DOI:10.11918/202509069

自动驾驶中的世界模型:综述与展望

殷鸿博,田大新

(北京航空航天大学 交通科学与工程学院,北京 100074)

摘 要:在自动驾驶系统向通用智能演进的过程中,世界模型作为一种可对环境进行内在建模、推演与预测的认知引擎,正成为突破传统感知决策范式瓶颈、应对长尾场景的关键技术路径。为系统梳理世界模型在自动驾驶领域的研究进展与关键问题,探讨其推动通用智能驾驶落地的技术路径,文中对自动驾驶世界模型的研究现状与发展趋势开展了系统性综述。首先,阐明了世界模型的基本概念及其在自动驾驶中的核心功能,归纳了其主流技术架构,进而对比分析了各类范式的优势与不足。其次,总结了世界模型在3大关键应用方向的最新进展——未来场景生成与理解、端到端驾驶策略学习、数据驱动的闭环仿真系统,揭示了其在提升系统前瞻性与交互理解能力方面的实际价值。最后,系统整理了世界模型的评估指标与公开数据集的适用范围,为后续分析其技术挑战奠定了基础。结果表明:尽管世界模型在多尺度时空表征与复杂场景生成方面已取得阶段性突破,但在物理规律遵从性、安全可信推理、长时序稳定性及轻量化部署等方面仍存在显著挑战。据此,建议未来研究应重点关注高效计算架构、长时程生成一致性、不确定性建模及融合物理知识的自监督表征,以推动世界模型在各类交通场景中有效发挥作用。

关键词:智能交通;自动驾驶;世界模型;人工智能;闭环仿真

中图分类号:U463.6;TP18 **文献标志码**:A **文章编号**:0367-6234(2025)12-0165-14

World models in autonomous driving: A review and outlook

YIN Hongbo, TIAN Daxin

(School of Transportation Science and Engineering, Beihang University, Beijing 100074, China)

Abstract: Towards to general intelligentization of autonomous driving systems, the world models as a cognitive engine that internally models, infers, and predicts the environment, is becoming a critical technical pathway to break bottlenecks in traditional perception-decision paradigms and address long-tail scenarios. To synthesize the research progress and key issues of the world models in autonomous driving, and explore their technical routes for advancing the implementation of general intelligent driving, the research status and development trends in autonomous driving are reviewed. Firstly, the basic concept of world models and their core functionalities in autonomous driving are clarified, mainstream technical architectures are summarized, and the merits and drawbacks of various paradigms are comparatively analyzed. Secondly, the latest progress of world models in three key application directions are summarized including of future scene generation and understanding, end-to-end driving policy learning, and data-driven closed-loop simulation systems, and practical value in enhancing the system's forward-looking capabilities and interaction understanding is revealed. Thirdly, the evaluation metrics of world models and the application scopes of public datasets are organized, which lays a foundation for the subsequent analysis of their technical challenges. Overall, despite achieving phased breakthroughs in multi-scale spatiotemporal representation and complex scene generation, the world models still face the challenges in adhering to physical laws, safe and credible reasoning, long-term temporal stability, and lightweight deployment. Accordingly, it is suggested that future research should focus on efficient computing architectures, long-term generation consistency, uncertainty modeling, and self-supervised representation integrated with physical knowledge, so as to promote the effective function of world models in various traffic scenarios.

Keywords: intelligent transportation systems; autonomous driving; world models; artificial intelligence; closed-loop simulation

收稿日期:2025-09-16;**录用日期**:2025-10-13;**网络首发日期**:2025-11-14
网络首发地址:https://link.cnki.net/urlid/23.1235.T.20251113.1726.008
基金项目:国家自然科学基金(62571014,52202391,62432002);京津冀基础研究合作专项课题(F2024201070)
作者简介:殷鸿博(2000—),男,博士研究生;田大新(1980—),男,教授,博士生导师
通信作者:田大新,dtian@buaa.edu.cn

自动驾驶（autonomous driving，AV）作为未来交通体系的核心战略支撑，已被全球学术界与产业界广泛关注。自动驾驶是指车辆在极少或无需人类干预的条件下，通过集成先进传感器、定位与地图系统、高性能计算与决策模块，以及执行控制单元，实现环境感知、路径规划与自主导航驾驶[1-2]。大量研究显示，人为驾驶失误仍是交通事故的主要诱因，自动驾驶不仅有望通过减少人为失误来大幅降低交通事故和伤亡率[3-5]，同时被认为可以有效缓解交通拥堵、提升效率，并为智慧出行、物流运输及城市规划创造巨大的经济价值[6]。

然而，尽管技术快速发展，实现真正意义上的高级别自动驾驶仍面临多重挑战与瓶颈。当前的核心问题主要体现在感知、决策与场景泛化等方面。首先，车辆需要在动态、复杂且充满噪声的现实环境中，对来自激光雷达、毫米波雷达、摄像头等多模态传感器的高频数据进行实时融合与解析。系统不仅需对动态物理世界的状态进行准确感知与深层理解[7-8]，还需预测未来场景演化，以应对传感器失效、光照突变或积雪覆盖的道路路面等极端情况。其次，自动驾驶系统的安全性高度依赖实时决策。突发障碍、非结构化交通行为及不确定异常行为必须在毫秒级被系统响应[9]，周围交通参与者（含行人与其他车辆）的潜在意图与未来行为需被准确建模，以避免瞬时响应误差危及行车。此外，长尾场景与罕见高危事件的数据稀缺问题长期制约着系统的可靠性评估。罕见事件通常决定整体安全上限，但由于此类数据难以采集，且其在低频高危情境下的泛化能力不足，难以通过有限测试里程全面验证其安全性安全[10-11]。

传统的自动驾驶系统采用模块化部署策略，其中各功能模块（如感知、预测和规划[12]）被独立开发并集成到车辆中。由于模块间相互独立，系统难以实现全局协同优化，信息割裂与误差累积问题也被长期存在并制约整体性能提升[13]。近年来，数据驱动的端到端学习方法被提出，通过神经网络直接建立从传感器输入到控制输出的映射关系，在一定程度上缓解了模块间信息损失。但其“黑箱”特性缺乏可解释性，导致复杂交互场景中的因果推理、长期规划与策略泛化仍难以被有效处理[14]。尤其在长尾与未知场景下，依赖真实世界数据驱动的端到端模型在物理世界的理解能力不足，高风险情境的泛化依然受限[15]。

因此，实现智能、安全的自动驾驶系统被认为需要具备类似人类驾驶员的“内在世界模型”能力，即外部环境应能在系统内部被建模、想象与推演。换言之，系统不仅需被动响应当前感知结果，还应能够在“脑海”中构建环境动态表征，预测未来变化、推演潜在后果，从而实现更高层次的认知与决策。

在此背景下，借鉴认知科学中预测编码（predictive coding）理论的人类认知机制思想，世界模型（world models）应运而生。其核心目标不再是将高维传感器输入直接映射到驾驶动作，而是学习从多传感器物理观测中压缩时空表征与动态演化规律的生成式内在模型[16]。本文基于世界模型的自动驾驶系统能够在低维、抽象的“想象空间”中进行高效的“行为推演”，在无需真实环境交互的情况下预测自身及其他交通参与者的未来状态，从而实现更安全、更高效且具备因果推理能力的决策。通过给定历史观测，世界模型可以同时建模场景内在物理表征和预测未来演化进程，将感知 - 预测 - 规划整合为统一网络进行联合学习和推理，利用环境表征来全局优化行为规划。世界模型范式的提出，为自动驾驶系统在复杂动态环境中的稳健性、可解释性与泛化能力提升提供了新的研究方向与技术路径。

近年来，随着生成式建模、强化学习[17]和自监督学习等前沿技术的快速发展，世界模型在自动驾驶领域不仅是学术界的研究热点，也已成为全球顶尖科技公司和自动驾驶企业争相布局的核心战略方向。2023 年 6 月，特斯拉首次阐述了其世界模型的构想，为其后续发布的 FSD V12 版本奠定了理论基础。同年，英国自动驾驶公司 Wayve 推出自驾世界模型 GAIA - 1[18]，能够理解重要驾驶概念并生成高度逼真和多样的驾驶场景。中国厂商也在加速跟进世界模型的搭建。2024 年 7 月，蔚来开发并部署了多元自回归生成式的驾驶世界模型，具备全量理解数据、具有长时序推演和决策能力。商汤绝影提出基于多模态大模型的“开悟”世界模型，能够理解真实世界的物理规则、交通规则，生成场景视频时间最长为 150 s、分辨率可达 1 080 P。本文系统性地梳理了自动驾驶世界模型这一前沿领域的国内、外研究进展与现状，围绕其核心技术范式、关键应用领域与未来发展挑战展开深入论述。首先，本文追溯了世界模型的概念起源与理论框架；其次，重点综述了当前 3 种主流技术路径，即基于循环状态空间模型的潜在动力学范式、基于联合嵌入预测架构的表征学习范式以及基于扩散模型的生成式范式，并从规划效率、表征能力和生成质量等维度，分析对比了各类方法的优势与局限性。在此基础上，本文进一步探讨了世界模型在未来场景生成与理解、端到端自动驾驶、数据驱动的闭环仿真三大核心应用领域的最新进展与落地案例。最后，本文结合当前研究瓶

颈,对该领域未来可能的发展方向进行了总结与展望,指出了其在模型部署效率、长时程预测一致性、安全保证、表征有效性以及物理规律融合等方面面临的关键挑战与潜在解决路径。

1　自动驾驶世界模型

1.1　世界模型的定义

“世界模型”的概念最早可以追溯到认知科学与神经科学领域。在 19 世纪,德国物理学家赫尔曼·冯·亥姆霍兹就提出了“大脑是预测机器”的观点,认为生物体的感知并非对外部信号的被动接收,而是一个主动的、基于内部模型进行预测并与感觉输入进行比较、修正的过程。在机器学习领域,LeCun 等[19]在 2018 年首次对世界模型进行了系统性地阐述与实现。通用世界模型架构旨在模仿人脑建构外部物理世界和预判决策过程,包含 3 个解耦的关键组件。

1.1.1　感知模块

感知模块,其功能类似于生物体的感官系统,是模型与物理世界交互的唯一接口。该模块的核心任务是接收来自多模态传感器的高维、原始数据流(如摄像头图像、激光雷达点云等),并利用深度编码器,如变分自编码器(variational autoencoders, VAE)、掩码自编码器(masked autoencoders, MAE)等,将其编码为低维的、结构化的潜向量或特征表征,从而捕捉到当前世界状态的关键几何、语义及动态信息,为后续下游模块提供一个统一、抽象且易于处理的场景表征。

1.1.2　记忆模块

记忆模块,其功能类似于生物体的海马体结构,是整合、记忆与预测时序动态的核心。该模块的核心功能是学习环境的时间动态性,即世界状态随时间演化的内在规律。它发挥着承上启下的关键作用:接收由感知模块编码的当前世界状态,并结合控制器的行为指令,来预测未来的世界状态。具体而言,模块要解决的核心问题是建模物理世界的状态转移函数 $P(S_{t+1}|S_t,a_t)$,其中:S_t 为在时刻 t 的世界状态,a_t 为智能体执行的动作。通过序列模型不断积累历史知识,学习环境的时间动态性,从而赋予智能体在抽象潜空间中进行预测推演的能力。

1.1.3　控制器

控制器,是世界模型架构的决策核心,其功能类似于大脑的运动皮层,负责根据对世界当前状态的理解和对未来的预期,来决定并输出具体的驾驶动作。决策规划过程是在感知模块和记忆模块共同构建的内部模拟世界中进行,使得智能体能够在数千次的模拟试错中快速优化其行为策略网络,而无需承担真实世界中的任何风险。

通过整合这些模块世界模型能够模拟人类对外部物理环境的时空认知,并设想自身行为对潜在未来世界的影响,从而弥补机器智能与具身智能之间的认知鸿沟,为更复杂的自主驾驶系统提供了方向。

如图 1 所示,区别于通用世界模型,自动驾驶的世界模型更加侧重于对道路场景的结构化理解。它将来自多视角相机图像、激光雷达点云以及高精地图等感知与先验信息共同映射至一个紧凑且抽象的潜在场景空间,并以此为统一表示实现感知与预测的一体化融合[20]。在这一潜在空间中,模型进一步学习一个可微分的动态状态转移机制,用以捕捉交通场景的物理演化规律,并对多主体的未来行为展开长时程推演。该推演不仅是生成式与概率性的,能够覆盖多种可能场景分支,而且在建模过程中严格遵循真实世界的几何与动力学约束。因此,它为构建具备全局时空感知、交互建模与未来推演能力的统一计算框架奠定基础,并为鲁棒的多智能体预测、闭环规划控制及长尾场景泛化提供了理论支撑与实现范式。

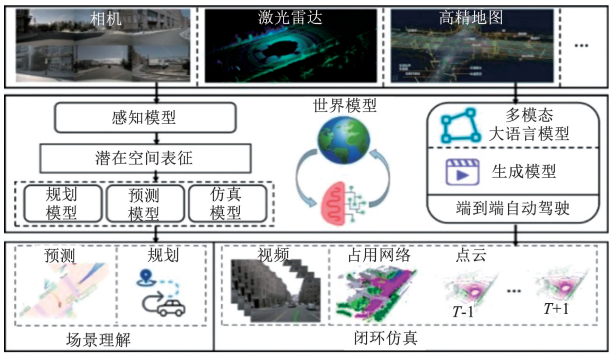


图 1　自动驾驶世界模型框架

Fig. 1　Framework diagram of autonomous driving world models

1.2　自动驾驶世界模型的问题描述

对于自动驾驶任务,准确预判自车运动及周围环境的未来演化状态至关重要。为解决这一核心挑战,世界模型的核心任务可以被形式化为一个基于历史观测序列的、条件性的时空动态联合预测问题。其根本目标是学习一个能够捕捉并推演整个驾驶场景时空演化规律的转移函数 w。

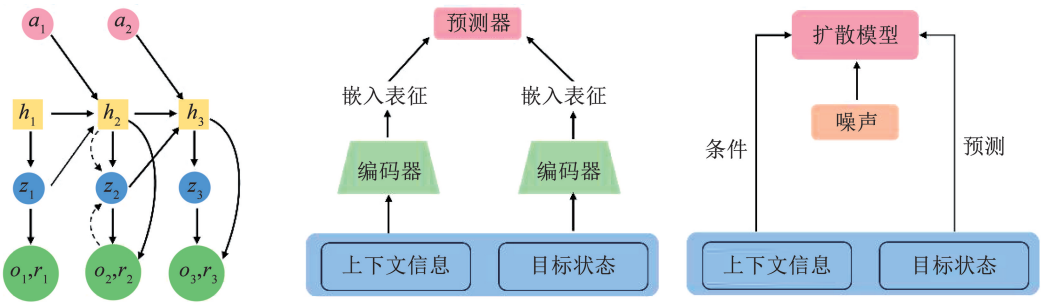
具体而言,给定自动驾驶系统过去 t 个时间步内对物理世界的观测 O(通常为采集到的多模态传感器序列),模型输出下一时刻的场景潜在表征 z_{T+1} 以及自车轨迹预测 τ_{T+1}。w 直接建模自车决策与周围交通演化之间的耦合动态,这一过程可以表示为

$$(z_{T+1},\tau_{T+1}) = w(O_T,\cdots,O_{T-t}) \tag{1}$$

由此,自动驾驶世界模型的首要任务是预测未来物理世界的状态,强调在快速变化、结构复杂的场景中捕捉潜在交互、随机行为以及不确定性。第 2 个核心任务是智能体的行为规划,其目标是为自车生成最优且可执行的轨迹,需要同时满足安全约束、动态障碍物规避、交通法规遵循以及实时自适应等要求。

2 世界模型技术范式

世界模型因其能够在潜空间中统一刻画场景几何、主体交互、动态演化与决策可行域而受到广泛关注。在深度学习和生成式 AI 技术发展的推动下,主流的世界模型架构(图 2)可以归纳为 3 种主要技术路径:1)基于循环状态空间模型的潜在动力学范式,直接在低维潜在空间中实现高效预测与规划;2)基于联合嵌入预测架构的表征学习范式致力于在不依赖像素级重构的条件下学习世界的抽象规律;3)基于扩散模型的生成式范式追求对未来观测进行极致逼真的像素级模拟。接下来,本文将对这 3 种技术架构进行系统性梳理,并横向对比其优势与局限性,见表 1。



(a) 循环状态空间架构 (b) 联合嵌入预测架构 (c) 扩散生成架构

图 2 世界模型技术架构示意

Fig. 2 Technical architecture diagram of world models

表 1 不同技术架构横向对比

Tab. 1 Horizontal comparison of different technical architectures

技术架构	潜在动力学范式	表征学习范式	生成范式
核心思想	低维潜在空间显式预测环境动态	抽象表征空间学习高层语义	迭代去噪生成未来观测
优点	规划效率高 决策友好 样本利用率高	表征能力强 模型轻量 泛化性好	生成质量高 多样性好 可控性强
局限性	长序列保真度低 物理一致性弱 表征能力受限	规划能力弱 动态建模弱 难以量化不确定性	计算开销大 采样效率低 存在物理幻觉
代表工作	PlaNet[21],Dreamer 系列[23-25],TD-MPC[26]	I-JEPA[27],V-JEPA[28]	Sora[31],Diffusion planner[35]

2.1 基于循环状态空间模型的潜在动力学范式

在真实物理场景中,智能体需要对部分可观测的世界状态进行准确建模和预测。基于循环状态空间模型[21](recurrent state-space model, RSSM)的潜在动力学范式提供了一种在低维潜在空间中高效表征和预测动态环境的方法。与直接在高维观测空间(如像素空间)进行建模和预测相比,能够显著降低计算复杂度并提高模型的泛化能力。其核心思想是将环境的状态分解为确定性状态和随机状态。前者通过循环神经网络[22](recurrent neural network, RNN)维护历史信息中的连续信息,捕捉时序依赖性;后者则通过随机变量建模环境随机性和不确定性。这种混合结构可确保强大的学习和预测能力,从而适应现实世界动态的不可预测性,同时保持信息连续性。

具体来说,首先将观测和动作序列表示为连续过程($o_0,a_1,o_1,a_2,o_2,\cdots,a_T,o_T$),即智能体会在观测到 o_t 后采取动作 a_{t+1},并接收下一时刻观测 o_{t+1}。接着,RSSM 通过以下过程对观测和状态转移进行建模:

$$p(o_{0:T}\mid a_{1:T}) = \int \prod_{t=0}^{T} p(o_t \mid z_{\leqslant t}, a_{\leqslant t}) p(z_t \mid z_{<t}, a_{\leqslant t}) \mathrm{d}z_{0:T} \tag{2}$$

式中 $z_{0:T}$ 为随机潜在状态,其近似后验概率为

$$q(z_{0:T}\mid o_{0:T}, a_{1:T}) = \prod_{t=0}^{T} q(z_t \mid z_{<t}, a_{\leqslant t}, o_t) \tag{3}$$

式中,先前的随机潜在状态 $z_{<t}$ 和动作 $a_{\leqslant t}$ 会随时间步迭代更新,RSSM 采用共享的门控循环单元(gate recurrent unit,GRU)将 $z_{<t}$ 和 $a_{\leqslant t}$ 映射为确定性状态编码 h_t,即

$$h_t = \mathrm{GRU}(h_{t-1}, \mathrm{MLP}(\mathrm{concat}[z_{t-1}, a_t])) \quad (4)$$

然后分别计算得到下一状态的先验概率分布 $p(z_t \mid z_{<t}, a_{\leqslant t})$、似然概率分布 $p(x_t \mid z_{<t}, a_{\leqslant t})$ 和后验概率分布 $q(z_t \mid z_{<t}, a_{\leqslant t}, o_t)$。训练优化目标是最大化证据下界(evidence lower bound,ELBO),即

$$\log p(o_{0:T} \mid a_{1:T}) \geqslant E_q\Big[\sum_{t=0}^{T} \log p(o_t \mid z_{\leqslant t}, a_{\leqslant t}) - L_{\mathrm{KL}}(q(z_t \mid z_{<t}, a_{\leqslant t}, o_t), p(z_t \mid z_{<t}, a_{\leqslant t}))\Big] \quad (5)$$

该目标函数由两部分构成,前一项为重构损失,衡量从后验分布中采样的潜在状态能否通过解码器重构出原始观测 o_t;后一项为动力学正则化,通过 KL 散度来度量先验预测与后验认知之间的分布差异,衡量对潜在空间动态演化的模拟性能。

2018 年,Hafner 等[21]在其开创性工作 PlaNet 中,首次成功证明 RSSM 架构可以从高维像素空间学习环境潜在动态特性,并直接在内在空间中进行在线规划,从而解决了复杂的连续控制任务,为后续研究奠定了坚实的基础。但是,PlaNet 在线规划效率低下,难以迁移到复杂任务。此后,Hafner 等[23-25]继续提出 Dreamer 系列工作,将在线规划转为生成低维潜在空间轨迹,从而避免计算耗时。Dreamer[23]首先利用真实交互数据训练好 RSSM 架构世界模型,然后在潜空间中生成大量虚拟未来轨迹,采用演员-评论家(actor-critic)强化学习算法训练策略网络和价值函数,极大提升样本效率。为了更好地建模世界的离散与多模态特性,DreamerV2[24]引入了离散潜在变量来替代连续的高斯变量,并采用直通梯度技术进行训练。这一改进使得模型能够更精确地重构图像细节,有效避免了 VAE 常见的细节丢失问题,从而能够处理更复杂的视觉环境。在雅达利(Atari)规划基准上首次超越顶尖无模型强化学习方法。为进一步构建一个通用的、可扩展的世界模型,DreamerV3[25]则通过对模型参数和训练目标的精心设计与调整,无需针对特定任务进行超参数微调,便能在极其广泛的任务领域,构建了能够支撑自动驾驶等复杂多变场景的基础模型。与此同时,研究人员也在探索利用 RSSM 架构世界模型进行决策的其他方法,例如,TD-MPC[26]便是一个典型代表。它巧妙地将时间差分学习与 MPC 相结合,直接在 RSSM 学习到的潜在动力学模型上,利用模型梯度进行在线轨迹优化,为高效决策提供了另一条优雅且有竞争力的技术方案。

基于 RSSM 的潜在动力学范式在低维潜在空间中进行多步更新,避免了反复调用高成本的渲染或物理仿真,极大地加速了策略学习和规划过程。同时潜在状态更新比高维观测空间更平滑、噪声更小,便于进行基于梯度的优化或采样式规划。然而,RSSM 类模型依然存在长期预测保真度低、表征能力不足以及物理与因果一致性弱等问题。

2.2　基于联合嵌入预测架构的表征学习范式

与致力于在潜在空间中显式建模完整动态的 RSSM 范式不同,基于联合嵌入预测架构(joint-embedding predictive architecture, JEPA)[27]的表征学习范式最早在 2018 年由 LeCun 提出,出发点是认为智能体对世界的理解与预测,无需以生成高保真度的像素级细节为前提。传统的世界模型(如基于变分自编码器的范式)为实现对高维观测的精确重构,必须分配大量的建模能力与计算资源去捕捉与高级语义和核心动态无关的随机细节(例如背景中树叶的摇曳、水面的波光粼粼)。这种对像素级保真度的过度追求,可能导致模型陷入对低层纹理信息的过度拟合,反而忽略了对世界抽象结构与因果关系的深层理解。

为规避这一问题,JEPA 采用在表征空间内直接进行预测的策略。首先,给定上下文信息(如一段视频或历史状态)x 和目标信息 y(如未来的某一帧图像),通过内容编码器 f_θ 和目标编码器 $f_{\bar{\theta}}$ 嵌入到表征空间得到 s_x 和 s_y。接着,预测器 g_φ 接收来自编码器的表征,并结合描述目标块位置的掩码隐变量 z_y,输出对目标表征的预测值 $\hat{s}_y = g_\varphi(s_x, z_y)$。

最后,模型优化基于预测表征 $\hat{s}_y$ 与真实目标表征 s_y 在嵌入空间中的 L_2 距离,对于单个目标块损失函数表示为

$$\boldsymbol{L}_{\varphi,\theta}(y) = \| g_\varphi(s_x, z_y) - s_y \|_2^2 \quad (6)$$

式中:预测器的参数 φ 和内容编码器的参数 θ 通过梯度优化学习,而目标编码器的参数 $\bar{\theta}$ 基于 θ 的指数移动平均值更新。

传统的自监督学习方法(如对比学习、掩码自编码器)在学习场景表征时,往往会引入不必要的偏见,为避免对高频噪声细节的无效建模,Assran 等[27]提出奠基性工作 I-JEPA,首次系统性地将 JEPA 架构应用于大规模图像数据预训练。该工作通过可见的上下文块表征,去预测多个被遮蔽的目标块在表征空间中的抽象表示,而非预测其原始像素。成功地学习到了强大的且具有语义可解释性的视觉表征,且在多项下游任务中表现出色,为其后续

向动态世界模型的演进奠定了坚实基础。紧接着，Bardes 等[28]将这一思想自然地扩展到了视频领域，以构建能够理解和预测物理世界动态的模型。通过利用一段历史视频帧作为上下文，去预测未来某个时间步长内多个被遮蔽的时空区域（video patches）的抽象表征，高效地学习到关于物体运动、交互等关键动态的抽象规律，同时避免对背景等非关键元素的无效预测。这标志着 JEPA 架构正式成为一种能够理解和预测世界如何随时间演化的世界模型，为自动驾驶等序列决策任务提供了新的建模思路。但是，先前工作学习到的场景表征依然仅适用于特定的训练环境，泛化能力受到限制。Zhou 等[29]提出将视觉表征学习与世界动态学习解耦的世界模型，从而继承预训练视觉编码器的泛化性，无需微调实现零样本学习。该工作揭示了基于不同模态的预训练基础模型，能够构建通用世界模型。

基于 JEPA 的表征学习范式能够高效鲁棒地学习通用世界的高层语义，且自监督的特性使其能够利用海量的无标签数据预训练。但是由于不直接建模状态随时间连续演化的过程，难以直接支撑闭环规划任务。

2.3 基于扩散模型的生成式范式

近年来，扩散模型（denoising diffusion probabilistic models，DDPM）[30]因其在图像生成任务中展现出的前所未有的高保真度与多样性，已成为生成式人工智能领域的颠覆性技术。受此启发，学术界开始探索将其强大的生成能力应用于世界模型的构建，旨在克服传统生成式范式（如 VAE）中存在的图像模糊、细节丢失以及长期预测“幻觉”等问题。

扩散模型本质上是一种学习数据分布的生成模型，其过程包含两个阶段：前向加噪过程与反向去噪过程。

1）前向加噪过程是一个固定的、无需学习的马尔可夫链。在 T 个离散的时间步中，通过向原始数据分布 x_0 有控制地注入高斯噪声 $\boldsymbol{\epsilon} \sim \boldsymbol{N}(0,\boldsymbol{I})$，得到加噪后的数据分布 x_t，这一过程可以表示为

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t} \tag{7}$$

式中，$\alpha_t = 1-\beta_t = \prod_{i=1}^{t}\alpha_i$，其中 β_t 为每个时间步的噪声强度。

2）反向去噪过程从纯粹的高斯噪声 $x_t \sim \boldsymbol{N}(0,\boldsymbol{I})$ 出发，通过逆转上述加噪过程让深度神经网络学会迭代地从噪声中采样出结构化的、符合原始数据分布的样本。

为了构建基于扩散模型的世界模型，核心在于实现其条件化生成能力。具体而言，模型必须能够根据给定的当前状态 s_t 与智能体采取的动作 a_t 作为条件，来精准地生成对下一时刻状态 s_{t+1} 的预测。因此，在给定加噪后的目标分布 x_0，条件信息 $c=(s_t,a_t)$ 和时间步编码 t 的前提下，模型训练一个条件化去噪网络 $\boldsymbol{\epsilon}_\theta$ 来学习潜在环境的复杂动态，即蕴含在状态 - 动作对 (s_t,a_t) 到下一状态 s_{t+1} 转移规律中的所有信息。模型的优化目标是最小化预测噪声与真实噪声之间的差异，其损失函数通常定义为

$$\boldsymbol{L}(\theta) = E_{t,x_0,c,\boldsymbol{\epsilon}}|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}x_0+\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon},t,c)|^2 \tag{8}$$

相较于通过离散隐变量序列建模环境动态，该方法可以最大程度保留关键场景细节，因此成为主流研究范式。OpenAI 发布的文生视频大模型 Sora[31]，展现出长时程、物理一致性视频生成的能力，涌现出物理世界动态的深刻隐式理解，是将扩散模型成功应用于世界模型构建的开创性工作之一。Xiang 等[32]首次将自然语言作为动作指令引入视频世界模型，允许实时控制生成未来世界，打通了语言与物理世界的桥梁，实现互动式场景生成和长时序推理。为了实现高度可控的多视角视频生成，Li 等[33]引入明确定义道路结构、车道线以及交通参与者的鸟瞰图布局作为强条件引导，设计跨视角注意力机制学习不同视角在内容和几何关系的一致性，从而增强时空连贯性和可控制性。此外，研究人员也在探索如何将基于扩散模型的世界模型直接用于交互决策和闭环规划。为避免视觉细节丢失影响决策，Alonso 等[34]以历史帧与动作为条件，在低维连续潜在空间扩散生成轨迹片段，结合强化学习策略训练首次超越传统离散潜变量建模方法。针对安全规划中的多目标优化问题，Zheng 等[35]提出基于分类器引导的扩散策略，将安全驾驶偏好条件作为控制条件，让运动输出与人类价值观对齐。

尽管基于扩散模型的世界模型生成质量高，能够模拟逼真且多样化的未来场景，但是也面临计算开销巨大、物理交互幻觉大等问题。

3 世界模型在自动驾驶中的应用

随着生成式人工智能[36-38]的发展，世界模型在自动驾驶领域的应用受到广泛关注。通过在统一的潜在空间内对多模态感知信息、多主体交互动态、物理运行规律与内在不确定性进行协同建模，世界模型能够赋能自动驾驶系统在未来场景生成理解、端到端自动驾驶和闭环模拟仿真等任务上的性能表现，见图 3。

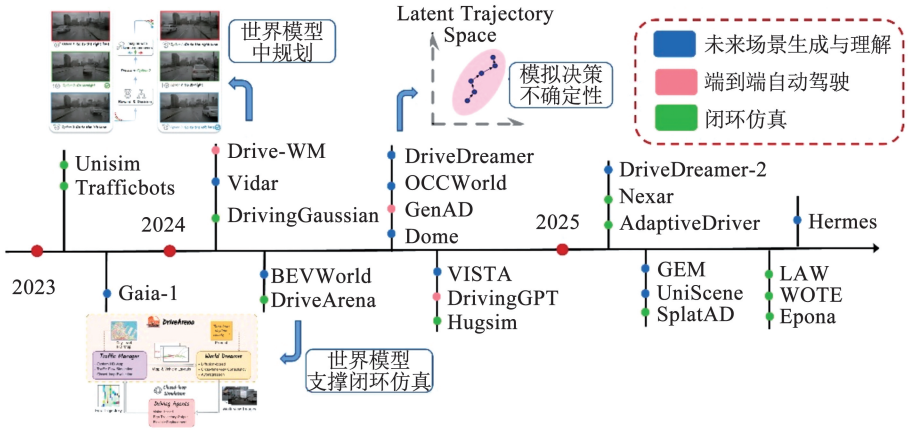


图 3　世界模型在自动驾驶中的应用概述

Fig. 3　Applications overview of world models in autonomous driving

3.1　未来场景生成与理解

对未来场景的生成与深层理解,是驱动自动驾驶系统从被动式的“感知 – 反应”模式,向具备前瞻性与预判能力的主动式“认知 – 决策”模式跃迁的核心基石。传统的自动驾驶系统往往依赖于对当前环境的瞬时观察,其决策逻辑可被简化为“看到什么,就反应什么”。这种模式在处理结构化道路环境时尚可应对,但在面对城市中心、交叉路口等动态交互频繁且潜在意图模糊的复杂场景时,则显得捉襟见肘,难以保障行驶的安全与效率。

为了突破这一瓶颈,研究人员率先探索了基于二维视觉表征的世界模型。Hu 等[18]开创性地将场景演化描述为“下一令牌预测”任务,通过自回归模型,将文本提示、历史图像与驾驶动作统一编码,生成时空动态序列。在此基础上,为了实现与规划意图的高度对齐,Wang 等[39]提出的 DriveDreamer 将场景显式解耦为静态背景与动态前景,利用细粒度的条件控制(如规划路径)来引导生成过程,从而为下游任务提供可控的扩展数据。DriveDreamer – 2[40]则进一步引入大语言模型的文本提示机制,赋予用户根据自然语言指令定制多样化场景的能力,显著增强了模型的交互性。随着研究的深入,模型的焦点转向了保真度、长时序一致性与在线规划能力。例如,Wang 等[41]通过视角因式分解的自回归方法解决了多视角场景的生成一致性问题,并首次将世界模型用作轨迹评分器,实现了在“想象”的未来中进行在线规划。Gao 等[42]则通过在隐空间引入动态实例先验,提升了模型对长时序、高分辨率场景的演化能力,并支持了轨迹、导航指令等多维度的动作控制。Hassan 等[43]的工作则将可控性推向了新的高度,通过引入行人位姿、深度图等多种控制条件,构建了一个强泛化的多模态世界模型,能够同时支持场景编辑、多模态输出与位姿修正等复杂任务。

尽管基于二维图像的方法取得了显著进展,但其固有的深度与几何信息缺失问题限制了模型的物理真实感。因此,另一条重要的技术路线转向了三维空间表征。Zheng 等[44]率先将复杂的时空动态预测问题重构为大规模三维体素序列的生成任务,通过预测驾驶空间中每个体素的未来占据状态,系统性地构建了首个动态场景三维占据世界模型。为解决离散化编码带来的细节损失,Gu 等[45]提出占用变分自编码器,将稠密体素压缩至连续表征空间,从而保留了精细的空间语义。此外,直接利用激光雷达(LiDAR)点云数据因其精确的几何信息,在动态场景捕捉和三维几何演化上展现出独特优势。

Yang 等[46]在预训练阶段引入视觉点云预测任务,通过三维渲染将历史视觉输入投影至几何空间,实现了场景语义、3D 结构与时间动态的协同学习。为了整合三维场景的理解与生成,Zhou 等[47]首次引入“世界查询”(world query)机制,构建了一个能够统一处理多种三维任务的框架。Li 等[48]利用高斯泼溅技术构建以三维占用为表征的统一生成框架,首次实现高保真统一生成语义占用、视频和激光雷达等不同预测结果。最终,为了克服单一模态的局限性,多模态融合成为构建全面世界表征的主流趋势,Zhang 等[49]提出里程碑工作 BEVWorld,创新性地将来自多视角摄像头、激光雷达等异构传感器的信息端到端地融合到一个统一的鸟瞰视角(Bird's-eye-view,BEV)表征中。它不仅能够感知当前,更能基于历史观测,生成未来多个时间步的 BEV 表征,其中包含了动态物体的轨迹、静态道路的结构以及场景的整体语义。

总结而言,未来场景生成的研究路径呈现出一条清晰的演进脉络。在表征层面,经历了从二维图像到三维占据栅格,再到多模态融合 BEV 的深化,其目标是构建一个日益全面、精确且物理一致的世

界状态表示。在生成范式上,技术从早期的无条件自回归生成,发展到由规划路径、文本指令乃至精细化姿态等多维度信号引导的条件可控生成,显著提升了模型的实用性与交互能力。在核心能力上,世界模型的功能已从单纯的“场景预测器”,扩展为能够支持在线规划、可控数据增广乃至作为通用任务接口的“世界引擎”。尽管在长时程一致性、计算效率与尾部事件的生成可信度上仍面临挑战,但这一系列进展无疑证明,高保真的未来场景生成与理解能力,是实现高级别自动驾驶系统鲁棒性、适应性与安全性的关键所在。

3.2 端到端自动驾驶

早期端到端自动驾驶模型接收传感器信号并直接映射为控制信号[50-52],这种“黑箱回归”范式因其内部决策逻辑不透明而备受诟病,泛化能力亦受限制。与此不同,世界模型作为一种内置的、可学习的动态环境模拟器,为端到端架构注入了深刻的变革。它通过支持时序建模、多可能性推演与人机交互,将原有系统重塑为一个基于“感知 - 模拟 - 规划”的、更为透明和鲁棒的决策新范式,显著提升了系统的可解释性、可控性与数据效率。

在这一新范式下,基于世界模型的生成式架构将复杂的驾驶任务,重构为一个基于多模态观测序列的自回归预测问题。该架构通过将图像、激光雷达、高精地图、历史轨迹与车辆运动状态等异构信息统一编码为离散令牌(token)序列,能够联合生成未来多视角的场景表征与车辆的行驶航点。Zheng 等[53]提出的生成式端到端框架 GenAD 是其中的代表性工作,它利用变分自编码器在结构化隐空间中捕捉未来轨迹的概率分布,从而模拟不确定性决策行为,显著提升了驾驶的安全性和舒适性。在此基础上,闭环世界模型的核心优势在于其“在想象中规划”(planning in imagination)的能力:模型于紧凑的潜在空间中快速“想象”并评估未来轨迹的长期奖励,从而在与其他交通参与者的动态博弈中寻得更优解。为将世界模型的推演能力显式地应用于规划,Wang 等[41]通过构建多视角世界模型,首次将其用作在线轨迹评分器,实现了在生成的未来场景中对候选路径的实时评估与择优。Li 等[54]则进一步发展了该思想,利用 BEV 世界模型进行在线轨迹评估,再次验证了该方法在复杂动态场景下进行端到端驾驶决策的有效性。

为应对大规模世界模型带来的高昂计算成本,研究人员也致力于提升模型的效率与表征能力。Li 等[55]通过引入轻量级的潜在世界模型(latent world model),在压缩的表征空间中进行高效的未来推演与策略学习,从而降低了对真实世界驾驶数据的依赖并显著减小了计算复杂度。在生成式架构方面,最新的研究工作积极借鉴多模态大模型的进展。例如,Chen 等[56]与 Zhang 等[57]均采用了先进的单一自回归或扩散模型架构,将感知、预测与规划任务无缝整合,用统一的模型实现世界建模、场景理解与运动规划,为构建通用驾驶智能体展现了清晰的技术路径。

尽管基于世界模型的端到端驾驶已取得显著进展,但该技术路线仍面临两大亟待解决的开放性挑战:有限上下文窗口的束缚与长时程预测的一致性难题。首先,当前基于 Transformer 的架构受限于其输入序列长度,难以处理超长时程的历史信息,这限制了模型对长期因果关系的理解。其次,在长达数秒乃至更长时间的闭环推演中,模型生成的场景容易出现物理失真、逻辑矛盾或细节漂移等问题,即“想象”出的世界会逐渐偏离真实物理规律。如何突破上下文窗口的限制,并确保模型在长时间推演中保持物理与逻辑的高度连贯性,是决定该技术能否在复杂多变的真实世界中规模化应用的关键。未来的研究方向可能包括:探索以 Mamba 为代表的新型状态空间模型架构以实现更高效的长时程记忆,设计层次化的世界模型以在不同时空尺度上进行推演,以及引入物理约束或因果推理机制来增强生成场景的真实感与逻辑一致性。

3.3 数据驱动的闭环仿真

闭环仿真是验证与迭代自动驾驶系统的核心环节,其关键在于能否在虚拟环境中安全、高效地复现真实世界的复杂性与不确定性。传统自驾仿真平台,如 CARLA[58],虽提供了可控的测试环境,但其依赖人工建模和图形学渲染的范式,在场景多样性、物理交互真实度以及与现实世界的“模拟 - 现实鸿沟”(sim-to-real gap)上暴露了固有局限。为突破此瓶颈,研究人员尝试利用世界模型构建数据驱动的新型仿真范式,其核心目标是创造出兼具高保真度与高可扩展性的数字孪生世界。

近期,三维高斯溅射(3D Gaussian splatting, 3DGS)技术[59]以其卓越的渲染质量与实时性能,成为该领域的主流方案。在这一技术浪潮下,Hess 等[60]开创性地利用车载多模态传感器数据流,构建了能够支撑实时渲染的 3DGS 场景表征,并通过对传感器特有伪影(如图像卷帘效应、激光雷达信号衰减)的精细建模与渲染优化,显著提升了重建场景的跨模态一致性。为将此技术应用于更大规模的

城市场景,Zhou 等[61]进一步提出了增量式静态场景构建与复合动态高斯图等策略,成功实现了对长轨迹驾驶场景的实时渲染,为大规模路测数据的仿真复现奠定了技术基础。

在高质量场景表征之上,研究界与产业界相继推出了功能强大的神经仿真器。NVIDIA 提出的 UniSim[62]和 Waymo 开发的 NeuroNCAP[63]能够将真实路采数据(Log)重建为可编辑、可交互的仿真环境。UniSim 允许用户通过自然语言指令对场景进行编辑(如增删物体、修改行为),从而高效地生成针对安全关键事件的“反事实”仿真。NeuroNCAP 则聚焦于感知系统的安全性评估,通过在仿真中生成对抗性实例,对感知算法进行系统性的压力测试以挖掘其潜在缺陷。学术界亦有诸多贡献,例如 Zhou 等[64]提出 Hugsim 支持对驾驶行为进行高效编辑,以模拟不同驾驶策略下场景的演化。Yang 等[65]则构建了一个统一的仿真基准平台,它不仅支持神经渲染,还兼容传统交通模型,为各类自动驾驶算法的综合性能评估提供了基准。

近期研究的重点转向了对多智能体交互与适应性行为的建模,以探索物理仿真引擎。Zhang 等[66]利用条件变分自编码器(conditional variational autoencoder, CVAE)学习不同驾驶风格下的交通流特性,能够生成符合物理规律、风格多样且与场景上下文紧密耦合的多智能体轨迹。针对跨城区交通行为的泛化性问题,Vasudevan 等[67]提出了一种自适应世界模型,该模型利用图神经网络直接推断其他智能体的控制器参数,从而构建出能够反映不同城市驾驶风格的反应式世界模型。更进一步地,Zhou 等[68]采用解耦扩散模型分别生成场景静态元素与动态智能体行为,且动态智能体的生成过程能够以被测车辆的规划为条件,从而模拟能够对被测车辆行为做出合理响应的交通流。

综上所述,世界模型正深刻地变革着自动驾驶闭环仿真的技术范式。从利用 3DGS 等技术实现对物理世界的高保真数字孪生,到构建可编辑、可交互的神经仿真器,再到生成具有社会智能的对抗性与适应性交通流,数据驱动的神经仿真不仅有效弥补了传统仿真器的“现实鸿沟”,更为自动驾驶系统提供能够进行大规模、高效率、深度交互的闭环仿真器。

4　评估指标与数据集

4.1　常用评估指标

为了综合评估世界模型应用于各种与驾驶有关任务上的性能表现,研究人员通常根据要解决的特定任务选择专门的指标。目前广泛使用的指标及含义描述见表 2。

表 2　世界模型性能评估指标

Tab. 2　Performance evaluation indexes of world models

评价指标	任务目标	指标含义描述
IS (inception score)	二维/三维场景生成	反映预测图像在特征空间中的区分度与多样性
FID (fréchet inception distance)	二维/三维场景生成	反映预测图像与真实图像在特征分布空间中的距离
P (precision)	二维/三维场景生成	衡量预测图像与真实图像在特征分布空间中高密度区域的重叠程度
R (recall)	二维/三维场景生成	衡量预测图像与真实图像在特征分布空间中覆盖范围的一致性
FVD (fréchet video distance)	视频生成	预测视频与真实图像在特征分布空间中的距离
CD (chamfer distance)	点云生成	预测点云与真实点云之间的平均双向最近邻距离
mIoU (mean IoU)	占用网格生成	所有类别交并比(IoU)数值的平均值
IoU (intersection over union)	占用网格生成	预测占据体与真实占据体之间交并比的比值
L2 (L2 displacement error)	开环路径规划	预测轨迹和真实轨迹之间的 L2 距离
CR (collision rate)	开环路径规划	沿预测轨迹与其他物体发生碰撞的频率
RC (route completion)	闭环路径规划	自车完成路线距离的百分比
DS (driving score)	闭环路径规划	平均路线完成图和交通违章数量的综合得分

4.1.1　Inception 分数(inception score,IS)

IS 衡量生成样本的清晰度与类别多样性,通过计算每个生成样本预测类别分布 $p(y\mid x)$ 与生成样本整体类别分布 $p(y)$ 的 KL 散度,并取指数,即

$$\mathrm{IS}=\exp\left(E_{x\sim p_g}\left[D_{\mathrm{KL}}(p(y\mid x)\|p(y))\right]\right) \tag{9}$$

式中:x 为生成样本,y 为标签类别,p_g 为生成分布。

4.1.2　Fréchet Inception 距离(fréchet inception distance,FID)

FID 衡量生成图像与真实图像在特征分布空间中的距离为

$$\mathrm{FID}=|\mu_r-\mu_g|_2^2+\mathrm{tr}(\Sigma_r+\Sigma_g-2(\Sigma_r\Sigma_g)^{1/2}) \tag{10}$$

式中：μ_r、Σ_r 分别为真实图像特征均值和协方差，μ_g、Σ_g 分别为生成图像特征均值和协方差。

4.1.3 Fréchet 视频距离（fréchet video distance，FVD）

FVD 为生成视频与真实视频在特征分布空间的距离，计算方式与 FID 相同。

4.1.4 Chamfer 距离（chamfer distance，CD）

CD 为预测点云与真实点云之间的平均双向最近邻距离，反映三维几何相似性，即

$$\mathrm{CD}(P,Q) = \frac{1}{|P|}\sum_{p\in P}\min_{q\in Q}\|p-q\|_2^2 + \frac{1}{|Q|}\sum_{q\in Q}\min_{p\in P}\|q-p\|_2^2 \tag{11}$$

式中：$p\in P$ 为预测场景的三维点云，$q\in Q$ 为真实场景的三维点云。

4.1.5 交并比（intersection over union，IoU）

CD 衡量预测占用网格与真实占用网格的重叠程度，用于四维占用网格生成任务，即

$$\mathrm{IoU} = \frac{|O\cap G|}{|O\cup G|} \tag{12}$$

式中：O 为预测占用网格面积，G 为真实占用网格面积。

4.1.6 平均交并比（mean intersection over union，mIoU）

mIoU 衡量所有类别的交并比平均值，衡量四维占据网格任务中预测与真实标签的重合程度，即

$$\mathrm{mIoU} = \frac{1}{C}\sum_{c=1}^{C}\frac{|O\cap G|}{|O\cup G|} \tag{13}$$

式中 C 为场景中语义类别数。

4.1.7 L2 位移误差（L2 displacement error，L2）

L2 反映预测轨迹与真实轨迹在每个时间步的欧氏距离，反映轨迹预测精度，即

$$\mathrm{L2} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\|\hat{P}_i^t - P_i^t\|_2 \tag{14}$$

式中：N 为样本数，T 为时间步数。

4.1.8 碰撞率（collision rate，CR）

CR 为沿预测轨迹与其他物体发生碰撞的频率，用于评价轨迹预测或规划的安全性，即

$$\mathrm{CR}(t) = \frac{\sum_{i=1}^{N} I_i}{N} \tag{15}$$

式中 I_i 为在第 i 个轨迹点处，自车是否和其他车辆发生碰撞。

4.1.9 驾驶分数（driving score，DS）

DS 综合路线完成率与交通违章数量得出的评分，用于闭环驾驶任务整体性能评估，即

$$\mathrm{DS} = \frac{1}{N}\sum_{i=1}^{N} R_i P_i \tag{16}$$

式中：N 为驾驶路线的数量，R_i 为第 i 条路线完成的百分比，P_i 为第 i 条路线的平均碰撞惩罚。

4.2 运动规划任务中的指标性能对比

以 nuScenes 数据集上的运动规划任务为例，表 3 汇总了部分基于世界模型方法的评估指标与结果。与端到端自动驾驶算法 ST-P3[69] 和 UniAD[50] 相比，世界模型在引入更丰富的监督信号后，显著提升了自动驾驶系统的规划精度与整体安全性能。

表 3 世界模型方法在 nuScenes 数据集上运动规划任务的性能对比

Tab. 3 Performance comparisons of world models methods for motion planning tasks evaluated based on the nuScenes datasets

方法	年份	输入模态	L2 位移误差/m				碰撞率/%			
			1 s	2 s	3 s	Avg.	1 s	2 s	3 s	Avg.
ST-P3[69]	2022	图像	1.33	2.11	2.90	2.11	0.23	0.62	1.27	0.71
UniAD[50]	2023	图像	0.48	0.96	1.65	1.03	0.05	0.17	0.71	0.31
UniAD + DriveWorld[70]	2024	图像	0.34	0.67	1.07	0.69	0.04	0.12	0.41	0.19
Drive-OccWorld[71]	2025	图像	0.32	0.75	1.49	0.85	0.05	0.17	0.64	0.29
DriveWM[41]	2023	图像	0.43	0.77	1.20	0.80	0.10	0.21	0.48	0.26
Epona[57]	2025	图像	0.61	1.17	1.98	1.25	0.01	0.22	0.85	0.36
FSDrive[72]	2025	图像	0.14	0.25	0.46	0.28	0.03	0.06	0.21	0.10
OccWorld[73]	2024	Occ 占用	0.43	1.08	1.99	1.17	0.07	0.38	1.35	0.60
RenderWorld[74]	2025	Occ 占用	0.35	0.91	1.84	1.03	0.05	0.40	1.39	0.61
OccLLaMA[75]	2024	Occ 占用	0.37	1.02	2.03	1.14	0.04	0.24	1.20	0.49

4.3 数据集

自动驾驶世界模型的训练与评估依赖于大规模、高保真、多模态的时空数据支撑。这类数据集通常涵盖丰富的场景语义、动态交通要素以及复杂的

人车交互关系,为模型学习环境表征、动态演化规律和决策预测提供基础。随着自动驾驶领域研究的不断深入,已涌现出一批具有代表性的世界模型相关公开数据集与仿真基准,见表 4。

表 4　典型世界模型数据集

Tab. 4　Typical world models datasets

数据集	年份	采集城市	采集方式	场景数量/k	任务
Carla	2017	未知	车载摄像头—激光雷达		闭环路径规划
KITTI	2012	1	车载摄像头—激光雷达	22	二维/三维场景生成
Waymo Open Motion	2020	3	车载摄像头—激光雷达	1	二维/三维场景生成,开环路径规划
nuScenes	2020	2	车载摄像头—激光雷达	1	二维/三维场景生成,开环路径规划
OpenScenes	2023	4	车载摄像头—激光雷达		占用网格生成与预测
OpenDV - 2K	2024	≥244	车载摄像头—描述性文本	2	二维场景生成与理解、开环路径规划
DrivingDojo	2024	9	车载摄像头—描述性文本	18	二维场景生成与理解

4.3.1　Carla

由西班牙巴塞罗那自治大学发布,用于自动驾驶研究的闭环仿真平台。数据通过虚拟环境合成,包含城市道路、乡村、交叉口等多种交通环境,可自定义天气、光照、交通密度等条件,适用于端到端自动驾驶的闭环测试。但与真实采集数据间存在显著差距。

4.3.2　KITTI

由卡尔斯鲁厄理工学院和丰田美国技术研究院联合发布,是自动驾驶领域早期基准数据集,包含立体图像、光流、激光雷达点云等多模态传感器数据。适用于二维/三维场景生成任务,但场景单一、交通参与者密度有限,难以覆盖复杂交互场景。

4.3.3　Waymo Open Motion

由 Waymo 公司发布,主要采集于美国城市(如旧金山、山景城等),数据规模超 1 000 h,提供高精度地图及动态交互信息,适用于行为预测理解以及世界模型长期场景生成。基于 Waymo 数据集标注的 Occ3D-Waymo 可以支撑四维占用网格生成与预测任务,涵盖 17 种语义类别。

4.3.4　nuScenes

由 Motional 公司在新加坡与波士顿两地城市采集,包含丰富的感知语义标注,适用于场景理解、短时场景预测和开环评测,但场景时长有限,驾驶行为单一且交互简单。

4.3.5　OpenScenes

由清华大学、上海人工智能实验室等机构联合发布,数据集提供跨模态同步数据(图像、点云、雷达、V2X 信息、地图等),支持长时间连续片段与多车视角,专为世界模型训练和评测设计。

4.3.6　OpenDV - 2K

由北京智源研究院与合作高校联合发布,是目前规模最大、场景最丰富的世界模型数据集之一。数据集结合互联网高质量驾驶视频,包含 2 000 + h 的多模态驾驶视频、语义标注、驾驶操作、轨迹与事件级标签,兼容视觉与语言描述,专门面向视觉语言世界模型,适用于视频预测、模拟与规划以及泛化性研究。

4.3.7　DrivingDojo

由中国科学院自动化所和美团联合发布,是首个专为自动驾驶世界模型研究设计的长尾数据集,包含丰富的驾驶行为、多智能体的交互以及世界知识,适用于高质量视频级场景生成与预测。

5　未来展望

尽管基于世界模型的自动驾驶技术取得显著进展,在长尾场景理解、决策可解释性以及闭环仿真验证展现出巨大潜力。然而,当前研究仍面临一系列严峻挑战,包括模型计算复杂度与车载平台实时性要求的矛盾、长时程预测的保真度衰减、对未来不确定性的量化不足以及数据驱动与物理规律一致性的保证等。为推动该领域向更高阶的认知智能与产业化落地迈进,未来研究可重点围绕以下方向展开探索。

5.1　高效计算与实时部署

当前,世界模型(尤其是基于 Transformer 和扩散模型的架构)往往伴随着巨大的计算开销和内存占用,如何在算力受限、功耗敏感的车载计算平台上部署是一个巨大的算法与工程挑战。未来的研究需要探索模型量化、剪枝、蒸馏等轻量化技术,并发展适用于车载硬件的新型高效模型架构(如状态空间模型 Mamba),以在模型性能与部署效率之间取得最佳平衡。

5.2　长时程预测的保真度与一致性

世界模型的核心能力在于“想象”未来,然而,

随着预测时程的增加,生成的场景容易出现物理失真、逻辑矛盾或细节漂移等问题,逐渐偏离真实物理规律。未来研究还需深度融合物理先验与因果表征,并探索能够抑制误差累积的新型序列模型架构,以从根本上提升长时程预测的逻辑一致性与物理真实性。

5.3 不确定性建模与安全保证

当前的世界模型在量化自身预测的不确定性、生成多样且合理的未来可能性方面仍有不足。未来研究可结合概率生成模型,如贝叶斯神经网络、能量模型等方法,使其能够为每一种推演结果提供可靠的不确定性估计,从而为下游的风险评估与鲁棒规划模块划定清晰的安全边界。

5.4 自监督学习的表征有效性

为摆脱对大规模人工标注数据的依赖,自监督学习已成为训练世界模型的主流范式。然而,其核心挑战在于设计对下游驾驶任务真正有效的自监督表征。未来的研究重心必须转向设计更具因果洞察的自监督范式,探索如对比学习、信息解耦等技术,迫使模型聚焦于场景中的驾驶关键要素,从而提升其表征的有效性与泛化能力。

5.5 数据驱动与物理规律的内在统一

世界模型作为数据驱动的生成式模型,其通过从海量观测数据中学习环境的隐式物理表征与动态演化规律。当面临分布外或对抗性场景时,其推演过程易生成与物理常识相悖的结果,未来的研究还须探索如何将经典物理学原理(如运动学约束、非完整约束、碰撞体积等)作为一种强先验,从而提升世界模型在长尾分布场景下的泛化能力与可靠性。

6 结 论

1)当前,基于世界模型的自动驾驶技术已成为学术界与产业界关注的核心方向,其统一表征能力和生成式推演能力正在重塑传统自动驾驶的技术栈,推动系统从“被动感知”走向“主动建模与推演”,对提升极端与长尾场景下的可靠性具有重要意义。

2)本文系统回顾了世界模型的发展脉络,总结了其在自动驾驶中的核心任务与主要技术范式,涵盖循环状态空间模型类潜在动力学建模、联合嵌入范式下的统一表征学习,以及扩散模型驱动的生成式推演,为理解其内部运行机理及能力边界提供了结构化视角。

3)本文从规划效率、表征能力、生成质量、计算开销与可控性等多个维度,深度剖析并横向对比了各类技术范式的核心优势及其固有局限性。

4)本文基于对典型应用的梳理,指出世界模型已在未来场景生成、端到端闭环控制与数据驱动仿真验证等应用方向展现出可观潜力,并形成初步落地态势;然而其真正迈向大规模应用仍面临计算成本、评测体系不完备,以及泛化鲁棒性不足等制约因素。

5)综合分析可见,世界模型正从“数据驱动的未来预测”向“结合物理推理的可解释闭环生成”。未来的发展方向将聚焦于高效计算架构、长时程生成一致性、不确定性建模以及融合物理知识和自监督学习的表征形式,以实现具备通用性、可验证性与可部署性的下一代智能驾驶体系。

参考文献

[1] On-Road Automated Driving Committee. Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles: SAE J3016 – 2021[S]. Warrendale: SAE International, 2021. DOI: 10.4271/j3016_202104

[2] FAGNANT D J, KOCKELMAN K. Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations[J]. Transportation Research Part A: Policy and Practice, 2015, 77: 167. DOI: 10.1016/j.tra.2015.04.003

[3] 布莱恩·瓦利留斯, 王德政. 人工智能时代自动驾驶中的注意义务[J]. 哈尔滨工业大学学报(社会科学版), 2024, 26(1): 42
VALERIUS B, WANG Dezheng. The duty of care about autonomous driving in the age of artificial intelligence[J]. Journal of Harbin Institute of Technology (Social Sciences Edition), 2024, 26(1): 42. DOI: 10.16822/j.cnki.hitskb.2024.01.003

[4] SOLTANI A, AFSHARI S, AMIRI M A. Time-series projecting road traffic fatalities in Australia: insights for targeted safety interventions [J]. Injury, 2025, 56(3): 112166. DOI: 10.1016/j.injury.2025.112166

[5] 陈淑琬, 赵鹏飞, 刘丹丹, 等. 2005—2021 年中国道路交通事故死亡趋势分析[J]. 疾病监测, 2025, 40(1): 133
CHEN Shuwan, ZHAO Pengfei, LIU Dandan, et al. Incidence trend of road traffic accident death in China, 2005—2021[J]. Disease Surveillance, 2025, 40(1): 133. DOI:10.3784/jbjc.202403260201

[6] ZHAO Jingyuan, ZHAO Wenyi, DENG Bo, et al. Autonomous driving system: a comprehensive survey[J]. Expert Systems with Applications, 2024, 242: 122836. DOI: 10.1016/j.eswa.2023.122836

[7] OMEIZA D, WEBB H, JIROTKA M, et al. Explanations in autonomous driving: a survey[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(8): 10142. DOI: 10.1109/TITS.2021.3122865

[8] GRIGORESCU S, TRASNEA B, COCIAS T, et al. A survey of deep learning techniques for autonomous driving[J]. Journal of Field Robotics, 2020, 37(3): 362. DOI: 10.1002/rob.21918

[9] PADEN B, ČÁP M, YONG S Z, et al. A survey of motion planning and control techniques for self-driving urban vehicles[J]. IEEE Transactions on Intelligent Vehicles, 2016, 1(1): 33. DOI: 10.1109/TIV.2016.2578706

[10] KOOPMAN P, WAGNER M. Autonomous vehicle safety: an interdisciplinary challenge[J]. IEEE Intelligent Transportation Systems Magazine, 2017, 9(1): 90. DOI: 10.1109/MITS.2016.2583491

[11] 邓伟文, 李江坤, 任秉韬, 等. 面向自动驾驶的仿真场景自动生成方法综述[J]. 中国公路学报,2022, 35(1): 316
DENG Weiwen, LI Jiangkun, REN Bingtao, et al. A survey on automatic simulation scenario generation methods for autonomous driving[J]. China Journal of Highway and Transport, 2022, 35(1):

316. DOI: 10.19721/j.cnki.1001-7372.2022.01.027
[12] 马艳丽, 董方琦, 秦钦, 等. 基于行车风险场的自动驾驶接管风险评估模型[J]. 哈尔滨工业大学学报, 2024, 56(9): 106
MA Yanli, DONG Fangqi, QIN Qin, et al. Risk evaluation model of autonomous driving takeover based on driving risk field[J]. Journal of Harbin Institute of Technology, 2024, 56(9): 106. DOI:10.11918/202211073
[13] YIN Hongbo, TIAN Daxin, LIN Chunmian, et al. V2VFormer++: multi-modal vehicle-to-vehicle cooperative perception via global-local transformer[J]. IEEE Transactions on Intelligent Transportation Systems, 2024, 25(2): 2153. DOI: 10.1109/TITS.2023.3314919
[14] 陈妍妍, 田大新, 林椿眄, 等. 端到端自动驾驶系统研究综述[J]. 中国图象图形学报, 2024, 29(11): 3216
CHEN Yanyan, TIAN Daxin, LIN Chunmian, et al. Survey of end-to-end autonomous driving systems[J]. Journal of Image and Graphics, 2024,29(11): 3216. DOI: 10.11834/jig.230787
[15] YIN Hongbo, TIAN Daxin, LIN Chunmian, et al. CUDA-X: unsupervised domain-adaptive vehicle-to-everything collaboration via knowledge transfer and alignment[J]. IEEE Transactions on Neural Networks and Learning Systems, 2025, 36(8): 14144. DOI: 10.1109/TNNLS.2025.3539358
[16] 任磊, 董家宝, 曾宪超, 等. 数字族谱:驱动工业具身智能世界模型[J]. 中国科学: 信息科学, 2025, 55(7): 1748
REN Lei, DONG Jiabao, ZENG Xianchao, et al. Digital genealogy: empowering industrial embodied intelligence world model[J]. Scientia Sinica (Informationis), 2025, 55(7): 1748. DOI: 10.1360/SSI-2025-0093
[17] 薛怡然, 吴锐, 刘家锋, 等. 组合动作空间深度强化学习的人群疏散引导方法[J]. 哈尔滨工业大学学报, 2021, 53(8): 29
XUE Yiran, WU Rui, LIU Jiafeng, et al. Crowd evacuation guidance based on combined action-space deep reinforcement learning[J]. Journal of Harbin Institute of Technology, 2021, 53(8): 29. DOI: 10.11918/202101029
[18] HU A, RUSSELL L, YEO H, et al. GAIA-1: a generative world model for autonomous driving[EB/OL]. 2023: arXiv: 2309.17080. https://arxiv.org/abs/2309.17080
[19] HA D, SCHMIDHUBER J. World models[EB/OL]. 2018: arXiv: 1803.10122. https://doi.org/10.48550/arXiv.1803.10122
[20] GUAN Yanchen, LIAO Haicheng, LI Zhenning, et al. World models for autonomous driving: an initial survey[J]. IEEE Transactions on Intelligent Vehicles, 2025: 1. DOI: 10.1109/tiv.2024.3398357
[21] HAFNER D, LILLICRAP T, FISCHER I, et al. Learning latent dynamics for planning from pixels[EB/OL]. 2018: arXiv: 1811.04551. https://arxiv.org/abs/1811.04551
[22] ZAREMBA W, SUTSKEVER I, VINYALS O. Recurrent neural network regularization[EB/OL]. 2014: arXiv: 1409.2329. https://arxiv.org/abs/1409.2329
[23] HAFNER D, LILLICRAP T, BA J, et al. Dream to control: learning behaviors by latent imagination[EB/OL]. 2019: arXiv: 1912.01603. https://arxiv.org/abs/1912.01603
[24] HAFNER D, LILLICRAP T, NOROUZI M, et al. Mastering atari with discrete world models[EB/OL]. 2020: arXiv: 2010.02193. https://arxiv.org/abs/2010.02193
[25] HAFNER D, PASUKONIS J, BA J, et al. Mastering diverse control tasks through world models[J]. Nature, 2025, 640(8059): 647. DOI: 10.1038/s41586-025-08744-2
[26] HANSEN N, WANG Xiaolong, SU Hao. Temporal difference learning for model predictive control[EB/OL]. 2022: arXiv: 2203.04955. https://arxiv.org/abs/2203.04955
[27] ASSRAN M, DUVAL Q, MISRA I, et al. Self-supervised learning from images with a joint-embedding predictive architecture[C]// 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver: IEEE, 2023: 15619. DOI: 10.1109/cvpr52729.2023.01499
[28] BARDES A, GARRIDO Q, PONCE J, et al. Revisiting feature prediction for learning visual representations from video[EB/OL]. 2024: arXiv: 2404.08471. https://arxiv.org/abs/2404.08471
[29] ZHOU Gaoyue, PAN Hengkai, LECUN Y, et al. DINO-WM: world models on pre-trained visual features enable zero-shot planning[EB/OL]. 2024: arXiv: 2411.04983. https://arxiv.org/abs/2411.04983
[30] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models[EB/OL]. 2020: arXiv: 2006.11239. https://arxiv.org/abs/2006.11239
[31] BROOKS T, PEEBLES B, HOLMES C, et al. Video generation models as world simulators[EB/OL]. OpenAI, 2024. https://openai.com/research/video-generation-models-as-world-simulators
[32] XIANG Jiannan, LIU Guangyi, GU Yi, et al. Pandora: towards general world model with natural language actions and video states[EB/OL]. 2024: arXiv: 2406.09455. https://arxiv.org/abs/2406.09455
[33] LI Xiaofan, ZHANG Yifu, YE Xiaoqing. DrivingDiffusion: layout-guided multi-view driving scenarios video generation with latent diffusion model[C]//Computer Vision-ECCV 2024. Cham: Springer Nature Switzerland, 2025: 469. DOI: 10.1007/978-3-031-73229-4_27
[34] ALONSO E, JELLEY A, MICHELI V, et al. Diffusion for world modeling: visual details matter in atari[EB/OL]. 2024: arXiv: 2405.12399. https://arxiv.org/abs/2405.12399
[35] ZHENG Yinan, LIANG Ruiming, ZHENG Kexin, et al. Diffusion-based planning for autonomous driving with flexible guidance[EB/OL]. 2025: arXiv: 2501.15564. https://arxiv.org/abs/2501.15564
[36] AWAIS M, NASEER M, KHAN S, et al. Foundation models defining a new era in vision: a survey and outlook[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, 47(4): 2245. DOI:10.1109/tpami.2024.3506283
[37] CAO Hanqun, TAN Cheng, GAO Zhangyang, et al. A survey on generative diffusion models[J]. IEEE Transactions on Knowledge and Data Engineering, 2024, 36(7): 2814. DOI: 10.1109/TKDE.2024.3361474
[38] 高超, 杨莹, 陈世超, 等. 多模态模型驱动的具身智能研究综述[J]. 智能感知工程, 2025, 2(2): 1
GAO Chao, YANG Ying, CHEN Shichao, et al. Survey on multimodal model-driven embodied intelligence research[J]. Intelligent Perception Engineering, 2025, 2(2): 1. DOI: 10.3969/j.issn.2097-4965.2025.02.001
[39] WANG Xiaofeng, ZHU Zheng, HUANG Guan, et al. DriveDreamer: towards real-world-drive world models for autonomous driving[C]// Computer Vision-ECCV 2024. Cham: Springer, 2025: 55. DOI: 10.1007/978-3-031-73195-2_4
[40] ZHAO Guosheng, WANG Xiaofeng, ZHU Zheng, et al. DriveDreamer-2: LLM-enhanced world models for diverse driving video generation[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2025, 39(10): 10412. DOI: 10.1609/aaai.v39i10.33130
[41] WANG Yuqi, HE Jiawei, FAN Lue, et al. Driving into the future: multiview visual forecasting and planning with world model for autonomous driving[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2024: 14749. DOI: 10.1109/CVPR52733.2024.01397
[42] GAO Shenyuan, YANG Jiazhi, CHEN Li, et al. Vista: a generalizable driving world model with high fidelity and versatile controllability[EB/OL]. 2024: arXiv: 2405.17398. https://arxiv.org/abs/2405.17398
[43] HASSAN M, STAPF S, RAHIMI A, et al. GEM: a generalizable ego-vision multimodal world model for fine-grained ego-motion, object dynamics, and scene composition control[C]//2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2025: 22404. DOI: 10.1109/cvpr52734.2025.02087
[44] ZHENG Wenzhao, CHEN Weiliang, HUANG Yuanhui, et al. Occworld: learning a 3D occupancy world model for autonomous driving[C]//Computer Vision-ECCV 2024. Cham: Springer, 2024: 55. DOI:10.1007/978-3-031-72624-8_4
[45] GU Songen, YIN Wei, JIN Bu, et al. DOME: taming diffusion

model into high-fidelity controllable occupancy world model[EB/OL]. 2024: arXiv: 2410. 10429. https://arxiv. org/abs/2410. 10429

[46] YANG Zetong, CHEN Li, SUN Yanan, et al. Visual point cloud forecasting enables scalable autonomous driving[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2024: 14673. DOI: 10. 1109/cvpr52733. 2024. 01390

[47] ZHOU Xin, LIANG Dingkang, TU Sifan, et al. HERMES: a unified self-driving world model for simultaneous 3D scene understanding and generation[EB/OL]. 2025: arXiv: 2501. 14729. https://arxiv. org/abs/2501. 14729

[48] LI Bohan, GUO Jiazhe, LIU Hongsi, et al. UniScene: unified occupancy-centric driving scene generation[C]//2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2025: 11971. DOI: 10. 1109/CVPR52734. 2025. 01118

[49] ZHANG Yumeng, GONG Shi, XIONG Kaixin, et al. Bevworld: A multimodal world model for autonomous driving via unified bev latent space[EB/OL]. 2024: arXiv: 2407. 05679. https://doi. org/10. 48550/arXiv. 2407. 05679

[50] HU Yihan, YANG Jiazhi, CHEN Li, et al. Planning-oriented autonomous driving[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver: IEEE, 2023: 17853. DOI: 10. 1109/CVPR52729. 2023. 01712

[51] YIN Hongbo, TIAN Daxin, ZHOU Jianshan. RCTracker: an efficient roadside cooperative tracker for real-world smart transportation[C]//Advances and Applications in SmartRail, Traffic, and Transportation Engineering. Singapore: Springer, 2025: 105. DOI: 10. 1007/978 - 981 - 96 - 7441 - 1_10

[52] JIANG Bo, CHEN Shaoyu, XU Qing, et al. VAD: vectorized scene representation for efficient autonomous driving[C]//2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris: IEEE, 2023: 8306. DOI: 10. 1109/ICCV51070. 2023. 00766

[53] ZHENG Wenzhao, SONG Ruiqi, GUO Xianda, et al. GenAD: generative end-to-end autonomous driving[C]//Computer Vision-ECCV 2024. Cham: Springer, 2025: 87. DOI: 10. 1007/978 - 3 - 031 - 73650 - 6_6

[54] LI Yingyan, WANG Yuqi, LIU Yang, et al. End-to-end driving with online trajectory evaluation via BEV world model[EB/OL]. 2025: arXiv: 2504. 01941. https://arxiv. org/abs/2504. 01941

[55] LI Yingyan, FAN Lue, HE Jiawei, et al. Enhancing end-to-end autonomous driving with latent world model[EB/OL]. 2024: arXiv: 2406. 08481. https://arxiv. org/abs/2406. 08481

[56] CHEN Yuntao, WANG Yuqi, ZHANG Zhaoxiang. DrivingGPT: unifying driving world modeling and planning with multi-modal autoregressive transformers[EB/OL]. 2024: arXiv: 2412. 18607. https://arxiv. org/abs/2412. 18607

[57] ZHANG Kaiwen, TANG Zhenyu, HU Xiaotao, et al. Epona: autoregressive diffusion world model for autonomous driving[EB/OL]. 2025: arXiv: 2506. 24113. https://arxiv. org/abs/2506. 24113

[58] DOSOVITSKIY A, ROS G, CODEVILLA F, et al. CARLA: an open urban driving simulator[EB/OL]. 2017: arXiv: 1711. 03938. https://doi. org/10. 48550/arXiv. 1711. 03938

[59] KERBL B, KOPANAS G, LEIMKÜHLER T, et al. 3D Gaussian splatting for real-time radiance field rendering[J]. ACM Transactions on Graphics, 2023, 42(4): 1. DOI: 10. 1145/3592433

[60] HESS G, LINDSTRÖM C, FATEMI M, et al. SplatAD: real-time lidar and camera rendering with 3d gaussian splatting for autonomous driving[C]//2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2025: 11982. DOI: 10. 1109/cvpr52734. 2025. 01119

[61] ZHOU Xiaoyu, LIN Zhiwei, SHAN Xiaojun, et al. Driving Gaussian: composite Gaussian splatting for surrounding dynamic autonomous driving scenes[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2024: 21634. DOI: 10. 1109/CVPR52733. 2024. 02044

[62] YANG Ze, CHEN Yun, WANG Jingkang, et al. UniSim: a neural closed-loop sensor simulator[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver: IEEE, 2023: 1389. DOI: 10. 1109/CVPR52729. 2023. 00140

[63] LJUNGBERGH W, TONDERSKI A, JOHNANDER J, et al. NeuroNCAP: photorealistic closed-loop safety testing for autonomous driving[C]//Computer Visio-nECCV 2024. Cham: Springer, 2025: 161. DOI: 10. 1007/978 - 3 - 031 - 73404 - 5_10

[64] ZHOU Hongyu, LIN Longzhong, WANG Jiabao, et al. HUGSIM: a real-time, photo-realistic and closed-loop simulator for autonomous driving[EB/OL]. 2024: arXiv: 2412. 01718. https://arxiv. org/abs/2412. 01718

[65] YANG Xuemeng, WEN Licheng, MA Yukai, et al. DriveArena: a closed-loop generative simulation platform for autonomous driving[EB/OL]. 2024: arXiv: 2408. 00415. https://arxiv. org/abs/2408. 00415

[66] ZHANG Zhejun, LINIGER A, DAI Dengxin, et al. TrafficBots: towards world models for autonomous driving simulation and motion prediction[C]//2023 IEEE International Conference on Robotics and Automation (ICRA). London: IEEE, 2023: 1522. DOI: 10. 1109/ICRA48891. 2023. 10161243

[67] VASUDEVAN A B, PERI N, SCHNEIDER J, et al. Planning with adaptive world models for autonomous driving[C]//2025 IEEE International Conference on Robotics and Automation (ICRA). Atlanta: IEEE, 2025: 14938. DOI: 10. 1109/icra55743. 2025. 11127860

[68] ZHOU Yunsong, YE Naisheng, LJUNGBERGH W, et al. Decoupled diffusion sparks adaptive scene generation[EB/OL]. 2025: arXiv: 2504. 10485. https://arxiv. org/abs/2504. 10485

[69] HU Shengchao, CHEN Li, WU Penghao, et al. St-p3: end-to-end vision-based autonomous driving via spatial-temporal feature learning[C]//Computer Vision-ECCV 2022. Cham: Springer, 2022: 533. DOI: 10. 1007/978 - 3 - 031 - 19839 - 7_31

[70] MIN Chen, ZHAO Dawei, XIAO Liang, et al. DriveWorld: 4D pre-trained scene understanding via world models for autonomous driving[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2024: 15522. DOI: 10. 1109/CVPR52733. 2024. 01470

[71] YANG Yu, MEI Jianbiao, MA Yukai, et al. Driving in the occupancy world: vision-centric 4D occupancy forecasting and planning via world models for autonomous driving[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2025, 39(9): 9327. DOI: 10. 1609/aaai. v39i9. 33010

[72] ZENG Shuang, CHANG Xinyuan, XIE Mengwei, et al. FutureSightDrive: thinking visually with spatio-temporal CoT for autonomous driving[EB/OL]. 2025: arXiv: 2505. 17685. https://arxiv. org/abs/2505. 17685

[73] ZHENG Wenzhao, CHEN Weiliang, HUANG Yuanhui, et al. OccWorld: learning a 3D occupancy world model for autonomous driving[C]//Computer Vision-ECCV 2024. Cham: Springer, 2025: 55. DOI: 10. 1007/978 - 3 - 031 - 72624 - 8_4

[74] YAN Ziyang, DONG Wenzhen, SHAO Yihua, et al. RenderWorld: World model with self-supervised 3D label[C]//2025 IEEE International Conference on Robotics and Automation (ICRA). Atlanta: IEEE, 2025: 6063. DOI: 10. 1109/ICRA55743. 2025. 11127609

[75] WEI Julong, YUAN Shanshuai, LI Pengfei, et al. OccLLaMA: an occupancy-language-action generative world model for autonomous driving[EB/OL]. 2024: arXiv: 2409. 03272. https://arxiv. org/abs/2409. 03272

(编辑 张 红)