中图法分类号： 文献标识码：A 文章编号：1006-8961(XXXX)XX-0001-28

论文引用格式：Zhao Yao, Li Jia, Jin Yi, Wei Yunchao, Zhao Yifan, Zhang Hui, Wang Xu, Qu Mengxue, Zeng Yuqiao, Wang Wenzhuang. XXXX. A Survey on Intelligent Generation of Traffic Data Towards Advanced Smart Driving: Models, Systems, and Evaluation. Journal of Image and Graphics, XX(XX):0001-0028(赵耀, 李甲, 金一, 魏云超, 赵一凡, 张慧, 王旭, 瞿梦雪, 曾宇乔, 王文状. XXXX. 面向高阶智驾的交通数据智能生成:模型、系统与评测综述. 中国图象图形学报, XX(XX):0001-0028)[DOI:10.11834/jig.250644]

面向高阶智驾的交通数据智能生成:模型、系统与评测综述

赵耀[1],李甲[2*],金一[1],魏云超[1],赵一凡[2],张慧[1],王旭[1],瞿梦雪[1],曾宇乔[1],王文状[2]

1. 北京交通大学,北京市,100044; 2. 北京航空航天大学,北京市,100191

摘 要：随着高阶智能驾驶对多模态感知、预测与决策的依赖不断提升,真实交通数据在极端天气、长尾场景与隐私敏感环境下面临采集成本高、覆盖不足和标注困难等瓶颈,难以支撑系统规模化训练与验证。如何高效生成具备真实感与可控性的交通数据,以提升系统在极端情形下的可靠性,已成为亟待解决的关键问题。基于此,本文对面向高阶智驾的交通数据智能生成技术开展系统综述,旨在把握研究进展并指引工程化实践。首先,我们引入模型—系统—评测的典型流程,定义并分析了当前面临的数据稀缺、跨模态对齐、条件可控、场景一致性与闭环验证等核心挑战;随后,围绕扩散模型、生成对抗网络、神经辐射场/三维高斯泼溅、世界模型与多模态大模型等技术脉络,系统梳理了代表性生成方法及其在智能座舱、单车智驾与基于车路协同感知的多车协同感知三大应用方向中的关键应用与技术要点;最后,提出了一套覆盖感知—预测—控制闭环度量与传感器物理一致性的多层次评测框架,并讨论了构建兼具真实性与多样性的工程化数据引擎的若干实践建议。本文提及的算法、数据集和评估指标已汇总至https://github.com/fayewong666999/higher-level-smart-driving-data-generation。本文力图为高阶智能驾驶的数据体系构建、评测规范与后续技术演进提供系统参考。

关键词：模型-系统-评测;高阶智能驾驶;数据智能生成;智能座舱;单车智驾;多车协同感知;多层次评测

A Survey on Intelligent Generation of Traffic Data Towards Advanced Smart Driving: Models, Systems, and Evaluation

Zhao Yao[1], Li Jia[2*], Jin Yi[1], Wei Yunchao[1], Zhao Yifan[2], Zhang Hui[1], Wang Xu[1], Qu Mengxue[1], Zeng Yuqiao[1], Wang Wenzhuang[2]

1. Beijing Jiaotong University, Beijing, 100044; 2. Beihang University, Beijing, 100191

Abstract: As higher-level smart driving continues to advance, it increasingly relies on multimodal perception, predictive modeling, and intelligent decision-making. However, the acquisition of real-world traffic data faces substantial challenges, especially under extreme weather conditions, rare or long-tail scenarios, and privacy-sensitive contexts. These challenges manifest as high data collection costs, insufficient scenario coverage, and labor-intensive labeling processes, which collectively hinder the ability to support large-scale training, validation, and deployment of smart driving systems. Consequently, the efficient generation of traffic data that simultaneously exhibits realism, controllability, and diversity has

收稿日期:2025-12-23;修回日期:2026-01-31

*通信作者:李甲 jiali@buaa.edu.cn

基金项目:国家自然科学基金(62132002, 62571027, 92470203, 62203040)

Supported by:National Natural Science Foundation of China (62132002, 62571027, 92470203, 62203040)

emerged as a crucial research problem, with significant implications for both safety and system reliability in extreme or unforeseen driving conditions. To address these pressing challenges, this paper presents a comprehensive survey of intelligent traffic data generation techniques specifically targeting high-level smart driving applications. The survey aims to provide a structured overview of the state-of-the-art methods, identify the core technical bottlenecks, and outline best practices for translating research into engineering applications. We organize the discussion along three interrelated dimensions—models, systems, and evaluation—forming a holistic perspective that links generative algorithms with practical deployment considerations and quantitative assessment protocols. We begin by introducing a model-system-evaluation workflow that clarifies the central technical challenges faced in traffic data generation. Key issues include data scarcity, which limits the capacity to model rare events; cross-modal alignment, which ensures consistent mapping between visual, spatial, and textual modalities; conditional controllability, enabling flexible generation under user-specified constraints; scene consistency, which preserves realistic spatial and temporal correlations; and closed-loop validation, which evaluates generated data in the context of perception-prediction-control feedback loops. These challenges not only reflect fundamental research questions but also directly impact the robustness and generalizability of smart driving systems. Following the problem definition, we systematically review representative generative techniques that have been applied in this domain. Our survey covers several prominent methodological families, including diffusion models, generative adversarial networks (GANs), neural radiance fields (NeRF) and 3D gaussian splatting (3DGS), world models, and multimodal foundation models. For each category, we discuss the underlying principles, highlight recent advances, and examine their suitability for generating high-fidelity, controllable traffic data. Special attention is given to the integration of spatial and temporal priors, multi-agent interactions, and semantic guidance, which are critical for ensuring that synthetic data faithfully reflect real-world driving dynamics. We further categorize applications into three principal domains: intelligent cockpits, single-vehicle autonomy, and V2X-based cooperative perception. In intelligent cockpits, generated data can support driver assistance systems and human-machine interface evaluation, enabling the study of driver behavior and risk perception under controlled yet realistic conditions. For single-vehicle smart driving, synthetic data facilitate model training for perception, prediction, and planning modules, especially in scenarios that are rare or safety-critical, such as pedestrian crossing at night or severe weather conditions. In multi-vehicle cooperative perception leveraging V2X communication, generated datasets allow for systematic exploration of sensor fusion strategies, information sharing protocols, and distributed decision-making under varying network and environmental conditions. Across these domains, we identify key technical considerations, including sensor modality alignment, occlusion handling, and fidelity of dynamic interactions among agents. In addition to algorithmic techniques and application scenarios, evaluation and benchmarking constitute an essential component of intelligent traffic data generation. We propose a multi-level evaluation framework that spans perception-prediction-control closed-loop metrics, physical consistency of sensors, and scenario diversity measures. This framework not only assesses the visual realism of synthetic data but also quantifies their utility in downstream smart driving tasks, ensuring that generated datasets contribute meaningfully to system reliability and safety validation. Moreover, we discuss engineering practices for constructing scalable data engines that balance realism, diversity, and controllability, offering practical insights into data augmentation strategies, hybrid synthetic-real data pipelines, and scenario generation workflows. In summary, this survey provides a structured and comprehensive reference for researchers and practitioners working on traffic data generation for advanced smart driving. By integrating perspectives from models, systems, and evaluation, it highlights both the progress made and the remaining challenges in generating high-quality, controllable, and task-relevant traffic data. We envision that such insights will support the development of robust data ecosystems, inform the establishment of standardized evaluation protocols, and ultimately accelerate the safe deployment of high-level smart driving technologies in real-world environments.

Key words: model-system-evaluation; high-level smart driving; intelligent data generation; intelligent cockpit; vehicle-centric smart driving; multi-vehicle cooperative perception; multi-level evaluation

0 引 言

随着高阶智能驾驶系统对多模态感知、预测与决策能力的依赖不断增强，真实交通数据在复杂场景下的获取难度也显著提升。特别是在恶劣天气、高动态交通、长尾事件或涉及隐私敏感的场景中，数据采集不仅成本高昂，而且覆盖面有限，标注工作也存在巨大挑战。这些因素直接制约了自动驾驶系统的规模化训练与验证，使得模型在极端或罕见场景下的表现难以得到充分保障。为应对这一瓶颈，如何高效生成兼具真实性与可控性的交通数据，已成为支撑高阶智能驾驶技术发展的关键环节。在这一背景下，生成式数据不仅能够补充现实数据的不足，还可系统构造难以自然获取的长尾和极端驾驶场景，为感知、预测和决策模块提供丰富的训练样本。通过模拟不同光照、天气、道路状况以及多主体交互情境，生成数据可以有效评估系统在高风险环境下的鲁棒性和安全性，从而为算法优化和闭环验证提供可重复、可量化的支撑。随着生成对抗网络（Karras 等，2019；Karras 等，2020）、扩散模型（Rombach 等，2022）等生成技术的发展，这些方法逐步成为推动高阶智能驾驶数据生成创新的关键力量，高阶智能驾驶体系包括智能座舱（intelligent cockpit）、单车智驾（vehicle-centric smart driving）及车路协同感知（vehicle-to-everything，V2X）等核心模块。

智能座舱（intelligent cockpit）作为高阶智能驾驶体系中车辆内部的人机协同核心，同样受到隐私敏感、极端状态难复现和乘员差异显著等因素的制约，导致真实多模态数据难以规模化获取，从而限制了驾驶员状态监测、车内决策辅助与自然交互等关键能力的模型训练。因此，基于生成模型的多模态座舱数据构建正成为弥补数据短板、提升座舱智能化水平的重要方向。围绕这一需求，当前国际与国内在语音、视觉与系统状态等方面开展了多种探索。语音方向中，传统“自动语音识别（automatic speech recognition，ASR）-大语言模型（large language model，LLM）-文本到语音合成（text-to-speech，TTS）”流水线难以保持情绪与语气等副语言信息，且存在较高延迟，而端到端语音语言模型，如 GSLM（Lakhotia 等，2021）、AudioLM（Hassid 等，2024）、AudioPaLM（Rubenstein 等，2023）等，虽提升了生成质量，但仍受到车内噪声和阵列差异的影响。视觉方向上，GAN 与扩散模型推动了人脸与动作生成的发展（Karras 等，2019；Karras 等，2020；Mokady 等，2023），然而受限于座舱空间狭小与频繁遮挡，动作约束、视角一致性以及光照建模仍不充分。系统状态方面，由于设备耦合度高，传统功能调用效率低且容错性弱（Zhang 等，2024），现有基于状态预测的尝试尚未形成覆盖多设备、多状态的体系化生成框架。这表明，智能座舱数据生成仍缺乏兼顾多模态一致性、物理约束与规模化能力的成熟体系。

单车智驾（vehicle-centric smart driving）作为高阶智能驾驶系统的核心研究与应用形态，其任务体系覆盖从感知—预测—决策—控制的完整闭环，涉及对多源信息的高精度理解与实时响应。该体系性能高度依赖大规模、多样化且高质量的交通场景数据，包括图像、视频与点云等多模态数据来源。然而，现实世界复杂交通环境的数据往往难以全面采集，尤其在恶劣天气、多光照及高动态交通要素交互下，不确定性行为频发，使长尾（long-tail）与极端场景（corner case）数据的获取成本高昂且覆盖不足。这一数据鸿沟直接限制了感知模型的泛化能力与稳健性，成为制约单车智驾系统性能持续提升的关键瓶颈。现有研究尝试通过场景语义驱动、空间结构驱动以及多模态条件联合驱动的数据生成模型，在保证物理合理性与语义一致性的前提下扩展训练数据覆盖，但多模态一致性、时空连续性与罕见场景覆盖仍存在挑战。

车路协同感知（vehicle-to-everything，V2X）技术通过车辆与路侧设施等多智能体的信息共享，显著拓展单车智能的感知边界，有效应对遮挡、盲区及远距离目标探测等关键问题。然而，大规模高质量的真实 V2X 数据获取面临严峻困难：多智能体需在同一场景下同步采集数据，导致现有协同数据集在规模、场景多样性及参与智能体数量上存在明显局限。为突破这一瓶颈，基于生成式技术构建逼真、多样的 V2X 数据成为重要研究方向，包括将单车数据集转化为协同数据集，以及通过重建与合成技术直接生成多视角感知数据，从而为 V2X 算法的训练与测试提供可持续的数据支撑。目前，该领域逐渐形成三类主流技术路径：基于扩散模型的生成式方法、基于高斯泼溅的重建编辑式方法，以及基于测试与对抗生成的构造式方法，共同推动 V2X 数据生成技术向

更高效、更逼真与更具挑战性方向发展。

高阶智能驾驶系统需要在复杂开放环境中长期稳定运行，对训练与测试数据的质量提出远高于传统视觉任务的要求。随着生成式模型和仿真平台的发展，合成交通数据成为补齐真实采集难以覆盖场景的重要途径：一方面可在可控条件下系统构造极端工况、罕见交互及高风险场景；另一方面，通过编辑与域迁移技术，可以在保护隐私前提下扩展多模态、多视角、多时序数据（Goodfellow 等，2014；Rombach 等，2022；Hu 等，2023）。然而，在缺乏系统化评测体系的情况下，即便生成能力不断提升，合成数据在可信性、可用性与安全性上的充分性仍难以得到有效论证。现有研究在评测方面存在三方面不足：其一，常用图像/视频生成指标多针对视觉逼真度和分布距离（Heusel 等，2017；Bińkowski 等，2018；Unterthiner 等，2018），难以反映数据在自动驾驶下游任务中的实际效用，也难以刻画安全性、物理合理性与隐私风险（Dwork 和 Roth，2014）；其二，大量工程实践依赖人工目检或经验规则，对评测结论客观性与可复现性保障不足，难以支撑跨系统、跨数据集与跨方法的公平比较；其三，开环指标与闭环表现之间缺乏系统联系，模型在合成数据集上的性能提升未必转化为真实道路上的安全收益，缺乏统一闭环评测协议与多维能力基准（Jia 等，2024）。针对高阶智驾这一安全敏感场景，构建系统化合成数据评测方法体系已成为生成技术落地应用的前提。本文从数据及其表征出发，逐层连接到任务表现与闭环可驾驶性，构建三层评测框架：其一，面向合成数据与真实数据的统计接近性及主观可接受性，包括人工主观评价（等级打分、对偶比较/A/B 测试、专家检验等）、统计一致性指标（均值/方差、均方误差（mean squared error，MSE）、KS 检验（Kolmogorov - Smirnov test，KS）、ECDF 差异（empirical cumulative distribution function，ECDF）、结构相似性（structural similarity index measure，SSIM）、余弦相似度（cosine similarity）等）及可视化分析方法（箱线图、小提琴图、Q - Q 图等）；其二，面向大规模合成数据集与复杂任务场景，综述基于预训练表征与任务模型的自动指标体系，包括 Fréchet Inception Distance（FID）、Kernel Inception Distance（KID）、Fréchet Video Distance（FVD）、Fréchet Pointcloud Distance（FPD）等基准模型指标，以及检测/分割 平均精度（mean Average Precision，mAP）、均值交并比（mean Intersection-over-Union，mIoU）、轨迹误差 平均位移误差（average displacement error，ADE）/最终位移误差（final displacement error，FDE）等任务或用例特定指标，分析其在分布距离、时空一致性与任务效用方面的优势与局限，并讨论如何结合场景标签、语义属性及风险类别进行细粒度评估；其三，面向实际驾驶闭环行为，构建兼顾可驾驶性、安全性、效率、舒适性、隐私及物理一致性的多维评测体系。

综上，随着智能座舱、单车智驾及 V2X 数据生成技术的逐步发展，基于生成模型的数据构建已成为高阶智能驾驶系统研究的重要方向。鉴于此，本文围绕典型高阶智能驾驶数据生成方法及其在智能座舱、单车智驾和 V2X 协同感知场景中的应用，进行了系统的调研与分析，希望为相关领域的研究人员提供参考，促进高阶智能驾驶数据生成技术的规范化、系统化发展。本文整体结构如图 1 所示，第 1 节介绍面向智能座舱的数据生成，包括基于语音/文本和人脸、人体动作的数据生成方法；第 2 节讨论面向单车智驾的数据生成模型，包括场景语义驱动、空间结构驱动及多模态条件联合驱动方法；第 3 节介绍基于 V2X 协同感知的数据生成方法，包括扩散模型、重建编辑式和对抗生成方法；第 4 节围绕高阶智驾的数据生成评测方法，涵盖基础评测、自动评测及闭环多维评测；第 5 节介绍典型数据与工具；最后总结面向高阶智驾的交通数据智能生成中目前尚存在的问题，并展望相关技术与应用的发展趋势。

1 面向智能座舱的数据生成

智能座舱作为汽车的“第三生活空间”，其智能化水平取决于对驾驶员与乘员状态的精准感知以及自然、流畅的人机交互。然而，获取用于训练模型的大规模真实数据面临隐私保护、极端状态采集风险及乘员多样性等挑战，导致现有数据集在规模与场景覆盖方面存在不足。为此，基于生成模型的数据合成逐渐成为重要方向，能够在保障隐私的同时高效生成多模态交互数据，为智能座舱感知与交互模型提供丰富样本。

本文从可能涉及的多个方向来回顾潜在的数据生成方法：一是基于语音和文本的生成，用于模拟驾驶员指令、系统提示及自然语言对话；二是基于图像

与视频的生成，用于人脸表情、人体动作和手势的模拟；三是基于系统状态的生成，用于多设备、多功能环境下的交互行为建模和状态预测。通过对这些方向的系统回顾，可以为未来多模态、以人为中心的智能座舱数据生成提供参考与研究启示。

1.1 基于语音与文本的数据生成

在智能座舱中，语音与文本是最早被用于生成的数据模态。通过自回归模型、扩散模型或大型语言模型，系统能够生成驾驶员指令、系统提示及对话内容，从而用于语音交互与状态监控模型的训练（Radford等，2023）。这类方法的核心在于将语音或文本作为生成约束，利用深度学习模型生成高质量的语音波形或文本序列。近年来，大型语言模型因其强大的文本理解与生成能力受到广泛关注，但人类的自然交互通常依赖语音，因此研究逐渐从纯文本模型转向基于语音的模型（Xie & Wu，2024；Fang等，2024）。

一种常见的实现方式是采用“自动语音识别

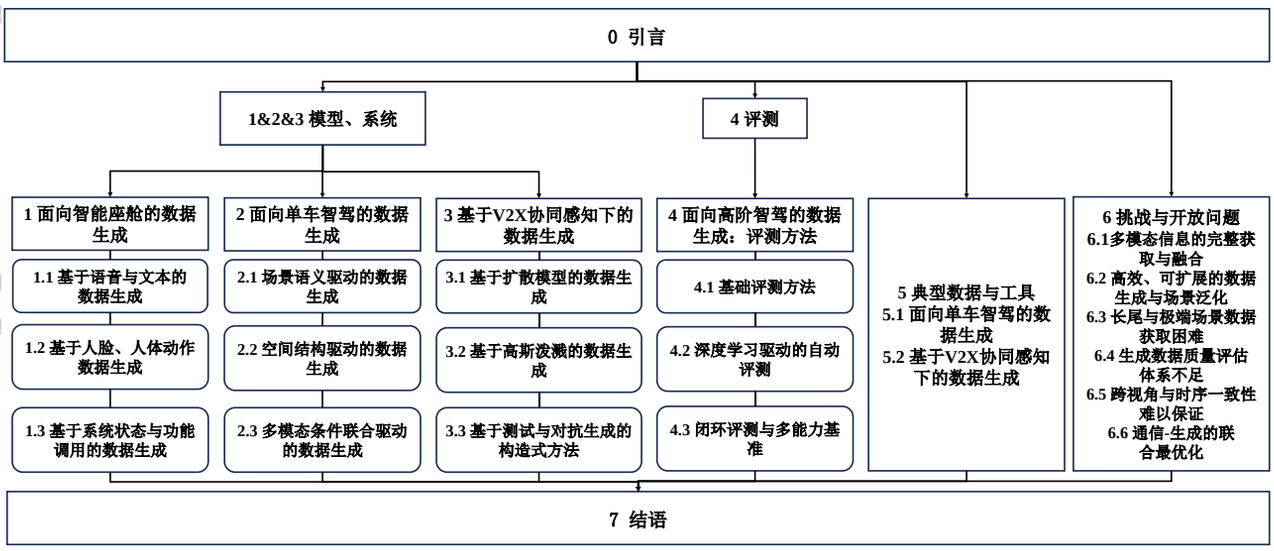


图1 本综述的章节安排

Fig. 1 The chapter arrangement of this review

（automatic speech recognition，ASR）+ LLM + 语音合成（text to speech，TTS）”的流水线结构（Radford等，2023），即将输入语音经ASR转录为文本，由LLM生成文本响应，再由TTS将其合成为语音输出。虽然这一方法结构简单，但在实际应用中存在明显局限。首先，语音信号不仅包含语义信息，还承载着诸如音高、音色与语调等副语言特征，而这些在仅使用文本的处理中会完全丢失，从而削弱模型对语气、情感及意图的理解能力（Le等，2020）。其次，ASR、LLM与TTS的顺序操作导致系统延迟较高，因模块结构复杂、计算量大而难以实现实时响应。最后，模块化的设计使得各阶段误差容易累积，ASR转录错误会影响LLM的文本生成，而LLM输出的不确定性也可能导致TTS合成质量下降。因此，构建能够直接从语音到语音的统一架构成为提升语音交互质量的关键方向。

为解决上述问题，端到端的语音语言模型（SpeechLMs）逐渐成为研究热点（Lakhotia等，2021）。SpeechLMs不再依赖文本中间层，而是直接对语音波形进行建模，通常包含语音分词器、语言模型与标记转语音合成器三部分。不同模型的主要差异在于其对语音标记的处理与建模方式。早期的生成式口语语言模型（generative spoken language model，GSLM，Lakhotia等，2021）是该领域的开创性工作，通过冷启动训练验证了端到端语义建模的可行性。随后，TWIST和AudioPaLM（Rubenstein等，2023）等模型引入持续预训练策略，利用预训练文本大模型（如LLaMA、PaLM）的参数进行初始化，并在交错的语音与文本标记上训练，显著提升了语音理解与生成能力。SPIRIT-LM（Nguyen等，2024）进一步在语音与文本混合建模方面取得进展，增强了跨模态协同表达的能力。为捕获更丰富的声学与副语言特征，Moshi（Defossez等，2024）提出多序列生成机制，能够并行生成语义与语音标记，在语义理解与

语音质量之间实现平衡。最终，SpeechGPT（Zhang等，2023）与SpeechGPT-Gen（Zhang等，2024）通过指令微调（instruction tuning）强化了模型的指令遵循与多轮语音对话能力。

1.2 基于人脸、人体动作数据生成

智能座舱数据生成的另一重要方向是视觉模态，包括驾驶员和乘员的人脸表情、身体姿态、手势动作，以及车内环境视频。现有方法尚未专门针对座舱场景，但可借鉴人脸生成、人体动作捕捉和视频生成技术来模拟座舱多模态交互。

在人脸生成方面，主要技术路线集中在生成对抗网络GANs（Goodfellow等，2014）和新兴的扩散模型（Diffusion Models）。以StyleGAN（Karras等，2019；2022）系列为代表的GAN架构，凭借其解耦的潜在空间（Latent Space）和分层生成能力，在生成高保真、可控人脸图像方面具有核心优势。研究人员利用GAN反演（GAN Inversion，Zhu等，2021）技术，将真实的乘员图像映射到StyleGAN的潜在空间，从而实现对特定个体面部特征（如表情、年龄）的精确语义编辑。对于座舱视频序列，时间连贯性至关重要，StyleGAN等架构通过引入等变性（Equivariance）来提升帧间的平滑度和细节运动的真实性。此外，扩散模型以其在文本到图像生成中的优越性能，正被应用于更复杂的文本条件人脸编辑（Mokady等，2023）任务，这为通过自然语言指令生成乘员的复杂情绪或注意力状态提供了新的可能性。在座舱应用中，挑战在于确保微表情等生物特征的真实性以及在车内复杂照明和视角变化下的模型鲁棒性。

人体动作生成则专注于模拟乘员在受限座舱空间内的姿态变化和交互动作。人体动作生成通常采用自回归序列模型（如基于Transformer的架构）或于变分自编码器（variational autoencoder，VAE，Holden等，2017）的方法来生成时序连续的动作序列。近年来，研究重点转向了条件动作生成，即根据文本、音频或场景上下文生成动作（Guo等，2022）。在座舱场景中，这转化为意图驱动动作的生成，如根据意图“伸手拿水杯”生成自然的动作序列。至关重要的是，这些模型必须考虑场景约束和物理真实性，确保

表1 面向单车智驾的三类数据生成方法（场景语义驱动、空间结构驱动与多模态条件联合驱动）所采用的生成模型、控制条件及输出模态总览表

Table1 Overview of Three Categories of Data Generation Methods for Vehicle-Centric Smart Driving (Scene-Semantic Driven, Spatial-Structure Driven, and Multi-Modal Condition Jointly Driven): Generative Models, Control Conditions, and Output Modalities

数据生成方式	模型名称	生成模型架构	控制输入	生成输出
场景语义驱动	BigDatasetGAN（Li等，2022）	GAN	文本	街景图像+掩码
	DiffuMask（Wu等，2023）	LDM	文本	街景图像+掩码
	DatasetDM（Wu等，2023）	LDM	文本	街景图像+标注
	SeeDiff（Park等，2024）	LDM	文本	街景图像+边框
空间结构驱动	Pix2Pix（Isola等，2017）	GAN	掩码	街景图像
	OASIS（Sushko等，2021）	GAN	掩码	街景图像
	ControlNet（Zhang等，2023）	LDM	边缘图与位姿等多样化控制条件	街景图像
	FICGen（Wang等，2025）	LDM	边框	街景图像
多模态条件联合驱动	GAIA-1（Hu等，2023）	LDM	车辆控制+文本	街景视频
	BEVGen（Swerdlow等，2024）	ARM	BEV鸟瞰图	多视角街景图像
	Magicdrive（Gao等，2023）	LDM	BEV鸟瞰图+3D边框+相机位姿+文本	多视角街景图像
	Drive-WM（Wang等，2024）	LDM	车辆控制+文本	多视角街景视频

生成的动作序列(如驾驶员转身、乘员拿取头顶物品)不会违反座舱的物理空间限制。新兴的运动扩散模型(motion diffusion model, MDM)(Tevet等, 2022)显著提升了条件人体动作序列的质量和多样性,为在座舱中生成高保真且多样化的交互数据提供了前沿解决方案(Sofianos等, 2021)。

1.3 基于系统状态与功能调用的数据生成

智能座舱作为高度集成的多设备环境,其数据生成面临着特有的挑战,即需要智能体(Agent)协调紧密耦合的子系统。传统的功能调用(function calling, FC)方法通常以无状态方式运作,需要多次探索性调用来建立对环境的感知。这种方式不仅会导致效率低下和高延迟,而且在API调用失败时,代理难以进行错误恢复,因为它们缺乏对全局状态的宏观理解。为了应对这些挑战,研究者提出了VehicleWorld(Zhang等,2024)这一虚拟座舱环境,它是针对汽车领域的首个综合性环境,集成了30个模块、250个API和680个属性,并能在代理执行期间提供实时状态信息,支持对车辆代理行为的精确评估。基于对系统状态的建模与分析,研究发现直接状态预测在环境控制方面比功能调用更有效。因此,基于状态的功能调用(state-based function call, SFC)被提出。SFC方法显式地维护对系统当前状态的感知,通过处理用户查询和当前系统状态,直接预测目标系统状态,并生成最少且高效的转换代码来实现状态的直接转换。实验结果表明,SFC显著优于传统FC方法,在执行准确性方面表现更佳,并能减少延迟,通常需要的交互轮次和输出令牌更少。进一步分析发现,FC和SFC具有互补优势:SFC凭借其对全局环境的感知擅长设备选择,而FC则受益于高级API的封装,在复杂设备状态下能更有效地操作多个设备属性。因此,结合SFC的设备选择能力和FC的API调用优势的FC+SFC混合方法,能够实现最高的端到端准确性。这一研究方向推动了智能座舱数据生成从单纯的指令序列向基于系统状态转换的高效、高准确率的交互数据生成范式转变。

2 面向单车智驾的数据生成

单车智驾(vehicle-centric smart driving)作为高阶智能驾驶系统的核心研究与应用形态,其任务体系覆盖从感知—预测—决策—控制的完整闭环,涉及对多源信息的高精度理解与实时响应。该体系的性能高度依赖于大规模、多样化且高质量的交通场景数据,包括图像、视频与点云等多模态数据来源。然而,现实世界中复杂交通环境的数据往往难以全面采集,尤其在恶劣天气、多光照及高动态交通要素的交互下,不确定性行为频发,使得长尾(long-tail)与极端场景(corner case)数据的获取成本高昂且覆盖不足。这一数据鸿沟直接限制了感知模型的泛化能力与稳健性,成为制约单车智驾系统性能持续提升的关键瓶颈。

为缓解这一数据稀缺问题,交通数据智能生成技术近年来受到广泛关注,并被视为突破数据瓶颈、支撑单车智驾发展的重要途径。该类技术通过引入生成对抗网络(generative adversarial networks, GANs)、潜在扩散模型(latent diffusion model, LDM)、自回归生成模型(autoregressive generative model, AGM)等新兴人工智能生成内容(artificial intelligence generated content, AIGC)方法,能够构建高保真、多样性和语义空间可控的交通数据样本,涵盖多视角图像与时序视频等多种形式,从而有效弥补真实数据采集的不足,显著提升单车智驾在复杂交通场景下的识别鲁棒性与泛化能力。根据控制条件的差异,面向单车智驾的数据生成技术可大致分为三类:场景语义驱动的方法、空间结构驱动的方法以及多模态条件联合驱动的方法。表1系统列举了面向单车智驾的三类方法所采用的生成模型、控制条件及输出模态。三类方法分别基于不同类型的控制信号进行生成约束,从而生成兼具语义多样性与结构一致性的交通视觉数据,为单车智驾感知模型及其下游任务提供高保真且具互补性的多模态训练样本。

2.1 场景语义驱动的数据生成

场景语义驱动的数据生成模型聚焦根据单条场景文本提示同时生成交通图像及其真值标注,如图2所示。其核心在于利用场景文本提示中蕴含的目标类别与交互关系等语义信息,以此约束并控制生成交通场景的语义内容。早期相关研究主要基于生成对抗网络范式,该范式由生成器与判别器两部分神经网络组成:生成器负责合成接近真实数据分布的交通场景图像,而判别器用于区分真实图像与生成图像。典型方法如Li等人(2022)在DatasetGAN(Zhang等人,2021)的基础上提出BigDatasetGAN,通

过在BigGAN的潜在编码空间中实现类别多样性控制,从而扩展了ImageNet级别的数据生成能力。

然而,基于GAN的场景语义驱动方法在复杂街景生成中仍面临显著挑战。一方面,GAN训练过程存在不稳定性,生成器与判别器间的博弈易导致梯度消失,特别是在交通场景的高复杂度条件下难以收敛;另一方面,受模式坍塌等问题影响,GAN生成样本的多样性受到限制。

为克服GAN在街景图像生成中存在的局限,研究者逐渐转向训练更稳定、生成保真度更高的潜在扩散模型。浙江大学的Wu等人(2023)提出DiffuMask,该方法利用LDM中固有的文本实体与潜在特征之间的注意力信息,定位合成街景图像中的像素级目标区域,从而根据场景文本语义同时生成街景图像与语义掩码对。然而,DiffuMask在单一语义类别生成方面存在局限。为解决该问题,Nguyen等人(2024)提出Dataset Diffusion,该方法通过类别提示附加和类别提示交叉注意力实现多语义目标的高质量分割数据生成。Yoshihashi等人(2024)进一步指出DiffuMask在掩码质量和可扩展性方面存在不足,并提出Attn2mask。该方法利用潜在空间中不稳定的交叉注意力分数图作为弱监督信号,引入可靠性感知的鲁棒训练以提升合成掩码质量,并通过基于低秩适配(low-rank adaptation, LoRA)微调方法实现向智驾场景的有效生成迁移。此外,Park等人(2024)提出SeeDiff,将潜在空间中的场景语义与图像特征之间的交叉注意力作为街景目标定位的初始种子,随后类似于种子分割中的区域扩张,通过多尺度自注意力迭代地将种子区域扩散至整个街景目标类别区域。

上述场景语义驱动的数据生成方法无需任何真实街景样本或额外训练,主要依赖潜在扩散模型的预训练知识,即场景语义与图像特征之间的内在关联关系。该方法能够同时合成街景图像及其伪掩码标注,具有一定的合成效率优势。然而,由于潜在空间中的注意力映射存在不稳定性,例如“car”类别的目标区域可能被误映射到“bus”类别特征,从而降低伪掩码标注的准确性。

为缓解这一问题,研究者开始尝试利用少量真实样本进行监督训练以校正不稳定的关联关系。Wu等人(2023)提出DatasetDM,使用不到1%的真实数据训练真值解码器,将潜在空间中的不稳定映射解码为街景中多样化、高质量的语义掩码标注。基于此,Yi-Ge(2025)提出FlexDataset,通过组合式生成范式扩展生成能力,使其适用于显著目标检测、深度估计等多种下游任务。Wang等人(2025)提出FreeGen,在无需真实样本的情况下,通过两阶段训练策略利用扩散模型内在预训练知识自校正文本与街景目标之间的注意力映射。为去除低质量生成样本,Tang等人(2025)提出SDS方法,通过对比语言-图像预训练(contrastive language-image pre-training, CLIP)相似性度量及类平衡标注相似性过滤,实现高质量生成样本的保留。

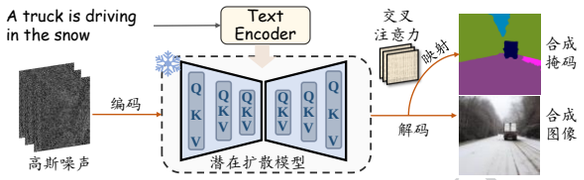


图2 场景语义驱动的数据生成管道

Fig. 2 Scene semantic-driven data generation pipeline

基于场景语义驱动的数据生成方法能够利用文本提示指导街景图像的生成,但图像与文本之间的内在关联关系仍存在不稳定性。这种不稳定性主要表现为潜在空间中的注意力映射可能无法准确将文本中描述的目标类别与图像区域对应,从而导致生成图像中的目标位置或形状偏离文本提示的语义意图。此外,由于这些方法通常依赖预训练的潜在扩散模型,其生成的街景数据在视觉风格、光照条件和目标分布上与真实复杂交通场景存在显著域偏移。这种偏移进一步加剧了生成图像与其对应真值标注之间的不对齐问题,尤其在多目标、遮挡或极端天气条件下更为突出。

2.2 空间结构驱动的数据生成

尽管场景语义驱动的数据生成方法取得了一定进展,但其细粒度可控能力较弱,无法精准调控生成交通要素的空间位置和对应语义类别。为了应对该挑战,空间结构驱动的数据生成方法从智能驾驶感知任务的逆过程出发,采用边框、掩码以及深度图等视觉真值作为条件来合成交通图像,从而实现更加精细的可控数据生成,如图3所示。此类方法通常以生成图像与真值标签联合构成配对数据,既能在视觉层面保持与真实世界的结构一致性,又能直接服务于交通场景中的下游感知、分割、检测等任务,为模型提供更具一致性和可控性的训练样本。

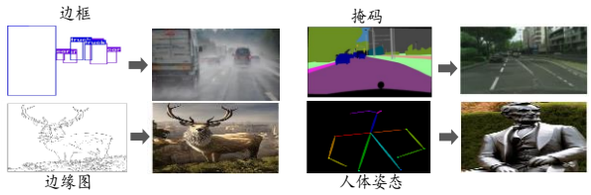


图3　空间结构驱动的图像数据生成

Fig. 3　Image data generation driven by spatial structure

类似地，早期的空间结构驱动数据生成方法多基于生成对抗网络范式。伯克利人工智能实验室的Isola等人(2017)提出Pix2Pix，通过条件GAN实现从语义掩码到交通场景图像的像素级映射，初步实现了可控生成。在此基础上，英伟达公司的Wang等人(2018)进一步提出Pix2PixHD，通过引入由粗到细的生成器和多尺度判别器结构，实现高分辨率街景图像的生成。受到StyleGAN(Karras等人，2019)启发，Sun等人(2019)提出LostGAN，引入面向对象实例的布局感知特征归一化机制，能够根据可重配置的布局和风格生成高质量、多样化图像。为了进一步提升基于GAN的街景图像生成质量，Sushko等人(2021)提出语义图像合成方法OASIS，通过空间与语义感知的判别器反馈，为生成器与判别器提供更强的监督，从而生成与语义标签图高度一致的街景图像。

潜在扩散模型凭借高保真的图像合成能力与稳定的训练过程，逐渐取代了GAN在图像生成任务中的主导地位。近年来，研究者开始尝试将空间真值标签引入扩散模型中，以实现更高质量、更可控的街景语义图像生成。Zheng等人(2023)率先提出基于潜在扩散模型的空间结构驱动数据生成方法LayoutDiffusion，通过布局融合模块与目标感知交叉注意力实现精确的可控图像生成。随后，Li等人(2023)提出GLIGEN，在开放场景下实现可控生成。该方法通过门控注意力机制将目标边界框的空间位置信息注入到冻结的潜在扩散模型中，在保留预训练生成知识的同时实现开集条件控制。为进一步统一多种空间结构输入(如边缘图、语义掩码及人体姿态)，斯坦福大学的Zhang等人(2023)提出ControlNet，其核心思想是冻结潜在扩散模型的U-Net主干，通过零卷积模块逐步注入多样化的控制条件，以生成高保真语义图像。在此基础上，Gao等人(2023)提出SCP-Diff，在街景图像合成中引入空间-类别联合先验，以缓解图像质量不足和语义不一致问题。

与上述空间结构驱动方法不同，微软的Yang等人(2023)提出ReCo，将目标边界框的空间坐标以文本形式嵌入文本编码器，并引入一组可学习的位置信标，实现区域可控的图像生成。受ReCo的编码思想启发，香港科技大学的Chen等人(2024)提出GeoDiffusion，将几何条件与相机视角编码为文本提示，并对文本引导的潜在扩散模型进行微调，用于生成服务于交通目标检测的数据样本。

尽管空间结构驱动的数据生成研究已取得显著进展，但在面对目标密集、遮挡严重的交通场景时，仍容易出现目标缺失、边界溢出及多目标合并等属性泄露问题，影响生成数据的结构一致性与语义完整性。为缓解这一问题，Zhou等人(2024)引入了多实例生成(multi-instance generation，MIG)任务，并提出空间结构驱动的数据生成方法MIGC。受“分而治之”思想启发，MIGC将整体生成过程拆解为若干实例级子任务，每个子任务独立负责单个交通目标的渲染，从而显著提升复杂场景中的实例区分度与可控性。随后，Meta公司的Wang等人(2024)提出InstanceDiffusion，进一步允许以多种形式灵活指定实例位置，如点、边框、实例掩码及线条等，为实例级图像生成提供了更具通用性的控制接口。此外，为了在雾、雨、雪等恶劣天气条件下生成高质量交通图像，北京航空航天大学的Wang等人(2025)提出FICGen，通过频率激发与上下文解耦机制将频域先验知识注入潜在空间，有效缓解前景目标与背景生成过程中的耦合与纠缠，从而生成具备更高物理一致性与真实感的退化交通图像。

基于空间结构驱动的数据生成方法能够精准构建具备几何约束的交通场景，实现目标实例在空间布局上的合理性与一致性。然而，此类方法通常需要在语义和空间层面与控制输入保持高度对齐，以确保生成内容的结构合理与语义一致。同时，当前研究普遍缺乏系统化的质量评估机制，过度依赖于弗雷歇Inception距离(Fréchet Inception Distance，FID)、核Inception距离(Kernel Inception Distance，KID)等感知层面的统计指标，难以全面反映生成数据在任务适应性与语义保真度方面的真实表现。更为关键的是，这类方法往往需要结合下游感知或决策任务进行间接验证，导致评估过程耗时且成本高昂，限制了其实用性与大规模应用潜力。

2.3 多模态条件联合驱动的数据生成

近年来，单车智驾数据生成方法旨在集成多模态异构信息来驱动多视角图像与时序视频的生成，如文本描述、相机位姿、LiDAR点云、鸟瞰视角（bird's eye view，BEV）路网图与3D边框等，以学习智驾全景的内在空间逻辑与时序语义变化，进而构建鲁棒性与泛化性能优异的智驾世界模型，支撑单车智驾场景生成、决策规划等任务。

得益于扩散模型及其衍生物（如潜在扩散模型）的变革性生成能力，单车智驾数据生成模型致力于将多模态控制条件编码进潜在空间，实现跨视角时序生成的可控性与一致性建模。早期研究主要聚焦于BEV鸟瞰视角下的多视角静态图像生成。Swerdlow等人（2024）提出一种基于自回归生成范式的BEVGen，根据交通场景下的BEV布局合成高保真且空间一致的多视角环视图像。为了进一步提升前背景的生成精度，Yang等人（2024）提出BEVControl，以BEV草图为控制输入，通过两阶段训练策略合成多视角街景图像，并设计了一套分层次评估协议，以公平评测合成交通场景中前景目标与背景几何的生成质量。为弥补高度信息的缺失，香港中文大学的Gao等人提出MagicDrive（2023），联合相机位姿、BEV路网图、3D边框与文本等多模态条件来生成多视角街景图像。其核心创新在于跨视角注意力模块，通过相邻视角间的特征交互，有效保证了多视角间的几何与语义一致性。此外，MagicDrive针对物体与道路图采用独立的编码策略，从而实现更精准的三维控制能力。在此基础上，Zhang等人（2025）提出基于多视角布局几何信息引导的生成方法PerLDiff，通过利用环视布局先验引导真实视角交通图像的掩码生成，实现目标级别的精确控制。为进一步增强对智驾场景的三维语义建模，Li等人（2025）提出基于双端条件扩散与占据射线采样（occupancy ray sampling，ORS）引导的生成方法DualDiff，通过引入三维占据表示实现对前景与背景的全面语义控制。

然而，上述方法虽然在多视角图像生成上取得了显著进展，但仍局限于静态场景的生成，缺乏对动态时序变化与连续帧间一致性的建模，难以支撑高阶单车智驾对时序预测与全景理解的需求。随着深度学习和生成式建模技术的进一步发展，视频生成驱动的智驾世界模型不断涌现，其核心目标在于同时建模多视角间与跨时间维度的全局一致性，从而学习可泛化的时空世界表示。针对这一趋势，Kim等人（2021）提出基于潜空间动态建模的DriveGAN，通过编码器与图像生成器构建潜在表示，并由动态引擎学习帧间时序变化，实现驾驶视频的可控生成。随后，极佳科技的Wang等人（2024）提出DriveDreamer，一个基于真实智驾视频与人类驾驶行为的多视角时序视频生成模型，能够基于历史观察预测未来驾驶行为，进一步提升生成结果的行为一致性。百度的Li等人（2024）提出DrivingDiffusion，从跨视角一致性、跨帧一致性以及生成质量三方面出发，采用三级级联策略实现高保真的多视角街景视频生成。同年，Wen等人（2024）提出全景视频生成模型Panacea，采用四维（4D）注意力机制与ControlNet结构，并通过两阶段训练策略显著增强了多视角与时序间的一致性。不同于前述多阶段生成策略，Wang等人（2024）提出Drive-WM，以端到端方式联合空间与时间建模，实现高保真多视角时序视频生成，并可依据驾驶行为生成多样化的未来轨迹。为进一步扩展生成数据的多样性与规模，Huang等人（2025）提出SubjectDrive，通过引入外部交通数据源，以可持续扩展的方式生成语义更丰富的多视角街景视频，从而提升模型对复杂交通场景的覆盖能力。为了进一步提升生成视频的分辨率与时长，Gao等人（2025）在MagicDrive的基础上提出了MagicDrive-V2，一种基于时空扩散Transformer的高分辨率长视频生成方法。该方法通过混合分辨率与多阶段训练策略，实现了视频质量的渐进式提升。

在时序空间建模与多条件可控生成的基础上，当前世界模型的研究趋势正由单一模态驱动逐步向多模态融合与生成理解统一转变。Hu等人（2023）提出面向智驾的生成式世界模型GAIA-1，通过引入大语言模型，融合文本、图像与动作等多模态信息，实现对智驾场景的生成与未来事件的精准预测。为进一步在三维空间中学习世界模型，Zheng等人（2023）提出OccWorld，引入三维占据表示（3D occupancy representation）替代常见的三维边框与BEV分割图，以实现更细粒度的交通场景建模。为增强模型对物理世界规律与动态变化的理解，Wang等人（2024）提出WorldDreamer，采用自回归Transformer结构结合多模态条件输入，促进不同模态间的交互建模。在DriveDreamer的基础上，Zhao等人（2024）

提出DriveDreamer-2，同样引入大语言模型以实现用户可自定义的驾驶视频生成，并结合高精地图引导学习自车轨迹与道路结构的映射关系。为了将状态生成、动作控制与奖励评估无缝集成至统一框架中，Li等人（2025）提出OmniNWM，首次实现RGB、语义图、度量深度图与三维语义占据的像素级对齐联合生成，为端到端的智驾世界建模提供了统一范式。

此外，一些研究聚焦于面向单车智驾的3D/4D交通场景生成。Gao等人（2024）提出MagicDrive3D，通过单目深度初始化与可变形高斯泼溅建模，有效缓解了跨视角间的曝光差异问题。随后，Zhao等人（2024）提出DriveDreamer4D，利用智驾世界模型的先验知识增强四维驾驶场景表示，从而生成符合真实交通规则的轨迹视频。英伟达的Lu等人（2025）进一步开发了InfiniCube，一种结合高精地图、车辆边界框及文本描述的动态3D场景生成方法，可实现大规模、可控的智驾场景合成，为自动驾驶模型的训练与测试提供了重要支持。

基于多模态条件联合驱动的单车智驾数据生成方法能够综合利用图像、文本、点云、语义标签等多源信息，从而在视觉、语义与几何层面实现更强的跨模态一致性。然而，由于不同模态间在视角、帧率及时间尺度上的差异，此类方法常面临跨视角与跨帧率不一致性问题，导致生成的街景内容在时空连续性与语义协同上存在偏差。此外，高分辨率、长时序视频的生成对算力与存储提出极高要求，限制了其实时性与可扩展性。当前研究虽尝试引入轻量化网络结构与高效跨模态对齐机制以缓解这一问题，但如何在保证生成质量与物理一致性的同时实现高效推理与多模态融合，仍是亟待突破的核心挑战。

3 基于V2X协同感知下的数据生成

车联网（vehicle-to-everything，V2X）协同感知通过整合车辆、路侧单元等多智能体的感知信息，能够有效突破单车智能的感知局限，解决遮挡、远距离探测等关键难题，是迈向高阶自动驾驶的必由之路。然而，获取用于训练协同感知模型的大规模真实世界数据极具挑战，因为它要求多个装备传感器的智能体同时出现在同一场景中，导致现有协同数据集在规模、场景多样性和智能体数量上严重受限。为了突破这一数据瓶颈，利用生成式技术来创造逼真、多样的V2X数据已成为一个重要的研究方向。这些方法旨在将易得的单车数据集“升级”为协同数据集，或直接从重建的世界模型中合成多视角数据，从而为V2X算法的训练与测试提供近乎无限的燃料。当前，该领域主要呈现出三种技术路径：一是基于扩散模型的生成式方法，二是基于高斯泼溅的重建编辑式方法，三是基于测试与对抗生成的构造式方法。

3.1 基于扩散模型的数据生成

基于扩散模型的生成方法将V2X数据生成定义为一个条件生成问题，其核心框架是利用前向加噪与反向去噪的深度学习模型，根据已知的自车视角数据（如LiDAR点云、语义地图），生成场景中任意新视角的感知数据。这类方法面临的核心挑战在于如何解决信息鸿沟（从有限视角推断未知区域）、确保生成结果的物理真实性（符合传感器物理特性），以及保持跨视角的语义与几何一致性。

Pan等人（2025）提出的TYP通过一种两阶段训练策略来应对上述挑战。第一阶段在大规模单车数据集上训练一个仅以物体边界框为条件的扩散模型，奠定强大的场景生成先验；第二阶段引入模拟协同数据，通过轻量级适配器将自车数据作为额外条件注入，学习从自车到参考视角的映射。为弥合模拟与真实数据的域差异，中间还引入了领域自适应模块。实验证明，该方法能成功将Waymo等大型单车数据集转换为名为“ColWaymo”的协同版本，用其预训练的协同感知模型在下游任务中表现出色，验证了该路径通过“数据转换”进行低成本扩展的可行性。

当前基于扩散模型的数据方法主要依赖离线生成，且在训练时通常假设理想通信条件。其生成质量严重依赖于条件信息的完整性与准确性。未来的研究方向包括：探索在不完美通信条件（如数据压缩、丢包）下的鲁棒生成，研究面向生成任务的高效语义通信策略，以及开发能预测未来状态的动态生成模型以补偿通信延迟。

3.2 基于高斯泼溅的数据生成

基于高斯泼溅的生成方法则遵循一条“重建-编辑-渲染”的技术路线。其核心框架是先利用多视角图像和LiDAR数据，通过高斯泼溅技术对真实V2X场景进行高保真三维重建，获得一个显式且可编辑的神经场景表示；然后在此基础上，对动态物体进行移除、增加或轨迹修改，最后重新渲染出新的多视角

数据。该方法的主要挑战在于实现大规模动态场景的高效与高保真重建、支持灵活且物理合理的场景编辑，以及确保生成数据的多视角同步与标注精准性。

Jagtap等人（2025）率先将高斯泼溅引入V2X动态场景重建并提出V2X-Gaussians。它针对V2X场景中视角稀疏、重叠少的核心挑战，提出了两大创新：V2X驱动的高斯变形场网络 和V2X感知的交叉射线致密化。前者通过在单次训练迭代中同时融入自车和路侧单元的视图，实现了多智能体协同优化动态高斯模型；后者通过计算动态兴趣区域，在多视角交叉的射线区域对高斯模型进行联合致密化，显著提升了动态物体的重建质量。该方法在仅需周期性交换少量数据（~561.8 KB）的情况下，就能实现超越单智能体方法的重建效果。

Xu等人（2025）则将该技术路径推向成熟，构建了一个完整的V2X数据合成框架CRUISE。它采用改进的Street Gaussians进行场景重建，通过引入自车掩码、外观解耦和多种几何约束损失，有效分解静态背景与动态车辆。在编辑阶段，它利用外部生成工具（如TRELLIS）创建3D车辆资产，并借助大语言模型（如GPT-4o）生成符合交通规则的车辆轨迹，从而实现可控的大规模场景合成。研究表明，CRUISE生成的数据不仅能提升单车、路侧及协同3D检测的性能，更能显著改善协同3D跟踪的效果，这得益于其提供的精准、平滑的标注。此外，它还能主动生成各类极端案例（如严重遮挡场景），并天然提供完全同步的多视角数据，解决了真实数据采集中的常见问题。

然而，这类方法的重建质量受限于初始输入数据，且对极端天气（如大雨）等复杂情况的处理能力较弱。此外，当前它主要作为一个离线的数据工厂，尚未与在线系统集成。未来需要提升重建方法的鲁棒性以应对复杂天气与光照，探索实时轻量化重建的可能性，以及研究如何将该框架用于在线仿真与测试，形成闭环的V2X系统开发流程。

3.3 基于测试与对抗生成的构造式方法

基于测试与对抗生成的构造式方法将V2X数据生成定义为一个优化问题，其核心思想是从已有的真实或仿真场景出发，通过系统性地施加扰动或搜索关键配置，构造出能够暴露系统缺陷或提升其鲁棒性的挑战性场景。这类方法面临的核心挑战在于如何确保生成场景的真实性（符合物理和交通规则）、挑战性（能有效触发系统错误），以及处理V2X系统特有的多智能体协同关系。

Guo等人（2025）提出了自动化测试场景生成工具V2XGen，它基于蜕变测试理论，设计了一系列场景变换算子（如插入、删除、缩放、平移、旋转），通过操纵场景中的实体来生成新的测试用例。为了解决V2X多视角一致性的核心挑战，它引入了多智能体视角变换和高保真虚拟LiDAR模拟，确保物体在不同车辆的视角下呈现正确且一致的形态。为了提升测试效率，它还提出了一个适应性引导策略，其设计的适应度函数能优先生成那些更容易引发遮挡感知错误 和远距离感知错误 的场景。实验证明，V2XGen能有效发现多种协同感知模型的故障，并且用其生成的数据重新训练模型，可以显著提升模型在原始测试集上的性能。

Xiang等人（2025）则专注于生成对抗性场景。它采用一个新颖的两阶段优化框架。第一阶段进行对抗性协作伙伴搜索，通过利用中间融合模型中的可学习注意力权重，识别出哪些车辆的组合作为协作方会导致整体感知性能最差。第二阶段进行对抗性位姿扰动搜索，使用黑盒优化算法（如贝叶斯优化）对多个车辆的位姿进行微小但物理可行的扰动，以进一步恶化感知性能。研究表明，V2XP-ASG生成的对抗场景能显著降低各种融合策略模型的精度，并且用这些“难样本”进行微调，能大幅提升模型在正常场景和未知挑战性场景上的泛化能力。

当前基于测试与对抗生成的方法主要依赖于仿真器或高质量的先验数据，其生成场景的多样性和复杂性受限于初始种子和优化空间。未来的研究方向包括：探索更高效的搜索策略以处理更大的状态空间，研究如何将通信限制（如延迟、丢包）纳入对抗因素，以及开发能同时针对感知、预测、规划全栈系统的端到端对抗场景生成技术。

4 面向高阶智驾的数据生成：评测方法

合成交通数据的质量，直接影响自动驾驶在“感知—预测—规划—控制”全链路的可用性与稳定性。其中核心的风险在于域间差异与域内差异：前者指训练域与部署/评测域之间的分布偏差（如合成与真

实、跨城市与道路形态、跨传感器配置与标定状态），后者指同一目标域内部不同子分布的差异（如昼夜/季节/天气、传感器老化）。当合成数据在统计分布、传感器物理属性或多智能体交互结构上偏离目标真实域，域间鸿沟被放大；而当覆盖不足或分布失衡时，域内的子场景会出现显著性能分化。

表2　评测方法总览表

Table 2　Overview of Evaluation Methods

评测类型	方法类别	代表性指标/方法
数据层	人工评测	专家打分/成对偏好
数据层	统计一致性	均值/方差/偏度/峰度；KS 统计量
数据层	可视分布分析	直方图/箱线图/小提琴图/Q－Q图
表征层	特征分布距离	FID/KID/MMD
表征层	预训练模型判别	判别器得分/特征可分性
效用层	TSTR（合成训、真实测）	GAN-train / TSTR 精度、mAP、mIoU、ADE/FDE
效用层	TRTS（真实训、合成测）	GAN-test / TRTS 指标
任务层	检测/分割	mAP、mIoU
任务层	可驾驶性	Driving Score/Route Completion/碰撞率/红灯率
任务层	交通规则/多样性	SCR、TRV、MASD
安全与合规	安全性	碰撞/急减速/越界/最小时距等
安全与合规	隐私与合规	重识别率、成员推断率
物理合理性	物理可行性	几何穿透性、不稳定率、动力学可行性

此外，单纯提升视觉渲染保真度（图像/点云的外观保真）并不必然带来任务性能改进；决定性因素往往来自样本的多样性与覆盖度。这既包括类别与实例的长尾充分性，也包括场景与地理的广度（城区、郊区、高速等），环境条件的层级化覆盖（昼夜、季节、天气、光照与能见度），行为与交互的复杂度（并线、礼让、侵占、异常机动），以及动态要素与稀有/极端场景的出现频度；同时还应覆盖传感器与物理属性的变化（相机/激光雷达参数、噪声与回波/反射率分布、标定与装车位姿差异）。只有在这些维度上达到足够的代表性与均衡性，合成数据才更有可能在真实部署中稳定转化为指标提升。

因此，本文提出面向合成数据的综合测评框架（见表2），旨在提供全面、可比且工程可用的评测依据，为后续方法选择与落地实践建立统一参照（Song等，2024；Liu等，2024，Liu等，2025）。

4.1　基础评测方法

4.1.1　人工评测

人工评测指由真实评审者（普通用户或领域专家）依据明确的任务目标与使用情境，对生成结果的可理解性、可信度、语义与时空一致性、可用性）。

量表评分：采用分级量表对若干维度打分（如真实感、语义合理性、时空一致性）。量表可以使用李克特量表（Likert scale）设计，最早由 R. Likert 提出并给出构造与计分方法，便于统计分析与报告一致性（Likert，1932）。

成对偏好/排序（A/B 测试）（Bradley and Terry，1952）：在同一输入与相同展示条件下，同时呈现两份候选结果，匿名标记为 A 与 B，评审者据预设准则选择更优者。核心思想是用直接对比来放大细微质量差异，并以胜率/偏好次数或由多轮对比汇总的简单排序来评估模型优劣。

专家核查：由具备领域经验的评审者对关键场景元素（如交通规则、物理接触关系）进行一致性核查与质检。

人工评测以真实人类判断为标准，能覆盖自动指标难以度量的语境理解、常识一致性与可用性感知，适合用于关键场景质检与模型间细微差异对比；同时可与自动指标联用，作为口径校准与决策依据。其主要不足在于：主观性与一致性问题（评审间偏

差、复现性弱）、成本与时效（招募与质控开销高、难以大规模与高频迭代）、采样与呈现偏差（任务/样本选择影响结论）、以及可比性与迁移性有限（不同任务与界面设置下的结果难横向对齐）。

4.1.2 统计指标

统计指标用于量化两组数据在“分布形态—变量关系—误差表现”等层面的接近程度。下列度量在交通多模态数据（图像、视频、点云、轨迹）中具有可操作性。

均值与方差：给定随机变量X（如像素强度、点云反射强度或BEV栅格占据概率），当合成样本均值$\bar{x}_{\text{syn}}$与真实样本均值$\bar{x}_{\text{real}}$、合成样本无偏方差s^2_{syn}与真实样本无偏方差s^2_{real}差距较大时，往往预示渲染强度标定或噪声模型失配（例如LiDAR回波强度/噪声谱），此类统计在传感器级真实感评估中被广泛使用。

皮尔逊相关系数（Rodgers & Nicewander，1988）：用于衡量多变量依赖结构，如RGB通道/深度与强度/车速与转向等。皮尔逊相关系数：

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}. \tag{1}$$

此系数用于衡量线性相关性，$r\in[-1,1]$，式中x_i,y_i为两变量观测，$\bar{x}$，$\bar{y}$为各自均值。

Kolmogorov – Smirnov（KS）统计量（Stephens，1974）：比较两组一维样本的经验分布函数（empirical cumulative distribution function，ECDF）差异。设真实样本$x_{1:n}$的ECDF为：

$$F_n(t)=\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{x_i\leqslant t\} \tag{2}$$

合成样本$y_{1:m}$的ECDF为：

$$G_m(t)=\frac{1}{m}\sum_{j=1}^{m}\mathbf{1}\{y_j\leqslant t\} \tag{3}$$

两样本KS统计量为：

$$D_{n,m}=\sup_a|F_n(a)-G_m(a)| \tag{4}$$

式中，F_n,G_m分别为两组样本的ECDF。n,m为两组样本量；a为阈值变量（在样本空间上取上确界sup）。KS为非参数检验，适用于像素强度、点云距离、速度等标量变量的一致性判别。

结构相似性指数（SSIM）（Wang等，2004）：一种感知质量评估指标，衡量两幅图像在亮度、对比度和结构信息方面的相似度。此类统计可以量化自动驾驶系统在不同情况下生成图像的质量。

余弦相似度（cosine similarity）（Singhal等，2001）：余弦相似度常用于衡量不同数据（如图像、视频帧、传感器数据等）之间的相似性。它通常用于评估生成数据与真实数据之间的相似度（如深度学习生成的数据）。

统计指标为真实域与生成域数据的可量化对照提供了自下而上的基线：从一阶/二阶矩（均值、方差）检视强度与噪声标定，到相关性（皮尔逊）刻画变量联动关系，再到分布一致性衡量ECDF差异，以及在表征空间用结构/角度相似评估感知质量与高维特征接近度。它们计算高效、易复现；但也有各自的局限性。

4.1.3 可视分布分析

可视化分布分析通过图形化手段，直观展示感知、决策和控制数据的分布特征（如场景分布、传感器模态一致性、决策行为合理性和极端工况下的离散度），帮助分析和优化自动驾驶系统的性能（Routray等，2024）。其核心目标是将多源异构数据（如来自不同传感器的多模态数据、算法输出的决策数据和执行器信号的控制数据）转化为可解读的视觉语言，精确匹配数据生成的场景覆盖需求。

具体来说，通过箱线图、小提琴图、直方图、Q-Q图等图表，可以直观地展示数据间的关联与差异。

箱线图（Box Plot）：展示数据的中位数、四分位范围、须线与异常点，便于比较不同场景或类别的集中趋势与离散程度。它是展示不同场景分布差异的有效工具，特别适用于分析数据是否存在明显的异常值或极端值（McGill，Tukey and Larsen，1978）。

小提琴图（Violin Plot）：在箱线图的基础上叠加了核密度曲线，能够揭示数据的多峰特征和长尾分布，适用于展示复杂的数据分布，帮助分析可能存在的长尾现象和极端情况。在自动驾驶的复杂场景中，小提琴图可以帮助揭示决策行为的多样性和分布的复杂性（Hintze and Nelson，1998）。

直方图/核密度曲线：适用于一维变量（如像素强度、点云距离、速度、加速度等），能够有效展示数据的频率分布。通过这些图表，能够清晰分析不同数据特征的分布状态，发现数据的集中或分散趋势。例如，通过直方图分析传感器数据的分布，可以评估

传感器的覆盖范围和有效性。

Q-Q 图（Wilk and Gnanadesikan，1968）：用于对比两种数据分布的分位数是否接近直线，进而判断数据是否符合某些特定分布（如正态分布）。这一图表尤其适用于验证数据是否来自相同的分布，或检查数据是否符合某些理论模型。这对于评估自动驾驶中模型预测的合理性和准确性尤为重要。

通过这些图表，能够直观、精确地展示数据的分布特征，帮助识别系统中的潜在问题，并验证评测体系的有效性。这些分析不仅支持自动驾驶系统性能优化，还为基准数据集的构建提供了量化依据。

4.2 深度学习驱动的自动评测

4.2.1 基准化模型指标

基准模型指标旨在在语义表征空间中衡量合成数据相对真实数据的保真度与覆盖性，避免仅做像素级统计的局限（Heusel 等，2017）。其基本思想是利用在真实数据上训练好的、参数固定的基准表征模型 f_θ 作为特征提取器，把真实数据与合成数据共同投影到特征空间，并在该空间比较两者分布差异。为覆盖不同场景需求，本文选取业内广泛使用的代表性方法进行介绍，包括基于高斯近似的 Fréchet Inception Distance（Fréchet Inception Distance，FID）（Heusel 等，2017），在小样本更稳健的核 Inception 距离（Kernel Inception Distance，KID）（Bińkowski 等，2018），面向视频时空一致性的 Fréchet 视频距离（Fréchet Video Distance，FVD）（Unterthiner 等，2018），以及针对点云模态的 Fréchet 点云距离（Fréchet Pointcloud Distance，FPD）（Shu D W 等，2019）。

设真实数据集为 $\mathcal{D}_r=\{x_i\}_{i=1}^{n}$，合成数据集为 $\mathcal{D}_s=\{\tilde{x}_j\}_{j=1}^{m}$。选定冻结的预训练模型 $f_\theta(\cdot)$：在图像任务中，可取 Inception 模型（Szegedy 等，2016）；在视频任务中，可取 I3D 模型（Carreira 等，2017）；在点云任务中，可取 PointNet 模型（Qi 等，2017）。所提取的样本特征向量被定义为：

$$z_i=f_\theta(x_i)\in\mathbb{R}^d,\qquad \tilde{z}_j=f_\theta(\tilde{x}_j)\in\mathbb{R}^d \tag{5}$$

式中，d 表示特征维度；$\{z_i\}$ 和 $\{\tilde{z}_j\}$ 分别代表真实和合成数据的特征向量集合，它们共同构成了后续计算基准模型指标的特征空间。

1） Fréchet Inception Distance

FID（Heusel 等，2017）评估指标在 Inception 模型提取的特征上，以高斯分布近似两个特征向量集合的分布 $\mathcal{N}(\mu_r,\Sigma_r)$、$\mathcal{N}(\mu_s,\Sigma_s)$，定义如下：

$$D_{fid}=\left\|\mu_r-\mu_s\right\|_2^2+\mathrm{Tr}\left(\Sigma_r+\Sigma_s-2\left(\Sigma_r^{1/2}\Sigma_s\Sigma_r^{1/2}\right)^{1/2}\right) \tag{6}$$

式中，μ_r,μ_s 为真实和合成数据的特征向量的均值，Σ_r,Σ_s 为真实和合成数据的特征向量的协方差矩阵，$\mathrm{Tr}(\cdot)$ 为迹算子。FID 的值越小表示两分布越接近，该指标被用于评价生成图像与真实图像在预训练模型特征空间上的分布相似性。

FID 基于预训练表征，具备计算简单、效率高的优势，并且能在语义层面比较真实与合成分布，在实践中与主观感受有一定相关性。关于局限性，由于其数值强依赖所选特征提取器与预处理，跨领域/跨模态时可比性有限。同时，高斯近似只利用特征的均值与协方差，小样本下的估计可能存在偏差，并且难以刻画多峰或长尾等复杂差异。

2） Kernel Inception Distance

KID 用核最大均值差异（maximum mean discrepancy，MMD）在 Inception 特征空间比较两分布，常取多项式核 $k(u,v)=(1/d(u^\top v)+1)^3$。其无偏估计形式为：

$$D_{kid}=\mathrm{MMD}^2\left(\{z_i\},\{\tilde{z}_j\}\right) \tag{7}$$

MMD^2 为最大均值差异为最大均值差异的无偏估计；d 为特征维度；u,v 为两域特征向量。

KID 优势在于不依赖高斯近似、对小样本更稳健，能捕捉到超出均值与协方差的高阶差异。在不足方面，其数值对核函数及其超参数较为敏感。同时，计算过程需成对求核，样本规模很大时较为耗时，与主观质量的相关性可能存在偏差。

3） Fréchet Video Distance

FVD（Unterthiner 等，2018）用于衡量视频生成结果与真实数据在时空语义表征上的分布差异。其思路与 FID 相同，在固定、以大规模动作识别数据训练良好的时空网络（I3D）上，对真实与生成的视频片段提取嵌入，分别拟合多元高斯分布，并计算两者的 Fréchet 距离。不同于逐帧图像指标，I3D 表征将帧内外观与帧间运动/时序关联共同编码，从而使 FVD 同时反映画面质量与时间一致性。实践中通常对每个视频抽取若干片段、逐片段计算后再聚合（如均值/中位数），并配套统一的预处理（分辨率、裁剪、时长）以确保可比性。其尤其适合评估关注运动真实性、动作连贯与时序稳定性的任务，如视频生成、

视频合成/编辑与视频驱动人像等。

FVD优势在于刻画运动与跨帧连贯性，相比逐帧图像指标更贴近视频主观感受。尽管具备上述优势，但其数值依赖所选时空特征器及其训练数据集与预处理策略，跨领域时可比性受限；同样采用高斯近似，仅利用均值与协方差，对多峰分布与长时依赖不够敏感，并对视频长度与采样方式均较为敏感。

4） Fréchet Pointcloud Distance

FPD（Shu D W等，2019）用于在预训练点云表征空间比较真实与生成点云的分布差异。其做法是选取在大规模点云任务上训练良好的编码器（如PointNet），先对点云做标准化（居中、尺度统一，必要时对齐姿态）并重采样为固定点数，将样本映射为嵌入向量。随后，分别对真实与生成嵌入拟合多元高斯分布，计算两者的Fréchet距离作为度量。得益于点云编码器对点集置换的天然鲁棒性以及对局部几何—整体形状的联合表征，FPD能够在分布层面反映生成形状与真实形状的一致性。实践上常按类别/场景分别计算并汇总，并需在编码器、点数、归一化与采样策略等方面保持固定、统一，以确保评估指标结果的可比性。

FPD优势在于对采样噪声与局部扰动更为鲁棒，能综合反映形状与语义几何信息，较逐点距离更能体现分布层面的接近，且实现简洁、适合大规模评估。此外，需要指出的是，其对点云的归一化与对齐、点数与采样策略较为敏感，在稀疏或遮挡严重的点云上稳定性表现较差。

总体来看，基准模型指标相较于仅在像素空间做统计，其基于预训练表征的比较能在更贴近任务语义的层面开展，可减弱低层纹理、噪声与分辨率差异的干扰，突出对象类别、形状、场景布局与运动等关键结构特征，使评估对光照、视角、风格等外观变化更加不变且更具鲁棒性。与此同时，基准模型指标也存在一些共性不足，其数值强依赖所选表征器及其训练语料与预处理流程，跨域迁移时可比性与稳健性受限；对评测协议与超参数较为敏感。

4.2.2 任务/应用特定指标

本节从任务可用性出发提出一组互补的评测维度：感知（检测/分割/跟踪）、开环规划/控制、风险与安全、隐私合规、物理一致性与传感器层校核。它们分别刻画识别质量、离线决策误差、稀有事件风险、数据/模型的合规边界，以及几何—动力学—成像物理的一致性。感知度量与现有数据集协议一致，复现性强，但对域间/域内分布偏移及长尾类别更敏感；开环指标计算成本低，但与闭环行为的对应关系受分布漂移影响（Alahi等，2016；Bansal等，2019）；安全指标直接反映风险，但需要充足且分层均衡的样本；隐私指标要求在给定(ε,δ)的前提下评估效用变化（Dwork 和 Roth，2014）；物理与传感器校核为生成/发布的下限约束。下文分别给出定义、符号与使用说明。

1） 感知任务指标

感知任务的评估遵循现有数据集通用协议，覆盖检测交并比（intersection over union，IoU）、平均精度（average precision，AP）、平均精度均值（mean average precision，mAP）、平均交并比（mean intersection over union，mIoU）。该类指标与训练目标一致，便于定位误差来源并与工程侧评测直接对齐（Lin等，2014；Bernardin 和 Stiefelhagen，2008；Guo等，2017）。

目标检测（2D/3D）（Lin等，2014）：主流目标检测基于平均精度均值（mAP）。对某一类别，先由真值框与预测框的交并比（IoU）定义匹配关系。在给定阈值（如IoU≥0.5）下计算精确率－召回率曲线$p(r)$，其面积即平均精度（AP）。mAP为对类别的算术平均。实际实现多为分段插值的数值积分，对所有类别取算术平均得mAP。

语义/实例分割：语义/实例分割最常用评价指标为平均交并比（mIoU）。

2）规划/控制指标

开环评测在日志回放（log-replay）下进行，模型不与环境交互，常用三类量化指标：

轨迹L2误差（ADE/FDE）（Alahi等，2016）：

$$E_{ADE}=\frac{1}{T}\sum_{t=1}^{T}\left\|\hat{y}_t-y_t\right\|_2,\qquad E_{FDE}=\left\|\hat{y}_T-y_T\right\|_2 \tag{8}$$

式中，y_t为GT位置（2D/3D），$\hat{y}_t$为预测，T为预测步数；$\|\cdot\|_2$为L2距离（Bansal，Krizhevsky and Ogale，2019）。

3）安全指标（Safety-centric）：

碰撞率（Collision Rate）：碰撞率是通过统计单位样本或单位里程内发生的碰撞事件频次来衡量的。具体而言，碰撞事件包括所有形式的交通碰撞（如轻微碰撞和严重碰撞）。这一指标有助于评估自动驾驶系统在实际驾驶过程中碰撞的风险，并为系统的安全性优化提供依据。

$$CR = \frac{N_{\text{coll}}}{N_{\text{eval}}} \times 100\% \tag{9}$$

N_{coll} 为发生碰撞的样本/片段/场景数(与评测口径一致)；N_{eval} 为参与统计的总样本数。该指标直接反映风险事件发生频度；作为稀有事件，需要足够样本规模并按场景稀有度分层。对于“违规事件”类度量(红灯、越道、逆行等)，可在每类事件上给出计数或距离归一化频度。

违规分数(IS)：违规分数通过统计自动驾驶系统在完成路线过程中发生的违规行为次数来计算。违规行为包括交通规则的违反(如闯红灯、不遵守车道规定等)和系统操作不当等。违规分数的高低反映了系统在遵守交通规则和规范方面的表现。

路线完成率(RC)：已完成路线距离占总路线距离的百分比。设 R_i 为汽车在第 i 条路线中的完成率，则路线完成率通过计算所有路线完成率的平均值得到。其中 N 为路线总数。

$$RC = \frac{1}{N}\sum_{i}^{N} R_i \tag{10}$$

能力分项：在并道/路口/礼让/应急等能力维度上分别统计成功率或违规率，降低总分的方差、提高可诊断性。

4)隐私与合规(用于含车牌/人脸/地理轨迹的合成数据)

差分隐私(DP)机制(Liu 等，2024)：属于隐私合规相关的技术实现，可通过该架构生成满足差分隐私要求的合成数据，避免原始数据隐私泄露。若生成流程引入 DP 保护，则需给出 (ε,δ) 参数，其中 ε (隐私预算)是隐私保护强度的核心指标；δ(松弛因子)是指允许隐私保护约束存在“极小概率的例外”，用于降低 ε 的取值。并满足：

$$\Pr\left[\mathcal{M}(D)\in S\right] \leqslant e^{\varepsilon}\Pr\left[\mathcal{M}(D')\in S\right]+\delta. \tag{11}$$

式中，D, D' 为只相差一条记录的数据集，M 为生成机制。

5)物理一致性与传感器层校核(Physical Plausibility)

面向高阶自动驾驶的数据生成，需要同时满足：几何—接触可行性、车辆动力学可实现性与舒适性、以及传感器成像物理的一致性。下面给出可直接落地的指标与公式，并给出符号含义与依据。

几何穿透/接触一致性(Collision & Penetration)：以场景的有符号距离场(signed distance field, SDF)或网格碰撞检测为基础，统计生成物体之间、物体与地面/墙体之间是否存在几何穿透。对第 i 个刚体，在其表面均匀采样，定义平均穿透深度：

$$p_i = \frac{1}{|\Omega_i|}\sum_{x\in\Omega_i}\max(0,-\phi(x)). \tag{12}$$

Ω_i 为物体 i 表面采样点集；$\phi(x)$ 为点到其他几何的 SDF 值(内部为负)；p_i 为平均“负 SDF”深度。若 $p_i > \tau_p$(阈值，单位 m)，则判为穿透体。据此给出两级比率：

$$\begin{aligned} C_O &= \frac{1}{N}\sum_{i=1}^{N}\gamma\left[p_i>\tau_p\right], \\ C_S &= \frac{1}{M}\sum_{s=1}^{M}\gamma\left[\max_{i\in s}p_i>\tau_p\right]. \end{aligned} \tag{13}$$

式中，τ_p 为判定阈值(m)；N 为被检物体数；M 为场景/帧数，γ 是指示函数(indicator function)，它把一个布尔判断转成数值：条件为真取 1，否则取 0。上述做法与基于SDF的穿透判别、对支撑/碰撞关系进行物理可行性检查的通行实现一致(使用SDF或网格相交、法向和接触对等信息进行判别)。

稳定性(Stability under Gravity)：将场景导入刚体物理引擎，在重力、库伦摩擦系数 μ_s 与接触约束下仿真 T 秒，记第 i 个物体位姿从 $(\mathrm{t}_i^0,\mathrm{q}_i^0)$ 演化为 $(\mathrm{t}_i^T,\mathrm{q}_i^T)$(q 为四元数)。位姿扰动定义为

$$\Delta r_i = \left\| t_i^T - t_i^0 \right\|_2, \qquad \Delta\theta_i = \arccos\left(\left|\left\langle q_i^T, q_i^0\right\rangle\right|\right). \tag{14}$$

若 $\Delta r_i > \tau_t$ 或 $\Delta\theta_i > \tau_r$ 则判为不稳定体 $I_i = 1$，否则 $I_i = 0$。据此给出

$$I_O = \frac{1}{N}\sum_{i=1}^{N} I_i, \qquad I_S = \frac{1}{M}\sum_{s=1}^{M}\gamma\left[\sum_{i\in s} I_i > 0\right]. \tag{15}$$

$t_i^{0/T}$ 为仿真前/后平移向量；$q_i^{0/T}$ 为四元数姿态；$\langle\cdot,\cdot\rangle$ 为四元数内积；Δr 与 $\Delta\theta$ 的阈值分别为 τ_t, τ_r；T 为仿真时长(s)；μ_s 为静摩擦系数；$I_i = \gamma[\Delta r_i > \tau_t$ 或 $\Delta\theta_i > \tau_r]$。若几何与接触关系不合理则会在短时仿真后发生明显位姿漂移，可用作三维场景生成的物理一致性度量与排序。

感知类度量与训练目标及工程评测口径一致，便于复现并能直接定位误差来源；在类别不均衡、遮挡或跨域情况下，单一平均值容易掩盖子场景差异，需配合距离/尺寸/可见度的分层统计与校准指标。开环误差(如 ADE/FDE)成本低、分析清晰，适合快速比较同协议下的方法；当未来具有多模态可能或评测分布与部署分布不一致时，误差与闭环性能的

对应关系会减弱，应同时报告安全相关事件并说明复现协议。安全度量与部署风险对齐、解释直观，但属于稀有事件，估计方差较大且对仿真配置与接触判定敏感，宜提供分层结果与置信区间。物理一致性与传感器层校核为数据生成与发布的下限约束，能够高效剔除几何或成像物理不一致的样本，但不刻画策略层行为；因此应与开环/闭环结果结合。隐私与合规则明确可发布边界；在给定(ε,δ)下开展对比可以避免因隐私预算不同导致的效用不可比，但更严格的隐私预算通常伴随可用性下降（Dwork和Roth，2014）。

4.3 闭环评测与多能力基准

闭环评测（closed-loop）关注可驾驶性、安全性与效率等综合表现，是对开环误差与感知质量的补充：开环指标便于低成本比较，但与部署环境存在分布差异时，其与闭环表现的相关性会减弱。近期基准（例如Bench2Drive）强调以多能力任务组构造评测集合，并在统一的路线与记分协议下报告指标，以提升可比性与复现性（Jia等，2024）。

4.3.1 指标体系

1）路线完成度（Route Completion，RC）：刻画车辆沿预定路线行驶的比例。对第i条路线，记该路线被分段为M个路段，完成与否的指示量为$R_{i,m}\in\{0,1\}$。则

$$RC_i=\frac{1}{M}\sum_{m=1}^{M}R_{i,m},\qquad RC=\frac{1}{N}\sum_{i=1}^{N}RC_i. \tag{16}$$

式中，N为评测路线数。

2）综合驾驶分（Driving Score，DS）：在路线完成度基础上，乘以各类交通违规与碰撞的惩罚系数。记路线i的违规类别集合为K，第k类违规的惩罚系数为$P_{i,k}\in(0,1]$，则

$$DS=\frac{1}{N}\sum_{i=1}^{N}\left(RC_i\cdot\prod_{k\in\mathcal{K}}P_{i,k}\right). \tag{17}$$

违规类型与惩罚：典型违规涵盖：碰撞（行人/两轮/车辆/静态）、闯红灯、逆行、越线、超速、驶离可行驶区域等。每类事件的$P_{i,k}$取值与触发条件在评测附录中有明确表，报告时需一并列出以保证可比性。

3）安全-效率-平顺三要素

安全：除DS之外，常同时报告情景碰撞率（scenario collision rate，SCR）与交通规则违规率（traffic rule violation rate，TRV）。SCR定义为：在每个采样情景中，发生碰撞的参与体占比的平均值；TRV通常覆盖“驶离可行驶区域”“闯信号灯”等规则。

效率（Efficiency）衡量车辆是否“积极而不过激地”完成任务。常见做法是对每段路计算有效速度比例：若车辆速度长期低于道路限速的一定比例，则视为效率不足。设第i条路线平均速度为v_i，限速为$v_i^{\max}$，则

$$E_i=\min\left(1,\frac{\bar{v}_i}{\alpha v_i^{\max}}\right),\qquad 0<\alpha<1. \tag{18}$$

整体效率取均值。实践中还会定义“停滞（stuck）”阈值（如速度长期低于限速的一定比例触发惩罚），细节需与评测配置保持一致。

平顺（Smoothness/Comfort）既要抑制高频抖动，也要避免卡顿。一种稳定做法是将整条路线划分为若干可行驶片段（free-vehicle segments，FVS），仅在片段内统计纵向/横向加加速度（jerk）或转向/制动变化率等量纲一致的平顺指标；当车辆处于阻塞（如被其他交通体“卡住”）时，该片段不计入平顺分。形式化地，设路线被划分为S_i个FVS片段，第s段的瞬时平顺代价为$c_{i,s}(t)$，则

$$S_i=1-\frac{1}{S_i}\sum_{s=1}^{S_i}\frac{1}{T_{i,s}}\int_0^{T_{i,s}}c_{i,s}(t)dt. \tag{19}$$

式中，$T_{i,s}$为片段时长；$c_{i,s}(t)$可取归一化的$|\text{jerk}|$或转向/制动变化率的加权求和。

4.3.2 多能力拆解与覆盖

闭环表现与具体能力密切相关。为避免路线级平均值掩盖薄弱环节，可将评测拆解为合流/并线、礼让与右让先、交通信号/标志遵循、无保护转弯、紧急制动/避障、狭窄会车/借道等能力集合，并在每个能力子集上分别统计*RC*/*DS*与安全事件（Jia等，2024）。设能力集合为A，能力$a\in A$的样本占比为q_a，目标占比为p_a（通常设为均匀或按实际交通分布），则覆盖一致性可用：

$$C(A)=1-\frac{1}{2}\sum_{a\in A}|q_a-p_a| \tag{20}$$

进行度量（值越大表示与目标分布越一致）。在相同总样本数下，提高对稀有或高风险能力的覆盖可降低SCR/TRV的估计方差，提高评测稳定性；多能力拆解还便于定位“失效来源”，与感知/开环误差进行对照分析（Jia等，2024；Liu等，2024）。

4.3.3 闭环评测协议

评测目标意在保持训练—评测协议一致、只替

换数据生成策略的前提下，闭环可驾驶性（SR/RC/DS）与三要素（安全-效率-平顺）是否实质改善。为此建议遵循以下最小可复现协议：

1）系统封装与冻结：固定被测系统的模型结构、训练超参、输入输出接口，仅替换训练数据中的合成部分，避免训练策略差异造成混淆。

2）仿真设置对齐：在评测端锁定地图、天气、交通流与随机种子；传感器组配（相机/LiDAR内外参、刷新频率、延迟/噪声）与车辆动力学模型（限速、轮胎、制动）应与训练端一致。真实感的传感器级建模能显著缩小感知域的仿真与现实差距，例如在LiDAR渲染中引入ray-drop学习与真实资产库，可将检测/分割性能逼近真实数据，并支持安全关键场景的端到端测试。

3）能力覆盖与配额：按多能力维度等额抽取路线（如每个主能力至少若干条），保证样本均衡与统计功效。

4）指标与触发阈值：沿用公开的*DS/RC*定义与违规惩罚表；效率阈值、停滞检测、红灯判定等阈值必须与评测配置一致；平顺指标采用“片段化—积分”的口径。

5）统计报告：给出总体均值±方差、分能力均值、违规事件直方图与按路线长度分层的DS箱线图；同时报告SCR/TRV以区分“撞/不撞”和“守/不守规”的变化来源。

6）与开环的关联：在相同路线集合上同步报告开环ADE/FDE与闭环RC/DS/SCR的相关分析，说明可能的分布差异与失配来源。

本章从“统计一致性、任务效用、闭环可驾驶性、安全与合规、物理/传感器一致性、隐私”六个层面构建了合成交通数据的评测框架：以FID/FVD/FPD等表征距离进行分布诊断；以mAP/mIoU/MOTA与ADE/FDE衡量感知与开环性能；以RC/DS联合SCR/TRV、效率与平顺刻画闭环表现；以SDF穿透率、稳定性与传感器回波/曝光一致性作为离线硬约束；在固定参数条件下评估隐私预算对效用的影响。综合分析表明，统计指标成本低但难以直接外推到部署；任务与开环指标可快速分化方法但受分布漂移影响；闭环与安全能给出可上线证据但依赖统一协议与足够样本，需综合各类评测方法，实现可比、可复现的综合能力水平测评。

5 典型数据与工具

5.1 面向单车智驾的数据生成

典型单车智驾数据生成基准数据集见表3，具体包括：

（1）Cityscapes由德国汽车制造商Daimler与多所高校联合发布，是一个专注于城市街景语义理解的高质量视觉数据集。该数据集覆盖50个城市的不同天气与时间条件，主要用于语义分割、实例分割及深度估计等任务。数据集共包含5,000张高精度像素级标注图像（其中训练集2,975张，验证集500张，测试集1,525张），另提供约20,000张弱标注图像用于半监督或迁移学习研究。图像分辨率为2048×1024像素。数据可通过https://www.cityscapes-dataset.com获取，示例图如图4所示。

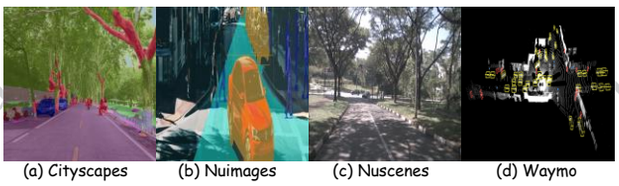
（2）nuScenes是由Motional（前nuTonomy）与Aptiv联合发布的多模态自动驾驶感知数据集，覆盖真实城市驾驶环境，包含激光雷达、毫米波雷达、六个车载摄像头以及惯性测量单元等多种传感器信息。该数据集提供了完整的3D标注信息，用

于多传感器融合、3D目标检测、跟踪与场景理解等任务。数据总时长约55小时，包含1,000个独立场景（每个约20秒），约140K帧图像与1.4 M个3D标注实例。图像分辨率为1600×900像素，支持RGB相机、点云与雷达信号的跨模态对齐。数据可通过https://www.nuscenes.org/download获取。

（3）nuImages由Motional发布，是一个大型驾驶场景二维图像数据集，其传感器布局与nuScenes保持一致，但专注于2D图像标注任务。该数据集旨在弥补真实驾驶场景中长尾类别、极端天气和复杂交通背景下2D视觉数据的稀缺问题，适用于目标检测、实例分割、语义分割及跨帧视觉理解等研究任务。数据集共包含约93,000张标注图像，其中训练集（train）约67,279张，验证集（val）约16,445张，测试集（test）约9,752张。图像分辨率为1600×900像素，数据结构与devkit的使用方式与nuScenes保持一致。数据可通过https://www.nuscenes.org/download获取。

（4）Waymo Open Dataset由Waymo发布，是目前规模最大、精度最高的多模态自动驾驶感知数据集之一，广泛用于3D检测、跟踪、预测及规划研究。

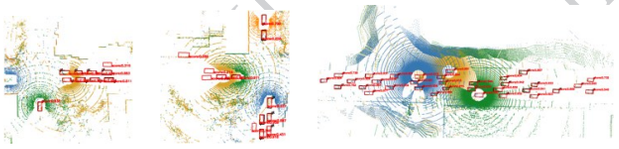
该数据集采集于美国多个城市，包含 1，150 个驾驶场景（每个约 20 秒），合计超过 12 百万帧 LiDAR 点云与 10 百万帧多视角高清图像。每辆采集车配备 5 个 LiDAR 与 5 个高清摄像头，图像分辨率最高可达 1920×1280 像素。数据按照任务划分为训练集、验证集与测试集，约占 60%/20%/20%。数据可通过 https://waymo.com/open 获取。



（a）Cityscapes image；（b）Nuimages image；（c）Nuscenes image；（d）Waymo image）

图 4　典型智驾数据集中的图像示例

Fig. 4　Image examples from typical intelligent driving dataset. (



（a）V2X-Sim 点云示例　（b）OPV2V 点云示例

（c）DAIR-V2X 点云示例　（d）V2V4Real 点云示例

（（a）Point cloud examples in V2X-Sim　（b）Point cloud examples in OPV2V

图 5　协同感知数据集中的点云示例

Fig. 5　Point cloud examples in collaborative perception datasets.

5.2　基于 V2X 协同感知下的数据生成

典型协同感知数据生成基准数据集见表 4，具体包括：

（1）V2X-Sim 是纽约大学 Li 等人（2021）设计的模拟多智能体感知的数据集，由交通模拟软件 SUMO 和 CARLA 模拟器共同生成，数据格式遵循 nuScenes 标准。V2X-Sim 配备了 RGB 摄像头、LiDAR（激光雷达）、GPS（全球定位系统）和 IMU（惯性测量单元），共收集 100 个场景，每个场景包含 2~5 辆车，总 10，000 帧数据。数据集按 8，000/1，000/1，000 的比例划分为训练集/验证集/测试集。V2X-Sim 的基准测试支持检测、追踪和分割 3 个关键感知任务，所有任务均采用 BEV 表示法，并在 2D BEV 中生成结果。

（2）OPV2V 是加利福尼亚大学洛杉矶分校 Xu 等人（2025）为 V2V（车对车）协同感知设计的模拟数据集，其通过 OpenCDA 框架和 CARLA 模拟器收集。该数据集提供了完整的配置文件以重现实验环境，包含 11464 帧的激光雷达点和 RGB 摄像头数据。此外，OPV2V 还提供了名为 Culver City 的真实模拟测试集，对于评估模型的泛化能力极具价值。该数据集支持三维物体检测和鸟瞰视图语义分割的基准测试，目前主要关注车辆这一类型的物体。

（3）DAIR-V2X 是由清华大学 Yu 等人（2022）采集的首个来自真实场景的大规模协同感知数据集。其中，DAIR-V2X-C 数据集特别适用于 V2I 协同研究，提供了 VIC3D 基准，以探索 V2I 物体检测任务。障碍物目标 3D 标注属性全面，标注 15 类道路常见障碍物目标，数据涵盖晴天、雨天、雾天、白天/夜晚、城市道路/高速公路等丰富场景。

（4）V2V4Real 是由加利福尼亚大学洛杉矶分校 Xu 等人（2023）牵头采集的首个大规模真实世界多模态 V2V 感知数据集。该数据集在俄亥俄州哥伦布市使用特斯拉和福特车型收集，覆盖了 410 km 的道路。数据集包含 20，000 帧激光雷达数据以及超过 240，000 个三维边界框注释。此外，V2V4Real 还为三维目标检测、目标追踪以及领域适应这 3 项协同感知任务提供了基准测试。

6　挑战与开放问题

6.1　多模态信息的完整获取与融合

智能座舱交互涉及语音、文本、视觉和系统状态等多种模态。当前生成模型在单一模态下已有较好性能，但如何在保证语义、声学、副语言、动作与环境状态等多模态信息一致性和完整性的前提下进行联合生成，仍是一个开放问题。尤其是语音与文本生成存在信息丢失、视觉模态生成在微表情和动作连贯性上存在难点，而系统状态生成需要精确反映多设备间的交互与约束。如何实现跨模态高保真生

成，同时兼顾交互自然性和物理合理性，是未来的重要研究方向。

6.2 高效、可扩展的数据生成与场景泛化

智能座舱涉及的驾驶场景和乘员行为高度多样，真实数据难以覆盖所有情况，尤其是极端驾驶状态或罕见交互行为。现有生成方法在生成效率、

延迟和可扩展性方面仍有限，难以快速适应新的座舱环境或新功能。同时，生成模型的泛化能力和稳定性仍需提升，以保证在不同座舱布局、光照条件、个体差异下生成数据的有效性和可靠性。因此，如何构建可扩展、高效且具有良好泛化能力的数据生成系统，是智能座舱研究中的核心开放问题。

6.3 长尾与极端场景数据获取困难

高阶单车智驾系统需要在多样化、复杂的交通环境下保持高鲁棒性和泛化能力。然而，现实交通场景呈现明显的长尾分布，大部分常规场景数据易于获取，但极端情况如夜间突发障碍、恶劣天气或密集交叉口等稀有事件的数据却极其有限。这种数据稀缺限制了生成模型在少见事件上的学习能力，也影响下游感知、预测与决策任务的表现。如何有效生成和扩展长尾与极端场景数据，以提升模型在稀有情况中的泛化能力，仍是当前单车智驾数据智能生成领域亟需解决的核心问题。未来研究可探索基于条件控制、场景增强及多模态融合的策略，以在稀缺事件上获得更全面的训练样本，缓解长尾与极端场景数据短缺带来的限制。

6.4 生成数据质量评估体系不足

生成数据在为单车智驾系统提供训练样本时，其多样性、真实性和代表性至关重要。然而，现有评估方法主要依赖诸如 FID、KID 等分布距离指标，或依赖下游任务表现进行评测，这类方法存在可靠性不足、主观性高、成本昂贵且耗时长的问题，难以全面反映生成数据在空间结构、语义合理性、尤其在复杂交通场景下，生成数据可能出现目标形变、遮挡不合理或时间不连续等问题，从而影响下游模型的训练效果。未来研究需要构建系统化、可量化且高效的评估方法，例如结合多模态一致性指标、特征可分性、同时探索生成与评估闭环机制，使生成模型能够在评估反馈下持续优化，提升生成数据的有效性与可用性。

6.5 跨视角与时序一致性难以保障

在多视角和时序交通场景生成中，保证生成数据在不同摄像头视角及连续帧之间的一致性是核心难题。现有方法在融合图像、点云、BEV 布局、文本描述等多模态条件时，往往存在条件对齐不充分、噪声累积及帧间跳变等问题，导致生成数据出现空间扭曲或语义冲突。这不仅影响数据可靠性，也制约下游感知与预测模型性能的提升。此外，缺乏端到端闭环奖励机制，使得生成过程与质量评估之间缺少直接反馈和互促机制，难以实现生成数据的自适应优化。未来研究应探索跨视角、跨帧一致性的统一建模策略，并结合闭环奖励或自监督反馈，使生成与评估能够互相驱动，从而提升单车智驾数据生成技术的可控性、稳定性和泛化能力。晴天）和特定城市道路布局下训练，其生成的数据可能无法有效泛化到其他域。例如，在晴朗天气下训练的模型，在生成雨、雪、雾等恶劣天气下的数据时，可能无法准确模拟传感器衰减（如 LiDAR 点云稀疏、相机图像模糊）和复杂的照明变化。同样，在一个城市学习的道路拓扑、交通标志和建筑风格，在另一个法规、布局迥异的城市可能完全不适用，导致生成的数据无法适配真实应用场景。

更进一步，挑战来源于协同规则与交互行为层面的域差异。V2X 的核心是协作，而协作规则和交

表 3 典型单车智驾数据生成基准总览表

Table 3 Overview Table of Typical single-vehicle intelligent driving data Generation benchmarks

数据集名称	数据规模	分辨率	数据集链接
Cityscapes	5K 精标图像	2048×1024	https://www.cityscapes-dataset.com
nuImages	93K 图像	1600×900	https://www.nuscenes.org/download
nuScenes	1000 场景 （约 5.5 小时视频）	1600×900	https://www.nuscenes.org/download
Waymo Open	1150 场景 （每个场景约 20s）	1920×1080	https://waymo.com/open

通参与者的行为模式深受当地交通法规、文化习惯甚至驾驶伦理的影响。例如,在“让行”规则不同的路口,车辆间的交互逻辑会完全不同。当前的数据生成方法,无论是基于场景编辑还是条件生成,都严重依赖于训练数据中隐含的交互模式,难以创造性地生成符合未知地域交通法规的、合理的多智能体行为序列。这导致生成的数据在语义和逻辑层面缺乏真实性,无法为需要在全球范围内部署的自动驾驶系统提供有效的、泛化的训练支持。

表 4　典型车路协同数据生成基准总览表

Table 4　Typical Vehicle-Infrastructure Collaborative Perception Data Generation Benchmarks Overview

数据集	模式	来源	传感器	帧数	链接
V2X-Sim	V2X	模拟	激光雷达(32线),相机(1600x900)	10k	https://ai4ce.github.io/V2X-Sim/
OPV2V	V2V	模拟	激光雷达(64线),相机(800x600)	12K	https://mobility-lab.seas.ucla.edu/opv2v/
DAIR-V2X	V2I	真实	激光雷达(车端40线,路端300线),相机(1920x1080)	39k	https://air.tsinghua.edu.cn/DAIR-V2X/
V2V4Real	V2V	真实	激光雷达(32线)	20K	https://mobility-lab.seas.ucla.edu/v2v4real/

6.6　通信—生成的联合最优化

基于V2X协同感知的数据生成,其最终目的是为了提升真实的V2X系统性能。尽管基于扩散模型与高斯泼溅的生成方法在离线数据生成上取得了显著进展,但若要将数据生成技术真正应用于在线V2X系统或生成更具现实性的训练数据,通信链路带来的约束成为一个不可回避的核心挑战。当前研究大多在理想通信假设下进行,忽视了真实V2X环境中存在的带宽限制、延迟、丢包及量化失真等非理想因素,导致“通信”与“生成”环节相互割裂,面临一系列联合优化难题。

首先,通信的不可靠性导致生成模型输入条件恶化,构成了首要挑战。在真实场景中,生成模型接收到的并非完整、无损的协同数据,而是经过高度压缩、可能存在丢包或带有信道噪声的特征。这种不完美的输入条件与生成模型在理想数据下训练的假设严重不符,不仅可能导致生成质量下降,更可能引发模型“幻觉”,产生事实上不存在的交通元素,为自动驾驶系统带来潜在的安全风险。

其次,通信与生成的联合设计范式亟待探索。传统的通信策略以最大化信道容量或数据传输效率为目标,而未考虑其对下游生成任务的影响。然而,对于生成模型而言,不同信息其“生成价值”各异。例如,关于被遮挡区域的关键信息即使数据量小,也应在通信调度中被优先保障。因此,如何构建以生成任务效用为导向的新型通信协议与资源分配机制,实现从“保数据”到“保感知”的范式转变,是一个前沿且复杂的优化问题。

最后,面向闭环系统的因果性与延迟挑战将问题进一步复杂化。当生成技术被用于在线系统进行实时感知补全或作为闭环仿真的一部分时,通信延迟使得生成模型所依赖的往往是过去时刻的场景状态。这就要求生成模型不仅要完成空间的补全,还需具备一定的时间预测能力,以补偿通信引入的滞后效应。这超越了当前绝大多数静态生成模型的能力范围,对生成范式提出了从“空间生成”到“时空预测”的更高要求。

综上所述,通信与生成的深度融合是未来V2X数据生成技术发展的关键方向。推动通信感知的生成模型与生成任务导向的通信协议的协同设计与联合优化,是构建能够在真实、复杂通信环境下稳定运行的高性能V2X系统的必由之路。

7　结　语

本文围绕高阶智能驾驶系统的数据生成问题,系统梳理了面向智能座舱(intelligent cockpit)、单车智驾(vehicle-centric smart driving)及V2X协同感知的数据生成技术及应用。针对智能座舱多模态数据匮乏所带来的建模瓶颈,本文从语音与文本生成、人脸与人体动作生成以及系统状态生成三类技术进行了系统分析,构建了涵盖声学、视觉及交互逻辑的座舱数据生成框架,展示了生成模型在语义表达、人物形象与系统操作方面提供高质量模拟数据的能力,为提升座舱感知、理解与人机交互能力提供了可行路径,并为构建以人为中心的智能座舱数据体系奠定了基础。在单车智驾数据生成方面,本文总结了

场景语义驱动、空间结构驱动及多模态条件联合驱动的生成模型，分析了这些方法在保证物理合理性和语义一致性的前提下，如何扩展训练数据覆盖、提升感知模型泛化能力及系统稳健性。在V2X协同感知数据生成部分，本文梳理了基于扩散模型的生成方法、基于高斯泼溅的重建编辑方法及基于测试与对抗生成的构造方法，阐释了其在多智能体同步数据采集、长尾和极端场景生成中的价值，为覆盖感知、预测和规划的闭环系统测试提供了坚实的数据支撑。针对高阶智驾的数据评测问题，本文构建了多层次评测框架，包括基础评测（人工评测、统计指标与可视分布分析）、深度学习驱动的自动评测（基准化模型指标与任务/应用特定指标）以及闭环评测与多能力基准，形成了可解释、可量化且与工程实践紧密结合的评测体系，为生成模型的优化和迭代提供了方法支撑。此外，本文对典型数据集和工具进行了整理，为单车、座舱及V2X场景的数据获取和生成提供了参考。综合分析表明，生成式数据在补充现实数据不足、构建极端与长尾场景、提升多模态一致性及系统稳健性方面发挥了核心作用，为高阶智能驾驶系统在感知、预测、决策及闭环验证中提供了可持续的数据支撑，并为构建可扩展、可控且高质量的数据生成体系奠定了理论与实践基础，为未来智能座舱、单车智驾和V2X协同感知技术的发展提供了重要参考和方法指导。

致谢：本文由中国图象图形学学会交通视频专业委员会组织撰写。本综述得到国家自然科学基金项目“极端环境视觉流感知与理解理论及方法”、国家自然科学基金面上项目“跨模态可信的行人再识别关键技术研究”、国家自然科学基金青年基金“复杂交通场景下基于局部特征交互的协同感知方法研究”、以及国家自然科学基金重大研究计划“重点支持项目”：“非完美标注下智能座舱多模态大模型训练研究”的资助，在此表示感谢。

参考文献（References）

Alahi A, Goel K, Ramanathan V, Robicquet A, Fei-Fei L and Savarese S. 2016. Social LSTM: Human trajectory prediction in crowded spaces [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE: 961 - 971 [DOI: 10.1109/CVPR.2016.110]

Bansal M, Krizhevsky A and Ogale A S. 2018. ChauffeurNet: Learning to drive by imitating the best and synthesizing the worst [EB/OL].

Bernardin K and Stiefelhagen R. 2008. Evaluating multiple object tracking performance: the CLEAR MOT metrics. EURASIP Journal on Image and Video Processing, 2008: 246309 [DOI: 10.1155/2008/246309]

Bińkowski M, Sutherland D J, Arbel M and Gretton A. 2018. Demystifying MMD GANs [C]//Proceedings of the 6th International Conference on Learning Representations. Vancouver: Curran Associates Inc.: 4783 - 4818 [DOI: 10.48550/arXiv.1801.01401]

Bradley R A and Terry M E. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. Biometrika, 39(3/4): 324 - 345 [DOI: 10.2307/2334029]

Brock A, Donahue J and Simonyan K. 2018. Large-scale GAN training for high-fidelity natural image synthesis [EB/OL]. [2018-9-28]. https://arxiv.org/abs/1809.11096.

Caesar H, Bankiti V, Lang A H, Vora S, Liong V E, Xu Q, et al. 2020. nuScenes: A multimodal dataset for autonomous driving//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE: 11618-11628. [DOI: 10.1109/CVPR42600.2020.01164]

Carreira J and Zisserman A. 2017. Quo vadis, action recognition? A new model and the Kinetics dataset [C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE: 4724 - 4733 [DOI: 10.1109/CVPR.2017.502]

Chen K, Xie E Z, Chen Z, Wang Y B, Hong L Q, Li Z G, et al 2023. GeoDiffusion: Text-prompted geometric control for object detection data generation//Proceedings of 2024 International Conference on Learning Representations (ICLR).

Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, et al. 2016. The Cityscapes dataset for semantic urban scene understanding//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE: 3213 - 3223. [DOI: 10.1109/CVPR.2016.350]

Defossez A, Mazaré L, Orsini M, Royer A, Pérez P, Jégou H, et al. 2024. Moshi: a speech-text foundation model for real-time dialogue [R]. Kyutai: Technical Report. Available: https://arxiv.org/pdf/2410.00037

Deng J, Dong W, Socher R, Li L J, Li K and Li F F. 2009. ImageNet: a large-scale hierarchical image database//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA: IEEE: 248-255 [DOI: DOI: 10.1109/CVPR.2009.5206848].

Duan Y Q, Guo X D, Zhu Z, Wang Z, Wang Y K and Lin C T. 2024. MaskFuser: Masked Fusion of Joint Multi-Modal Tokenization for End-to-End Autonomous Driving [EB/OL].

Dwork C and Roth A. 2014. The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3 - 4): 211 - 407 [DOI: 10.1561/0400000042]

Fang Q, Guo S, Zhou Y, Ma Z, Zhang S, and Feng Y. 2024. Llama-Omni: Seamless speech interaction with large language models[EB/OL]. [2026-01-07]. https://arxiv.org/abs/2409.06666.

Gao H A, Gao M J, Li J J, Li W Y, Zhi R, Tang H, et al. 2024. SCP-Diff: Spatial-categorical joint prior for diffusion-based semantic image synthesis//European Conference on Computer Vision. Cham: Springer Nature Switzerland: 37 - 54 [DOI: 10.1007/978-3-031-73411-3_3]

Gao R Y, Chen K, Li Z H, Hong L Q, Li Z G and Xu Q. 2024. MagicDrive3D: Controllable 3D generation for any-view rendering in street scenes[EB/OL].[2024-05-23]. https://arxiv.org/abs/2405.14475.pdf

Gao R Y, Chen K, Xiao B, Hong L Q, Li Z G and Xu Q. 2025. MagicDrive-V2: High-resolution long video generation for autonomous driving with adaptive control//Proceedings of the IEEE/CVF International Conference on Computer Vision: 28135 - 28144.

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.]

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. 2014. Generative Adversarial Networks [C]// Advances in Neural Information Processing Systems (NeurIPS). [DOI: 10.1007/978-981-33-6048-8_1]

Guo A, Gao X, Fang C, Tian H, Sun W, Mu Y, 等 2025. Generate Realistic Test Scenes for V2X Communication Systems [EB/OL]. [2025-06-09]. https://arxiv.org/pdf/2506.07419.pdf

Guo Y, Wang H, Hu Q, Liu H, Liu L and Bennamoun M. 2020. Deep learning for 3D point clouds: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(12): 4338 - 4364 [DOI: 10.1109/TPAMI.2020.3005434]

Hassid M, Remez T, Nguyen T A, Gat I, Conneau A, Kreuk F, et al. 2024. Textually pretrained speech language models[C]//Advances in Neural Information Processing Systems, 36. [DOI: 10.48550/ARXIV.2305.13009]

Heusel M, Ramsauer H, Unterthiner T, Nessler B and Hochreiter S. 2017. GANs trained by a two time-scale update rule converge to a local Nash equilibrium [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017). Long Beach: Curran Associates Inc.: 6629 - 6640 [DOI: 10.5555/3295222.3295408]

Hintze J L and Nelson R D. 1998. Violin plots: A box plot - density trace synergism. The American Statistician, 52(2): 181 - 184 [DOI: 10.1080/00031305.1998.10480559]

Ho J, Jain A and Abbeel P. 2020. Denoising diffusion probabilistic models [EB/OL].[2020-06-19]. https://arxiv.org/abs/2006.11239

Holden D, Saito S, and Komura T. 2017. Phase-functioned neural networks for character control[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). [DOI: 10.1145/3072959.3073663]

Hu A, Russell L, Yeo H, Murez 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 E, Shen Y L, Wallis P, Allen-Zhu Z Y, Li Y Z, Wang S, et al. 2022. LoRA: Low-rank adaptation of large language models//International Conference on Learning Representations, 1(2): 3.

Huang B Y, Wen Y Q, Zhao Y C, Hu Y S, Liu Y F, Jia F, et al. 2025. SubjectDrive: Scaling generative data in autonomous driving via subject control//Proceedings of the AAAI Conference on Artificial Intelligence [DOI:10.1609/aaai.v39i4.32376]

Isola P, Zhu J Y, Zhou T H and Efros A. 2017. Image-to-image translation with conditional adversarial networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [DOI: 10.1109/CVPR.2017.632]

Jagtap A D, Song R, Sadashivaiah S T and Festag A. 2025. V2X-Gaussians: Gaussian Splatting for Multi-Agent Cooperative Dynamic Scene Reconstruction [C]// IEEE Intelligent Vehicles Symposium. IEEE. 1033-1039 [DOI: 10.1109/IV64158.2025.11097436]

Jia X, Yang Z, Li Q, Zhang Z and Yan J. 2024. Bench2Drive: Towards Multi-Ability Benchmarking of Closed-Loop End-to-End Autonomous Driving [C]//Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2024), Datasets and Benchmarks Track. Vancouver: NeurIPS: 1 - 18 [DOI: https://doi.org/10.52202/079017-0025]

Karras T, Aittala M, Laine S, Härkönen E, Hellsten J, Lehtinen J, et al. 2022. Alias-free generative adversarial networks[C]//Advances in Neural Information Processing Systems (NeurIPS). [DOI: 10.54254/2753-8818/2025.22669]

Karras T, Laine S and Aila T. 2019. A style-based generator architecture for generative adversarial networks//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) [DOT: 10.1109/CVPR.2019.00453]

Karras T, Laine S, Aittala M, Hellsten J, Maaronen J, Lehtinen J, et al. 2020. Analyzing and improving the image quality of StyleGAN [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). [DOI: 10.1109/cvpr42600.2020.00813]

Kim S W, Philion 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].

Lakhotia K, Kharitonov E, Hsu W-N, Adi Y, Polyak A, Bolte B, et al.

2021. On generative spoken language modeling from raw audio[J]. Transactions of the Association for Computational Linguistics, 9: 1336 - 1354. [DOI: 10.1162/tacl_a_00430]

Le H, Pino J, Wang C, Gu J, Schwab D, and Besacier L. 2020. Dual-decoder transformer for joint automatic speech recognition and multilingual speech translation[C]//Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain (Online): International Committee on Computational Linguistics: 3520 - 3533. [DOI: 10.18653/V1/2020.COLING-MAIN.314]

Li B H, Ma Z, Du D L, Peng B R, Liang Z J, Liu Z Q, et al. 2025. OmniNWM: Omniscient driving navigation world models[EB/OL]. [2025-10-21]. https://arxiv.org/abs/2510.18313.pdf

Li D Q, Ling H, Kim S W, Kreis K, Barriuso A, Fidler S, et al. 2022. BigDatasetGAN: Synthesizing ImageNet with pixel-wise annotations//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE: 21330 - 21340 [DOI: 10.1109/CVPR52688.2022.02064]

Li H T, Yang Z, Qian Z Z, Zhao G P, Huang Y Q, Yu J, et al. 2025. DualDiff: Dual-branch diffusion model for autonomous driving with semantic fusion//Proceedings of the IEEE International Conference on Robotics and Automation (ICRA 2025. IEEE. [DOI: 10.1109/ICRA55743.2025.11128068]

Li X F, Zhang Y F and Ye X Q.2024. DrivingDiffusion: Layout-guided multi-view driving scenarios video generation with latent diffusion model//European Conference on Computer Vision. Milan, Italy: Springer:469-485[DOI:10.1007/978-3-031-73229-4_27]

Li Y H; Liu H T, Wu Q Y, Mu F Z, Yang J W, Gao J F, et al 2023. GLIGEN: Open-set grounded text-to-image generation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 22511 - 22521 [DOI: 10.1109/CVPR52729.2023.02156]

Li Y, Ma D, An Z, Wang Z, Zhong Y, Chen S, 等 2022. V2X-Sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving[J]. IEEE Robotics and Automation Letters. 7 (4): 10914-10921) [DOI: 10.1109/LRA.2022.3192802]

Likert R. 1932. A technique for the measurement of attitudes. Archives of Psychology, 140: 1 - 55

Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P and Zitnick C L. 2014. Microsoft COCO: Common objects in context [C]//European Conference on Computer Vision. Zurich: Springer: 740 - 755 [DOI: 10.1007/978-3-319-10602-1_48]

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, 钟芳栝, 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].

Lu Y F, Ren X C, Yang J W, Shen T C, Wu Z J, Gao J, et al. 2025. InfiniCube: Unbounded and controllable dynamic 3D driving scene generation with world-guided video models//Proceedings of the IEEE/CVF International Conference on Computer Vision: 27272 - 27283.

Mokady R, Hertz A, Aberman K, Pritch Y, Cohen-Or D. 2023. Null-text inversion for editing real images using guided diffusion models [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). [DOI: 10.1109/cvpr52729.2023.00585]

Nguyen Q, Vu T, Tran A and Nguyen K. 2023. Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation//Advances in Neural Information Processing Systems (NeurIPS).

Nguyen T A, Muller B, Yu B, Costa-Jussà M R, Elbayad M, Popuri S, et al. 2024. Spirit-LM: Interleaved spoken and written language model[EB/OL].[2026-01-07]. https://arxiv.org/abs/2402.05755.

Pan T Y, Jeon S, Fan M, Yoo J, Feng Z, Campbell M, 等 2025. Transfer Your Perspective: Controllable 3D Generation from Any Viewpoint in a Driving Scene[C]//Proceedings of the Computer Vision and Pattern Recognition Conference. 12027-12036 [DOI: 10.1109/CVPR52734.2025.01123]

Park J H, Jo K and Baik S. 2025. SeeDiff: Off-the-shelf seeded mask generation from diffusion models//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). [DOI: 10.1609/aaai.v39i6.32686]

Patashnik O, Wu Z, Shechtman E, Cohen-Or D, Lischinski D. 2021. StyleCLIP: Text-driven manipulation of StyleGAN imagery [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). [DOI: 10.1109/iccv48922.2021.00209]

Petrovich M, Black M J, Varo G. 2022. TEMOS: Generating diverse human motions from textual descriptions [C]//Proceedings of the European Conference on Computer Vision (ECCV). [DOI: 10.1007/978-3-031-20047-2_28]

Qi C R, Su H, Mo K and Guibas L J. 2017. PointNet: Deep learning on point sets for 3D classification and segmentation [C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE: 77 - 85 [DOI: 10.1109/CVPR.2017.16]

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. 2021. Learning transferable visual models from natural language supervision//Proceedings of the International Conference on Machine Learning. PMLR: 8748 - 8763.

Radford A, Kim J W, Xu T, Brockman G, McLeavey C, and Sutskever I. 2023. Robust speech recognition via large-scale weak supervision [C]//Proceedings of the International Conference on Machine Learning. PMLR: 28492 - 28518. [DOI: 10.48550/ARXIV.2212.04356]

Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, et al. 2021.

Zero-shot text-to-image generation//Proceedings of the International Conference on Machine Learning. PMLR: 8821 - 8831.

Rombach R, Blattmann A, Lorenz D, Esser P and Ommer B. 2022. High-resolution image synthesis with latent diffusion models//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). [DOI: 10.1109/CVPR52688.2022.01042]

Rubenstein P K, Asawaroengchai C, Nguyen D D, Bapna A, Borsos Z, Quitry F d C, et al. 2023. AudioPaLM: A large language model that can speak and listen[EB/OL].[2026-01-07]. https://arxiv.org/abs/2306.12925.

Savva M, Kadian A, Maksymets O, Zhao Y, Wijmans E, Jain B, et al. 2019. Habitat: A platform for embodied AI research[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).[DOI: 10.1109/iccv.2019.00943]

Shu D W, Park S W and Kwon J. 2019. 3D point cloud generative adversarial network based on tree structured graph convolutions [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul: IEEE: 3858 - 3867 [DOI: 10.1109/ICCV.2019.00396]

Song J, Meng C and Ermon S. 2020. Denoising Diffusion Implicit Models [EB/OL].[2020-10-06]. https://arxiv.org/abs/2010.02502

Sun P, Kretzschmar H, Dotiwalla X, Chouard A, Patnaik V, Paul Tsui P, et al. 2020. Scalability in Perception for Autonomous Driving: Waymo Open Dataset//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE: 2443-2451. [DOI: 10.1109/CVPR42600.2020.00252]

Sun W and Wu T F. 2019. Image synthesis from reconfigurable layout and style//Proceedings of the IEEE/CVF International Conference on Computer Vision: 10530-10539 [DOI: 10.1109/ICCV.2019.01063]

Sushko V, Schönfeld E, Zhang D, Gall J, Schiele B and Khoreva A. 2020. You only need adversarial supervision for semantic image synthesis//Proceedings of the European Conference on Computer Vision. Cham: Springer Nature Switzerland: 302 - 318. [DOI: 10.1007/978-3-030-65414-6_39]

Szegedy C, Vanhoucke V, Ioffe S, Shlens J and Wojna Z. 2016. Rethinking the Inception architecture for computer vision [C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE: 2818 - 2826 [DOI: 10.1109/CVPR.2016.308]

Tang H, Yu S Y, Pang J and Zhang B F. 2025. A training-free synthetic data selection method for semantic segmentation//Proceedings of the AAAI Conference on Artificial Intelligence, 39(7): 7229 - 7237. [DOI: 10.1609/aaai.v39i7.32777]

Tevet G, Raab S, Gordon B, Shafir Y, Cohen-Or D, Bermano A H. 2022. Motion diffusion model (MDM): Generating human motions with diffusion models[C]//Proceedings of the International Conference on Learning Representations (ICLR). [DOI: 10.1109/iros55552.2023.10342382]

Unterthiner T, van Steenkiste S, Kurach K, Marinier R, Michalski M and Gelly S. 2018. Towards Accurate Generative Models of Video: A New Metric & Challenges [EB/OL]. [2018-12-03].

Wang T C, Liu M Y, Zhu J Y, Tao A, Kautz J and Catanzaro. B. 2018. High-resolution image synthesis and semantic manipulation with conditional GANs//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [DOI: 10.1109/CVPR.2018.00917]

Wang W Z, Ma M C, Chen Y, Xia C Q, Liang Z B and Li J. 2025. FreeGen: Bridging visual-linguistic discrepancies towards diffusion-based pixel-level data synthesis//Proceedings of the AAAI Conference on Artificial Intelligence, 39(8): 7916 - 7924. [DOI: 10.1609/aaai.v39i8.32853]

Wang W Z, Zhao Y F, Ma M C, Liu M, Jiang Z L, Chen Y, et al. 2025. FICGen: Frequency-inspired contextual disentanglement for layout-driven degraded image generation//Proceedings of the IEEE/CVF International Conference on Computer Vision: 19097 - 19107.

Wang X D, Darrell T, Rambhatla S S, Girdhar R and Misra I. 2024. InstanceDiffusion: Instance-level control for image generation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 6232 - 6242 [DOI: 10.1109/CVPR52733.2024.00596]

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 X F, Zhu Z, Huang G, Wang B Y, Chen X Z and Lu J W. 2024. WorldDreamer: Towards general world models for video generation via predicting masked tokens [EB/OL].[2024-01-18]. https://arxiv.org/abs/2401.09985.pdf

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 2024IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 14749-14759 [DOI: 10.1109/CVPR52733.2024.01397]

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]

Wilk M B and Gnanadesikan R. 1968. Probability plotting methods for the analysis of data. Biometrika, 55(1): 1 - 17 [DOI: 10.1093/biomet/55.1.1]

Wu W J, Zhao Y Z, Chen H, Gu Y C, Zhao R, He Y F, et al. 2023.

DatasetDM: Synthesizing data with perception annotations using diffusion models//Advances in Neural Information Processing Systems (NeurIPS).

Wu W J, Zhao Y Z, Shou M Z, Zhou H and Shen C H. 2023. DiffuMask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models//Proceedings of the IEEE/CVF International Conference on Computer Vision. IEEE: 1206 - 1217 [DOI: 10.1109/ICCV51070.2023.00117]

Xiang H, Xu R, Xia X, Zheng Z, Zhou B and Ma J. 2023. V2XP-ASG: Generating Adversarial Scenes for Vehicle-to-Everything Perception [C]// IEEE International Conference on Robotics and Automation. IEEE. 3584-3591 [DOI: 10.1109/ICRA48891.2023.10161384]

Xie Z and Wu C. 2024. Mini-Omni: Language Models Can Hear, Talk While Thinking in Streaming[EB/OL].[2026-01-07]. https://arxiv.org/abs/2408.16725.

Xu H, Zhang S, Li P, Ye B, Chen X, Gao H A, 等 2025. Cruise: Cooperative reconstruction and editing in v2x scenarios using gaussian splatting[EB/OL].[2025-07-24]. https://arxiv. org/pdf/ 2507.18473. pdf [J]. arXiv preprint arXiv: 2507.18473.

Xu R S, Xia X, Li J L, Li H Z, Zhang S, Tu Z Z, 等. 2023. V2v4real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition [DOI: 10.1109/CVPR52729.2023.01318]

Xu R, Xia X, Li J, Li H, Zhang S, Tu Z, 等 2023. V2V4Real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 13712-13722. [DOI: 10.1109/CVPR52729.2023.01318]

Xu R, Xiang H, Xia X, Han X, Li J and Ma J. 2022. OPV2V: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication [C]//International Conference on Robotics and Automation. IEEE, 2583-2589 [DOI: 10.1109/ICRA46639.2022.9812038]

Yang J, Chen J, Yin Z, Chen S, Wang Y, Guo Y, et al. 2024. VehicleWorld: A highly integrated multi-device environment for intelligent vehicle interaction [C]//Proceedings of the IEEE Intelligent Vehicles Symposium (IV). [DOI: 10.18653/v1/2025. findings-emnlp.23]

Yang K R, Ma E H, Peng J B, Guo Q, Lin D and Yu K C. 2023. BEVControl: Accurately controlling street-view elements with multi-perspective consistency via BEV sketch layout [EB/OL].[2023-08-03]. https://arxiv.org/abs/2308.01661

Yang Z Y, Wang J F, Gan Z, Li L J, Lin K, Wu C F, et al. 2023. ReCo: Region-controlled text-to-image generation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 14246 - 14255 [DOI: 10.1109/CVPR52729.2023.01369]

Yi-Ge E and Shawn L. 2025. FlexDataset: Crafting annotated dataset generation for diverse applications//Proceedings of the AAAI Conference on Artificial Intelligence, 39(9): 9481 - 9489. [DOI: 10.1609/aaai.v39i9.33027]

Yoshihashi R, Otsuka Y, Doi K, Tanaka T and Kataoka H. 2024. Exploring limits of diffusion-synthetic training with weakly supervised semantic segmentation//Proceedings of the Asian Conference on Computer Vision: 2300 - 2318.

Yu H B, Luo Y Z, Shu M, Huo Y Y, Yang Z B, Shi Y F, 等. 2022. Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [DOI: 10.1109/CVPR52688.2022.02067]

Yu H, Luo Y, Shu M, Huo Y, Yang Z, Shi Y, 等 2022. DAIR-V2X: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 21361-21370 [DOI: 10.1109/CVPR52688.2022.02067]

Zhang D, Li S, Zhang X, Zhan J, Wang P, Zhou Y, et al. 2023. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities [C]//Findings of the Association for Computational Linguistics: EMNLP 2023. Singapore: Association for Computational Linguistics: 15757 - 15773 [DOI: 10.18653/v1/2023.findings-emnlp.1055]

Zhang D, Zhang X, Zhan J, Li S, Zhou Y, and Qiu X. 2024. SpeechGPT-Gen: Scaling chain-of-information speech generation [EB/OL].[2026-01-07]. https://arxiv.org/abs/2401.13527.

Zhang J H, Sheng H L, Cai S J, Deng B, Liang Q, Li W, et al. 2025. PerlDiff: Controllable street-view synthesis using perspective-layout diffusion models//Proceedings of the IEEE/CVF International Conference on Computer Vision.

Zhang L M, Rao A Y and Agrawala M. 2023. Adding conditional control to text-to-image diffusion models//Proceedings of the IEEE/CVF International Conference on Computer Vision: 3836 - 3847 [DOI: 10.1109/ICCV51070.2023.00355.]

Zhang Y X, Ling H, Gao J, Yin K X, Lafleche J, Barriuso A, et al. 2021. DatasetGAN: Efficient labeled data factory with minimal human effort//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE: 10145 - 10155 [DOI: 10.1109/CVPR46437.2021.01001]

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

Zhao G S, Ni C J, Wang X F, Zhu Z, Zhang X Y, Wang Y D, et al. 2024. DriveDreamer4D: world models are effective data machines for 4D driving scene representation [EB/OL].[2025-04-15]. https://arxiv.org/pdf/2410.13571.pdf

Zheng G C, Zhou X P, Li X W, Qi Z A, Shan Y and Li X. 2023. LayoutDiffusion: Controllable diffusion model for layout-to-image generation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 22490 - 22499 [DOI: 10.1109/CVPR52729.2023.02154]

Zheng W Z, Chen W L, Huang Y H, Zhang B R, Duan Y Q and Lu J W. 2024. OccWorld: 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]

Zhou D W, Li Y, Ma F, Zhang X T and Yang Y. 2024. MIGC: Multi-Instance Generation Controller for Text-to-Image Synthesis//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 6818 - 6828 [DOI: 10.1109/CVPR52733.2024.00651]

作者简介

赵耀，男，教授，主要研究方向为数字媒体信息处理与智能分析、生成式人工智能与安全。E-mail：yzhao@bjtu.edu.cn

李甲，通信作者，男，教授，主要研究方向为视觉内容感知理解与生成。E-mail：jiali@buaa.edu.cn

金一，女，教授，研究方向为多模态数据融合感知、交通视频语义理解、可信行为分析、多媒体大数据隐私保护。Email：yjin@bjtu.edu.cn

魏云超，男，教授，主要研究方向为图像/视频分割和物体检测、多模态数据建模、生成式人工智能。E-mail：yunchao.wei@bjtu.edu.cn

赵一凡，男，副教授，主要研究方向为计算机视觉、多模态内容解析生成、虚拟现实等。E-mail：zhaoyf@buaa.edu.cn

张慧，女，副教授，主要研究方向为多模态感知，具身智能。Email：huizhang1@bjtu.edu.cn

王旭，男，讲师，研究方向为机器视觉与智能AI算法。Email：xu.wang@bjtu.edu.cn

瞿梦雪，女，博士研究生，主要研究方向为多模态感知推理。E-mail：qumengxue@bjtu.edu.cn

曾宇乔，男，博士研究生，研究方向为多模态融合检测、图像融合。Email：yuqiaozeng@bjtu.edu.cn

王文状，男，博士研究生，主要研究方向为视觉内容可控生成与评估。E-mail：wz_wang@buaa.edu.cn