对自动驾驶任务链条的研究综述不断推进,为理解感知—预测—决策模块之间的协同机制提供了重要参考。例如,陈妍妍等人(2024)系统总结了端到端自动驾驶系统的整体结构与关键实现方式;胡学敏等人(2024)梳理了仿真到现实环境下的决策迁移问题,并分析了多源不确定性对决策模块的影响;刘旖菲等人(2021)则聚焦于基于视觉感知的运动规划任务,深入探讨了感知—控制闭环

中规划策略的端到端优化路径。这些研究成果为本文后续围绕感知与行为数据生成的技术路径分析奠定了理论基础。

在感知、预测与规划任务中,数据生成技术发挥了关键作用,其目标是构建广泛且高度逼真的虚拟测试数据。这些数据需准确模拟真实场景中的复杂交互与动态变化,并适应极端驾驶条件和多模态输入的特性。生成模型通过扩展建模能力支持复杂场景合成,优化方法则通过高效生成策略提升数据的精度与多样性,而语义驱动技术通过深度嵌入交通规则与场景上下文信息,进一步增强数据生成的现实感与适用性。感知、预测和规划三大任务的协同发展,以及数据生成技术在各任务中的深度结合,标志着自动驾驶技术向更高层次发展的无限可能性。

1.2 数据生成技术

传统的真实路测方法在成本、覆盖范围和场景重现性方面存在显著局限,难以满足长尾和极端场景下的系统验证需求。为应对这一挑战,数据生成技术被广泛应用于构建多样化、可控、低成本的测试样本,用于提升自动驾驶系统在稀缺场景下的鲁棒性与安全性。近年来,随着深度学习能力的不断进步,数据生成方法逐渐形成了3类主要技术路径:1)基于生成模型的方法,侧重于学习数据分布并进行高质量样本合成;2)基于优化的方法,通过设定目标函数引导生成过程以满足特定约束条件;3)语义驱动的方法,强调从高层语义或规则中抽象生成逻辑一致的测试数据。本节将依次介绍这3类主流路径的基本原理、代表方法及其在自动驾驶数据生成中的典型应用,构建起完整的技术演进图景。

1.2.1 基于生成模型的生成方法

基于生成模型的方法是当前自动驾驶数据生成领域中最具代表性和广泛应用的一类技术路径。其基本思想是通过学习真实数据的潜在分布,合成具有多样性和高质量的样本,以模拟复杂、稀缺甚至未曾出现的驾驶场景。近年来,随着深度生成建模能力的显著提升,自动驾驶相关研究中广泛采用3类主流生成模型:变分自编码器(variational autoencoder, VAE)、生成对抗网络(generative adversarial network, GAN)和扩散模型(diffusion model)。本文从技术原理与应用特性的角度对这3类方法进行归类与分析,分别介绍其核心机制、典型应用,以及其

在自动驾驶测试数据生成中的优势与局限。

VAE 是一类概率生成模型(Kingma 和 Welling, 2014),通过在编码器—解码器结构中引入变分推断,有效学习数据的潜在表示。在自动驾驶场景中,VAE 能够生成低频的长尾场景,例如行人突然出现或极端天气等稀有事件,其优势在于建模潜在空间结构和生成分布受控的样本。VAE 结构稳定、易于训练,适用于需要生成多样性较强、结构清晰的场景。但其生成图像的精细度通常低于其他方法,存在一定模糊性问题。

GAN 由生成器和判别器组成(Goodfellow 等, 2014a),通过对抗训练方式逼近真实数据分布。GAN 在合成真实感较强的图像和动态驾驶场景方面表现突出,特别适用于城市交通、交互密集等复杂环境建模(Kuefler 等, 2017)。同时 GAN 也可以通过基于条件的扩展,实现天气、时间、道路类型等属性的控制,有助于合成特定测试条件下的数据。然而,GAN 训练过程不稳定,存在模式崩溃等问题,生成样本多样性控制能力相对有限。

扩散模型是一类通过逐步去噪生成数据的生成模型,该方法从随机噪声中重建出样本数据,能够兼顾生成质量与物理一致性(Sohl-Dickstein 等, 2015; Ho 等, 2020)。扩散模型近年来在图像合成任务中表现优越,生成结果精度高、细节真实,稳定性好,适用于构造稀缺、高复杂度、高真实性的感知样本。例如模拟暴雨、夜间强光反射等极端条件下的图像,对系统的极限识别能力测试具有重要意义。其主要挑战在于生成速度较慢、计算资源需求较高。

综上所述,3 类主流生成模型在自动驾驶数据生成中的适用性各具特点。VAE 结构简单、分布建模能力强,适用于生成具备潜在语义结构的稀缺场景;GAN 在视觉真实感方面表现突出,适合模拟动态交通环境,但训练稳定性不足;扩散模型在生成精度、稳定性和可控性方面具有明显优势,适合用于构建物理一致性高的测试样本。选择适当的生成模型类型需结合具体应用需求,综合考虑生成质量、多样性、控制精度及资源开销等因素。

除了上述 3 类概率生成模型,近年来以 Transformer(Vaswani 等, 2017)为代表的递归生成方法也在自动驾驶数据生成任务中展现出广泛应用,特别是在跨模态场景建模与时序数据生成方面。Transformer 结构通过自注意力机制建模长距离依赖,已

被用于鸟瞰图驱动的图像合成、语义到图像生成以及驾驶视频预测等任务中。这类方法具备强大的表示能力和可控生成能力,正逐步成为生成模型体系的重要补充路径。

1.2.2 基于优化的生成方法

基于优化的生成方法通过引入反馈机制,对数据生成过程进行目标驱动的调整与改进,适用于自动驾驶中生成具有挑战性且目标导向明确的测试样本。相较于直接建模数据分布的方法,该类方法更关注生成样本对系统性能的影响。本文将其分为两类典型技术路径:1)强化学习(reinforcement learning, RL),通过系统交互反馈优化生成策略,用于构造触发系统边界行为的关键场景;2)模仿学习(imitation learning, IL),通过对人类或专家行为的建模,优化生成策略以复现复杂驾驶行为与交互过程。以下将分别介绍这两类方法的原理及其在自动驾驶测试中的应用特点。

强化学习(Mnih 等, 2015; Lillicrap 等, 2015)是一类以最大化累积奖励为目标的优化方法,该方法通过设计合适的奖励函数,引导生成代理不断探索能有效触发系统误判、故障或极端行为的输入或情境。在自动驾驶数据生成中,强化学习能够生成具有高风险、低概率特性的长尾场景和极端场景。例如,通过设定奖励函数鼓励生成器探索雨雪天气、紧急避障、高速并道等复杂条件下的场景,从而逼近系统的能力边界(Sallab 等, 2017)。强化学习还能够通过状态和动作的交互动态调整生成策略,使得生成的数据与自动驾驶系统的能力弱点高度匹配。例如,在高交通流环境中,强化学习通过持续优化奖励机制生成紧急制动或复杂并道的场景,为验证系统的极限性能提供了精准的测试数据。强化学习方法的优点在于生成行为具有高针对性、适用于黑盒系统测试;但其训练成本高、结果稳定性依赖奖励函数设计,且生成场景的物理或语义合理性难以严格控制。

模仿学习(Pomerleau, 1989; Ross 等, 2011)通过学习专家行为策略以指导数据生成过程,在自动驾驶任务中广泛用于模拟人类驾驶者或真实交通参与者的行为演化路径,包括拟合专家轨迹数据和从专家行为中推理隐含的目标函数等。在数据生成任务中,模仿学习可用于生成更贴近真实交通语义的行为场景,如模拟驾驶员对突发事件的规避反应、多车

协同行为等。此外,模仿学习还具有生成多样化数据的独特能力,通过学习不同交通风格和密度下的行为,生成涵盖广泛分布的测试数据,从而为自动驾驶系统的泛化能力提供保障。与强化学习相比,其优势在于生成行为合理性强,能提升测试场景的语义一致性与可信度;但存在对高质量专家数据依赖性强、泛化能力受限的问题,难以覆盖少见或极端行为模式。

总体而言,强化学习和模仿学习作为两种主流的优化生成策略各具优势。强化学习侧重于任务导向与挑战性样本的生成,适用于发现系统的潜在弱点与鲁棒性瓶颈;而模仿学习更关注生成行为的真实还原与语义一致性,适用于构造覆盖典型行为逻辑的高可信度测试样本。在实际应用中,可根据测试目标选择不同优化策略,或将两者融合以兼顾生成样本的可控性与真实性。

1.2.3 语义驱动的生成方法

语义驱动的生成方法是一类通过解析和操控场景布局信息来生成具有逻辑一致性、现实意义的自动驾驶测试数据的技术。近年来,随着自动驾驶领域对多模态语义信息理解和场景生成需求的增长,基于世界模型(world model)的生成方法逐渐成为热点研究方向。“世界模型”的定义仍在不断讨论中,主要可以分为两种核心视角:理解世界和预测未来。这种双重视角赋予了语义驱动生成方法全新的能力,使其能够在高度复杂的自动驾驶场景中展现强大的语义表达与生成性能。

从“理解世界”的视角来看,语义驱动生成方法依托于对场景中关键语义元素的解析与建模,包括道路布局、交通规则、车辆行为、行人动态和环境条件等。传统的语义驱动方法以场景布局图(semantic map)为基础,通过对这些元素的调控实现场景生成。然而,世界模型通过更高层次的抽象和语义理解,结合大模型的强大能力,实现了对复杂语义关系的全面建模。例如,利用视觉和激光雷达数据训练大模型,提取语义特征后再生成复杂场景,可以快速覆盖长尾事件或高挑战性驾驶场景。这种基于语义的生成能力不仅提高了场景的真实感,还增强了对复杂交通环境的适应性。

另一方面,从“预测未来”的视角出发,世界模型通过学习语义元素间的动态关系,能够生成时间演化一致性更高的场景。例如,基于交通流预测生成

拥堵场景,或者根据车辆轨迹预测复杂交互行为。这种动态生成能力使语义驱动方法不仅能描述当前的场景状态,还能生成可能的未来情境,用于测试自动驾驶系统在连续场景下的鲁棒性和反应能力。例如,通过语义扩展生成极端天气中的动态场景,可以测试系统在突发事件中的应对策略。这样的生成过程体现了世界模型在预测能力上的显著优势。

语义驱动生成方法还展现了强大的多模态融合能力,通过结合图像、文本、传感器等多模态数据,世界模型可以实现跨模态语义信息的转化。例如,从简单的文本描述“模拟雨天的高速公路拥堵场景”出发,结合语义特征生成复杂交通场景。这种多模态生成技术不仅提升了生成数据的多样性,还显著降低了生成过程对场景设计的复杂度。此外,语义驱动方法还能通过场景分解与组合技术,将复杂场景拆解为基本语义单元,再重新组合生成新的场景,从而进一步扩大生成数据的覆盖范围。

在实际应用中,语义驱动生成方法广泛用于生成复杂的长尾场景和极端场景。例如,通过调整场景布局中的关键特征(如非典型驾驶行为、异常交通流量或天气条件),可以生成车辆突然停车、行人无视交通规则横穿马路等具有挑战性的测试场景。这些方法特别适用于测试自动驾驶系统在极端条件下的鲁棒性。此外,语义驱动生成方法也能够利用真实数据与虚拟数据的语义融合,通过从真实驾驶数据中提取语义信息并与虚拟环境结合,生成贴近真实分布的测试数据。这种语义驱动的生成过程在提升数据真实性的同时,也增强了生成场景的多样性和灵活性。

语义驱动生成方法的优势在于其对场景元素的高度可控性和生成结果的逻辑一致性。通过引入世界模型,这些优势得到了进一步放大,使得生成的场景在复杂性和多样性上都实现了显著提升。然而,这种方法也面临一些挑战,例如构建和训练世界模型需要大量高质量、多模态数据,其标注成本较高;此外,生成的语义场景可能受到模型偏差的影响,从而导致某些长尾场景或极端情况的覆盖不足。

2 自动驾驶安全威胁分析

随着自动驾驶技术的不断演进,其在真实交通环境中所面临的安全性问题日益凸显。系统中感

知、预测与决策三大核心模块在面对多源不确定性、环境复杂性及输入扰动等因素时,表现出显著的鲁棒性挑战。为有效开展系统级安全验证,亟需深入分析典型威胁类型及其传导路径,并明确高质量安全评测数据的构建需求。如图2所示,本文归纳展示了数据生成技术在自动驾驶系统中从感知数据、

场景布局到行为交互的全链路中的典型应用位置,并阐明了这些方法是如何作用于环境感知和决策制定环节,进而提升测试数据的多样性、复杂性与挑战性的。在环境感知阶段,生成技术构建用于系统输入的多样化感知数据;在决策制定阶段,生成技术支持构建丰富的场景结构与交通要素。

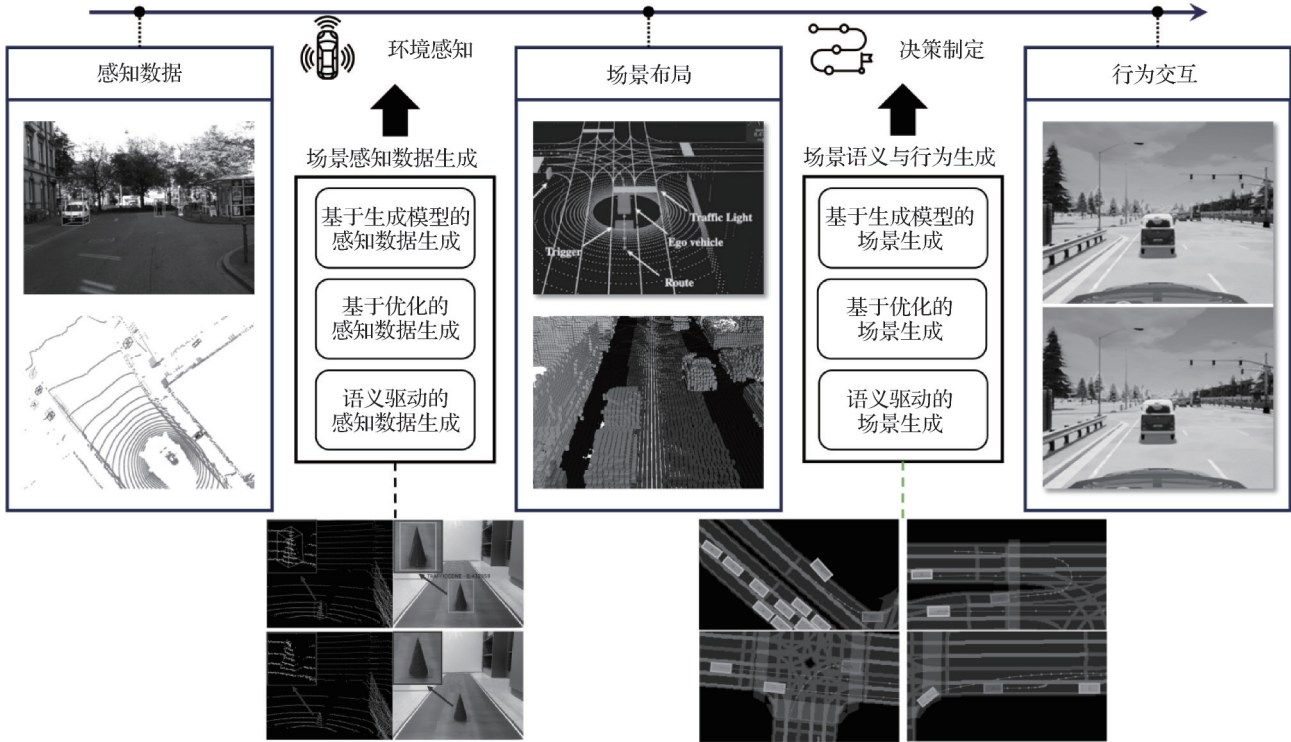


图2 数据生成技术在全链路中的典型应用位置

Fig. 2 Typical placement of data generation technology within the end-to-end pipeline

2.1 模块间耦合结构的安全脆弱性分析

自动驾驶系统作为高度模块化的智能体,其运行依赖感知、预测与决策模块的高效协同。然而,在非典型场景或系统能力边界附近,各模块间的语义传递及逻辑对接机制易出现误差放大与行为偏离问题。

1)感知—预测阶段。安全风险与误差放大。作为环境建模与动态推理的关键路径,感知—预测模块通过结构化语义传递实现环境理解的时空一致性。但模块间复杂的耦合结构可能导致语义表达受干扰,从而引发潜在的安全风险。其中,感知模块需对多模态传感器输入进行融合以提取动态目标状态,然而其输出易受传感器噪声、光照突变和遮挡干扰等自然因素影响,常出现微小抖动或误差。这些看似微小的误差在进入预测模块后,往往会被预测模型的运动外推机制非线性放大。此外,感知与预

测模块间在抽象层级、语义粒度与置信度表示上的天然差异,进一步加剧了语义传递的不稳定性。这些跨模块误差一旦发生,系统将从环境感知开始,进入误差链式传播路径,极大削弱预测环节的稳定性与可信度。

2)预测—决策阶段。行为模态分歧与策略失效。在完成环境动态推理后,系统进入将预测结果转化为实际控制指令的关键决策环节。这一过程涉及多模态轨迹信息的聚合与控制策略的生成,但模块间抽象层级的差异及语义鸿沟问题,可能导致行为模态分歧与策略失效。预测模块通常输出多个未来轨迹假设及其对应的置信度分布,而决策模块需从中筛选最优路径以生成操作性指令。然而,当预测结果中存在置信度模糊区间或轨迹假设间出现时空逻辑冲突时,决策模块可能因语义解读矛盾而生成逻辑不一致甚至物理不可行的指令序列。更严重

的是,当预测结果因感知偏差而存在偏移时,决策模块可能基于错误的环境状态估计生成路径规划。这种策略失效不仅会导致物理不可行的控制指令,还会通过置信度机制将预测模块对低频、高风险事件预测误差的影响放大至最终决策输出。具体而言,预测模块对罕见事件的轨迹预测往往伴随置信度低估或误标,而决策模块依赖置信度权重进行路径选择时,可能忽略潜在高风险轨迹,或过度保守地响应低概率事件。此类问题在交互密集或存在遮挡的场景中尤为突出,最终可能导致车辆表现出迟滞反应、策略冲突或控制指令无效等系统级失效模式。

2.2 典型安全威胁分析

在自动驾驶系统全流程运行中,主要面临以下两类高风险威胁:输入扰动与对抗攻击、长尾与极端场景下的失效风险。这些问题是系统稳定性与泛化能力的关键制约因素。

1)输入扰动与对抗攻击威胁。对抗输入作为近年来深度学习模型安全性研究中的核心议题,已被广泛证实可通过嵌入细微扰动,引导模型产生显著误判。对抗样本通过在输入数据中植入结构微扰,诱导深度学习模型在感知和决策任务中作出错误判断,成为自动驾驶系统安全的重要威胁之一。研究表明,感知模块是对抗攻击的主要目标,包括目标检测与分类等关键任务均存在较大脆弱性。例如,Liu等人(2019)提出的PS-GAN(perceptual sensitive GAN)物理攻击方法通过在交通标志表面附加对抗补丁,如图3所示,这种攻击使系统错误识别关键标志信息,并暴露出深度模型对输入分布变化的高度敏感性。该类攻击在低光照、高动态范围等复杂环境下攻击效果更为显著,具有隐蔽性强、可物理实现、部署成本低等特征,进一步放大了系统的潜在风险空间。

2)长尾与极端场景下的行为推断失效。长尾场景广泛存在于自然驾驶过程中,其特征表现为出现频率极低、事件分布不均,却往往伴随较高事故风

险。这类情境源于现实世界中稀有但高危的特殊驾驶状态,难以在训练数据中充分覆盖,致使系统在处理此类输入时的鲁棒性显著不足。长尾场景的挑战主要体现在数据数量稀缺、多样性不足与采集效率低下,传统数据采集方法难以涵盖这类稀有事件,进而限制了系统在关键情境下的性能表现。极端场景进一步超越了长尾场景的定义,指那些完全不在训练数据分布范围内的特殊情况,即典型的分布外(out-of-distribution, OOD)问题。此类场景包括自然灾害、突发事件以及系统内部故障,具有高度不可预测性和潜在的系统性风险,均可能超出模型的知识边界,导致系统无法做出合理决策。极端场景对自动驾驶系统的泛化能力与应急处理机制提出更高要求,也是现阶段系统级安全测试中最难覆盖的部分。

总体来看,对抗扰动和长尾与极端场景共同构成了自动驾驶系统在现实环境中面临的核心安全威胁。这些问题不仅挑战系统的鲁棒性与泛化能力,同时也为自动驾驶系统的边界测试提供了关键支撑。通过模拟这些极限情境,研究人员能够识别模型在决策边界附近的行为变化,捕捉系统在罕见条件下的潜在失效模式,进而发现传统测试未能覆盖的薄弱区域。对抗扰动的引入亦可作为高效、可控的验证手段,用于评估系统在微扰条件下的稳定性与鲁棒性。

2.3 安全评测数据生成

面对上述多元化的安全威胁,传统的验证与测试方法在覆盖范围、效率以及成本控制等方面均存在显著瓶颈,难以满足日益复杂的自动驾驶系统安全评估需求。特别是在长尾场景、极端事件及对抗扰动等高风险低概率情境下,依赖实车测试或人工构造的方式无法系统性复现关键风险因素,评估能力严重受限。为提升自动驾驶系统在非典型情境下的容错性验证与边界能力测试,近年来安全评测数据生成技术逐步成为研究焦点,其核心目标在于通过主动构造、可控生成与语义引导等策略,构建高覆盖、高挑战、高保真度的测试数据体系,支撑系统在极端条件下的鲁棒性检验与能力上限探索。

安全评测数据生成方法主要包括以下几类代表性技术路径。仿真驱动路径通过构建高保真度的虚拟环境,以规则建模、随机扰动、参数调控等手段生成具有可重复性和可配置性的多样化测试场景,显著提高了测试样本的生成效率和结构完整性。深度



图3 PS-GAN所生成的被攻击图像示例(Liu等,2019)

Fig. 3 Examples of attacked images generated by PS-GAN (Liu et al., 2019)

生成模型路径借助生成对抗网络、变分自编码器与扩散模型等方法对真实数据分布进行建模,能够合成语义一致、视觉真实的感知数据或行为轨迹样本,在图像、点云和语义图等多模态维度展现出良好表现。优化驱动路径则以系统性能指标为导向,采用强化学习、模仿学习或梯度优化算法,引导数据生成过程主动探测模型脆弱区域,实现“以攻促测”、“以错促优”的边界验证机制。语义驱动路径通过引入高层次抽象语义信息,如交通规则、行为逻辑或自然语言提示,对场景结构与动态演化过程进行约束,进而生成具备语义合理性和策略多样性的测试样本,支持多粒度、多尺度的系统验证需求。

从功能角度看,安全评测数据生成技术在测试样本的覆盖广度、内容复杂度与攻击能力方面均具备显著优势。一方面,该类技术可构造传统采集难以获得的关键性场景,有效拓展系统训练与测试空间,弥补长尾与极端条件下的数据缺口;另一方面,借助优化与语义引导策略,可构建模型敏感区域的高风险样本,主动激发系统的潜在失效行为,识别其在边界条件下的稳定性缺陷。因此,生成技术不仅是系统安全评估的重要补充,更是实现闭环式鲁棒性增强与决策可靠性提升的核心手段。

尽管该领域技术发展迅速,安全评测数据生成在实际应用中仍面临一系列关键挑战。一方面,生成数据的物理一致性与现实可部署性尚难充分保障,尤其在多模态数据融合、高动态交互建模、稀有情境精细刻画等方面仍缺乏统一建模框架与质量评价标准;另一方面,当前部分方法对计算资源依赖度高,生成效率有限,难以适配大规模系统级测试需求。此外,如何确保生成样本在语义层面与真实驾驶环境的分布一致,以及如何量化其在鲁棒性提升与失效风险发现中的实际贡献,亦是当前研究亟待突破的重要方向。

3 面向安全测试的场景感知数据生成

感知系统作为自动驾驶车辆获取环境信息的关键模块,其性能优劣直接影响后续预测与决策的安全性。在复杂多变的交通环境中,感知模块需从RGB相机、LiDAR和Radar等多模态传感器中提取并融合场景要素,以识别道路结构、交通参与者及其动态状态。然而,面对长尾事件、遮挡干扰和光照变

化等非理想条件,传统采集数据往往覆盖不足,难以全面评估感知系统在关键场景下的鲁棒性与边界行为。

为支撑系统性安全测试,近年来大量研究聚焦于面向感知模块的测试数据生成技术,旨在合成具备挑战性、多样性与语义可控性的感知样本,用以发现潜在风险与提升模型稳定性。如图4所示,该类技术主要包括基于生成模型的方法、基于优化的对抗样本构建方法以及语义驱动的感知合成方法,不仅能生成高保真图像和点云数据,还能在结构一致性与任务相关性方面提供支持。

3.1 基于生成模型的感知数据生成

在自动驾驶场景中,生成模型正逐渐成为感知数据合成的重要技术路径。相比传统仿真引擎依赖高精度建模与渲染流程,生成模型通过学习数据中的空间与语义结构,实现了更高效、灵活且可控的图像生成能力。近年来,GAN及其相关技术在三维感知场景数据生成方面取得显著进展,涵盖图像质量提升(Karras等,2019,2020,2021)、单图多视角生成(Wiles等,2020)、三维建模与多视角一致性生成(Nguyen-Phuoc等,2019;Schwarz等,2020;Niemeyer和Geiger,2021;Xu等,2021;Xu等,2022;Chan等,2021,2022;Xue等,2022;Shi等,2022;OrEl等,2022),以及光照与风格控制(Gu等,2021;Pan等,2021;Deng等,2022)等方向。围绕不同输入形式与任务目标,学术界也提出了一系列面向自动驾驶系统的生成模型方法,满足从静态街景到动态城市市场景等多样化的数据合成需求。

针对从抽象地图生成街景图像的问题,Zhang等人(2024a)提出了BerfScene,通过将鸟瞰图(bird's eye view,BEV)作为输入条件,引入具有空间等变性的神经辐射场生成机制,借助U-Net网络结构与Fourier编码,实现大尺度、多视角的一致图像渲染。类似地,Swerdlow等人(2024)提出的BEVGen进一步提升了BEV到图像的建模表达能力,其利用VQ-VAE(vector quantized variational autoencoder)与自回归Transformer构建图像的离散潜在空间,并结合相机方向嵌入与空间偏置机制,确保合成图像在几何与视觉上的连贯性。这两项工作均致力于从结构化布局出发,合成真实感强的街景图像,为感知系统训练提供多样、高效的场景数据。

除了布局驱动的图像生成,也有研究探索从视

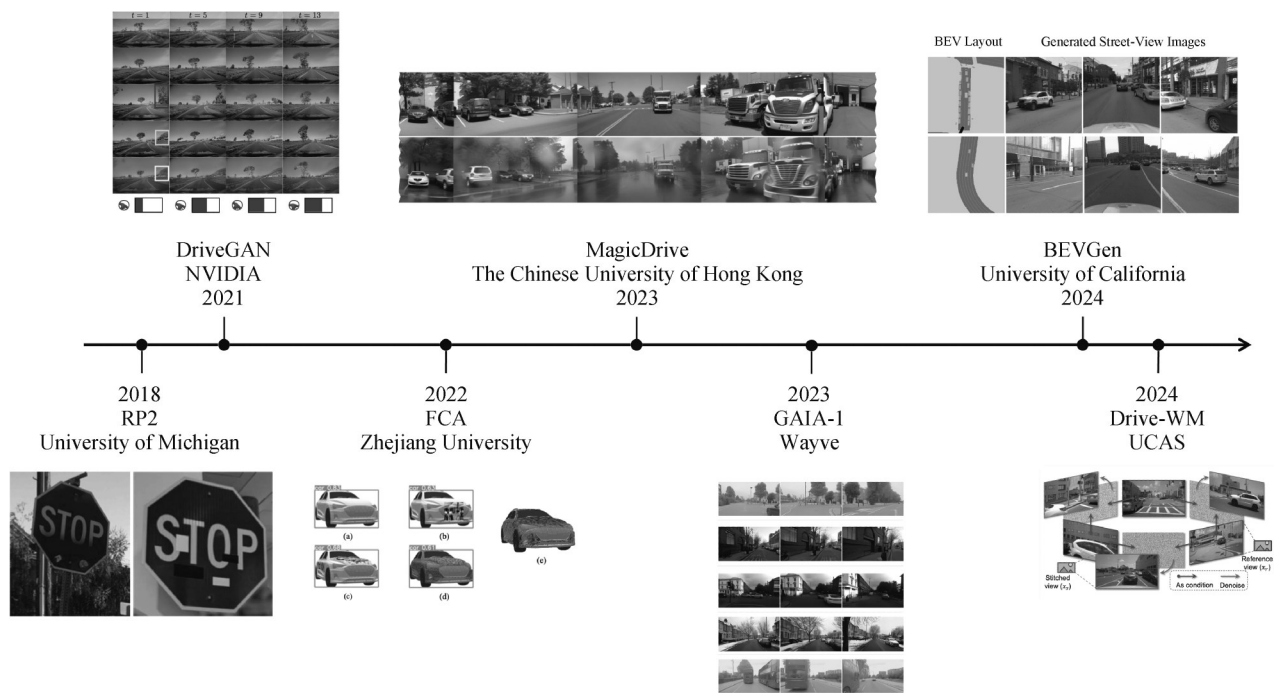


图 4 面向安全评测的多模态场景感知数据方法发展脉络

Fig. 4 The history of multi-modal perception data methods for safety evaluation

角变换中合成图像。Regmi 和 Borji(2018)提出基于条件生成对抗网络的跨视角图像生成方法,实现了从航拍图到街景图的互相转换,并通过语义引导提升生成质量。这类方法对于缓解多传感器系统中的视角盲区具有重要意义。与此同时, Han 等人(2024)提出的 GGS(generalizable Gaussian splatting)则聚焦于车道变换场景,通过从单车道图像合成邻接车道视角图像,解决了数据集中多车道样本稀疏的问题。GGS 结合虚拟车道生成、深度估计优化与扩散引导机制,在保持几何一致性的同时,提升了生成图像的感知质量。

为增强生成模型对场景语义与结构控制的能力, Xu 等人(2023)设计了 DisCoScene, 以 3D 包围盒表示的场景布局为输入,采用空间解耦的神经辐射场对前景与背景分别建模,并在潜空间中支持物体插入、移动与重建,从而生成结构可控的街景图像。与之不同, Yurtsever 等人(2022)则提出 HGNG(hybrid generative neural graphics)模型,融合传统图形渲染与 GAN 图像合成,通过使用真实渲染的前景与生成式背景进行组合,在保障几何一致性的同时提升视觉质量,适用于仿真平台中的真实感提升与图像增强任务。

此外,在动态场景合成方面, Kim 等人(2021)提

出了 DriveGAN,通过将街景图像与车辆控制信号作为输入,学习动作与环境变化之间的映射关系,从而合成可控的连续驾驶图像序列,适用于策略验证与闭环系统仿真。在此基础上, Chen 等人(2023)提出的 PVG(periodic vibration Gaussian)进一步引入 3D 高斯泼溅(3D Gaussian splatting)与周期振动机制,实现了动态城市环境中静态与动态元素的统一建模与实时渲染。该方法不仅显著提升了重建质量与渲染速度,还通过时序平滑与位置感知控制机制,有效支持复杂街景中的时空一致性建模,尤其适用于需要动态环境数据的感知系统训练与仿真测试任务。

综上所述,基于生成模型的感知数据生成方法已在自动驾驶领域呈现出多样化的发展趋势。从静态场景合成到动态过程建模,从 BEV 驱动到图像—动作耦合,这些方法共同构成了一条以数据驱动的方式服务于感知系统训练与验证的关键路径,为构建安全、高效、可泛化的自动驾驶系统奠定了坚实的合成数据基础。

3.2 基于优化的感知数据生成

基于优化的感知数据生成方法通过设定目标函数,并利用优化算法生成最符合预期的数据。这类方法通常用于生成具有特定属性的测试数据,例如安全关键数据或对抗性数据,用以验证自动驾驶系

统在极端情况下的性能。

Szegedy 等人(2013)首次发现了对抗样本的存在,研究表明在输入数据上添加微小扰动可以影响模型输出,导致其输出错误的结果。此后,针对数字世界中的对抗样本展开了一系列研究(Goodfellow 等, 2014b; Kurakin 等, 2016; Madry 等, 2017; Carlini 和 Wagner, 2017; Xiao 等, 2018),取得了显著成效。随着研究的深入,物理世界中的对抗样本研究也逐渐展开(Kurakin 等, 2018; Szegedy 等, 2016)。对抗样本的深入研究促使研究者关注自动驾驶视觉感知层面的攻防,进而催生了多种基于优化的感知数据生成方法。Athalye 等人(2018)提出了期望变换方法,通过模拟视角、亮度、图像比例等变化,确保对抗样本在不同图像变换分布下的鲁棒性,该方法广泛应用于物理对抗样本攻击。Eykholt 等人(2018)提出了针对道路标志识别系统的通用攻击算法 RP2 (robust physical perturbations)。Zhang 等人(2018)提出的 CAMOU 方法通过近似模仿模拟器将伪装应用于车辆,采用局部搜索算法找到最佳伪装,最小化检测分数,从而提出了一种生成对抗伪装的方法。Huang 等人(2020)提出了通用物理伪装攻击,使用一种通用对抗伪装来攻击同一目标类别下的物体。此后,研究进一步向通用物理伪装方向演进,Wang 等人(2021a)为了提高对抗性伪装的泛化能力,提出了双重注意力抑制攻击(dual attention suppression attack, DAS),旨在同时抑制模型和人类的注意力。Wang 等人(2022)进一步改进了先前工作中对抗伪装仅限于车顶、引擎盖和车门的不足,提出了将伪装完全覆盖在车辆上的方法,显著提升了对抗伪装的攻击性。为进一步增强伪装的现实效果与模拟精度,Suryanto 等人(2022)在可微变换攻击中使用了可微变换网络,结合神经渲染器的可微性和物理世界中的各种变换,提升了对抗伪装的有效性和泛化性。Sato 等人(2021)则创新性地提出将脏路补丁作为攻击载体,结合车辆运动模型和透视变换更新受攻击车辆的相机图像,并设计了可微的车道曲率目标函数,使用优化方法生成脏路补丁。

上述方法主要聚焦于图像模态下的对抗样本构造,而在多传感器融合的自动驾驶系统中,点云数据作为激光雷达输入的重要载体,同样面临对抗攻击的潜在威胁。Hallyburton 等人(2022)提出了一种视锥体攻击方法,通过保持图像与点云数据之间的语

义一致性来欺骗相机与激光雷达传感器的融合。进一步地,Tu 等人(2021)将对抗物体可微渲染为点云和图像输入,优化训练过程生成通用对抗物体,使感知系统无法检测到该物体,进而影响车辆的检测能力。在此基础上,Cao 等人(2021)引入变形交通锥作为三位攻击载体,利用基于梯度的三维形状和表面纹理优化生成物理对抗性物体,从而同时欺骗相机—激光雷达的传感器融合感知,使其无法检测到该物体,导致碰撞事故。

综上所述,基于优化的感知数据生成技术通过结合对抗攻击等方法,大幅提升了自动驾驶系统在复杂驾驶场景下的测试与评估能力。这些研究不仅丰富了自动驾驶安全测试的工具箱,也为探索更鲁棒的自动驾驶技术奠定了基础。

3.3 语义驱动的感知数据生成

语义驱动的感知数据生成方法通过引入语言、图像及其他高层次语义信息,为自动驾驶系统提供更精细的控制能力与更丰富的场景表达。这类方法不仅关注图像或场景本身的生成,更强调生成内容与目标任务之间的语义一致性,能够根据自然语言描述、任务需求或行为指令构建具备特定语义属性的交通场景,从而为系统的训练、测试与评估提供高度可控的模拟环境。

近年来,研究者围绕语义驱动的生成范式展开了大量探索。Hu 等人(2023)提出的 GAIA-1 (generative AI for autonomy)模型融合多模态输入与生成式建模技术,有效实现了对复杂驾驶场景的语义建构与动态生成,显著提升了场景的多样性与时间一致性。在此基础上,Wang 等人(2023)提出了一个基于世界模型的驾驶视频生成框架——DriveDreamer,能够模拟长时间、任务导向的驾驶行为。随后,Zhao 等人(2024a, b)相继提出 DriveDreamer-2 与 DriveDreamer4D,前者通过引入大语言模型 (large language model, LLM)解析文本语义,引导场景构建与行为规划;后者则结合时空感知模块与世界建模能力,支持生成空间与时间高度一致的4D驾驶视频。这一系列方法拓展了从语言到场景生成的路径,极大提升了系统对复杂、多模态驾驶语境的模拟能力。

除以语言为输入的方案外,部分研究进一步推动了语义、结构与几何信息的融合生成。Zhou 等人(2024)提出的 SimGen 框架通过现实数据与模拟器输出的混合建模,结合级联扩散网络以桥接仿真与

真实之间的语义差距。在生成控制层面, Gao 等人(2023)提出的 MagicDrive 结合几何控制与条件生成策略, 增强了生成街景的几何一致性与语义多样性。而针对多视角视频建模问题, Li 等人(2024)提出的 DrivingDiffusion 则引入语义布局与扩散模型相结合的方式, 实现了多视角视频中空间与语义的一致重建。

在融合语义与几何的生成路径基础上, 进一步地, 一些方法开始聚焦于高维感知空间中对时空一致性与多条件控制的深度融合。Pan 等人(2022)提出的 ISO-Dream 通过解耦光照与天气等干扰因素, 增强了世界模型对复杂环境的适应能力。面向大规模场景构建, Lin 等人(2023)开发的 InfiniCity 采用条件生成对抗网络与分块式生成机制, 结合局部与全局语义建模, 实现了大规模城市场景的生成与扩展。在提升跨模态与跨视角一致性方面, Gao 等人(2024a)在 MagicDrive3D 中引入了 BEV 地图、3D 对象与文本描述等多条件控制因子, 利用单目深度估计与可变形高斯表达, 有效解决了多视角间的曝光与结构差异。

在高维空间建模与多条件控制取得突破的基础上, 视频生成作为语义驱动感知数据生成的重要拓展方向, 也取得了显著进展。Wen 等人(2024)提出 Panacea 方法, 以 BEV 序列与文本提示为输入, 结合 4D 注意力机制与双阶段生成策略, 生成具备一致性与可控性的全景驾驶视频。面向多任务理解与行为建模, Wei 等人(2024)的 OccLLaMA 则融合占用推理、语言信息与行为控制, 构建出面向多任务场景理解与行为决策的统一生成模型。在生成结构方面, Xiang 等人(2024)提出 Pandora 模型, 基于混合自回归与扩散结构, 并通过指令微调, 实现了支持文本控制与动态交互的高保真视频生成。为提升长时预测能力, Deng 等人(2024)的 S4WM 则以 S4 神经网络为基础, 利用其长时记忆能力对世界状态进行建模, 增强了在动态复杂环境下的长期预测能力。Wang 等人(2024)开发的 Drive-WM 进一步结合时空建模能力, 生成多视图、高保真度的视频序列, 旨在提升自动驾驶系统的鲁棒性与场景预见能力。

综上所述, 语义驱动的感知数据生成正逐步成为连接高层认知与场景建构的重要技术路径。这类方法不仅提升了生成过程的控制粒度, 还显著增强了生成场景的动态一致性、语义多样性与行为合理

性。其发展为构建智能、泛化能力强的自动驾驶系统提供了全新的思路和更丰富的数据支持。

3.4 小结

本节聚焦自动驾驶感知系统的安全测试需求, 系统梳理了当前主流的感知数据生成方法, 重点归纳生成模型、优化算法与语义驱动 3 类技术路径, 并通过表 1 对典型方法的生成架构、输入形式及输出模态等核心特征进行对比分析。

基于生成模型的方法通过建模场景的几何语义特征实现图像或图像序列的合成, 此类技术在图像真实感、生成效率和场景构造灵活性方面表现尤为突出, 但其性能高度依赖训练数据的多样性和标注精度; 优化算法驱动的技术则侧重于通过构造对抗性扰动来拓展系统测试的覆盖广度与测试强度, 这类方法虽然具备明确的目标导向性, 却受限于复杂的优化过程, 且需在真实感与工程可行性间进行权衡; 语义驱动方法以文本描述、语义地图或轨迹控制为条件, 结合扩散模型或 GAN 实现多模态、多场景图像合成, 其优势在于能够实现从高层语义到视觉模态的直接映射, 在任务驱动测试与可控生成方面具有独特价值, 但同时也面临语义建模复杂性与跨模态对齐的双重挑战。

3 类技术在生成机制、控制粒度与场景类型等方面各具特点, 生成模型侧重数据分布建模与多样性生成, 优化算法聚焦鲁棒性测试中的边界场景构建, 语义驱动方法则提供高度可控的任务导向样本生成能力。未来研究可探索 3 类技术路径的深度融合, 在生成效率、真实感与语义一致性之间寻求更优平衡, 进而构建支持安全评测与策略泛化的高质量合成感知数据体系。

4 面向安全测试的场景布局生成

在自动驾驶系统的安全验证过程中, 场景布局与行为建模是连接感知输入与决策输出的重要中间层, 直接影响测试环境的真实性与挑战性。相较于感知层面的数据生成, 面向安全测试的场景布局生成更加关注从环境结构到参与者行为的语义建构过程, 旨在构造具有代表性、复杂性与可控性的动态交通情境, 以评估系统在关键交互和潜在风险下的鲁棒性。如图 5 所示, 相关技术通过对静态要素(如道路拓扑、交通标志、天气条件)与动态要素(如车辆与

表 1 感知数据生成的主干方法与输出模态/效果(本表列举了面向安全评测的场景感知数据生成的代表性方法)
Table 1 Main methods for generating perception data and output modalities/effects (this table lists representative methods for generating perception data for safety evaluation)

生成方式	方法名称	生成模型架构	输入	输出
生成模型	BEVGen(Swerdlow 等, 2024)	自回归模型	BEV	多视角街景图像
	(Regmi 和 Borji (2018)	GAN	航拍图像/街景图像	街景图像/航拍图像
	DisCoScene(Xu 等, 2023)	辐射场	场景中物体的位置和大小	街景图像
	DriveGAN(Kim 等, 2021)	VAE + GAN	街景图像 + 车辆控制	预测街景图像序列
优化算法	RP2(Eykholt 等, 2018)	\	含路标的街景图像	路标中包含对抗贴纸的街景图像
	FCA(Wang 等, 2022)	\	包含车辆的街景图像	车辆包含对抗纹理的街景图像
	Tu 等人(2021)	\	街景图像 + 点云	包含对抗网格的街景图像 + 点云
	Cao 等人(2021)	\	街景图像 + 点云	包含对抗交通锥的街景图像 + 点云
语义驱动	GAIA-1(Hu 等, 2023)	扩散模型	街景图像序列 + 车辆控制 + 文本	预测街景图像序列
	MagicDrive(Gao 等, 2023)	GAN	BEV + 3D 目标框 + 文本	多视角街景图像
	InfiniCity(Lin 等, 2023)	GAN	城市地图信息 + 车辆轨迹	任意城市中街景图像
	Drive-WM(Wang 等, 2024)	扩散模型	多视角街景图像序列 + 车辆控制 + 文本	多视角预测街景图像序列

注:优化算法类别中的方法并未使用生成模型架构,而是采用非生成式策略实现数据生成,在表中以“\”表示。

行人的轨迹、交互模式)的联合建模,生成可用于系统性测试的语义丰富场景。

该类数据生成方法不仅弥补了真实路测在高风险、低概率场景覆盖方面的不足,也为触发边界行为、揭示系统弱点提供了结构性支持。具体而言,场景布局生成关注交通场景的空间结构与语义组成,

行为生成则模拟交通参与者之间的动态演化与决策逻辑。近年来,从生成模型到优化搜索,再到语义驱动的方法体系逐步发展,为构建具备语义一致性与测试针对性的合成环境奠定了技术基础。本节将围绕3类主流技术路径展开讨论:基于生成模型的场景构建、基于优化的策略探索以及基于语义控制的

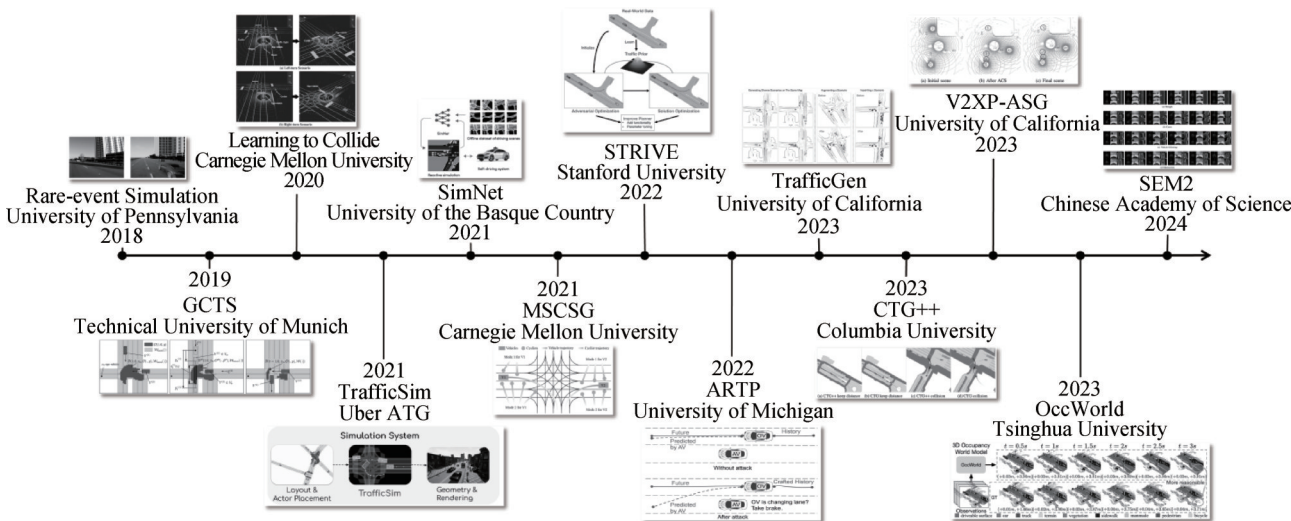


图 5 场景布局与行为生成方法发展脉络

Fig. 5 The history of scene layout and behavior generation methods

行为建模,系统梳理其在安全评测场景生成中的研究现状与关键挑战。

4.1 基于生成模型的场景布局生成

基于生成模型的场景生成技术通过深度学习方法从真实数据中提取特征,并生成具有高度真实性和多样性的场景。常见的生成模型包括生成对抗网络、变分自编码器和自回归模型等。它们能够高效捕捉场景的语义特征及其潜在关系,为自动驾驶系统提供了复杂且丰富的测试环境。

早期研究侧重于通过VAE对复杂交通场景进行整体建模与合成。Ding等人(2025)提出了一种结合领域知识的语义对抗场景生成方法SAG(semantically adversarial generation)。该框架通过树状结构的变分自编码器学习层次化的场景表示,并通过在树结构中施加语义规则来实现知识的显式整合,SAG能够生成遵循语义规则的自然场景,并有效地识别对抗不同最新3D点云分割模型的驾驶场景。在对场景可操控性的进一步拓展中,Rempe等人(2022)提出的STRIVE(stress-test drive)方法通过创新性地引入基于图的条件变分自编码器来学习交通运动的潜在表示,为自动驾驶车辆的规划系统提供了一种有效的事故倾向型驾驶场景生成手段。这种方法不仅能够自动生成具有挑战性且逼真的交通场景,而且通过优化过程在潜在空间中扰动真实世界的交通场景,以触发特定规划器产生例如碰撞的危险驾驶场景。与整体建模方式不同,Suo等人(2021)在其研究中提出了TrafficSim,这是一个基于隐含潜在变量模型的生成框架,能够为场景中的所有交通参与者共同生成社会一致性的行动计划,创造出逼真的多智能体交通场景。TrafficSim利用变分自编码器从真实世界的驾驶数据中学习,捕捉到复杂的交互模式和不规则的驾驶行为,进而生成多样化且高度真实的交通模拟样本。这一模型不仅增强了交通模拟的真实感和多样性,而且通过其生成能力,为自动驾驶技术的发展提供了一个强有力的工具,尤其是在模拟复杂交通情况和优化驾驶策略方面展现出巨大潜力。

除基于VAE的模型外,部分研究探索了自回归式建模机制,以提升生成结果的可控性与结构一致性。Ding等人(2020)提出了一种创新的自回归模型框架,用于生成自动驾驶领域中的安全关键场景。通过将交通场景分解为一系列可复用且独立的构建

块,并在这些构建块之间建立条件概率依赖关系,使得模型能够高效地学习和生成具有复杂相互依赖性的交通场景。这种方法提高了场景生成的真实性和多样性,而且通过条件概率的链式结构,使得模型能够根据输入状态(如路线和目标速度)动态调整场景参数,生成与特定任务算法相适应的安全关键场景。相较于Ding等人(2020)聚焦于构建块级的结构建模,Tan等人(2021)提出了SceneGen,进一步将自回归机制应用于场景细节层面的参与者插入与行为配置,该模型能够根据自动驾驶车辆的状态和高精地图自动插入不同类别的交通参与者,并合成它们的大小、方向和速度,如图6所示。SceneGen的核心是一个自回归模型,它能够模拟人类构建场景的过程,从而生成逼真的交通场景。通过在大规模自动驾驶数据集上的评估,SceneGen展现出在模拟真实交通场景分布上的优势,并且能够生成更具现实感的样本。

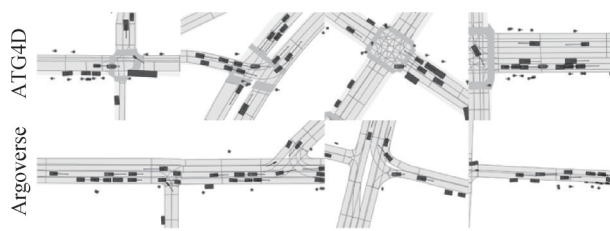


图6 自回归模型生成示意图(Tan等,2021)

Fig. 6 Illustration of autoregressive model generation
(Tan et al., 2021)

另一类研究则更加强调多智能体交互关系的建模,以应对自动驾驶系统在复杂社会交通中的策略评估需求。Feng等人(2023)提出了TrafficGen,这一方法基于自回归神经生成模型,具有编码器—解码器架构。在每次自回归迭代中,其首先使用注意力机制编码当前交通上下文,然后解码车辆的初始状态,并生成其长期轨迹。TrafficGen能够从真实世界的驾驶数据中学习,并生成逼真的交通场景,包括车辆的初始状态和长期轨迹。

在生成目标的进一步细化方面,Bergamini等人(2021)提出了Simnet,通过将驾驶事件的抽样建模为马尔可夫过程,通过两个生成器分别采样初始状态和参与者策略(转向和速度),其中采样初始状态的生成器为条件生成对抗网络,该网络以仅包含语义图的空场景为条件。在此基础上,Suo等人

(2023)提出了 MIXSIM 用于混合现实交通模拟,能够从真实数据中学习并预测交通代理的未来行为。该方法通过显式建模代理目标作为沿道路网络的路线,并学习一个反应性的路线条件策略,从而实现了真实世界场景的反应性重模拟和 what-if 情况的探索。该框架不仅能够重现原始场景的高级行为和交互,还能够通过改变每个代理的路线来扩展测试范围,包括代理行为的真实变化甚至安全关键的互动。此外,Ding 等人(2021)提出了一种能够适应不同风险水平并具有多模态分布特征的生成模型,从而得到基于流模型的多模态安全关键场景生成方法,用于评估决策算法的鲁棒性。该方法通过加权似然最大化和梯度基采样过程优化生成模型,有效地从任务算法和模拟器中查询生成安全关键场景。该生成模型特别关注于模拟真实世界中罕见但关键的交通场景,通过学习驾驶数据的多模态分布,能够生成具有不同碰撞概率的样本。

这些基于生成模型的技术在应对长尾场景和极端交通事件方面展现出巨大的潜力,但它们仍面临一些挑战,例如,如何在生成的多样性与场景合理性之间找到平衡,以及如何结合真实驾驶数据以增强模型的泛化能力。

4.2 基于优化的场景布局生成

基于优化的场景生成方法通过设定目标函数并利用优化算法构建具有挑战性或对抗性的驾驶场景,以用于系统性能边界的验证。这一类方法通常围绕安全性评估目标,设计使得系统性能指标恶化的扰动策略,从而在高风险、低概率情境下测试感知、预测和控制模块的鲁棒性。

早期研究多采用采样与演化类优化算法实现事故概率最大化或系统失败率估计。O' Kelly 等人(2018)实现了一个模拟框架,用于测试整个现代自动驾驶系统,包括使用深度学习感知和控制算法的系统。它使用自适应重要性抽样方法来加速罕见事件概率评估,估计在标准交通行为的基分布下事故发生的概率。这个框架也涉及到基于优化的场景生成,因为它通过优化方法加速罕见事件的模拟和评估。随后,一系列工作开始引入贝叶斯优化与梯度导向策略,提升优化效率与场景可解释性。Abeyisirigoonawardena 等人(2019)提出了一种自动化生成对抗性驾驶场景的方法,利用贝叶斯优化技术,在高保真模拟器中生成能够暴露自动驾驶软件缺陷的对

抗性场景,增加了与模拟行人和车辆发生碰撞的风险。除此之外,Klischat 和 Althoff(2019)则采用进化算法,从运动规划解空间中最小化临界路径,通过构造非线性搜索过程生成规划系统难以处理的临界情形。为了增强测试样本的策略多样性与系统适应能力,Chen 等人(2022)提出基于深度强化学习的自适应评估框架,结合非参数贝叶斯聚类对抗抗性策略进行建模,在保持场景复杂性的同时提升了搜索效率与策略可解释性。随着研究向更精细控制与物理可行性约束推进,梯度导向与可微优化方法逐渐成为主流路径。Cao 等人(2022)提出了 AdvDO 框架,以针对轨迹预测模型进行攻击,方法首先通过动态参数估计从轨迹历史中准确估计出车辆的动态特征,然后用估计的动态参数,框架能够高效地通过梯度优化过程生成对抗性轨迹。这些对抗性轨迹旨在以一种看似合理但实际上会误导轨迹预测系统的方式改变车辆的行驶路径。近年的研究进一步将优化目标从单车系统推广到多智能体交互与车路协同系统。Wang 等人(2021b)提出 AdvSim,采用黑盒对抗攻击方法模拟真实交通环境中的多智能体行为,生成复杂且逼真的安全关键场景,用于测试自动驾驶系统的安全性和鲁棒性。Zhang 等人(2022)提出通过白盒(梯度下降)或黑盒(粒子群优化策略)的对抗攻击框架巧妙地调整正常车辆轨迹,从而最大化预测误差,以模拟潜在攻击场景。为了确保这些对抗性轨迹的现实可行性,其在优化过程中加入了对物理属性的严格约束,以保证这些轨迹可以被实际的车辆复现。为增强优化过程中的物理一致性,Hanselmann 等人(2022)提出了 KING 方法,它使用运动学梯度来生成安全关键的驾驶场景。KING 通过迭代更新初始场景来诱导与自动驾驶代理的碰撞。随着车路协同系统的兴起,部分研究开始将优化对象扩展至多主体和通信网络维度。Xiang 等人(2023)提出了基于车路协同的全景自动驾驶场景生成框架 V2XP-ASG,通过构建对抗性协作图,智能地选择多个智能代理的组合,并同时对这些代理的姿态进行扰动,从而生成对抗性场景。在场景生成的部署可行性方面,Chen 等人(2024)引入 FREA 方法,通过最大可行区域建模与对背景车辆的行为控制,生成既具对抗性又满足物理可执行性的关键测试场景,提升了方法在真实测试闭环中的应用潜力。

基于优化的场景生成技术能够高效构建自动驾

驶车辆的极限工况场景,但其对优化算法的性能要求较高,且生成的场景需要与实际交通场景的统计分布保持一致,以确保实验结果的可信度。

4.3 语义驱动的场景布局生成

语义驱动的场景布局生成方法强调从高层语义信息出发建构测试环境,通过引入语言、目标任务、行为意图等外部约束,实现对交通场景的结构化控制。相较于数据驱动的模式生成与优化式扰动,语义驱动方法更适用于构造具备任务相关性、交互合理性与可解释性的测试样本,已逐步成为提高安全评测针对性与系统一致性的重要路径之一。现有研究大致沿着两条技术演化路线展开:1)基于语言或任务条件的语义场景生成,旨在实现自然语言或任务约束下的高层到低层映射控制;2)以世界模型或占用建模为核心的语义空间表示学习,用于统一交通参与者、动态环境与几何结构的信息表征。

在第1类方向中,研究者尝试将自然语言或任务级行为目标引入场景生成流程,从而实现用户意图驱动下的交通模拟控制。Zhang等人(2023)提出了一个基于运动预测和端到端驾驶的多智能体策略TrafficBots,通过引入导航目的地和行为风格,生成可配置和可扩展的交通参与者行为,为自动驾驶规划模块提供了一个定制化的模拟环境。在此基础上,为进一步提升语义条件的表达能力与建模粒度,Zhong等人(2023b)提出了CTG++方法,作为其前期GAN轨迹生成模型CTG(Zhong等,2023a)的扩展,利用大型语言模型将自然语言指令转换为可微损失函数,引导扩散模型生成符合用户查询的交通模拟。这项工作不仅展示了通过自然语言控制交通模拟的潜力,也为连接模拟与文本数据提供了新的途径。

与之互补的是第2类研究路径,即以空间占用预测或世界状态建模为核心,构建统一的场景语义表示。Zheng等人(2024)提出了一个3D占用世界模型,该模型不仅预测自动驾驶车辆的运动,还预测周围场景的演变,包括动态和静态元素。这种方法通过学习3D占用空间中的世界观来同时预测自我车辆的轨迹和场景的发展,展示了在复杂的交通场景中进行细粒度预测的能力。为进一步推进语义建模的通用性与可迁移性,Min等人(2023)提出了一个统一的预训练框架——UniWorld,通过预测4D几何占用作为世界模型的基础阶段,然后在下游任务中进行微调。该方法能够估计关于世界状态的缺失信

息,并预测世界的合理未来状态。这种方法的预训练过程无需标签,使得可以使用大量图像—激光雷达对来构建基础模型,显著减少了对昂贵的3D标注的依赖。

此外,部分研究关注于任务鲁棒性与语义滤波机制的引入,以增强场景生成在感知—决策链条中的可适配性。如Gao等人(2024b)提出的SEM2(semantic masked recurrent world model)模型利用语义掩码机制过滤非任务相关扰动,提高端到端自动驾驶系统在复杂场景中的泛化能力。

语义驱动的场景生成在增强自动驾驶系统的测试效率和适应性方面展现出巨大潜力。然而,其实现过程中仍面临高维语义条件的建模难题,以及如何从多模态数据中有效学习的技术挑战。

4.4 小结

本节系统梳理了面向安全评测的场景布局与行为生成技术,重点归纳生成模型、优化算法与语义驱动3类技术路径,并通过表2对典型方法的输入输出结构及生成机制进行对比分析。

生成模型方法以初始状态、控制信号与高精地图等结构化信息为输入,生成多智能体行为轨迹或完整交通场景。通过学习历史数据分布,可有效建模复杂交互模式与稀有事故场景,支持仿真训练与策略优化。但该类方法存在双重局限:一方面依赖数据质量,另一方面难以保证生成结果的物理合理性与语义一致性。优化算法驱动方法通过显式目标函数与搜索策略生成高危或对抗性测试场景。此类技术具备靶向测试优势,特别适用于鲁棒性验证与黑盒模型评估。然而受限于生成效率,且其与自然分布的偏差可能削弱评测的现实参考价值。语义驱动方法创新性地引入文本、交通规则、占用网格等高层语义作为控制信号,实现任务导向的场景构建。通过语言或语义条件引导多智能体行为生成,显著提升可控性与泛化能力。更进一步地,通过学习占用空间中的时空动态,推动世界模型层面的构建,支持细粒度行为预测与仿真建模。尽管具备高度可扩展性,但多模态对齐与语义映射的不确定性仍带来工程部署挑战。

综上所述,3类路径在方法结构与生成目标上形成互补格局,生成模型提供多样性与数据拟合能力,优化方法强化极限测试与脆弱性暴露,语义驱动则引领可控性、通用性与任务适配性的发展方向。

表2 场景布局与行为生成的主干方法与输出模态/效果(本表列举了面向安全评测的场景布局生成的代表性方法)

Table 2 Main methods and output modalities/effects for scene layout and behavior generation (this table lists representative methods for generating scene layouts for safety evaluation)				
生成方式	方法名称	生成模型架构	输入	输出
生成模型	STRIVE(Rempe等,2022)	VAE	初始状态+高清地图	安全关键场景
	Ding等人(2020)	自回归模型	初始状态+高清地图	安全关键场景
	TrafficSim(Suo等,2021)	GRU	初始状态+车辆控制+高清地图	交通场景
	SceneGen(Tan等,2021)	自回归模型	初始状态+高清地图	交通场景
	TrafficGen(Feng等,2023)	自回归模型	初始状态+高清地图	布局+轨迹
优化算法	O'Kelly等人(2018)	\	标准交通行为的分布	安全关键场景
	Klischat和Althoff(2019)	\	初始状态+高清地图	安全关键场景
	Chen等(2022)	\	初始状态+高清地图	安全关键场景
	Advdo(Cao等,2022)	\	历史轨迹+高清地图	安全关键场景
	Advsim(Wang等,2021b)	\	历史轨迹+点云+高清地图	点云表示的安全关键场景
	V2XP-ASG(Xiang等,2023)	\	街景图像+v2x数据	安全关键场景
	Trafficbots(Zhang等,2023)	世界模型	初始状态+交通信号+高清地图	交通场景
语义驱动	CTG++(Zhong等,2023b)	扩散模型	初始状态+高清地图+文本	安全关键场景
	SEM2(Gao等,2024b)	世界模型	街景图像+初始状态+高清地图	安全关键场景+街景视频
	Occworld(Zheng等,2024)	世界模型	占用网格+街景图像+点云+高清地图	3D占用预测
	Uniworld(Min等,2023)	世界模型	街景图像+点云	4D占用预测

注:优化算法类别中的方法并未使用生成模型架构,而是采用非生成式策略实现数据生成,在表中以“\”表示。

当前研究正呈现三大演进趋势,即从结构驱动转向语义控制,从静态布局拓展至动态交互,从三维空间延伸到四维时空建模。

5 挑战与未来研究方向

当前,面向自动驾驶的安全评测数据生成技术已取得显著进展,相关研究在提升数据多样性、模拟复杂场景以及识别系统潜在弱点方面展现出强大能力,逐步成为系统级安全验证的重要工具。然而,随着自动驾驶技术不断迈向高等级与实用化阶段,现有生成技术仍面临多方面挑战。首先,生成数据在语义合理性、物理可行性与现实交通条件下的适配性方面仍存在显著差距,导致其在真实部署中的参考价值有限。其次,部分方法过度聚焦于静态或单模块测试,忽略了自动驾驶系统感知、预测与决策等环节间的联动机制,难以构建具备动态响应能力与

闭环验证能力的测试流程。为进一步推动该领域的研究与落地应用,本文从评测数据的生成质量和高精度闭环测试平台的构建能力两个关键维度出发,系统梳理当前存在的核心问题,并提出面向未来的发展方向,力求为构建智能化、可信赖的自动驾驶安全测试体系提供理论支撑与技术参考。

5.1 自动驾驶安全评测数据的质量提升

5.1.1 真实性与多样性间的平衡

在自动驾驶数据生成中,真实性和多样性是两个核心目标,聚焦于不同的生成需求。真实性指的是生成数据在视觉、几何结构、物理属性等方面与现实世界的高度一致性。其目的是确保生成的数据能够准确反映实际环境中的物理规律与交互模式。然而,过度追求真实性可能导致模型过度拟合于常见场景和主流分布,难以有效覆盖稀有或极端事件,削弱系统对复杂、不确定环境的应对能力。

多样性则聚焦于生成数据在语义空间中的广

度,尤其强调生成低频事件、异常行为和长尾场景的能力。其目标是通过生成具有较大差异性和不常见特征的样本,来增强系统在面对复杂、极端条件下的适应性和鲁棒性。然而,过度追求多样性可能引发生成样本的语义偏移和物理不一致,导致生成的数据偏离真实环境的分布,使得生成样本与实际任务的需求不符。

因此,真实性与多样性之间的冲突构成了数据生成中的 trade-off 问题。追求高真实性的生成方法会限制生成数据的多样性,因为它要求样本严格遵循常见的、符合真实环境的模式,从而忽略了对稀有事件的建模;而追求多样性的生成方法则可能导致生成的样本与实际物理环境存在较大差异,使得系统无法有效接受这些样本,从而降低生成数据的有效性和应用价值。两者在目标上的相互制约,使得在实际应用中,生成过程难以在真实性和多样性之间达成最优平衡,这成为了数据生成技术面临的一个重要瓶颈。

更为复杂的是,自动驾驶系统中不同模块对真实性与多样性的定义与需求本身就存在差异。感知模块更关注通过低层物理特征的还原来反映真实性;而行为预测与决策模块则侧重时序逻辑与物理可行性的合理演化,强调语义行为的一致性与动态合理性。同一生成样本可能在感知层面高度真实,在决策层面却被视为语义异常。此外,多样性在各模块的需求上也有所不同。感知系统强调视觉要素的广泛性;而决策系统则更关注行为模式的多样性。这些差异意味着,即使是同一组生成数据,不同模块对其多样性的需求和评价标准可能截然不同。

因此,构建一个能够平衡真实性与多样性的生成目标体系,是未来数据生成研究的核心方向。一方面,应发展具备条件控制与任务适配能力的生成策略,使其能够在不同场景和任务条件下动态调整真实性与多样性的平衡机制;另一方面,可以引入多阶段语义引导的生成框架,从高层的语义布局到低层的物理建模进行分层优化,从而提高生成样本的任务适配性与系统协调性。

5.1.2 安全威胁数据的可解释性

尽管当前数据生成系统在模型性能与场景还原能力方面不断进步,但其在语义引导机制方面仍然薄弱,难以实现面向任务目标的有向生成。特别是在应对长尾风险场景、复杂交互行为与多主体决策

等高维语义任务时,现有生成策略缺乏对系统弱点的感知能力,导致样本在语义空间中呈现出随机分布,难以聚焦系统能力边界,影响测试效果与验证效率。

该问题的核心在于生成系统与自动驾驶系统之间缺乏明确的可解释性联通机制。系统在运行过程中所暴露出的决策失效、鲁棒性下降与行为异常,并未能及时转化为对生成策略进行调整的反馈信息。尤其是在端到端架构或基于预训练大模型的系统中,系统的行为决策往往缺乏足够的透明度和可追踪性,使得生成任务难以有明确的目标或方向,从而导致生成的内容与系统的实际需求之间脱节。此外,生成系统在语义空间中的表征不统一,也使得任务导向的样本构建过程难以有效进行控制和调整。

为缓解这一问题,未来研究应聚焦于从系统行为的因果路径中提取可供生成策略调度的信息。一方面,可结合中间表示、轨迹偏移、预测偏差等可解释手段,提取系统的语义薄弱区域;另一方面,应发展基于语言、图谱或大模型语义嵌入的生成策略调控机制,使得生成系统能够在多任务目标下实现动态语境适配。此外,可探索系统弱点驱动下的策略重构路径,通过因果性引导机制,将生成策略与语义风险控制结合,从而实现对高风险场景的主动建模与深度验证。

5.2 高精度仿真闭环测试环境

生成数据在自动驾驶系统中的有效性不仅依赖其表达能力与语义结构,还高度依赖其在系统测试与实际部署中的可融入性。当前技术体系尚未形成从数据生成、仿真验证到系统调优的高效闭环流程,在高保真仿真、系统反馈调度与物理部署适配等方面仍存在诸多障碍。

首先,现有生成策略大多集中于离线数据扩展,缺乏与系统测试的动态联动,生成数据与测试反馈之间无法形成有效的调控通路。系统在仿真测试中所表现出的性能瓶颈与失效路径,未能以结构化方式反馈至数据生成模块,造成生成样本在真实任务覆盖上的随机性与冗余,降低了测试效率与验证质量。要实现生成—测试—优化的闭环机制,需要打通生成策略、测试评估与系统调优之间的结构接口,并发展具备动态感知与策略调节能力的生成体系。具体而言,测试平台应具备对系统性能的语义解析能力,能够将测试输出转化为生成输入参数;生成系

统则应基于反馈信息动态调整目标空间,从而实现样本生成的任务驱动化与系统自适应增强。

其次,物理引擎驱动的仿真平台作为连接生成与验证的关键枢纽,在真实感建模、数据多模态支持、计算资源调度等方面仍存在不足。目前,已有研究者采用基于物理引擎驱动的数据生成方法。例如,Zhang 等人(2024b)提出了 LanEvil 评测基准,使用基于 Unreal Engine 4 物理引擎的 CARLA 自动驾驶仿真模拟器,构建了用于车道线检测的环境幻觉评测基准。不仅能够实现数据的快速生成采集,还支持场景的闭环测试,显著提升了评测效率和灵活性。然而现有平台在动态行为建模、交互过程演化与极端场景重构能力上表现有限,难以支持复杂交互、稀有事件及长时间演化过程的全面测试。此外,仿真平台在多模态数据协同建模方面仍缺乏一致性定义与标准化接口,阻碍了生成样本在系统各子模块间的统一适配与协同优化。

为应对上述挑战,未来应构建具备高保真物理一致性建模能力、支持多模态协同生成与语义可控调度的下一代仿真平台。平台应实现数据生成与测试评估的闭环联动,通过统一接口实现场景表达的标准化与跨模块验证流程的可追溯性。在实现层面,应引入基于物理约束的建模机制、并行化的场景生成策略及语义层级可控接口,以支撑高密度、低延迟、多场景的大规模系统测试。

最后,从虚拟空间向物理世界的生成样本迁移过程中,还需要应对生成数据与现实环境之间的差异,测试结果的可重复性以及验证流程的不足等问题。生成样本的现实适配应基于物理规则建模,通过对视觉效果、动力学行为和传感器感知等方面的高精度还原,提升生成数据在真实场景中的表现稳定性。同时,需构建将虚拟测试与物理部署结合的验证机制,采用实景回放、物理设备接口测试等方法,量化生成数据在真实系统中的有效性,从而推动数据生成技术从单一验证工具向系统优化引擎的转变。

6 结 语

随着基于深度学习的自动驾驶技术在实际场景中的不断推进,其安全性问题也日益凸显。由于真实环境的开放性与复杂性,如何有效生成高质量的

评测数据,以支撑自动驾驶系统在极端场景下的安全验证,已成为研究的核心议题。本文围绕自动驾驶工作流程中的关键环节,系统分析了当前所面临的安全挑战,并对相关的数据生成技术路径进行了全面综述。具体而言,本文重点探讨了生成模型、优化方法与语义驱动等3类主要技术在数据生成中的典型应用,归纳了各自的技术特点与适用场景。在此基础上,进一步指出了未来值得关注的发展方向,包括多模态数据融合、语义层级建模与系统反馈机制的引入,强调构建具备真实性与多样性的生成体系,以及依托高精度闭环仿真平台对于安全评测的重要意义。综上所述,安全评测数据生成作为自动驾驶系统可靠性保障的重要支撑,仍有大量值得深入探索的空间。未来,融合多模态数据与语义建模是重要方向,需要在理论建模与工程实现层面协同推进,为构建更安全、更可靠的自动驾驶系统奠定坚实基础。

参考文献(References)

- Abeyisirigoonawardena Y, Shkurti F and Dudek G. 2019. Generating adversarial driving scenarios in high-fidelity simulators//Proceedings of 2019 IEEE International Conference on Robotics and Automation (ICRA). Montreal, Canada: IEEE: 8271-8277 [DOI: 10.1109/ICRA.2019.8793740]
- Athalye A, Engstrom L, Ilyas A and Kwok K. 2018. Synthesizing robust adversarial examples//Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden: [s.n.]
- Bergamini L, Ye Y W, Scheel O, Chen L, Hu C, Del Pero L, Osinski B, Grimmett H and Ondruska P. 2021. SimNet: learning reactive self-driving simulations from real-world observations//IEEE International Conference on Robotics and Automation (ICRA). Xi'an, China: IEEE: 5119-5125 [DOI: 10.1109/ICRA48506.2021.9561666]
- Cao Y L, Wang N F, Xiao C W, Yang D W, Fang J, Yang R G, Chen Q A, Liu M Y and Li B. 2021. Invisible for both camera and LiDAR: security of multi-sensor fusion based perception in autonomous driving under physical-world attacks//Proceedings of 2021 IEEE Symposium on Security and Privacy. San Francisco, USA: IEEE: 176-194 [DOI: 10.1109/SP40001.2021.00076]
- Cao Y L, Xiao C W, Anandkumar A, Xu D F and Pavone M. 2022. AdvDO: realistic adversarial attacks for trajectory prediction//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 36-52 [DOI: 10.1007/978-3-031-20065-6_3]
- Carlini N and Wagner D. 2017. Towards evaluating the robustness of

- neural networks//IEEE Symposium on Security and Privacy. San Jose, USA: IEEE: 39-57 [DOI: 10.1109/SP.2017.49]
- Chan E R, Lin C Z, Chan M A, Nagano K, Pan B X, De Mello S, Gallo O, Guibas L, Tremblay J, Khamis S, Karras T and Wetzstein G. 2022. Efficient geometry-aware 3D generative adversarial networks//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 16102-16112 [DOI: 10.1109/CVPR52688.2022.01565]
- Chan E R, Monteiro M, Kellnhofer P, Wu J J and Wetzstein G. 2021. pi-GAN: periodic implicit generative adversarial networks for 3D-aware image synthesis//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 5795-5805 [DOI: 10.1109/CVPR46437.2021.00574]
- Chen B M, Chen X, Wu Q and Li L. 2022. Adversarial evaluation of autonomous vehicles in lane-change scenarios. *IEEE Transactions on Intelligent Transportation Systems*, 23(8): 10333-10342 [DOI: 10.1109/TITS.2021.3091477]
- Chen K Y, Lei Y H, Cheng H, Wu H R, Sun W C and Zheng S F. 2024. FREA: feasibility-guided generation of safety-critical scenarios with reasonable adversariality [EB/OL]. [2025-04-15]. <https://arxiv.org/pdf/2406.02983.pdf>
- Chen Y R, Gu C, Jiang J Z, Zhu X T and Zhang L. 2023. Periodic vibration Gaussian: dynamic urban scene reconstruction and real-time rendering [EB/OL]. [2025-04-15]. <https://arxiv.org/pdf/2311.18561.pdf>
- Chen Y Y, Tian D X, Lin C M and Yin H B. 2024. Survey of end-to-end autonomous driving systems. *Journal of Image and Graphics*, 29(11): 3216-3237 (陈妍妍, 田大新, 林椿昞, 殷鸿博. 2024. 端到端自动驾驶系统研究综述. *中国图象图形学报*, 29(11): 3216-3237) [DOI: 10.11834/jig.230787]
- Chia W M D, Keoh S L, Goh C and Johnson C. 2022. Risk assessment methodologies for autonomous driving: a survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(10): 16923-16939 [DOI: 10.1109/TITS.2022.3163747]
- Deng F, Park J and Ahn S. 2024. Facing off world model backbones: RNNs, transformers, and S4//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 72904-72930 [DOI: 10.5555/3666122.3669310]
- Deng Y, Yang J L, Xiang J F and Tong X. 2022. GRAM: generative radiance manifolds for 3D-aware image generation//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 10663-10673 [DOI: 10.1109/CVPR52688.2022.01041]
- Ding W H, Chen B M, Li B, Eun K J and Zhao D. 2021. Multimodal safety-critical scenarios generation for decision-making algorithms evaluation. *IEEE Robotics and Automation Letters*, 6(2): 1551-1558 [DOI: 10.1109/LRA.2021.3058873]
- Ding W H, Chen B M, Xu M J and Zhao D. 2020. Learning to collide: an adaptive safety-critical scenarios generating method//Proceedings of 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Las Vegas, USA: IEEE: 2243-2250 [DOI: 10.1109/IROS45743.2020.9340696]
- Ding W H, Lin H H, Li B and Zhao D. 2025. Semantically adversarial scene generation with explicit knowledge guidance. *IEEE Transactions on Intelligent Transportation Systems*, 26(2): 1510-1521 [DOI: 10.1109/TITS.2024.3510515]
- Eykholt K, Evtimov I, Fernandes E, Li B, Rahmati A, Xiao C W, Prakash A, Kohno T and Song D. 2018. Robust physical-world attacks on deep learning visual classification//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 1625-1634 [DOI: 10.1109/CVPR.2018.00175]
- Feng L, Li Q Y, Peng Z H, Tan S H and Zhou B L. 2023. TrafficGen: learning to generate diverse and realistic traffic scenarios//Proceedings of 2023 IEEE International Conference on Robotics and Automation (ICRA). London, United Kingdom: IEEE: 3567-3575 [DOI: 10.1109/ICRA48891.2023.10160296]
- Gao R Y, Chen K, Li Z H, Hong L Q, Li Z G and Xu Q. 2024a. MagicDrive3D: controllable 3D generation for any-view rendering in street scenes [EB/OL]. [2025-04-15]. <https://arxiv.org/pdf/2405.14475.pdf>
- Gao R Y, Chen K, Xie E Z, Hong L Q, Li Z G, Yeung D Y and Xu Q. 2023. MagicDrive: street view generation with diverse 3D geometry control//Proceedings of 2023 International Conference on Learning Representations (ICLR). [s.l.]: [s.n.]
- Gao Z Y, Mu Y, Chen C, Duan J L, Luo P, Lu Y F and Li S E. 2024b. Enhance sample efficiency and robustness of end-to-end urban autonomous driving via semantic masked world model. *IEEE Transactions on Intelligent Transportation Systems*, 25(10): 13067-13079 [DOI: 10.1109/TITS.2024.3400227]
- Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y. 2014a. Generative adversarial nets//Proceedings of the 28th International Conference on Neural Information Processing Systems. Cambridge, USA: MIT Press: 2672-2680 [DOI: 10.5555/2969033.2969125]
- Goodfellow I J, Shlens J and Szegedy C. 2014b. Explaining and harnessing adversarial examples [EB/OL]. [2025-04-15]. <https://arxiv.org/pdf/1412.6572.pdf>
- Gu J T, Liu L J, Wang P and Theobalt C. 2021. StyleNeRF: a style-based 3D-aware generator for high-resolution image synthesis [EB/OL]. [2025-04-15]. <https://arxiv.org/pdf/2110.08985.pdf>
- Guo J Y, Kurup U and Shah M. 2020. Is it safe to drive? an overview of factors, metrics, and datasets for driveability assessment in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 21(8): 3135-3151 [DOI: 10.1109/TITS.2019.2926042]
- Hallyburton R S, Liu Y P, Cao Y L, Mao Z M and Pajic M. 2022. Security analysis of Camera-LiDAR fusion against Black-Box attacks on

- autonomous vehicles//31st USENIX Security Symposium (USENIX Security 22). Boston, USA: USENIX Association; 1903-1920
- Han H S, Zhou K X, Long X X, Wang Y S and Xiao C X. 2024. GGS: generalizable Gaussian splatting for lane switching in autonomous driving//Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, Pennsylvania, USA: AAAI [DOI: 10.1609/aaai.v39i3.32344]
- Hanselmann N, Renz K, Chitta K, Bhattacharyya A and Geiger A. 2022. KING: generating safety-critical driving scenarios for robust imitation via kinematics gradients//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 335-352 [DOI: 10.1007/978-3-031-19839-7_20]
- Ho J, Jain A N and Abbeel P. 2020. Denoising diffusion probabilistic models//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: #574
- Hu A, Russell L, Yeo H, Murex Z, Fedoseev G, Kendall A, Shotton J and Corrado G. 2023. GAIA-1: a generative world model for autonomous driving [EB/OL]. [2025-04-15]. <https://arxiv.org/pdf/2309.17080.pdf>
- Hu X M, Huang T Y, Yu Y L, Ren J J, Xie W and Chen L. 2024. Decision technologies of simulation to reality for autonomous driving: a survey. *Journal of Image and Graphics*, 29(11): 3173-3194 (胡学敏, 黄婷玉, 余雅澜, 任佳佳, 谢微, 陈龙. 2024. 仿真到现实环境的自动驾驶决策技术综述. *中国图象图形学报*, 29(11): 3173-3194) [DOI: 10.11834/jig.230780]
- Huang L F, Gao C Y, Zhou Y Y, Xie C H, Yuille A L, Zou C Q and Liu N. 2020. Universal physical camouflage attacks on object detectors//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 717-726 [DOI: 10.1109/CVPR42600.2020.00080]
- 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 [DOI: 10.1016/j.tra.2016.09.010]
- Karras T, Aittala M, Laine S, Härkönen E, Hellsten J, Lehtinen J and Aila T. 2021. Alias-free generative adversarial networks//Proceedings of the 35th International Conference on Neural Information Processing Systems. Red Hook, USA: Curran Associates: 852-863 [EB/OL]. [2025-04-15]. https://proceedings.neurips.cc/paper_files/paper/2021/file/076ccd93ad68be51f23707988e934906-Paper.pdf
- Karras T, Laine S and Aila T. 2019. A style-based generator architecture for generative adversarial networks//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 4396-4405 [DOI: 10.1109/CVPR.2019.00453]
- Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J and Aila T. 2020. Analyzing and improving the image quality of styleGAN//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 8107-8116 [DOI: 10.1109/CVPR42600.2020.00813]
- Kim S W, Phillion J, Torralba A and Fidler S. 2021. DriveGAN: towards a controllable high-quality neural simulation//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 5816-5825 [DOI: 10.1109/CVPR46437.2021.00576]
- Kingma D P and Welling M. 2014. Auto-encoding variational bayes [EB/OL]. [2025-04-15]. <https://arxiv.org/pdf/1312.6114.pdf>
- Klischat M and Althoff M. 2019. Generating critical test scenarios for automated vehicles with evolutionary algorithms//2019 IEEE Intelligent Vehicles Symposium (IV). Paris, France: IEEE: 2352-2358 [DOI: 10.1109/IVS.2019.8814230]
- Kuefler A, Morton J, Wheeler T and Kochenderfer M. 2017. Imitating driver behavior with generative adversarial networks//IEEE Intelligent Vehicles Symposium (IV). Los Angeles, USA: IEEE: 204-211 [DOI: 10.1109/IVS.2017.7995721]
- Kurakin A, Goodfellow I J and Bengio S. 2016. Adversarial machine learning at scale [EB/OL]. [2025-04-15]. <https://arxiv.org/pdf/1611.01236.pdf>
- Kurakin A, Goodfellow I J and Bengio S. 2018. Adversarial examples in the physical world [EB/OL]. [2025-04-15]. <https://arxiv.org/pdf/1607.02533.pdf>
- Li X F, Zhang Y F and Ye X Q. 2024. DrivingDiffusion: layout-guided multi-view driving scenarios video generation with latent diffusion model//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 469-485 [DOI: 10.1007/978-3-031-73229-4_27]
- Li X Y, Ye Z H, Wei S K, Chen Z, Chen X T, Tian Y H, Dang J W, Fu S J and Zhao Y. 2023. 3D object detection for autonomous driving from image: a survey—benchmarks, constraints and error analysis. *Journal of Image and Graphics*, 28(6): 1709-1740 (李熙莹, 叶芝桢, 韦世奎, 陈泽, 陈小彤, 田永鸿, 党建武, 付树军, 赵耀. 2023. 基于图像的自动驾驶3D目标检测综述——基准、制约因素和误差分析. *中国图象图形学报*, 28(6): 1709-1740) [DOI: 10.11834/jig.230036]
- Lillicrap T P, Hunt J J, Pritzel A, Heess N, Erez T, Tassa Y, Silver D and Wierstra D. 2015. Continuous control with deep reinforcement learning [EB/OL]. [2025-04-15]. <https://arxiv.org/pdf/1509.02971.pdf>
- Lin C H, Lee H Y, Menapace W, Chai M L, Siarohin A, Yang M H and Tulyakov S. 2023. InfiniCity: infinite-scale city synthesis//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 22751-22761 [DOI: 10.1109/ICCV51070.2023.02085]
- Liu A S, Liu X L, Fan J X, Ma Y Q, Zhang A L, Xie H Y and Tao D C. 2019. Perceptual-sensitive GAN for generating adversarial patches//Proceedings of the 39th AAAI Conference on Artificial

- Intelligence. Philadelphia, Pennsylvania, USA: AAAI Press: 1028-1035 [DOI: 10.1609/aaai.v33i01.33011028]
- Liu Y F, Hu X M, Chen G W, Liu S H and Chen L. 2021. Review of end-to-end motion planning for autonomous driving with visual perception. *Journal of Image and Graphics*, 26(1): 49-66 (刘旂菲, 胡学敏, 陈国文, 刘士豪, 陈龙. 2021. 视觉感知的端到端自动驾驶运动规划综述. *中国图象图形学报*, 26(1): 49-66) [DOI: 10.11834/jig.200276]
- Lou G N, Deng Y, Zheng X, Zhang M S and Zhang T Y. 2022. Testing of autonomous driving systems: where are we and where should we go?//*Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. New York, USA: Association for Computing Machinery: 31-43 [DOI: 10.1145/3540250.3549111]
- Madry A, Makelov A, Schmidt L, Tsipras D and Vladu A. 2017. Towards deep learning models resistant to adversarial attacks[EB/OL]. [2025-04-15]. <https://arxiv.org/pdf/1706.06083.pdf>
- Min C, Zhao D W, Xiao L, Nie Y M and Dai B. 2023. UniWorld: autonomous driving pre-training via world models[EB/OL]. [2025-04-15]. <https://arxiv.org/pdf/2308.07234.pdf>
- Mnih V, Kavukcuoglu K, Silver D, Rusu A A, Veness J, Bellemare M G, Graves A, Riedmiller M, Fidjeland A K, Ostrovski G, Petersen S, Beattie C, Sadik A, Antonoglou I, King H, Kumaran D, Wierstra D, Legg S and Hassabis D. 2015. Human-level control through deep reinforcement learning. *Nature*, 518 (7540): 529-533 [DOI: 10.1038/nature14236]
- Nguyen-Phuoc T, Li C, Theis L, Richardt C and Yang Y L. 2019. HoloGAN: unsupervised learning of 3D representations from natural images//*Proceedings of 2019 IEEE/CVF International Conference on Computer Vision Workshop*. Seoul, Korea (South): IEEE: 2037-2040 [DOI: 10.1109/ICCVW.2019.00255]
- Niemeyer M and Geiger A. 2021. GIRAFFE: representing scenes as compositional generative neural feature fields//*Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Nashville, USA: IEEE: 11448-11459 [DOI: 10.1109/CVPR46437.2021.01129]
- O'Kelly M, Sinha A, Namkoong H, Duchi J and Tedrake R. 2018. Scalable end-to-end autonomous vehicle testing via rare-event simulation//*Proceedings of the 32nd International Conference on Neural Information Processing Systems*. Red Hook, USA: Curran Associates Inc.: 9849-9860 [DOI: 10.5555/3327546.3327650]
- OrEl R, Luo X, Shan M Y, Shechtman E, Park J J and Kemelmacher-Shlizerman I. 2022. StyleSDF: high-resolution 3D-consistent image and geometry generation//*Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New Orleans, USA: IEEE: 13493-13503 [DOI: 10.1109/CVPR52688.2022.01314]
- Pan M T, Zhu X M, Wang Y B and Yang X K. 2022. Iso-Dream: isolating and leveraging noncontrollable visual dynamics in world models//*Proceedings of the 36th International Conference on Neural Information Processing Systems*. Red Hook, USA: Curran Associates Inc.: 23178-23191
- Pan X G, Xu X D, Loy C C, Theobalt C and Dai B. 2021. A shading-guided generative implicit model for shape-accurate 3D-aware image synthesis//*Proceedings of the 35th International Conference on Neural Information Processing Systems*. Red Hook, USA: Curran Associates Inc.: 20002-20013
- Pomerleau D A. 1989. ALVINN: an autonomous land vehicle in a neural network//*Proceedings of the 2nd International Conference on Neural Information Processing Systems*. Cambridge, USA: MIT Press: 305-313 [DOI: 10.5555/2969735.2969771]
- Regni K and Borji A. 2018. Cross-view image synthesis using conditional GANs//*Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, USA: IEEE: 3501-3510 [DOI: 10.1109/CVPR.2018.00369]
- Rempe D, Phillion J, Guibas L J, Fidler S and Litany O. 2022. Generating useful accident-prone driving scenarios via a learned traffic prior//*Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New Orleans, USA: IEEE: 17284-17294 [DOI: 10.1109/CVPR52688.2022.01679]
- Ross S, Gordon G and Bagnell D. 2011. A reduction of imitation learning and structured prediction to no-regret online learning//*Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*. Ft. Lauderdale, USA: Springer: 627-635 [DOI: 10.48550/arXiv.1011.0686]
- Sallab A E, Abdou M, Perot E and Yogamani S. 2017. Deep reinforcement learning framework for autonomous driving. *Electronic Imaging*, 29 (19): 70-76 [DOI: 10.2352/ISSN.2470-1173.2017.19.AVM-023]
- Sato T, Shen J J, Wang N F, Jia Y H, Lin X and Chen Q A. 2021. Dirty road can attack: security of deep learning based automated lane centering under physical-world attack//*30th USENIX security symposium (USENIX Security 21)*. [s.l.]: USENIX Association: 3309-3326
- Schwarz K, Liao Y Y, Niemeyer M and Geiger A. 2020. GRAF: generative radiance fields for 3D-aware image synthesis//*Proceedings of the 34th International Conference on Neural Information Processing Systems*. Vancouver, Canada: Curran Associates Inc.: 20154-20166 [DOI: 10.5555/3495724.3497416]
- Shi Z F, Xu Y H, Shen Y J, Zhao D L, Chen Q F and Yeung D Y. 2022. Improving 3D-aware image synthesis with a geometry-aware discriminator//*Proceedings of the 36th International Conference on Neural Information Processing Systems*. New Orleans, USA: Curran Associates Inc.: 7921-7932
- Sohl-Dickstein J, Weiss E A, Maheswaranathan N and Ganguli S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics//*Proceedings of the 32nd International Conference on Machine Learning*. Lille, France: JMLR.org: 2256-2265 [DOI: 10.5555/3045118.3045358]

- Suo S, Regalado S, Casas S and Urtasun R. 2021. TrafficSim: learning to simulate realistic multi-agent behaviors//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 10395-10404 [DOI: 10.1109/CVPR46437.2021.01026]
- Suo S, Wong K, Xu J, Tu J, Cui A, Casas S and Urtasun R. 2023. MIXSIM: a hierarchical framework for mixed reality traffic simulation//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 9622-9631 [DOI: 10.1109/CVPR52729.2023.00928]
- Suryanto N, Kim Y, Kang H, Larasati H T, Yun Y, Le T T H, Yang H M, Oh S Y and Kim H. 2022. DTA: physical camouflage attacks using differentiable transformation network//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 15284-15293 [DOI: 10.1109/CVPR52688.2022.01487]
- Swerdlow A, Xu R S and Zhou B L. 2024. Street-view image generation from a bird's-eye view layout. IEEE Robotics and Automation Letters, 9(4): 3578-3585 [DOI: 10.1109/LRA.2024.3368234]
- Szegedy C, Vanhoucke V, Ioffe S, Shlens J and Wojna Z. 2016. Rethinking the inception architecture for computer vision//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 2818-2826 [DOI: 10.1109/CVPR.2016.308]
- Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow and Fergus R. 2013. Intriguing properties of neural networks [EB/OL]. [2025-04-15]. <https://arxiv.org/pdf/1312.6199.pdf>
- Tan S H, Wong K, Wang S L, Manivasagam S, Ren M Y and Urtasun R. 2021. SceneGen: learning to generate realistic traffic scenes//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 892-901 [DOI: 10.1109/CVPR46437.2021.00095]
- Tu J, Li H C, Yan X C, Ren M Y, Chen Y, Liang M, Bitar E, Yumer E and Urtasun R. 2021. Exploring adversarial robustness of multi-sensor perception systems in self driving//Proceedings of the 5th Conference on Robot Learning (CoRL 2021). London, UK: [s.n.]
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: MIT Press: 6000-6010 [DOI: 10.5555/3295222.3295349]
- Wang D H, Jiang T S, Sun J L, Zhou W E, Gong Z Q, Zhang X Y, Yao W and Chen X Q. 2022. FCA: learning a 3D full-coverage vehicle camouflage for multi-view physical adversarial attack//Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, Pennsylvania: AAAI Press: 2414-2422 [DOI: 10.1609/aaai.v36i2.20141]
- Wang J K, Liu A S, Yin Z X, Liu S C, Tang S Y and Liu X L. 2021a. Dual attention suppression attack: generate adversarial camouflage in physical world//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 8561-8570 [DOI: 10.1109/CVPR46437.2021.00846]
- Wang J K, Pun A, Tu J, Manivasagam S, Sadat A, Casas S, Ren M Y and Urtasun R. 2021b. AdvSim: generating safety-critical scenarios for self-driving vehicles//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 9904-9913 [DOI: 10.1109/CVPR46437.2021.00978]
- Wang X F, Zhu Z, Huang G, Chen X Z, Zhu J G and Lu J W. 2023. DriveDreamer: towards real-world-drive world models for autonomous driving//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer [DOI: 10.1007/978-3-031-73195-2_4]
- Wang Y Q, He J W, Fan L, Li H X, Chen Y T and Zhang Z X. 2024. Driving into the future: multiview visual forecasting and planning with world model for autonomous driving//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 14749-14759 [DOI: 10.1109/CVPR52733.2024.01397]
- Wei J L, Yuan S S, Li P F, Hu Q D, Gan Z X and Ding W C. 2024. OccLLaMA: an occupancy-language-action generative world model for autonomous driving [EB/OL]. [2025-04-15]. <https://arxiv.org/pdf/2409.03272.pdf>
- Wen Y Q, Zhao Y C, Liu Y F, Jia F, Wang Y H, Luo C, Zhang C, Wang T C, Sun X Y and Zhang X Y. 2024. Panacea: panoramic and controllable video generation for autonomous driving//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 6902-6912 [DOI: 10.1109/CVPR52733.2024.00659]
- Wiles O, Gkioxari G, Szeliski R and Johnson J. 2020. SynSin: end-to-end view synthesis from a single image//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 7465-7475 [DOI: 10.1109/CVPR42600.2020.00749]
- Xiang H, Xu R S, Xia X, Zheng Z L, Zhou B L and Ma J Q. 2023. V2XP-ASG: generating adversarial scenes for vehicle-to-everything perception//Proceedings of 2023 IEEE International Conference on Robotics and Automation (ICRA). London, United Kingdom: IEEE: 3584-3591 [DOI: 10.1109/ICRA48891.2023.10161384]
- Xiang J N, Liu G Y, Gu Y, Gao Q Y, Ning Y T, Zha Y H, Feng Z Y, Tao T H, Hao S B, Shi Y M, Liu Z Z, Xing E P and Hu Z T. 2024. Pandora: towards general world model with natural language actions and video states [EB/OL]. [2025-04-15]. <https://arxiv.org/pdf/2406.09455.pdf>
- Xiao C W, Li B, Zhu J Y, He W, Liu M Y and Song D. 2018. Generating adversarial examples with adversarial networks//Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, Sweden: AAAI Press [DOI: 10.5555/3304222.3304312]

- Xu X D, Pan X G, Lin D H and Dai B. 2021. Generative occupancy fields for 3D surface-aware image synthesis//Proceedings of the 35th International Conference on Neural Information Processing Systems. Red Hook, USA: Curran Associates Inc.: 20683-20695 [DOI: 10.5555/3540261.3541843]
- Xu Y H, Chai M L, Shi Z F, Peng S D, Skorokhodov I, Siarohin A, Yang C Y, Shen Y J, Lee H Y, Zhou B L and Tulyakov S. 2023. DisCoScene: spatially disentangled generative radiance fields for controllable 3D-aware scene synthesis//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 4402-4412 [DOI: 10.1109/CVPR52729.2023.00428]
- Xu Y H, Peng S D, Yang C Y, Shen Y J and Zhou B L. 2022. 3D-aware image synthesis via learning structural and textural representations//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 18409-18418 [DOI: 10.1109/CVPR52688.2022.01788]
- Xue Y, Li Y H, Singh K K and Lee Y J. 2022. GIRAFFE HD: a high-resolution 3D-aware generative model//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 18419-18428 [DOI: 10.1109/CVPR52688.2022.01789]
- Yurtsever E, Yang D F, Koc I M and Redmill K A. 2022. Photorealism in driving simulations: blending generative adversarial image synthesis with rendering. IEEE Transactions on Intelligent Transportation Systems, 23 (12): 23114-23123 [DOI: 10.1109/ITITS.2022.3193347]
- Zhang Q H, Xu Y H, Shen Y J, Dai B, Zhou B L and Yang C Y. 2024a. BerfScene: Bev-conditioned equivariant radiance fields for infinite 3D scene generation//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 6839-6849 [DOI: 10.1109/CVPR52733.2024.00653]
- Zhang Q Z, Hu S T, Sun J C, Chen Q A and Mao Z M. 2022. On adversarial robustness of trajectory prediction for autonomous vehicles//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 15138-15147 [DOI: 10.1109/CVPR52688.2022.01473]
- Zhang T Y, Wang L, Li H N, Xiao Y S, Liang S Y, Liu A S, Liu X L and Tao D C. 2024b. LanEvil: benchmarking the robustness of lane detection to environmental illusions//Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, Australia: Association for Computing Machinery: 5403-5412 [DOI: 10.1145/3664647.3680761]
- Zhang Y, Foroosh H, David P and Gong B Q. 2018. CAMOU: learning physical vehicle camouflages to adversarially attack detectors in the wild [EB/OL]. [2025-04-15]. <https://openreview.net/pdf?id=SJgEl3A5tm>
- Zhang Z J, Liniger A, Dai D X, Yu F and Van Gool L. 2023. TrafficBots: towards world models for autonomous driving simulation and motion prediction//Proceedings of 2023 IEEE International Conference on Robotics and Automation (ICRA). London, United Kingdom: IEEE: 1522-1529 [DOI: 10.1109/ICRA48891.2023.10161243]
- Zhao G S, Ni C J, Wang X F, Zhu Z, Zhang X Y, Wang Y D, Huang G, Chen X Z, Wang B Y, Zhang Y Y, Mei W J and Wang X G. 2024b. DriveDreamer4D: world models are effective data machines for 4D driving scene representation [EB/OL]. [2025-04-15]. <https://arxiv.org/pdf/2410.13571.pdf>
- Zhao G S, Wang X F, Zhu Z, Chen X Z, Huang G, Bao X Y and Wang X G. 2024a. DriveDreamer-2: LLM-enhanced world models for diverse driving video generation [EB/OL]. [2025-04-15]. <https://arxiv.org/pdf/2403.06845.pdf>
- Zheng W Z, Chen W L, Huang Y H, Zhang B R, Duan Y Q and Lu J W. 2024. OceWorld: learning a 3D occupancy world model for autonomous driving//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 55-72 [DOI: 10.1007/978-3-031-72624-8_4]
- Zhong Z Y, Rempe D, Chen Y X, Ivanovic B, Cao Y L, Xu D F, Pavone M and Ray B. 2023b. Language-guided traffic simulation via scene-level diffusion//Proceedings of the 7th Conference on Robot Learning (CoRL 2023). Atlanta, USA: PMLR: 144-177
- Zhong Z Y, Rempe D, Xu D F, Chen Y X, Veer S, Che T, Ray B and Pavone M. 2023a. Guided conditional diffusion for controllable traffic simulation//Proceedings of 2023 IEEE International Conference on Robotics and Automation (ICRA). London, United Kingdom: IEEE: 3560-3566 [DOI: 10.1109/ICRA48891.2023.10161463]
- Zhou Y S, Simon M, Peng Z H, Mo S C, Zhu H Z, Guo M Y and Zhou B L. 2024. SimGen: simulator-conditioned driving scene generation [EB/OL]. [2025-04-15]. <https://arxiv.org/pdf/2406.09386.pdf>

作者简介

刘江帆,男,博士研究生,主要研究方向为可信人工智能。

E-mail: liujiangfan@buaa.edu.cn

刘艾杉,通信作者,男,副教授,主要研究方向为对抗样本、模型鲁棒性、可信人工智能。E-mail: liuaishan@buaa.edu.cn

张天缘,男,博士研究生,主要研究方向为可信人工智能。

E-mail: zhangtianyuan@buaa.edu.cn

钟芳桢,男,本科生,主要研究方向为可信人工智能。

E-mail: zfz0411@buaa.edu.cn

岳鹏,男,本科生,主要研究方向为可信人工智能。

E-mail: yuep6@mail2.sysu.edu.cn

刘祥龙,男,教授,主要研究方向为计算机视觉与可信人工智能。E-mail: xlliu@buaa.edu.cn