Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2026, 15(5), 1188-1202
Published Online September 2026 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2026.155108

Hans汉斯

端到端自动驾驶规划方法研究综述：场景表示、规划生成与评估验证

禹　鑫

西华大学汽车与交通学院，四川 成都

收稿日期：2026年7月10日；录用日期：2026年9月2日；发布日期：2026年9月11日

摘　要

端到端自动驾驶规划通过减少人工规则和模块接口限制，使模型能够围绕最终驾驶任务进行联合优化。然而，场景表示、辅助任务和模型结构日益复杂，并不意味着其实际驾驶能力得到充分验证，场景信息能否有效支撑规划、不同技术机制如何生成驾驶行为，以及现有评估能否可靠反映模型能力，仍需系统分析。为此，本文围绕“场景表示－规划生成－评估验证”这一主线，对端到端自动驾驶规划研究进行综述。首先，界定端到端规划的任务边界与分类依据，归纳鸟瞰视角表示、占据表示、向量化场景表示以及语义增强与潜在世界状态表示，分析其从空间关系、三维占用、道路拓扑、高层语义和未来演化等方面对规划的支撑作用。其次，依据规划结果生成与改进的主导机制，将现有方法划分为模仿学习、多任务联合学习、生成模型和视觉－语言－动作模型四类，比较其规划依据、生成机制、输出形式及能力边界。随后，比较开环评估、闭环仿真评估和语言－动作一致性评估的评价对象、主要指标与适用场景，分析其在衡量离线规划质量、连续驾驶能力和语义－行为一致性方面的适用性与局限。最后，归纳现有研究的共性挑战，并展望面向闭环可信与可验证安全的发展方向。

关键词

端到端自动驾驶，场景表示，规划生成，视觉－语言－动作模型，闭环评估，安全验证

A Review of End-to-End Autonomous Driving Planning Methods: Scene Representation, Planning Generation, and Evaluation and Validation

Xin Yu

School of Automotive and Transportation Engineering, Xihua University, Chengdu Sichuan

文章引用: 禹鑫. 端到端自动驾驶规划方法研究综述: 场景表示、规划生成与评估验证[J]. 人工智能与机器人研究, 2026, 15(5): 1188-1202. DOI: 10.12677/airr.2026.155108

Received: July 10, 2026; accepted: September 2, 2026; published: September 11, 2026

Abstract

End-to-end autonomous driving planning reduces reliance on manually designed rules and module interfaces, enabling models to be jointly optimized toward the final driving task. However, increasingly complex scene representations, auxiliary tasks, and model architectures do not necessarily imply that actual driving capabilities have been adequately validated. Whether scene information can effectively support planning, how different technical mechanisms generate driving behaviors, and whether existing evaluation methods can reliably reflect model capability still require systematic investigation. To address these issues, this paper reviews end-to-end autonomous driving planning along the main line of "scene representation - planning generation - evaluation and validation". First, the task boundaries and classification criteria of end-to-end planning are defined. Major scene representations, including bird's-eye-view representation, occupancy representation, vectorized scene representation, semantic-enhanced representation, and latent world-state representation, are summarized, with emphasis on how they support planning through spatial relationships, three-dimensional occupancy, road topology, high-level semantics, and future evolution. Second, according to the dominant mechanisms used to generate and refine planning results, existing methods are categorized into four groups: imitation learning, multi-task joint learning, generative models, and vision-language-action models. Their planning basis, generation mechanisms, output forms, and capability boundaries are then compared. Subsequently, open-loop evaluation, closed-loop simulation evaluation, and language-action consistency evaluation are compared in terms of evaluation objects, major metrics, and applicable scenarios, and their applicability and limitations in measuring offline planning quality, continuous driving capability, and semantic-behavior consistency are analyzed. Finally, the common challenges faced by existing studies are summarized, and future research directions toward closed-loop trustworthiness and verifiable safety are discussed.

Keywords

End-to-End Autonomous Driving, Scene Representation, Planning Generation, Vision-Language-Action Model, Closed-Loop Evaluation, Safety Verification

1. 引言

自动驾驶规划需要依据车辆对周围环境的观测和行驶目标，持续生成安全、合规且可执行的驾驶行为。传统自动驾驶系统通常将环境感知、运动预测、决策规划和车辆控制划分为相对独立的模块，并通过预定义接口顺序连接，具有功能边界清晰、便于开发和故障追溯等优点。但人工接口可能造成规划相关信息损失，不同模块的优化目标也未必与最终驾驶任务一致，上游误差还可能在顺序处理过程中逐步传播，复杂道路语境和长尾场景也难以由有限规则充分覆盖。端到端方法通过减少人工接口限制，并围绕最终规划或控制目标进行联合优化，为提升自动驾驶系统的整体协同性提供了新的技术路径[1] [2]。

早期端到端规划主要依赖模仿学习，根据专家示范建立环境观测与驾驶行为之间的映射。条件模仿学习通过引入高层驾驶指令，缓解了相同场景观测对应多种驾驶行为的歧义[3]，但行为克隆仍受到数据分布偏置、因果混淆以及连续执行引起的状态分布偏移等问题影响[4]。随后，相关研究逐步引入结构化

场景表示以及感知、预测等辅助任务，使场景理解、未来预测和轨迹规划能够在统一框架中协同优化[5][6]；进一步又扩展到多模式轨迹分布、未来场景推演以及视觉－语言－动作联合建模，以增强对未来不确定性和复杂交通语义的表达。端到端规划的研究重点由直接行为拟合逐渐拓展到场景信息利用、未来演化建模和语义驱动的规划生成。

然而，模型结构和中间任务日益复杂，并不意味着实际驾驶能力已经得到充分验证。感知、预测或场景建模精度的提高，只能说明相应任务表现改善，不能证明这些信息真正参与了最终规划。已有研究表明，部分仅依赖自车状态的简单模型也可能获得具有竞争力的开环轨迹指标，视觉信息削弱也未必引起规划指标同步下降，说明现有离线评价可能难以准确反映外部场景信息对规划结果的实际作用[7]。同时，固定日志下的轨迹误差难以充分刻画连续执行中的状态偏移、误差累积、环境反馈和异常恢复。因此，端到端规划研究不仅需要关注模型采用了何种结构，还需要分析规划依据、生成机制与实际驾驶能力之间的关系。

为此，本文围绕“场景表示－规划生成－评估验证”三个相互关联的环节，对端到端自动驾驶规划研究进行系统梳理，重点分析不同场景信息如何提供规划依据、不同技术机制如何生成规划结果，以及开环、闭环与语言－动作一致性评估如何反映模型的实际驾驶能力。

2. 端到端自动驾驶规划的任务定义与研究框架

2.1. 任务定义与综述范围

端到端自动驾驶规划是指利用可学习模型，根据车辆当前及历史传感器观测、自车运动状态和导航信息，生成局部驾驶行为、未来轨迹、候选规划或车辆控制量，并围绕最终驾驶目标对相关计算过程进行联合优化。其输入通常包括多视角图像、激光雷达点云、自车状态和导航指令等信息，输出则可表现为路径点、连续轨迹、多模式候选轨迹或低层控制量。端到端并不意味着必须取消全部中间结构，模型既可以利用隐式特征直接生成规划结果，也可以通过 BEV、占据、向量化场景、未来状态预测或语言推理等显式信息辅助规划。该任务本质上是面向连续驾驶过程的时序决策，需要处理环境动态演化、未来状态不确定性以及自车与周围交通参与者之间的交互。

由于相同的传感器输入、场景表示或规划输出可能服务于不同技术机制，仅按照输入模态、网络结构、场景表示或输出形式分类，难以揭示各方法生成规划结果时的核心差异。因此，本文依据生成或改进驾驶行为、轨迹或控制量时所采用的主导技术机制，将现有方法概括为基于模仿学习、多任务联合学习、生成模型和视觉－语言－动作模型的四类方法，分别以专家行为学习、规划相关任务协同、未来轨迹与场景分布建模以及视觉－语言－动作联合推理为主要机制。上述类别并非严格互斥，对于同时采用多种技术的方法，本文根据其主要规划机制和核心创新确定归属。

本文聚焦由传感器观测生成局部驾驶行为、轨迹或控制量的端到端规划方法；对 BEV、占据、向量化、语义增强和潜在世界状态等研究，仅分析其对规划的信息支撑作用，不系统比较独立感知、建图或场景生成性能。本文不系统讨论传统全局路径规划、独立的低层控制方法，以及未直接服务于规划生成的单一感知方法。

2.2. 场景表示、规划生成与评估验证的关系

端到端自动驾驶规划虽可由传感器观测直接生成轨迹或控制量，但仍需提取和组织道路结构、交通参与者、空间占用、未来运动及潜在风险等任务相关信息。场景表示既可隐含于网络特征，也可显式表现为空间、结构化语义或潜在状态；规划生成在场景信息和导航目标基础上，通过专家行为学习、结构化约束、未来分布建模或语义推理，将其转化为局部驾驶行为、未来轨迹或控制量。场景表示的价值不

仅取决于独立精度，更取决于其能否保留并传递影响行为选择和规划安全的关键信息；规划结果接近专家轨迹只能说明模型具有一定的行为拟合能力，不能证明道路约束、交通交互和潜在风险已被充分利用。

评估验证用于判断场景表示和规划机制能否共同支撑可靠驾驶。开环评估比较模型输出与专家轨迹或控制量，适合衡量固定日志下的离线规划质量，但难以反映连续执行中的状态偏移和环境反馈；闭环评估通过持续执行模型输出，检验连续驾驶中的安全性、合规性和稳定性。对于视觉–语言–动作模型，还需评价视觉事实、语言推理与最终行为的一致性，以判断语言理解是否真正影响规划结果。

场景表示、规划生成与评估验证构成由信息组织、行为生成到能力检验的连续链路。评估中暴露的碰撞、越界、异常恢复失败或语言–动作不一致，可反向定位场景信息缺失、规划机制不足或评价标准缺陷，并指导数据补充、表示优化、模型改进和评估体系完善，三者关系如图 1 所示。

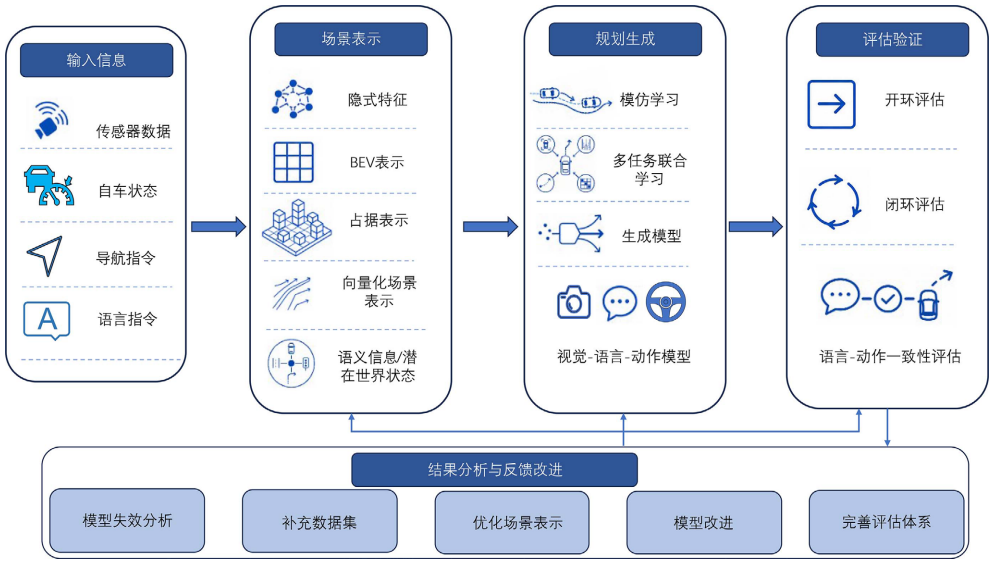

Figure 1. Overall framework for scene representation, plan generation, and evaluation and verification in end-to-end autonomous driving planning
图 1. 端到端自动驾驶规划中场景表示、规划生成与评估验证的总体框架

3. 面向端到端自动驾驶的场景表示方法

为增强规划对道路结构、空间占用、交通关系及未来演化的利用，现有研究逐步引入显式场景表示。本章从 BEV、占据、向量化以及语义增强与潜在世界状态四个方面分析其对规划的支撑作用。

3.1. BEV 表示

BEV 表示是端到端自动驾驶中较为基础的场景表示方式，其核心作用是将多视角图像或多传感器信息统一到与车辆运动平面一致的鸟瞰坐标系中。与透视图像相比，BEV 能够更直接地表达道路结构、交通参与者位置、自车周围空间关系及轨迹可能经过的区域，因而便于进行道路约束、碰撞判断和轨迹生成。BEVFormer 通过 BEV 查询、空间交叉注意力和时间自注意力聚合多摄像头及历史帧信息，为多视角场景构建统一的时空特征[8]。对于端到端规划而言，BEV 的价值不仅在于提升感知精度，更在于为道路、目标和自车运动关系提供统一的空间载体。

BEV 表示的可靠性主要取决于空间映射准确性和动态信息建模能力。纯视觉方法缺少直接深度信息，图像特征可能被投影到错误的鸟瞰位置，进而影响道路边界、目标位置和碰撞风险判断。相关研究通过

显式深度监督提高特征反投影的几何准确性[9]，并通过长期历史状态传播增强对目标速度、运动方向及遮挡状态的表达[10]。在多传感器系统中，BEV 还可作为相机和 LiDAR 等异构信息的共享空间，使图像语义与点云几何信息在统一坐标系中融合[11]。这些改进共同增强了 BEV 对规划所需空间结构和动态信息的表达能力。

在规划导向的端到端系统中，BEV 进一步由感知中间表示发展为连接感知、预测与规划的共享信息空间。UniAD 在统一 BEV 特征基础上组织跟踪、建图、运动预测、占据预测和规划任务，使前序任务产生的道路、目标及未来状态信息共同服务于自车轨迹生成[6]。ThinkTwice 则根据初步规划轨迹检索 BEV 关键区域特征，并预测动作条件下的未来 BEV 状态，以进一步修正轨迹和控制输出[12]。由此，BEV 不再只是场景重建结果，而是规划查询、未来预测和轨迹优化之间的信息交互基础。

3.2. 占据表示

占据表示以体素、三维网格或多平面特征描述空间位置的占用状态和语义属性，使模型不仅能够判断目标位于何处，还能够识别哪些空间被占用、哪些区域可通行以及规划轨迹可能面临的碰撞风险。TPVFormer [13]通过俯视、侧视和前视三个平面近似表达三维空间，OccFormer [14]则直接在三维体素空间中建模语义占据状态。占据表示可以视为 BEV 空间表达向三维空间安全约束的进一步扩展。

相较于基于三维检测框的目标表示，占据表示更适合描述非规则障碍物、未知目标和遮挡区域。车辆、行人等规则目标可以利用检测框近似表达，但施工设施、散落物、异形车辆及临时道路障碍通常难以用固定类别和规则边界完整描述。SurroundOcc [15]通过多摄像头图像预测周围三维空间的体素占用，以增强对任意形状目标和开放场景结构的表达；VoxFormer [16]则利用可见区域的稀疏体素查询补全不可见空间，为遮挡区域和潜在障碍物建模提供依据。这些方法使规划模型能够由“目标级避障”进一步扩展到“空间级风险判断”。

随着占据表示的发展，其表达内容也由几何占用逐渐扩展到语义和实例信息。单纯几何占据只能判断某一位置是否存在障碍，而语义占据能够区分车辆、行人、道路、建筑物和植被等不同类别，使规划模型根据对象属性采用差异化约束。PanoOcc [17]在统一三维占据空间中联合建模语义分割和目标实例信息，使占据表示同时包含空间状态、目标类别及实例结构。对于端到端规划而言，这类表示可以进一步转化为可通行区域、碰撞风险图或轨迹约束，为候选轨迹生成、筛选和安全校验提供依据。

3.3. 向量化场景表示

向量化场景表示将车道分隔线、道路边界、人行横道和中心线等道路元素编码为折线、点集或图结构，以较紧凑的形式保留实例语义、几何形态和连接关系。与密集网格或体素表示相比，向量化表示更直接地表达道路元素的行驶方向和拓扑关系，可为车道边界约束、路口连通性分析及交通规则关联提供结构化依据。

在线向量地图是该类表示的典型形式。VectorMapNet 直接从车载传感器观测中预测 BEV 空间内的稀疏折线集合，避免由栅格分割结果经启发式后处理生成地图元素，并使输出形式更接近运动预测和规划所使用的实例级地图[18]。针对开放曲线、闭合多边形等地图元素存在点序不唯一的问题，MapTR 将地图元素建模为点集及其等价排列，并通过层级查询同时编码实例级和点级信息，以并行方式预测地图元素[19]。由此，向量地图建模由“从栅格中提取曲线”进一步发展为对元素实例、几何形态和点集组织方式的统一学习。

面向连续驾驶，向量化表示还需兼顾时间一致性和关系完整性。StreamMapNet 通过传播历史 BEV 特征和地图查询，在连续帧中更新局部向量地图，以缓解遮挡、远距离感知不足和单帧预测波动带来的结

构缺失[20]。在此基础上，拓扑化表示进一步显式建模道路元素之间的关系。TopoMLP 联合完成三维车道中心线检测、二维交通元素检测以及车道–车道、车道–交通元素拓扑预测，使表示内容由单个元素的几何形态扩展到车道连通性和交通控制关系[21]。这些关系可为路口通行方向选择和交通规则约束提供结构依据。

3.4. 语义增强场景表示与潜在世界状态表示

语义增强场景表示与潜在世界状态分别面向高层交通语义和未来环境演化建模。前者利用语言和图结构组织交通规则、对象属性、行为意图与交互关系，后者通过世界模型学习场景随时间及自车动作变化的潜在状态，为风险分析、行为选择和候选轨迹评价提供依据。

在当前场景的语义增强方面，Talk2BEV 在 BEV 地图中为交通对象附加视觉–语言特征、属性和空间关系，使模型能够围绕对象语义、可供性及自车通行条件进行查询和推理[22]。VLP 则将预训练语言知识用于增强 BEV 中的局部目标特征和规划查询，使场景内容与自车状态及驾驶目标建立语义联系[23]。GraphPilot 进一步将道路结构、交通参与者和控制设施组织为场景图，并将其作为语言驾驶模型的结构化上下文[24]。这些方法使场景表示由对象位置和类别扩展到对象属性、空间关系、道路拓扑及规则约束。

在未来状态建模方面，OccWorld 将三维语义占据压缩为离散场景标记，并联合预测未来占据状态和自车运动[25]；DriveWorld 通过动态记忆与静态场景传播学习统一的四维时空表示，为感知、预测和规划任务提供预训练特征[26]。Vista 则从真实驾驶视频出发，引入道路结构、文本提示及不同粒度的驾驶动作，生成动作条件下的未来视觉场景[27]。世界状态表示正由当前环境编码扩展到未来场景演化和行为后果建模，可为候选行为分析、轨迹评价和风险预测提供依据。不过，语义增强表示仍可能受到视觉事实错误、关系构建偏差和语言幻觉影响，潜在世界状态也面临长期预测漂移、物理一致性不足和计算开销较高等问题，其闭环规划收益仍需进一步验证。

综上，不同场景表示在信息粒度和规划作用上各有侧重，实际系统可根据任务需求单独或组合使用。表 1 从典型表示形式、主要表达信息、规划支撑作用及特点与局限等方面进行比较。

Table 1. Comparison of main scenario representation methods for end-to-end autonomous driving planning
表 1. 面向端到端自动驾驶规划的主要场景表示方法对比

场景表示类型	典型表示形式	主要表达信息	对规划的支撑作用	主要特点与局限
BEV 表示	二维鸟瞰网格或 BEV 特征	道路结构、目标位置、空间关系及历史时序信息	统一多视角、时序及多传感器信息，支持轨迹生成、道路约束和碰撞判断	空间关系直观、便于融合；对深度估计和标定敏感，高度表达有限，计算开销较大
占据表示	三维体素、占据网格或三平面特征	空间占用、语义类别、物体形状、遮挡区域和可通行空间	提供三维占用、可通行区域和碰撞约束	适于非规则目标和遮挡建模；计算与存储开销较高
向量化场景表示	折线、点集、实例向量及拓扑图	车道线、道路边界、中心线、交通元素、动态目标及其连接关系	提供明确的道路方向、边界、拓扑和目标交互约束	表示紧凑、实例语义清晰；依赖道路元素和拓扑预测精度，对非规则障碍物和空间占用风险表达不足

续表

语义推理	语言描述、场景图和知识图谱	目标属性、空间关系、交通规则、行为意图和风险语义	支持复杂交通语境理解、规则推理、风险分析和规划依据解释	高层语义丰富；但可能受到视觉－语言对齐误差、场景事实构建误差和模型幻觉影响
潜在世界状态表示	连续或离散潜变量、场景标记及未来潜在特征	场景时序演化、未来占用、目标运动和动作条件下的状态变化	通过未来场景推演辅助候选轨迹生成、评估和行为后果预测	能建模未来演化和动作影响；但潜在状态可解释性较弱，物理一致性难以保证，训练和推理开销较高

4. 端到端自动驾驶的规划方法

本章按照生成规划结果时采用的主导技术机制，将现有方法划分为基于模仿学习、多任务联合学习、生成模型和视觉－语言－动作模型四类，重点分析各类方法的规划依据、核心生成机制、输出形式及能力边界。

4.1. 基于模仿学习的规划方法

基于模仿学习的规划方法利用专家驾驶数据学习从环境观测到未来轨迹、路径点或控制量的映射关系，是端到端自动驾驶由规则驱动转向数据驱动的重要技术路线。该类方法减少了对人工规则、代价函数和模块接口的依赖，使传感器输入与驾驶输出能够围绕专家行为进行统一优化。

早期方法主要采用行为克隆，通过监督学习使模型输出接近专家驾驶行为。然而，自动驾驶场景中的环境观测与驾驶动作并不总是简单的一一对应关系。在交叉口、分岔路口等场景中，相同的观测可能对应左转、右转或直行等不同意图，若仅依赖当前传感器输入，模型难以确定唯一合理的控制行为。条件模仿学习由此引入高层导航指令，使模型能够根据不同驾驶目标生成相应动作[3]。这表明，模仿学习式规划不仅需要拟合观测与动作之间的关系，还需结合导航意图和车辆状态消除行为歧义。

围绕行为克隆在复杂驾驶场景中的不足，后续研究主要从输入表征和输出形式等方面增强模仿学习式规划。在输入表征方面，单一视觉输入难以同时提供丰富语义信息和准确几何结构，多模态传感器融合因此成为提升端到端规划能力的重要方向。TransFuser 利用图像与 LiDAR 信息的互补性增强场景理解能力，从而提高模型在复杂交通场景中的闭环驾驶表现[28]。在输出形式方面，直接控制输出具有执行链路短的优势，但对未来运动趋势的表达能力有限；轨迹或路径点输出具有更好的前瞻性，却需要依赖底层控制器进一步跟踪。TCP 将轨迹预测与控制预测结合起来，在轨迹约束与控制执行之间建立联系[29]。

模仿学习在专家数据覆盖充分且训练与部署分布接近时，能够有效学习常规驾驶行为，但其能力边界受示范覆盖和因果混淆制约。Codevilla 等发现，行为克隆方法在密集动态交通中可靠性明显下降，并可能因训练数据中的“低速－不加速”伪相关出现过度停车和停车后难以重新起步的惯性问题[4]。这表明，该类方法在长尾交互和闭环状态分布偏移下仍缺乏稳定性。

4.2. 基于多任务联合学习的规划方法

多任务联合学习通过同时优化感知、建图、运动预测、占据预测和规划等任务，使道路边界、目标状态、未来运动和空间占用等中间信息共同约束轨迹生成[30]。与仅共享特征骨干的普通多任务学习不同，规划导向的多任务方法更强调中间结果通过特征传递、损失约束或安全控制实际参与规划过程，使端到

端规划由单纯拟合专家行为转向学习与驾驶任务相关的场景信息。

InterFuser 通过可解释中间任务增强规划与控制过程的安全约束。该方法融合多视角图像和激光雷达信息，在预测自车路径点的同时，生成目标密度图和交通规则状态[31]。前者用于表达周围交通参与者及其运动信息，后者用于描述交通灯、停止标志和路口等驾驶条件，二者进一步输入安全控制器以调整车辆速度和停车行为。由此，中间任务由辅助特征学习转化为可检查的形式直接参与最终驾驶行为生成。

多任务联合学习的关键不在于任务数量，而在于不同任务能否围绕最终规划目标形成有效协同。ST-P3 在统一时空特征中联合建模视觉感知、未来语义预测和轨迹规划，并利用未来占据和代价信息筛选候选轨迹[5]；UniAD 进一步以规划为最终目标，通过任务查询连接检测、跟踪、建图、运动预测、占据预测和规划，使前序任务生成的信息共同服务于自车轨迹优化[6]。这些方法表明，多任务端到端规划正在由多个任务的简单并列，转向规划目标驱动的信息组织和任务协同。

随着任务数量和中间表示不断增加，多任务系统也面临表示冗余、计算开销和顺序误差传递等问题。VAD 利用向量化场景表示组织道路结构和交通参与者运动，使规划能够直接使用实例级地图和目标关系[32]；SparseDrive 进一步以稀疏实例统一检测、跟踪、建图、预测和规划，减少密集鸟瞰网格中的冗余计算[33]。现有研究还从任务组织和信息流角度改进端到端架构，通过任务并行化和统一标记空间中的任务协同，缓解严格顺序结构带来的误差传播和效率瓶颈[34] [35]。

UniAD 的消融实验表明，简单并列多个任务的朴素多任务训练弱于规划导向的任务组织，且运动预测与占据预测同时参与时，规划误差和碰撞率才取得较优结果[6]。这说明辅助任务只有在信息能够有效传递至规划模块时才会产生收益，否则可能增加计算负担，并使感知或预测误差继续传播至轨迹。

4.3. 基于生成模型的规划方法

在交互场景中，相同观测可能对应让行、等待或变道等多种安全行为，周围交通参与者也可能因自车动作产生不同响应。基于生成模型的规划方法由此将未来轨迹、候选行为或潜在世界状态建模为可生成分布，以描述不同未来模式及其交互差异。

在轨迹生成方面，GenAD 通过实例中心场景表示建模地图、自车与交通参与者之间的关系，并在潜在轨迹空间中联合生成周围目标运动和自车规划结果[36]。DiffusionDrive 则利用多模式轨迹锚点构造初始分布，通过截断扩散和级联解码过程生成、细化并评价多条候选轨迹，在减少去噪步骤的同时兼顾轨迹质量、模式多样性和推理效率[37]。这类方法使端到端规划由单一专家轨迹回归，进一步转向对多种合理驾驶行为及其置信度的联合建模。

生成式规划还由轨迹分布建模扩展到未来世界状态推演。LAW 预测未来潜在场景特征，并利用时序连续性增强场景表示和轨迹预测[38]；World4Drive 进一步结合驾驶意图生成未来潜在状态，并利用世界模型评价和选择多模式候选轨迹[39]；SeerDrive 则将未来鸟瞰特征预测与轨迹规划进行迭代交互，使未来场景建模和自车规划能够相互修正[40]。三者体现了世界模型在规划中的作用由提供未来特征，发展到参与候选轨迹评价，并进一步扩展到场景演化与规划结果的双向优化。

生成模型能够表达多种未来状态和候选轨迹，但其能力边界不仅受模式覆盖和候选质量影响，还取决于评价与选择机制是否可靠。World4Drive 的消融实验表明，仅引入多模式驾驶意图而缺少世界模型评价时，规划误差和碰撞率反而上升；利用未来潜在状态对候选轨迹进行评价和排序后，规划性能才得到改善[39]。这说明生成多样性不能自动转化为规划安全，合理的场景演化建模与候选选择机制同样关键。

4.4. 基于视觉－语言－动作模型的规划方法

随着视觉－语言模型和多模态大模型的发展，相关研究由驾驶场景问答和行为解释[41] [42]逐步扩展

到规划生成。DriveVLM 通过场景描述、关键目标分析和层级规划生成自车轨迹，并结合三维感知与传统规划模块细化输出[43]；EMMA 统一编码环视图像、导航指令和自车历史状态，生成文本化未来轨迹[44]；Think-Driver 则通过场景理解与风险推理选择离散元动作，再由外部模块转换为可执行轨迹[45]。这些方法推动视觉–语言模型由场景理解走向规划输出，但部分结果仍依赖外部规划或执行模块。

VLM 与 VLA 的主要区别在于动作是否被纳入模型的原生输出空间。VLM 主要在语言空间中组织场景理解、文本化轨迹或高层行为，其中部分结果还需要执行模块进一步转换；VLA 则将轨迹标记或物理动作标记与视觉、语言标记共同建模，使模型能够在统一框架中直接学习由场景语义到规划动作的映射。

在原生动作生成方面，OpenDriveVLA 从多视角图像中构建全局场景、动态目标和地图等结构化视觉标记，通过层级视觉–语言对齐及目标–环境–自车交互建模，自回归生成轨迹标记[46]。AutoVLA 则将连续轨迹编码为包含短时位置与航向变化的物理动作标记并通过监督与强化微调实现直接动作生成和自适应推理，以兼顾复杂场景推理与规划效率[47]。二者表明，VLA 正在由文本化规划逐步转向空间语义、语言推理和物理动作的原生联合建模。

近期研究还关注语言条件能否真正作用于驾驶行为，以及模型如何从闭环失败中改进。SimLingo 通过动作想象(Action Dreaming)任务构造不同的指令–动作关系，使语言条件影响几何路径与速度规划[48]；CoReVLA 则利用闭环安全驾驶员接管数据和偏好优化，改进长尾安全关键场景中的决策[49]。这表明，VLA 研究已由动作表示设计进一步扩展到语言–动作对齐、闭环反馈和人类偏好学习。

视觉–语言–动作模型的能力边界主要体现在视觉事实理解、空间关系推理和语言–动作对齐。DriveBench 发现，部分视觉语言模型在视觉信息损坏或缺失时仍会生成表面合理的回答，并在遮挡、目标重叠或转向信号不明显时出现方向误判[50]；当此类判断作为 VLA 的动作生成条件时，错误可能沿“视觉事实–语言推理–规划动作”链路传播。因此，该类方法除轨迹误差和问答准确率外，还需结合语言–动作一致性与闭环驾驶结果验证其可靠性。

4.5. 混合规划机制的融合趋势

前述四类规划方法依据生成或改进规划结果时采用的主导机制进行划分，实际模型通常融合多种机制。模仿学习与多任务联合学习构成较为成熟的融合模式。ST-P3、UniAD、VAD 和 InterFuser 等方法在联合感知、预测、地图构建、占据预测和运动建模等任务的同时，仍采用专家轨迹监督最终规划结果[5] [6] [31] [32]。其中，模仿学习提供驾驶行为监督，多任务联合学习则将道路结构、动态目标状态和未来场景信息转化为规划约束。

生成式规划开始在模仿监督与多任务场景建模的基础上，引入对未来状态分布和多模式轨迹的显式建模。World4Drive 结合未来世界状态预测、候选轨迹生成和轨迹评分完成规划[39]；SeerDrive 联合建模未来 BEV 演化与自车轨迹，使场景预测和规划结果相互约束[40]。此类方法以专家轨迹监督行为合理性，以多任务场景建模提供结构化约束，并通过生成模型表达未来交通演化和驾驶行为的多种可能性，形成模仿监督、场景约束与多模式生成相结合的规划机制。

多任务联合学习也开始与视觉–语言–动作模型融合。OpenDriveVLA 利用结构化场景表示组织目标、地图和交通状态信息，并通过层级视觉–语言对齐与交互预测任务增强场景理解[46]。结构化场景表示能够补充视觉语言模型在空间定位和场景关系表达方面的不足，语言模型则为规划引入高层交通语义和行为推理能力。ORION 进一步以多任务场景信息支撑语言推理，并通过规划标记条件化生成式规划器，同时利用专家轨迹和安全损失约束最终输出，体现了模仿学习、多任务联合学习、生成模型和视觉–语言–动作模型在统一规划链路中的协同[51]。

总体上，混合规划正由多种机制的并列使用转向结构化场景、未来生成与语言推理的深层耦合，但

其收益取决于各类信息能否有效传递并共同约束最终规划，而非机制数量的简单叠加。

5. 端到端自动驾驶规划的评估方法

5.1. 开环评估方法

开环评估在固定驾驶日志或离线测试集上，根据传感器观测、自车状态和导航信息预测未来轨迹、路径点或控制量，并与数据中记录的专家行为进行比较。由于模型输出不改变车辆状态和后续观测，无需运行完整的车辆动力学与交通交互仿真，因而具有成本较低、复现性较强和便于统一比较等特点，主要用于训练迭代和模型初步筛选。

开环评估的对象随规划输出形式而变化。直接控制类方法通常评价转向、速度、加速度或制动等控制量；轨迹规划类方法主要评价未来路径点或自车轨迹；生成多条候选轨迹的方法还需考察候选集合对真实行为的覆盖程度及候选评分结果。对于联合感知、预测、建图和规划的多任务方法，还可同时报告各中间任务性能，TAD-E2E 数据集已在统一时空条件下提供相应模块真值和自车规划轨迹真值[52]。但中间任务指标主要反映相应任务的准确性，最终规划结果仍需通过动作或轨迹指标评价。

常用开环指标主要包括动作误差、轨迹误差和离线安全指标。动作误差通常采用平均绝对误差、均方误差或均方根误差，衡量模型输出控制量与专家动作之间的偏差；轨迹误差常采用 L2 误差、平均位移误差和终点位移误差，评价预测轨迹与专家轨迹之间的空间差异。对于多候选规划，可采用 minADE 和 minFDE 等指标衡量候选集合对记录行为的覆盖程度；部分研究还结合离线碰撞率、未命中率和可行驶区域约束，对预测轨迹的基本安全性与可行性进行检查。但单一专家轨迹并非复杂场景中的唯一合理解，接近专家轨迹也不能证明模型充分利用了道路结构和交通交互信息[7]；固定未来状态下的碰撞检查同样无法反映自车动作引起的交互反馈。因此，开环评估适合衡量离线拟合能力和进行相对比较，不能单独作为闭环安全能力的证据。

5.2. 闭环仿真评估方法

闭环仿真评估通过连续执行模型生成的轨迹或控制量，更新自车状态和后续环境观测，形成“感知–规划–控制–反馈”的循环过程。与固定日志下的一次性预测不同，该方法能够直接评价模型连续运行后的驾驶表现。其评价内容通常包括任务完成、安全性、交通规则遵守、驾驶效率和舒适性，常用指标包括路线完成率、碰撞率、违规次数、驾驶得分以及基于加速度、加加速度(jerk)和横摆角速度等变量计算的平稳性指标。

不同仿真平台和评价基准侧重的能力有所不同。CARLA 提供可配置的城市道路、车辆、行人、天气和传感器环境，并能够记录碰撞、驶入对向车道和侵入人行道等事件，适合开展可控条件下的交互式闭环测试[53]。nuPlan 结合真实驾驶日志、地图信息和仿真工具，从交通规则遵守、行驶进展、车辆动力学和舒适性等维度评价规划器[54]。Bench2Drive 则基于 CARLA 构建短路线和多类交互场景，分别测试并道、超车、紧急制动、让行和交通标志响应等驾驶能力，并通过成功率、驾驶得分、效率和舒适性评价端到端模型[55]。这些平台和基准使闭环评估由单一的路线完成评价逐渐扩展到对安全、合规、效率和复杂交互能力的综合检验。

闭环仿真能够揭示连续执行中的误差累积和环境反馈，但其结果受观测、执行和交互模型近似的限制。仿真–现实差距主要表现为传感器与环境建模偏差、车辆动力学与执行器响应近似；同时，仿真交通参与者的反应性和行为多样性受代理模型设定限制。García Daza 等发现，简化车辆模型在大曲率和高动态场景中更难准确复现真实运动[56]。因此，闭环仿真只能反映特定模型假设下的驾驶表现，仍需结合硬件在环和受控实车验证。

5.3. 语言理解与动作一致性评估方法

随着视觉语言模型和视觉–语言–动作模型被引入端到端自动驾驶，模型评估不再局限于轨迹误差、碰撞率和路线完成率等结果指标，还需要检验语言理解是否具有可靠的场景依据，并真正作用于驾驶行为。语言理解与动作一致性评估因此关注“视觉事实–语言推理–规划动作”之间的联系。

从评估内容看，可将其概括为三个层次。第一，视觉事实准确性，主要评价模型对交通对象、运动状态和空间关系的识别是否正确。NuScenes-QA 基于自动驾驶场景构建目标属性、状态和关系问答，并采用准确率评价模型的场景理解能力[57]。第二，推理与解释合理性，主要考察场景描述、风险判断和行为解释是否与视觉事实相符，推理过程是否连贯，以及解释是否与最终结论保持一致。DriveBench 通过正常、视觉损坏和纯文本输入对比感知、预测、规划和行为任务表现，检验模型是否真正利用视觉信息[50]。第三，语言–动作一致性，主要检验场景事实或语言指令发生变化时，模型的行为选择、规划轨迹或控制输出能否作出合理响应。SimLingo 通过动作想象任务构造不同指令–动作关系，以动作预测结果检验语言条件是否真实影响规划[48]。该方法表明，仅获得正确的问答结果不足以证明模型实现了语言与驾驶行为的有效对齐。

5.4. 评估体系的发展方向

现有开环、闭环和语言–动作一致性评估能够分别衡量离线规划质量、连续驾驶结果和语义–行为关系，但仍存在因果辨别能力不足、长尾风险覆盖有限以及规划依据和交互过程难以度量等问题。为弥补传统指标的不足，端到端规划评估正由平均性能比较转向对关键场景依赖、风险边界和规划过程可信度的进一步检验。

轨迹误差、碰撞率和驾驶得分主要评价模型产生的最终结果，难以判断规划行为真正依赖的关键场景因素。模型可能借助自车状态、历史轨迹或数据偏置获得较好的评价结果，却无法在关键条件变化时作出合理响应。因果干预与反事实评估可通过改变交通信号状态、目标状态、道路关系等条件，检验规划是否产生符合交通逻辑的响应，并通过无关背景扰动下的稳定性识别模型对数据偏置和虚假相关的依赖。但该类方法目前在干预变量选择、反事实场景真实性和评价标准方面仍缺少统一规范。

针对固定测试集难以暴露长尾失效的问题，对抗性场景生成通过调整交通参与者轨迹、速度和交互关系，主动搜索容易导致碰撞、异常制动或交互失败的测试条件。AdvSim 对交通参与者轨迹施加物理可行的扰动，并同步更新 LiDAR 观测和遮挡关系，从而生成面向感知、预测和规划完整链路的安全关键场景[58]。该类方法使评估由被动统计已有失效转向主动探测模型的能力边界，但还需同时保证生成场景的物理可行性、语义合理性、多样性和可复现性。

传统结果指标还难以揭示模型的规划依据，也不能充分衡量连续规划对交通交互的影响。可解释性评价通过检验场景证据、语言解释与规划动作之间的对应关系，判断模型是否依据与驾驶任务相关的信息生成规划结果。可预见性评价则关注连续规划是否稳定，以及自车行为能否被其他交通参与者合理预期。Chiva Gil 等利用规划分布与预测模型所估计行为分布之间的差异刻画可预见性，并以相邻规划周期的轨迹变化定义规划努力度[59]；其实验表明，提高可预见性有助于减少反复调整、交互犹豫和死锁。

6. 结论与展望

本文围绕“场景表示–规划生成–评估验证”三个相互关联的环节，对端到端自动驾驶规划研究进行了系统梳理。研究表明，BEV、占据、向量化、语义增强和潜在世界状态分别从空间关系、三维占用、道路拓扑、高层语义和未来演化等方面提供互补信息，其价值不只取决于独立表示精度，更取决于能否被规划机制有效利用。模仿学习、多任务联合学习、生成模型和视觉–语言–动作模型并非严格替代关

系，而是可交叉组合的主导机制，分别侧重专家行为学习、结构化任务协同、未来分布建模和语义–动作联合推理。开环、闭环及语言–动作一致性评估则分别检验离线拟合、连续驾驶和语义–行为关系，但单一评估方式难以完整反映规划可靠性。

针对规划依据难以验证、未来不确定性沿规划链路传播、闭环评价覆盖不足以及语言–动作错配和车端资源受限等问题，未来研究应围绕任务相关表示、可控规划生成和分层评估验证建立更加完整的技术体系。在场景表示与规划生成方面，应强化关键交通参与者、通行关系、潜在冲突和未来交互等规划相关信息的建模，联合估计场景状态、未来演化和候选轨迹的不确定性，并将车辆动力学、道路边界、碰撞风险和交通规则融入轨迹生成与独立安全校验。在评估验证方面，应形成开环快速筛选、风险导向闭环测试和受控实车验证相结合的分层体系，同时加强视觉事实、语言推理与规划动作的一致性检验。此外，还应利用事故、人工接管和闭环失效案例持续更新模型，并通过跨域适应、高效模型设计和软硬件协同优化降低部署开销，推动端到端规划走向闭环可信与可验证安全。

参考文献

[1] Chen, L., Wu, P., Chitta, K., Jaeger, B., Geiger, A. and Li, H. (2024) End-to-End Autonomous Driving: Challenges and Frontiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **46**, 10164-10183. https://doi.org/10.1109/tpami.2024.3435937

[2] Tampuu, A., Matiisen, T., Semikin, M., Fishman, D. and Muhammad, N. (2022) A Survey of End-to-End Driving: Architectures and Training Methods. *IEEE Transactions on Neural Networks and Learning Systems*, **33**, 1364-1384. https://doi.org/10.1109/tnnls.2020.3043505

[3] Codevilla, F., Muller, M., López, A., Koltun, V. and Dosovitskiy, A. (2018) End-to-End Driving via Conditional Imitation Learning. 2018 *IEEE International Conference on Robotics and Automation* (*ICRA*), Brisbane, 21-25 May 2018, 4693-4700. https://doi.org/10.1109/icra.2018.8460487

[4] Codevilla, F., Santana, E., Lopez, A. and Gaidon, A. (2019) Exploring the Limitations of Behavior Cloning for Autonomous Driving. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27 October-2 November 2019, 9329-9338. https://doi.org/10.1109/iccv.2019.00942

[5] Hu, S., Chen, L., Wu, P., Li, H., Yan, J. and Tao, D. (2022) ST-P3: End-to-End Vision-Based Autonomous Driving via Spatial-Temporal Feature Learning. In: *Computer Vision-ECCV* 2022, Springer Nature Switzerland, 533-549. https://doi.org/10.1007/978-3-031-19839-7_31

[6] Hu, Y., Yang, J., Chen, L., Li, K., Sima, C., Zhu, X., *et al*. (2023) Planning-Oriented Autonomous Driving. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 18-22 June 2023, 17853-17862. https://doi.org/10.1109/cvpr52729.2023.01712

[7] Zhai, J.T., Feng, Z., Du, J., *et al*. (2023) Rethinking the Open-Loop Evaluation of End-to-End Autonomous Driving in nuScenes. arXiv: 2305.10430. https://arxiv.org/abs/2305.10430

[8] Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., *et al*. (2022) BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers. In: *Computer Vision-ECCV* 2022, Springer Nature Switzerland, 1-18. https://doi.org/10.1007/978-3-031-20077-9_1

[9] Li, Y., Ge, Z., Yu, G., Yang, J., Wang, Z., Shi, Y., *et al*. (2023) BEVDepth: Acquisition of Reliable Depth for Multi-View 3D Object Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, **37**, 1477-1485. https://doi.org/10.1609/aaai.v37i2.25233

[10] Chang, M., Zhang, X., Zhang, R., Zhao, Z., He, G. and Liu, S. (2024) RecurrentBEV: A Long-Term Temporal Fusion Framework for Multi-View 3D Detection. In: *Computer Vision-ECCV* 2024, Springer Nature Switzerland, 131-147. https://doi.org/10.1007/978-3-031-73220-1_8

[11] Liu, Z., Tang, H., Amini, A., Yang, X., Mao, H., Rus, D.L., *et al*. (2023) BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation. 2023 *IEEE International Conference on Robotics and Automation* (*ICRA*), London, 29 May-2 June 2023, 2774-2781. https://doi.org/10.1109/icra48891.2023.10160968

[12] Jia, X., Wu, P., Chen, L., Xie, J., He, C., Yan, J., *et al*. (2023) Think Twice before Driving: Towards Scalable Decoders for End-to-End Autonomous Driving. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 18-22 June 2023, 21983-21994. https://doi.org/10.1109/cvpr52729.2023.02105

[13] Huang, Y., Zheng, W., Zhang, Y., Zhou, J. and Lu, J. (2023) Tri-Perspective View for Vision-Based 3D Semantic Occupancy Prediction. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 18-22 June 2023,

9223-9232. https://doi.org/10.1109/cvpr52729.2023.00890

[14] Zhang, Y., Zhu, Z. and Du, D. (2023) Occformer: Dual-Path Transformer for Vision-Based 3D Semantic Occupancy Prediction. 2023 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Paris, 2-3 October 2023, 9433-9443. https://doi.org/10.1109/iccv51070.2023.00865

[15] Wei, Y., Zhao, L., Zheng, W., Zhu, Z., Zhou, J. and Lu, J. (2023) SurroundOcc: Multi-Camera 3D Occupancy Prediction for Autonomous Driving. 2023 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Paris, 2-3 October 2023, 21729-21740. https://doi.org/10.1109/iccv51070.2023.01986

[16] Li, Y., Yu, Z., Choy, C., Xiao, C., Alvarez, J.M., Fidler, S., *et al*. (2023) VoxFormer: Sparse Voxel Transformer for Camera-Based 3D Semantic Scene Completion. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 18-22 June 2023, 9087-9098. https://doi.org/10.1109/cvpr52729.2023.00877

[17] Wang, Y., Chen, Y., Liao, X., Fan, L. and Zhang, Z. (2024) PanoOcc: Unified Occupancy Representation for Camera-Based 3D Panoptic Segmentation. 2024 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 16-22 June 2024, 17158-17168. https://doi.org/10.1109/cvpr52733.2024.01624

[18] Liu, Y., Yuan, T., Wang, Y., *et al*. (2023) VectorMapNet: End-to-End Vectorized HD Map Learning. *Proceedings of the* 40*th International Conference on Machine Learning*. *PMLR*, **202**, 22352-22369.

[19] Liao, B., Chen, S., Wang, X., *et al*. (2023) MapTR: Structured Modeling and Learning for Online Vectorized HD Map Construction. *International Conference on Learning Representations*, Kigali, 1-5 May 2023, 710-727.

[20] Yuan, T., Liu, Y., Wang, Y., Wang, Y. and Zhao, H. (2024) StreamMapNet: Streaming Mapping Network for Vectorized Online HD Map Construction. 2024 *IEEE/CVF Winter Conference on Applications of Computer Vision* (*WACV*), Waikoloa, 3-8 January 2024, 7356-7365. https://doi.org/10.1109/wacv57701.2024.00719

[21] Wu, D., Chang, J., Jia, F., *et al*. (2024) TopoMLP: A Simple Yet Strong Pipeline for Driving Topology Reasoning. *International Conference on Learning Representations*, Vienna, 7-11 May 2024, 13809-13820.

[22] Choudhary, T., Dewangan, V., Chandhok, S., Priyadarshan, S., Jain, A., Singh, A.K., *et al*. (2024) Talk2BEV: Language-Enhanced Bird's-Eye View Maps for Autonomous Driving. 2024 *IEEE International Conference on Robotics and Automation* (*ICRA*), Yokohama, 13-17 May 2024, 16345-16352. https://doi.org/10.1109/icra57147.2024.10611485

[23] Pan, C., Yaman, B., Nesti, T., Mallik, A., Allievi, A.G., Velipasalar, S., *et al*. (2024) VLP: Vision Language Planning for Autonomous Driving. 2024 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 16-22 June 2024, 14760-14769. https://doi.org/10.1109/cvpr52733.2024.01398

[24] Schmidt, F., Enzweiler, M. and Valada, A. (2025) GraphPilot: Grounded Scene Graph Conditioning for Language-Based Autonomous Driving. arXiv: 2511.11266. https://arxiv.org/abs/2511.11266

[25] Zheng, W., Chen, W., Huang, Y., Zhang, B., Duan, Y. and Lu, J. (2024) OccWorld: Learning a 3D Occupancy World Model for Autonomous Driving. In: *Computer Vision-ECCV* 2024, Springer Nature Switzerland, 55-72. https://doi.org/10.1007/978-3-031-72624-8_4

[26] Min, C., Zhao, D., Xiao, L., Zhao, J., Xu, X., Zhu, Z., *et al*. (2024) DriveWorld: 4D Pre-Trained Scene Understanding via World Models for Autonomous Driving. 2024 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vienna, 7-11 May 2024, 15522-15533. https://doi.org/10.1109/cvpr52733.2024.01470

[27] Gao, S., Yang, J., Chen, L., Chitta, K., Qiu, Y., Geiger, A., *et al*. (2024) Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability. *Advances in Neural Information Processing Systems* 37, Vancouver, 10-15 December 2024, 91560-91596. https://doi.org/10.52202/079017-2906

[28] Chitta, K., Prakash, A., Jaeger, B., Yu, Z., Renz, K. and Geiger, A. (2023) Transfuser: Imitation with Transformer-Based Sensor Fusion for Autonomous Driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **45**, 12878-12895. https://doi.org/10.1109/tpami.2022.3200245

[29] Wu, P., Jia, X., Chen, L., Yan, J., Li, H. and Qiao, Y. (2022) Trajectory-Guided Control Prediction for End-to-End Autonomous Driving: A Simple Yet Strong Baseline. *Advances in Neural Information Processing Systems* 35, New Orleans, 28 November-9 December 2022, 6119-6132. https://doi.org/10.52202/068431-0443

[30] Casas, S., Sadat, A. and Urtasun, R. (2021) MP3: A Unified Model to Map, Perceive, Predict and Plan. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), 19-25 June 2021, 14403-14412. https://doi.org/10.1109/cvpr46437.2021.01417

[31] Shao, H., Wang, L., Chen, R., *et al*. (2023) Safety-Enhanced Autonomous Driving Using Interpretable Sensor Fusion Transformer. *Proceedings of the* 6*th Conference on Robot Learning*. *PMLR*, **205**, 726-737.

[32] Jiang, B., Chen, S., Xu, Q., Liao, B., Chen, J., Zhou, H., *et al*. (2023) VAD: Vectorized Scene Representation for Efficient Autonomous Driving. 2023 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Paris, 1-6 October 2023, 8340-8350. https://doi.org/10.1109/iccv51070.2023.00766

[33] Sun, W., Lin, X., Shi, Y., Zhang, C., Wu, H. and Zheng, S. (2025) SparseDrive: End-to-End Autonomous Driving via Sparse

Scene Representation. 2025 *IEEE International Conference on Robotics and Automation* (*ICRA*), Atlanta, 19-23 May 2025, 8795-8801. https://doi.org/10.1109/icra55743.2025.11128800

[34] Weng, X., Ivanovic, B., Wang, Y., Wang, Y. and Pavone, M. (2024) PARA-Drive: Parallelized Architecture for Real-Time Autonomous Driving. 2024 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 16-22 June 2024, 15449-15458. https://doi.org/10.1109/cvpr52733.2024.01463

[35] Jia, X., You, J., Zhang, Z., *et al*. (2025) DriveTransformer: Unified Transformer for Scalable End-to-End Autonomous Driving. *International Conference on Learning Representations*, Singapore, 24-28 April 2025, 57196-57212.

[36] Zheng, W., Song, R., Guo, X., Zhang, C. and Chen, L. (2024) GenAD: Generative End-to-End Autonomous Driving. In: *Computer Vision-ECCV* 2024, Springer Nature Switzerland, 87-104. https://doi.org/10.1007/978-3-031-73650-6_6

[37] Liao, B., Chen, S., Yin, H., Jiang, B., Wang, C., Yan, S., *et al*. (2025) DiffusionDrive: Truncated Diffusion Model for End-to-End Autonomous Driving. 2025 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 11-15 June 2025, 12037-12047. https://doi.org/10.1109/cvpr52734.2025.01124

[38] Li, Y., Fan, L., He, J., *et al*. (2025) Enhancing End-to-End Autonomous Driving with Latent World Model. *International Conference on Learning Representation*s, Singapore, 24-28 April 2025, 34849-34866.

[39] Zheng, Y., Yang, P., Xing, Z., Zhang, Q., Zheng, Y., Gao, Y., *et al*. (2025) World4Drive: End-to-End Autonomous Driving via Intention-Aware Physical Latent World Model. 2025 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Honolulu, 19-23 October 2025, 28632-28642. https://doi.org/10.1109/iccv51701.2025.02659

[40] Zhang, B., Song, N., Li, J., Zhu, X., Deng, J. and Zhang, L. (2025) Future-Aware End-to-End Driving: Bidirectional Modeling of Trajectory Planning and Scene Evolution. *Advances in Neural Information Processing Systems* 38, San Diego, 2-7 December 2025, 11358-11383. https://doi.org/10.52202/085713-0345

[41] Xu, Z., Zhang, Y., Xie, E., Zhao, Z., Guo, Y., Wong, K.K., *et al*. (2024) DriveGPT4: Interpretable End-to-End Autonomous Driving via Large Language Model. *IEEE Robotics and Automation Letters*, **9**, 8186-8193. https://doi.org/10.1109/lra.2024.3440097

[42] Sima, C., Renz, K., Chitta, K., Chen, L., Zhang, H., Xie, C., *et al*. (2024) DriveLM: Driving with Graph Visual Question Answering. In: *Computer Vision-ECCV* 2024, Springer Nature Switzerland, 256-274. https://doi.org/10.1007/978-3-031-72943-0_15

[43] Tian, X., Gu, J., Li, B., *et al*. (2025) DriveVLM: The Convergence of Autonomous Driving and Large Vision-Language Models. *Proceedings of the* 8*th Conference on Robot Learning*. *PMLR*, **270**, 4698-4726.

[44] Hwang, J.J., Xu, R., Lin, H., *et al*. (2025) EMMA: End-to-End Multimodal Model for Autonomous Driving. *Transactions on Machine Learning Research*, 1-31.

[45] Zhang, Q., Zhu, M. and Yang, H.F. (2024) Think-Driver: From Driving-Scene Understanding to Decision-Making with Vision Language Models. *European Conference on Computer Vision Workshop on Autonomous Vehicles Meet Multimodal Foundation Models*, Milan, 29 September 2024, 1-14.

[46] Zhou, X., Han, X., Yang, F., Ma, Y., Tresp, V. and Knoll, A. (2026) OpenDriveVLA: Towards End-to-End Autonomous Driving with Large Vision Language Action Model. *Proceedings of the AAAI Conference on Artificial Intelligence*, **40**, 13782-13790. https://doi.org/10.1609/aaai.v40i16.38386

[47] Zhou, Z., Cai, T., Zhao, S., Zhang, Y., Huang, Z., Zhou, B., *et al*. (2025) AutoVLA: A Vision-Language-Action Model for End-to-End Autonomous Driving with Adaptive Reasoning and Reinforcement Fine-Tuning. *Advances in Neural Information Processing Systems* 38, San Diego, 2-7 December 2025, 31725-31761. https://doi.org/10.52202/085713-0942

[48] Renz, K., Chen, L., Arani, E. and Sinavski, O. (2025) SimLingo: Vision-Only Closed-Loop Autonomous Driving with Language-Action Alignment. 2025 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 11-15 June 2025, 11993-12003. https://doi.org/10.1109/cvpr52734.2025.01120

[49] Fang, S., Cui, Y., Liang, H., *et al*. (2025) CoReVLA: A Dual-Stage End-to-End Autonomous Driving Framework for Long-Tail Scenarios via Collect-and-Refine. https://arxiv.org/abs/2509.15968

[50] Xie, S., Kong, L., Dong, Y., Sima, C., Zhang, W., Chen, Q.A., *et al*. (2025) Are VLMs Ready for Autonomous Driving? An Empirical Study from the Reliability, Data, and Metric Perspectives. 2025 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Honolulu, 19-20 October 2025, 6585-6597. https://doi.org/10.1109/iccv51701.2025.00621

[51] Fu, H., Zhang, D., Zhao, Z., Cui, J., Liang, D., Zhang, C., *et al*. (2025) ORION: A Holistic End-to-End Autonomous Driving Framework by Vision-Language Instructed Action Generation. 2025 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Honolulu, 19-20 October 2025, 24823-24834. https://doi.org/10.1109/iccv51701.2025.02302

[52] Liu, C., Zhu, M., Zhang, Z., Song, L., Zhao, X., Luo, Q., *et al*. (2025) TAD-E2E: A Large-Scale End-to-End Autonomous Driving Dataset. 2025 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Honolulu, 19-20 October 2025, 26600-26609. https://doi.org/10.1109/iccv51701.2025.02469

[53] Dosovitskiy, A., Ros, G., Codevilla, F., *et al*. (2017) CARLA: An Open Urban Driving Simulator. *Proceedings of the* 1*st*

Annual Conference on Robot Learning. *PMLR*, **78**, 1-16.

[54] Caesar, H., Kabzan, J., Tan, K.S., *et al*. (2021) nuPlan: A Closed-Loop ML-Based Planning Benchmark for Autonomous Vehicles. *CVPR Workshop on Autonomous Driving*: *Perception*, *Prediction and Planning*, 19-25 June 2021, 1-5.

[55] 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. *Advances in Neural Information Processing Systems* 37, Vancouver, 10-15 December 2024, 819-844. https://doi.org/10.52202/079017-0025

[56] Daza, I.G., Izquierdo, R., Martínez, L.M., Benderius, O. and Llorca, D.F. (2023) Sim-to-Real Transfer and Reality Gap Modeling in Model Predictive Control for Autonomous Driving. *Applied Intelligence*, **53**, 12719-12735. https://doi.org/10.1007/s10489-022-04148-1

[57] Qian, T., Chen, J., Zhuo, L., Jiao, Y. and Jiang, Y. (2024) NuScenes-QA: A Multi-Modal Visual Question Answering Benchmark for Autonomous Driving Scenario. *Proceedings of the AAAI Conference on Artificial Intelligence*, **38**, 4542-4550. https://doi.org/10.1609/aaai.v38i5.28253

[58] Wang, J., Pun, A., Tu, J., Manivasagam, S., Sadat, A., Casas, S., *et al*. (2021) AdvSim: Generating Safety-Critical Scenarios for Self-Driving Vehicles. 2021 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), 19-25 June 2021, 9909-9918. https://doi.org/10.1109/cvpr46437.2021.00978

[59] Gil, R.C., Ornia, D.J., Mustafa, K.A. and Alonso Mora, J. (2025) Predictability Awareness for Efficient and Robust Multi-Agent Coordination. *International Joint Conference on Autonomous Agents and Multiagent Systems*, Detroit, 19-23 May 2025, 886-894. https://doi.org/10.65109/hhpp1298