自动化学报
ACTA AUTOMATICA SINICA

迈向驾驶智能体: 大模型时代端到端自动驾驶的认知架构演进与展望

刘瑞琪 方可 王远大 王晓 孙长银

Towards Driving Agents: Evolution and Outlook of Cognitive Architectures for End-to-end Autonomous Driving in the Era of Large Models

LIU Rui-Qi, FANG Ke, WANG Yuan-Da, WANG Xiao, SUN Chang-Yin

在线阅读 View online: https://doi.org/10.16383/j.aas.c260083

您可能感兴趣的其他文章

基于平行测试的认知自动驾驶智能架构研究

An Intelligent Architecture for Cognitive Autonomous Driving Based on Parallel Testing

自动化学报. 2024, 50(2): 356−371 https://doi.org/10.16383/j.aas.c220820

视觉语言动作模型综述: 从前史到前沿

Vision−Language−Action Models: From the Early Foundations to the State−of−the−Art

自动化学报. 2025, 51(9): 1922−1950 https://doi.org/10.16383/j.aas.c250417

基于大模型的具身智能系统综述

Embodied Intelligence Systems Based on Large Models: A Survey

自动化学报. 2025, 51(1): 1−19 https://doi.org/10.16383/j.aas.c240542

具身智能自主无人系统技术

Embodied Intelligence Autonomous Unmanned Systems Technology

自动化学报. 2025, 51(4): 762−777 https://doi.org/10.16383/j.aas.c240456

具身智能研究的关键问题: 自主感知、行动与进化

Key Problems of Embodied Intelligence Research: Autonomous Perception, Action, and Evolution

自动化学报. 2025, 51(1): 43−71 https://doi.org/10.16383/j.aas.c240364

面向具身操作的视觉语言动作模型综述

Survey of Vision−Language−Action Models for Embodied Manipulation

自动化学报. 2026, 52(1): 18−51 https://doi.org/10.16383/j.aas.c250394

迈向驾驶智能体: 大模型时代端到端自动驾驶的认知架构演进与展望

刘瑞琪[1] 方 可[1] 王远大[2,3] 王 晓[2,3,4,5] 孙长银[1,3,4]

摘 要 自动驾驶作为具身智能的关键应用方向, 正经历从模块化结构向端到端框架的深刻变革, 而大模型的爆发为其从"感知智能"向"认知智能"的跃迁提供了关键驱动力. 首先回顾自动驾驶技术的演进历程, 剖析传统模块化结构的局限与端到端范式的兴起, 指出现有的基于监督学习的端到端模型在可解释性与因果推理上的不足. 接着引出大模型尤其是多模态大语言模型赋能自动驾驶的前沿机制, 系统梳理大语言模型、视觉−语言模型、视觉−语言−动作模型的贡献. 与此同时, 也梳理了新兴世界模型的发展脉络及其为自动驾驶带来的环境推演能力. 进一步地, 探讨自动驾驶的训练范式与验证体系变革, 分析数据集从几何感知向认知理解的演变, 以及开环评估与闭环验证的方法. 最后, 总结当前技术面临的挑战, 并展望构建数据驱动、认知增强与虚实闭环的一体化自动驾驶系统的未来方向.

关键词 认知自动驾驶; 驾驶智能体; 端到端学习; 多模态大语言模型; 具身智能; 世界模型

引用格式 刘瑞琪, 方可, 王远大, 王晓, 孙长银. 迈向驾驶智能体: 大模型时代端到端自动驾驶的认知架构演进与展望. 自动化学报, 2026, **52**(10): 2152−2183

DOI 10.16383/j.aas.c260083 **CSTR** 32138.14.j.aas.c260083

Towards Driving Agents: Evolution and Outlook of Cognitive Architectures for End-to-end Autonomous Driving in the Era of Large Models

LIU Rui-Qi[1] FANG Ke[1] WANG Yuan-Da[2,3] WANG Xiao[2,3,4,5] SUN Chang-Yin[1,3,4]

Abstract As a pivotal application of embodied intelligence, autonomous driving is shifting from modular to end-to-end frameworks. The surge of large models provides a key driving force for its leap from perceptual intelligence to cognitive intelligence. The paper reviews the evolution of autonomous driving technology, dissects limitations of traditional modular structure and the rise of end-to-end paradigms, and points out the deficiencies in interpretability and causal reasoning of supervised learning-based end-to-end models. Then it introduces how large models, especially multimodal large language models, empower autonomous driving, systematically categorizing contributions of large language models, vision-language models, and vision-language-action models. It also sorts out the development of emerging world models and their capability for environmental reasoning in autonomous driving. Furthermore, it explores changes in training paradigms and validation systems for autonomous driving, analyzing dataset evolution from geometric perception to cognitive understanding, and evaluation methods from open-loop to closed-loop verification. Finally, it summarizes the challenges currently facing the technology and envisions future directions for integrated autonomous driving systems that are data-driven, cognitively enhanced, and based on a virtual-real closed loop.

Keywords cognitive autonomous driving; driving agent; end-to-end learning; multimodal large language model; embodied intelligence; world model

Citation Liu Rui-Qi, Fang Ke, Wang Yuan-Da, Wang Xiao, Sun Chang-Yin. Towards driving agents: Evolution and outlook of cognitive architectures for end-to-end autonomous driving in the era of large models. *Acta Automatica Sinica*, 2026, **52**(10): 2152−2183

收稿日期 2026-01-29 录用日期 2026-04-12
Manuscript received January 29, 2026; accepted April 12, 2026
国家自然科学基金 (62522601) 资助
Supported by National Natural Science Foundation of China (62522601)
本文责任编委 温广辉
Recommended by Associate Editor WEN Guang-Hui

1. 安徽大学人工智能学院 合肥 230601 2. 安徽大学机器人学院 合肥 230601 3. 自主无人系统技术教育部工程研究中心 合肥 230601 4. 安徽省无人系统与智能技术工程研究中心 合肥 230601 5. 光电信息获取与防护技术全国重点实验室 合肥 230601

1. School of Artificial Intelligence, Anhui University, Hefei 230601 2. School of Robotics, Anhui University, Hefei 230601 3. Engineering Research Center of the Ministry of Education for Autonomous Unmanned Systems Technology, Hefei 230601 4. Engineering Research Center of Anhui Province for Unmanned Systems and Intelligent Technology, Hefei 230601 5. State Key Laboratory of Opto-electronic Information Acquisition and Protection Technology, Hefei 230601

自动驾驶作为具身智能 (embodied intelligence) 的关键领域与核心课题, 正处于从特定场景辅助驾驶向全场景无人驾驶迈进的关键发展期. 长期以来, 实现安全、高效且具备人类智能水平的自动驾驶系统, 一直是学术界和工业界共同追求的核心目标. 在这一演进过程中, 自动驾驶系统的架构设计经历了深刻的范式转移: 从经典的模块化设计, 逐渐过渡到端到端 (end-to-end, E2E) 学习框架, 并随着近期多模态大语言模型 (multimodal large language model, MLLM) 的爆发, 迎来了向"通用智能体"跃迁的历史性机遇.

当前, 新一轮科技革命与产业变革深入发展, 自动驾驶与通用人工智能 (artificial general intelligence, AGI) 已成为引领未来智能交通、重塑人类出行范式的关键驱动力. 我国自动驾驶产业的政策环境正经历从"顶层规划"向"落地执行"的深刻转变. 以《智能汽车创新发展战略》[1] 确立国家战略地位为起点, 2023 年发布的《国家车联网产业标准体系建设指南》[2] 进一步完善了技术规范体系. 随后, 工业和信息化部等四部门发布的《关于开展智能网联汽车准入和上路通行试点工作的通知》[3], 以及"车路云一体化"试点城市的全面铺开[4–6], 标志着 L3 级自动驾驶正式迎来了商业化落地的政策窗口期. 这一完善的政策闭环, 为自动驾驶在真实复杂场景下的部署与迭代提供了坚实的制度保障. 与此同时,《中华人民共和国国民经济和社会发展第十四个五年规划和 2035 年远景目标纲要》[7] 也相应提出, 要积极发展自动驾驶和车路协同技术, 推动"智能网联汽车"产业化应用, 构建智慧交通体系.

在传统的自动驾驶技术栈中, 系统通常被解耦为 4 个核心子模块: 感知 (perception)、预测 (prediction)、规划 (planning) 和控制 (control). 这种模块化架构在早期发展中通过"分而治之"的策略降低了工程实现的复杂度, 但其串行流程不可避免地带来了"误差累积"效应[8], 且难以应对复杂的长尾场景. 为打破模块化架构的局限, 端到端自动驾驶框架应运而生. 自 Bojarski 等[9] 于 2016 年首次提出基于卷积神经网络 (convolutional neural network, CNN) 的端到端模型以来, 学术界近年来都在致力于构建从原始传感器输入到车辆控制信号的直接映射. 随着 Transformer[10] 架构的引入, 以 UniAD[8] 为代表的新一代端到端模型成功实现了感知、预测与规划的多任务集成, 在复杂场景下展现出超越模块化方法的性能.

尽管端到端方法在性能上取得了突破, 但由于其结构的限制, 往往缺乏对物理世界常识与社会交互规则的深层理解, 在可解释性方面依然面临不小的挑战. 近年来, 以 GPT-3[11] 为代表的基础模型研究表明, 通过扩大模型参数与训练数据规模, 深度神经网络能够涌现出惊人的上下文学习与推理能力. 这一颠覆性的发现推动了大语言模型 (large language model, LLM) 和多模态大模型, 如融合视觉输入的视觉–语言模型 (vision-language model, VLM) 与实现具身操作的视觉–语言–动作 (vision-language-action, VLA) 模型的飞速发展, 也为解决上述问题提供了一条全新的技术路径. 图 1 所示为模块化、端到端以及融合大模型的 3 种自动驾驶基本架构示意图. LLM 的应用, 使得自动驾驶系统不再是仅能处理数据的黑盒, 而是成为具备逻辑思维的智能体. 它不仅能够利用其通识知识库解决训练数据之外的长尾难题, 还能在面对复杂且未知的交通博弈时, 像人类司机一样进行逻辑推演与宏观规划. 这种将感知直觉与认知推理相结合的新范式, 不仅提升了系统在极端场景下的安全性, 更为自动驾驶的可解释性难题提供了可参考的解决方案. 与此同时, 具备时空预测能力的世界模型 (world model, WM) 正迅速发展, 其内生模拟功能能够基于对物理法则的理解进行反事实推演, 与大模型的语义逻辑形成关键互补. 图 2 为自动驾驶技术架构演进的发展脉络与代表性工作, 清晰地勾勒出自动驾驶技术从模块化、端到端到大模型与世界模型的演进时间线, 并展示了各阶段的关键工作.

同时, 近年来学术界涌现了多篇关于端到端自动驾驶与基础模型应用的高质量综述. 例如, 部分文献详细梳理了基于深度学习的端到端范式及其统一架构[12–14]; 另有一些研究则专门探讨了大语言模型[15] 或视觉–语言–动作模型[16] 在自动驾驶中的应用潜力. 尽管上述工作对各自领域进行了详尽剖析, 但是多局限于特定技术路线的探讨, 缺乏从"驾驶智能体"这一具身认知全系统视角对技术演进脉络的宏观梳理. 具体而言, 现有的端到端自动驾驶综述往往侧重于网络结构的映射关系, 如传感输入到规划输出, 而聚焦大模型的综述则多关注指令跟随与语义推理. 鲜有工作能够将旨在打破"语义理解与物理控制异构壁垒"的闭环大模型, 尤其是 VLA 模型, 与为智能体注入内生仿真与反事实推演能力的新兴世界模型, 深度整合在同一条从感知向认知跃迁的演进图谱中. 鉴于此, 本文旨在填补这一空白, 致力于梳理出一条数据驱动与认知增强相结合的一体化技术发展路线, 主要贡献如下.

1) 系统梳理自动驾驶从模块化到端到端、再到融合大模型的认知驾驶智能体的架构演进, 揭示系

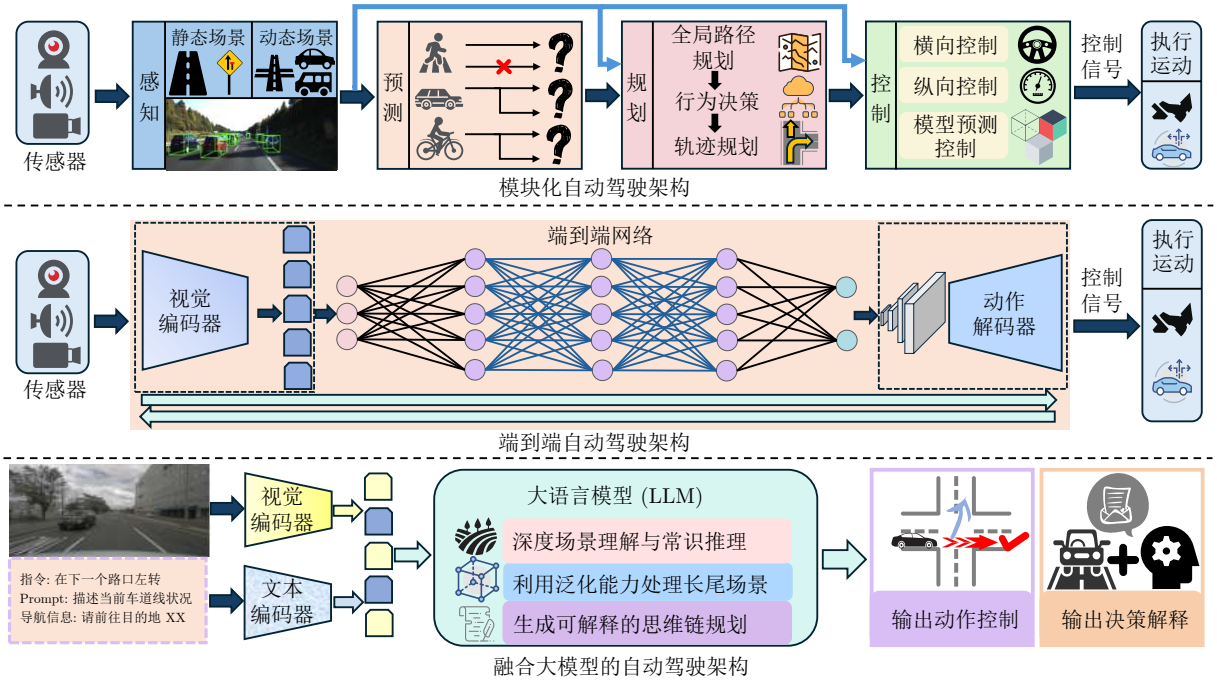


图 1 3 种自动驾驶架构示意图

Fig. 1 Schematic diagram of three autonomous driving architectures

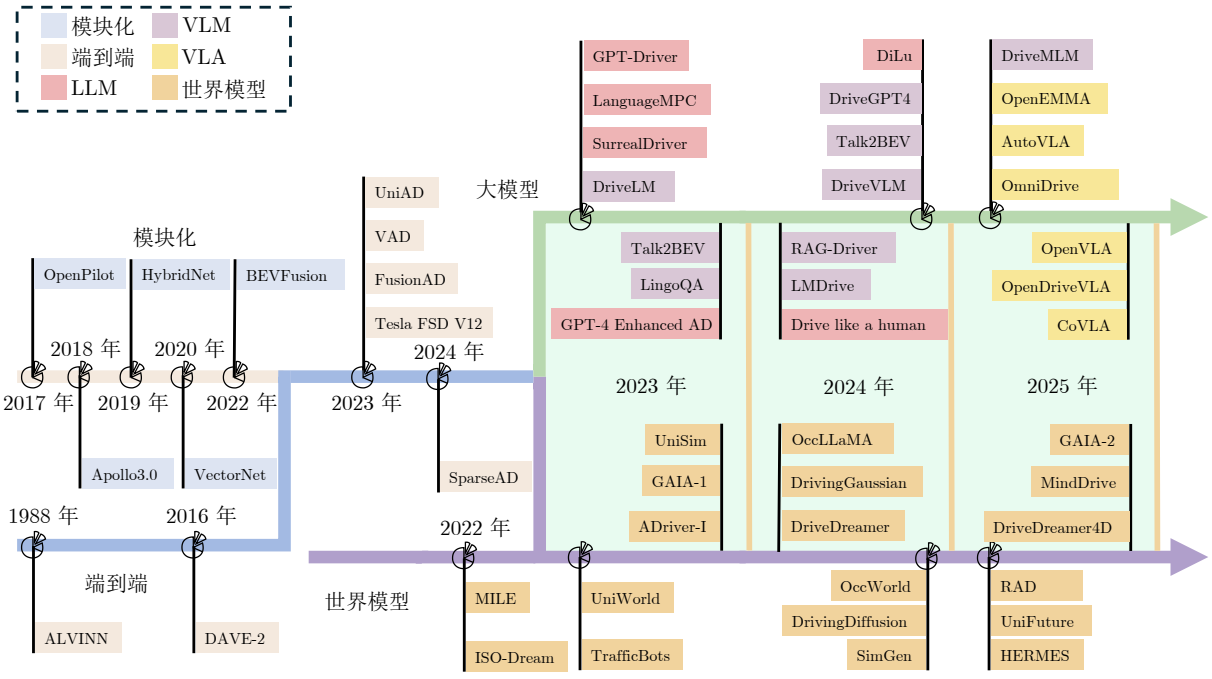


图 2 自动驾驶架构演进的发展脉络与关键工作

Fig. 2 Development trajectory and key works of autonomous driving architecture evolution

统从“拟合数据”向“理解世界”演变的内在逻辑, 在较长时间跨度下宏观地展现了自动驾驶的发展历程.

2) 系统性地整合 LLM 的逻辑推理、VLM 的多模态感知、VLA 的闭环控制及世界模型的推演能力, 构建面向自动驾驶的数据驱动与认知增强相结合的一体化技术发展路线.

3) 深入探讨伴随认知架构升级而来的数据演进——从几何感知到因果推理, 以及验证体系变革——从开环评估到虚实闭环, 为构建下一代具备类人决策能力的通用自动驾驶系统提供前瞻性的理

论参考与指引.

综上所述, 当前的自动驾驶研究正处于多模态融合与大模型驱动的交汇点. 本文的组织结构安排如下: 第 1 节回顾传统模块化自动驾驶的技术架构, 详细阐述感知、预测、规划及控制四大核心子模块的原理和发展, 并分析该架构在应对复杂场景时的局限性; 第 2 节深入探讨由模块化架构向端到端范式的演变, 梳理从早期行为克隆到结构化网络的演进脉络, 并剖析其向类人认知演进的内在逻辑; 第 3 节重点探讨大模型尤其是多模态大模型的发展及其给自动驾驶技术带来的深刻影响, 系统梳理大语言模型、视觉–语言模型与视觉–语言–动作模型的发展脉络及其在自动驾驶领域的应用; 第 4 节对世界模型这一具有强大推演能力的模型及其发展与在自动驾驶中的具体应用机制进行介绍; 第 5 节则聚焦于大模型时代下自动驾驶模型的训练范式与验证体系, 从训练、测试的开环与闭环 2 个维度阐述自动驾驶训练及评估体系的演变; 最后, 总结现有研究的挑战并对未来的研究方向进行展望.

1 传统模块化自动驾驶

传统的模块化方法将自动驾驶系统分解为不同的模块, 通常包括感知、预测、规划和控制四部分. 其中, 感知模块负责从摄像头、激光雷达等传感器中提取环境语义; 预测模块基于感知结果推演交通参与者的未来轨迹; 规划模块在时空约束下生成自车的行驶路径; 控制模块则将路径转化为底层的油门、刹车和转向指令. 下面详细介绍每个模块.

1.1 环境感知

感知的本质在于从物理环境中采集数据并解析高维语义信息. 构建一个高精度、全覆盖的感知体系, 是保障自动驾驶车辆在复杂交通流中实现后续决策的先决条件. 得益于近二十年来传感器硬件的迭代与机器学习技术的发展[17–19], 对静态对象, 如车道线、障碍物等的基础感知任务已取得阶段性进展.

为进一步精准重构物理世界的动态场景, 自动驾驶系统通常配备由数十种传感器组成的复杂阵列, 以复刻人类驾驶员的感官机能. 在环境感知层面, 该阵列主要集成了摄像头、激光雷达 (light detection and ranging, LiDAR)、毫米波雷达及超声波雷达等硬件设备[20]. 早期的感知架构主要分为两大类型: 一是以特斯拉为代表的依赖图像特征的纯视觉感知方案; 二是强调多源信息互补、结合视觉与雷达数据的多传感器融合方案. 纯视觉感知方案主要依赖摄像头获取丰富的语义信息, 具有成本低、硬件成熟度高等优势. 特别是在感知的实时性优化方面, Dai[21] 针对自动驾驶场景中车辆尺度变化大与遮挡严重的挑战, 提出一种基于级联回归的混合检测架构 HybridNet, 在保证高精度的同时显著降低了计算耗时. 然而, 该类方法本质上是试图从二维图像平面恢复丢失的三维几何信息, 这使得其在深度估计精度以及恶劣天气 (如雨雪、浓雾) 下的鲁棒性面临严峻挑战[17]. 为克服单一传感器的物理瓶颈, 多传感器融合方案应运而生. 通过引入激光雷达提供的精确几何点云与毫米波雷达的速度信息, 融合系统能够在特征层面实现多源信息的优势互补, 从而显著提升感知系统的安全冗余性[22].

多模态融合方法虽然解决了信息缺失问题, 但不同传感器之间的数据异构性与视角差异也带来了复杂的时空对齐难题, 即从 2D 图像到 3D 空间的转换能力较弱. 针对这一痛点, 近年来兴起的鸟瞰图 (bird′s-eye-view, BEV) 感知范式[23] 成为解决异构数据融合的关键突破口. Li 等[24] 的研究表明, 通过将多模态特征统一投影到以车辆为中心的 BEV 坐标空间中, 不仅有效消除了透视畸变, 还为后续的规划与控制模块提供了一个直观、统一的可解释性表征. 同时, 另一类方法试图通过 Transformer[10] 强大的注意力机制直接学习 2D 到 3D 的映射关系. 其中的代表性工作 BEVFormer[25] 创新性地构建了全 Transformer 架构, 利用空间交叉注意力与时间自注意力机制替代了显式深度估计, 在解决特征漂移问题的同时显著增强了对动态遮挡场景的感知能力. 同时, 针对多模态对齐瓶颈, BEVFusion[26] 是一种基于统一鸟瞰图表征的融合范式, 将视觉语义与雷达几何信息无损映射至共享特征空间. 该架构不仅在 nuScenes[27] 3D 目标检测与 BEV 地图分割任务上刷新了 SOTA 记录, 更在单一传感器失效, 如夜间视觉受限的场景下维持了高度的系统鲁棒性.

尽管现有感知方案已取得显著进展, 但在面对开放动态的真实世界时仍面临两大严峻挑战. 首先是长尾分布下的泛化难题[28]. 当前基于学习的方法高度依赖训练数据, 一旦遭遇训练集之外的场景, 模型往往难以适应. 其次是缺乏类人的深层场景理解能力. 现有方法本质上多局限于浅层的数据拟合与模式匹配, 缺少对环境的上下文推理, 导致系统难以像人类驾驶员一样形成对复杂交通态势的全局认知.

1.2 轨迹预测

作为连接上游环境感知与下游决策规划的关键枢纽, 轨迹预测模块负责根据其他交通参与者 (如

行人、车辆等) 的历史状态推演其未来的运动趋势. 在早期的自动驾驶方案中, 这一环节往往被简化或隐含在规划模块中. 随着自动驾驶向复杂城市场景延伸, 预测模块逐渐独立并弥补感知与规划之间的差距, 即实现了感知的“当前态”向规划所需的“未来态”的映射.

纵观技术演进脉络, 传统的预测方案主要遵循基于模型的技术路线, 其核心涵盖了物理动力学、意图理解以及行为交互等建模维度. 尽管这些基于模型的方法具备了一定的可解释性, 但在处理长时域轨迹推演及行为不确定性时往往存在缺陷. 近十年来, 数据驱动的深度学习范式逐渐占据主导地位[29–31]. 研究者广泛利用循环神经网络 (recurrent neural network, RNN)[32]、卷积神经网络 (convolutional neural network, CNN)[33] 以及图注意力网络 (graph attention networks, GAT)[34] 等来建模复杂的时序特征与交互关系, 从而实现从手工规则设计到数据中自动学习复杂行为模式的范式转变, 为处理高维、不确定的交通场景提供了更具表达力的解决方案. 在此基础上, 为进一步提升预测结果在复杂路网中的合规性与合理性, 引入高精度地图 (HD map) 作为强先验知识已成为主流趋势. 代表工作 VectorNet[35] 是一种矢量化的场景表征方法, 其整体框架如图 3 所示. VectorNet 不采用传统的图像渲染输入, 而是将各类地图元素 (如车道几何、斑马线和智能体历史轨迹) 近似为由有序向量组成的折线. 其核心设计包含一个层次化的图神经网络, 该网络首先通过折线子图 (polyline subgraphs) 模块来处理这些局部向量集合. 该子图将属于同一地理实体或轨迹的所有向量节点相互连接, 利用多层感知机与置换不变的聚合操作 (如最大池化) 来提取局部空间特征, 从而将每一条复杂的折线压缩编码为单一的高维特征节点. 随后, 系统利用全连接的全局交互图来建模所有环境要素与代理节点之间的高阶交互关系, 从而实现精准的轨迹预测, 有效约束了轨迹生成的空间分布. 此外, 考虑到真实交通环境中交通参与者 (如行人、周边车辆) 的意图的高度随机性与多样性, 如路口左转亦或是直行, 确定性的单一轨迹输出已难以满足安全需求. 因此, 近年来一些预测模型逐渐转向采用多模态概率输出范式[36], 不局限于寻求单一最优解, 而是基于场景上下文预测出多条符合路网拓扑结构与运动学约束的未来轨迹, 并为每条轨迹赋予相应的联合概率或置信度, 从而为下游规划模块提供更丰富的风险评估依据.

目前, 预测技术不再局限于对单个物体的轨迹推演, 而是强调对整体交通情境的理解. 车辆必须具备解读社会化交互行为及适应环境瞬变的能力, 这构成了当前研究的关键挑战——如何在复杂场景下实现相对全面的感知与预测. 对这一瓶颈的突破,

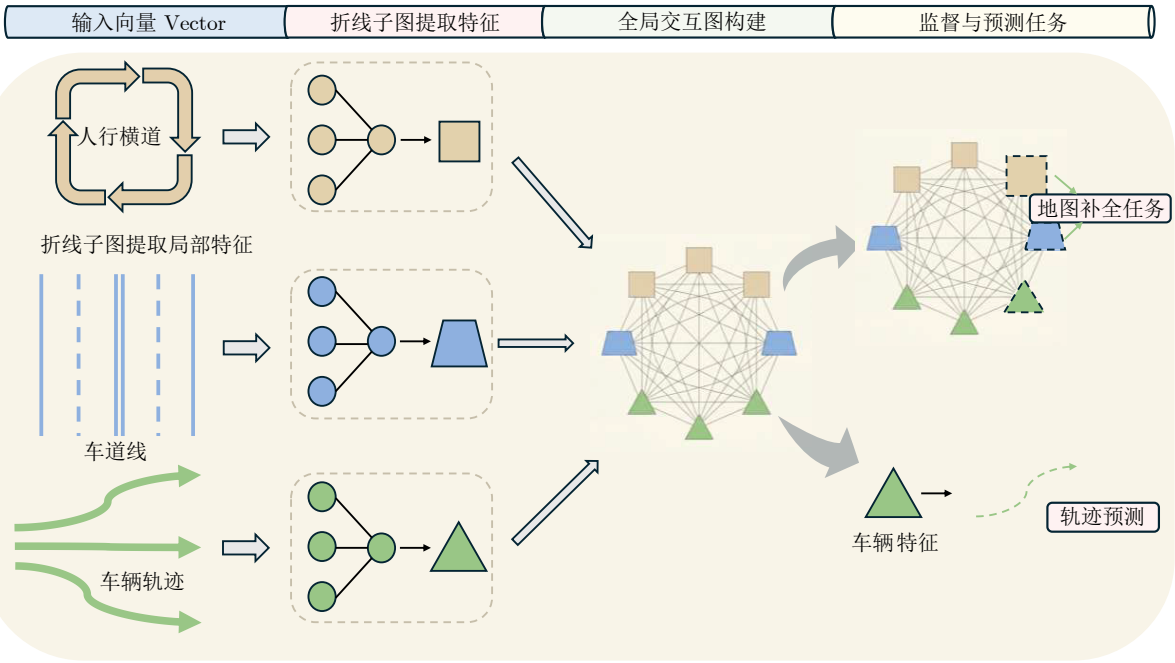


图 3 基于矢量化场景表征的轨迹预测框架 VectorNet

Fig. 3 VectorNet: A trajectory prediction framework based on vectorized scene representation

将标志着预测技术向更高精度与更强智能化的关键跨越.

1.3 决策规划

作为自动驾驶系统的核心组成部分, 决策规划模块负责对上游的感知与预测模块输出的异构信息进行深度整合, 并在满足车辆运动学约束的前提下, 制定出安全、高效且舒适的行驶指令[37]. 从控制论的角度来看, 规划本质上是一个在动态不确定的交通环境中寻求最优解的过程[38–40]. 为降低计算复杂度并实现任务解耦, 依据时空尺度与功能属性, 传统的规划架构通常可划分为 3 个递进的层级: 全局路线规划、行为决策以及轨迹规划.

首先是全局路线规划, 它主要处理静态环境信息以解决“如何到”的宏观导航问题. 其核心任务是基于拓扑路网, 利用 A* 或 Dijkstra 等传统图搜索算法[41], 在起点与终点间规划出全局最优路径. 在此基础上, 行为决策层作为连接宏观导航与微观控制的关键枢纽, 负责解决“如何开”的动态难题. 由于自动驾驶系统对实时性有较强要求, 该层级需要在毫秒级时间内对周围交通流进行意图推理, 并生成如“跟车”、“变道”或“超车”等高维策略, 这不仅高度依赖对环境的语义理解, 更直接反映了驾驶系统的智能水平与风格差异[42]. 最后, 轨迹规划层负责将上述抽象的行为指令转化为带有时间戳的具体几何路径点, 供下游控制模块执行. 在算法演进上, 这一领域正经历从早期的基于采样搜索 (如快速探索随机树 (rapidly-exploring random tree, RRT)[43]) 向基于数值优化方法的范式转移[38], 后者凭借数学理论在复杂受限空间中展现出了更高的计算效率与轨迹平滑度.

近年来, 随着人工智能技术的渗透, 基于学习的规划范式逐渐兴起. 根据其学习机制的不同, 主要可分为强化学习、模仿学习以及平行学习等类别[44]. 强化学习 (reinforcement learning, RL)[45] 将规划问题建模为马尔科夫决策过程 (Markov decision process, MDP)[46]; 模仿学习 (imitation learning, IL) 则致力于直接从人类专家示范中提取拟人化的驾驶策略, 训练神经网络拟合出从传感器输入到车辆底层控制的映射函数[47]; 而平行学习是一种基于平行系统理论[48] 的虚拟–现实交互学习框架, 通过人工系统与现实系统的持续交互与协同优化, 以解决自动驾驶端到端规划中数据稀缺、训练效率低和长尾场景鲁棒性不足的问题[44]. 此外, 将神经网络与经典规划器相结合的混合架构[49–51] 正成为提供实时解决方案的新趋势.

尽管现有体系已趋成熟, 但决策规划仍面临两大核心瓶颈: 一是在满足复杂运动学约束的同时, 需有效抑制由上游感知与预测模块引入的级联不确定性; 二是传统架构中各子模块间的相互独立限制了全局最优解的获取. 因此, 打破模块隔离, 构建更加紧密耦合的统一混合规划框架, 以提升系统在长尾场景下的鲁棒性, 已成为技术发展的主要突破方向, 也为后续端到端范式的研究打下了结构基础.

1.4 运动控制

作为传统模块化自动驾驶的末端执行环节, 运动控制构建了顶层数学模型与底层物理执行之间的桥梁[52–54]. 与上游模块不同, 该层级并不涉及对感知信息的语义理解, 而是承担着将规划出的离散时空轨迹转化为车辆具体指令 (如油门、制动及转向角) 的核心任务, 其根本目标是在最小化跟踪误差的同时, 兼顾车辆的动力学稳定性与乘坐舒适性[55].

在工程实践中, 控制任务通常被分解为纵向与横向 2 个维度. 具体而言, 纵向控制主要聚焦于速度维度的调节, 涵盖对动力、传动及制动系统的综合管控. 相比之下, 横向控制则致力于垂直于行进方向的侧向运动管理, 具体体现为对转向角的实时调整与侧滑修正. 早期的经典方法多采用 PID[56] 控制处理纵向线性部分, 或利用基于几何的 Stanley 控制器[57] 处理横向偏差. 然而, 在高速变道或路口转向等复杂场景下, 独立的横纵向控制难以满足实际需求, 促使技术演进向着横纵协同与动力学优化的方向发展. 其中, 模型预测控制 (model predictive control, MPC)[58] 凭借其处理多变量耦合约束的卓越能力, 成为主流方案. MPC 能够在一个有限的时域内预测车辆未来状态并优化控制序列, 从而实现对非线性车辆动力学的精准补偿, 提升车辆在动态环境中的控制鲁棒性与适应性.

值得注意的是, 尽管深度学习在感知领域占据主导, 但在控制层, 由于对安全性与可解释性的极高要求, 纯数据驱动方法的利用需相对谨慎. 增强学习型 MPC[59] 是一种融合传统方法可靠性与学习方法适应性的混合控制范式. 它利用高斯过程 (Gaussian process, GP)[60] 与神经网络在线修正物理模型的残差, 从而显著提升了控制器在车辆参数变化或环境扰动下的自适应能力.

虽然上述控制流程相对完善, 但如何弥合理想模型与真实物理世界之间的“Sim-to-Real (仿真–现实)”鸿沟仍是关键难题. 控制器如何在长尾场景下依然保持鲁棒性[61], 并同时适应不同乘客的个性化需求, 构成了当前面临的主要挑战. 这也要求控制

模块打破被动执行的角色, 与上游规划层建立更紧密的双向联动机制.

2 端到端自动驾驶范式

随着自动驾驶技术的发展, 传统模块化架构在处理复杂长尾场景时暴露出的累积误差与信息丢失问题, 逐渐成为限制系统性能上限的瓶颈. 作为应对, 端到端自动驾驶范式应运而生. 它试图打破模块间的壁垒, 将传统的感知、预测、规划、控制等模块整合于单一网络架构之中, 建立从原始传感器输入到车辆控制指令的直接映射, 从而实现全链路的联合优化与特征共享. 本节将深入探讨这一范式转移的内在逻辑: 首先梳理端到端架构从早期简单的行为克隆到多任务协同与结构化网络的演进脉络, 随后分析其如何涌现出类人的环境认知与决策能力, 为理解下一代自动驾驶技术提供理论框架.

2.1 端到端自动驾驶的演进

纵观端到端自动驾驶的发展, 呈现出清晰的从“原始信号直连”向“深层语义理解”演进的脉络. Tampuu 等[62] 将这一范式定义为一种利用单一神经网络直接将原始传感器数据映射为车辆控制指令的架构. 与传统模块化栈中各子任务被人为割裂不同, 现代 E2E 范式的核心优势在于构建“全微分学习系统”[63], 这种机制允许利用原始传感器数据对整个链路进行联合特征优化, 使模型在反向传播中自动习得对驾驶任务最关键的特征表示. 早期的端到端模型结构极其简洁[9], 主要采用单一摄像头输入, 再通过行为克隆生成转向控制命令, 但在城市复杂场景下的泛化能力受限. 随着数据规模、计算能力与模拟环境的提升, 研究开始融入多模态输入 (如摄像头+LiDAR+地图) 与更复杂的输出形式 (如轨迹预测、规划点生成等)[62, 64–65]. 而在架构上, 由纯卷积网络向 Transformer 结构的转变, 极大地提升了模型捕捉长时序与全局空间关系的能力[64]. 在输出的控制上, 由低层控制命令向中层轨迹或高层意图规划的迁移, 提升了系统的可解释性与安全性[66]. 沿着这一趋势, 针对单纯轨迹输出依赖 PID 控制器导致的动态性能瓶颈, Wu 等[67] 创新性地构建了双流多任务架构, 通过轨迹引导注意力机制将显式的规划特征融入控制分支, 实现了规划引导下的端到端直接控制优化. 此外, 训练范式也逐步从单传感器开环训练向闭环训练与仿真–现实迁移方向发展, 增强了系统在现实世界的稳健性[63].

端到端自动驾驶研究范式的雏形最早可追溯至 Pomerleau[68] 于 1988 年提出的 ALVINN 系统, 它首次验证了利用浅层神经网络直接映射原始传感器信号到车辆转向指令的可行性, 为后续基于深度学习的端到端模型奠定了核心思想基础. 2016 年, 以 NVIDIA 的 DAVE-2[9] 为代表的基于深度学习的研究, 利用 CNN 构建了从前视摄像头图像到转向角的直接映射, 其结构如图 4 所示.

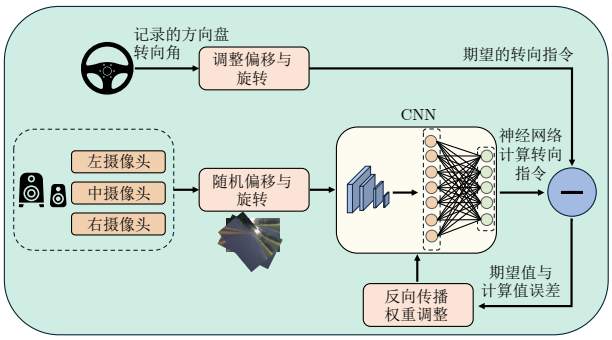


图 4 基于 CNN 的端到端自动驾驶系统 DAVE-2

Fig. 4 DAVE-2: A CNN-based end-to-end autonomous driving system

虽然其验证了“感知–动作”直接耦合的可行性, 但这种单一输入的“黑盒”特性导致系统缺乏对路口转向等多义性场景的判断能力. 针对这一局限, Codevilla 等[47] 提出条件模仿学习 (conditional imitation learning, CIL) 方法. 该方法将高级导航指令编码为独热向量并注入网络分支, 成功实现了端到端系统从被动到主动的跨越, 解决了交叉路口的决策歧义性问题.

然而, 随着场景复杂度的提升, 仅依赖 CNN 的简单架构已难以应对动态环境, 研究重心开始向增强时空建模能力与多模态融合转移. 一方面, 架构设计逐渐向 Transformer 进行转变. 如 NEAT[69] 引入神经感知场增强了对动态障碍物的响应, TransFuser[70] 首次利用 Transformer 在 BEV 空间融合了激光雷达与视觉数据, 通过全局上下文建模在 CARLA[71] 仿真任务中实现了 SOTA 性能, 而 FusionAD[72] 则进一步验证了多模态融合在同时优化预测与规划任务中的有效性. 另一方面, 为缓解端到端带来的可解释性问题, ST-P3[73] 等工作引入了显式中间特征输出, 使模型在保持端到端优化的同时能够输出可理解的感知与预测结果, 从而为实车部署提供了安全验证依据.

进入 2023 年后, 端到端研究呈现出高度集成化与认知化的趋势. 其中, 最具代表性的开篇之作 UniAD[8] 设计了一种以规划为导向的统一架构, 将全栈驾驶任务 (感知、预测、规划) 统一在一个基于 Transformer[10] 的架构中, 通过查询 (query) 作为核

心交互机制, 实现跨模块的特征共享与联合策略优化, 从而减少了传统模块化系统的信息损失与误差累积, 其框架结构如图 5 所示. 该框架始于多视角图像的输入, 经 BEV 编码器生成统一的 2D 鸟瞰图特征. 而后“动态物体跟踪”和“静态地图构建”分别利用跟踪查询和地图查询提取动态物体与静态道路元素. 这些查询携带着实例级语义与历史状态流入运动预测模块, 与自车查询进行多智能体交互, 实现联合运动预测. “占用栅格预测”则利用代理特征引导场景特征, 生成未来的占用栅格. 最终, 规划部分聚合经过全流程交互的自车查询、导航指令以及占用预测, 解码出安全可行驶的规划轨迹. 同期的 VAD[74] 则开创了基于矢量化场景表征的端到端范式, 通过将路况建模为完全矢量化的图结构, 在大幅提升推理速度的同时, 利用显式的几何约束显著增强了规划的一致性与安全性. 而针对端到端模型中稠密 BEV 特征计算开销庞大的痛点, Sun 等[75] 提出 SparseDrive 框架. 该框架采用全稀疏的场景表示范式, 通过构建对称的稀疏感知头与并行的运动规划器, 在不使用高计算复杂度的 BEV 编码器的同时, 实现了感知、预测与规划任务的高效联合优化, 并在 nuScenes[27]、Bench2Drive-base[76] 等基准上取得了具有竞争力的端到端规划性能. 而在时序维度的优化上, Zhang 等[77] 进一步提出历史预测与规划范式. 该范式创新性地通过多步查询机制对历史特征进行分时复用, 即当前帧服务于感知而未来帧服务于规划, 从而高效桥接了过去与未来, 为提升端到端系统的长时域连贯性提供了全新的解决思路. 与此同时, 基础模型 (foundation model) 的爆发给端到端带来了新的机遇, 赋予系统高级语义理解与指令解析能力, 而多模态模型则实现了视觉与语言联合建模的突破, 二者共同推动端到端驾驶从“感知执行”向“通用认知”迈进. CarLLaVA[78] 探索了多模态模型在闭环驾驶中的应用, 赋予车辆语义层面的理解力. 同样致力于认知增强的 ReAL-AD[79] 则聚焦于类人因果推理能力的构建, 通过引入显式的思维链 (chain of thought, CoT)[80] 机制, 在一定程度上缓解了端到端模型可解释性薄弱的问题. 而地平线提出的 AlphaDrive[81] 则引入“监督学习+强化学习”的两阶段训练策略, 标志着端到端系统正从单纯的模仿人类向具备逻辑推理能力的类人驾驶演进.

尽管端到端自动驾驶通过联合优化感知与规划模块展现出了超越传统模块化架构的潜力, 但在实际部署中仍面临着多模态融合困难、决策过程不可解释、因果混淆以及长尾场景鲁棒性不足等关键挑战[64]. 面对这些深层次的认知瓶颈, 单纯依赖数据堆叠已难以奏效, 而引入具备强大泛化能力与语义理解优势的基础模型, 将是突破现有系统上限、迈向通用自动驾驶的核心前沿方向, 这也为后来大模

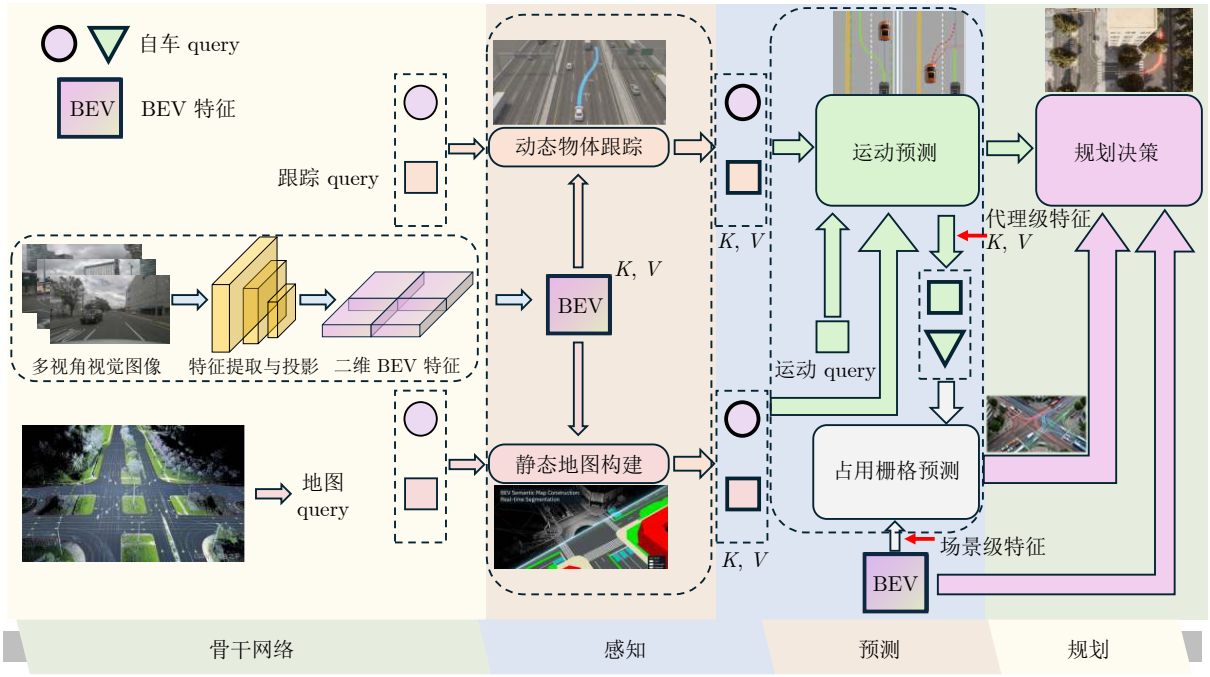


图 5　UniAD 框架概览

Fig. 5　Overview of the UniAD framework

型的应用打下了认知基础.

2.2 端到端范式的类人认知

端到端自动驾驶的目标是建立从传感器输入到控制指令的直接映射, 这一过程在本质上可类比于人类驾驶员在复杂环境中基于感知、经验与认知所作出的反应行为. 因此, 构建具有类人认知的端到端系统, 须对人类的认知结构与决策范式进行深度建模.

首先, 神经认知科学领域经典的"系统 1/系统 2"双重加工理论[82], 为设计新一代类人自动驾驶系统提供了核心的理论范式. 这一理论指出, 人类决策包含处理直觉与经验的"快系统"以及主导逻辑推演的"慢系统". 当前端到端架构正逐渐将这一分层机制引入模型设计. 以 UniAD[8] 为代表的"快系统"架构通过感知−决策的统一机制, 模拟人类在处理高频动态交互时的快速判断能力, 实现了类似于"肌肉记忆"的敏捷操控. 而针对长尾场景与复杂博弈, "慢系统"的逻辑推理则发挥了重要作用. Jia 等[83] 提出"先直觉、后修正"的类人范式 ThinkTwice, 它通过两阶段解码器首先生成粗略轨迹, 随后利用回顾机制聚焦关键时空区域进行逻辑校准, 有效提升了复杂场景下的决策鲁棒性. 而 DriveGPT4[84] 与 OccLLaMA[85] 等工作则引入了大模型, 赋予系统语义层面的理解力与因果推理能力. DriveVLM[86] 则创新性地构建了"快慢双通路"机制, 将传统端到端网络作为快通路处理常规驾驶, 而将多模态大模型作为慢通路处理突发事件并生成决策对应的解释, 显著提升了系统的可解释性.

再者, 模拟人类驾驶员对环境的内隐建模, 也是实现类人认知决策的关键. DriveDreamer4D[87] 等基于世界模型的框架, 通过构建高保真的动态场景表征, 使系统具备了在潜空间中进行"思维推演"的能力. 这种机制允许车辆在实际执行前, 对未来的多种可能性进行安全预演, 从而显著提升了其对罕见与高危场景的零样本泛化与应对能力.

最后, 针对现有模型难以维持长时序工作记忆的认知缺陷, 引入类人的记忆存取机制成为必然趋势. 现有的端到端框架往往依赖有限的历史帧堆叠, 难以处理长距离依赖. Graves 等[88] 提出的可微神经计算机 (differentiable neural computer, DNC) 为此提供了关键的理论支撑. 不同于静态规则库, DNC 允许网络通过可微的读写操作自主学习, 将关键环境特征存储至外部记忆矩阵并在决策时进行精准检索. 这种"神经网络+外部记忆"的混合计算范式, 有效解决了传统模型的遗忘问题, 为构建具备长时程推理能力的类人智能体奠定了基础.

综上所述, 端到端自动驾驶系统正朝着认知智能体演进, 逐步实现从单纯的"看懂环境"向具备直觉、逻辑与记忆的认知自动驾驶[89] 转变, 但由于其结构所导致的可解释性较弱的问题仍亟待解决.

3 大模型赋能自动驾驶

尽管端到端范式成功打通了感知与控制的特征链路, 但面对开放世界中无穷尽的长尾场景, 传统深度学习模型在逻辑推理与语义泛化上仍显乏力. 与此同时, 以 GPT[11] 为代表的基础模型的涌现, 标志着人工智能进入了大模型时代. 特别是多模态大语言模型的迅猛发展, 实现了跨模态知识的深度对齐与迁移. 通过引入大语言模型的推理能力、视觉−语言模型的多模态理解、视觉−语言−动作模型的闭环执行, 自动驾驶车辆被赋予了前所未有的认知大脑, 迅速从"感知智能"向"认知智能"跃迁. 本节将以上述结构梳理大模型的发展及其为自动驾驶带来的机遇与挑战.

3.1 大模型的发展历程

人工智能的演进是一个从规则驱动向数据驱动不断深化的过程, 而大模型的诞生则标志着这一进程迈入了以通用认知为核心的新阶段. 追溯其起源, 2006 年 Hinton 等[90] 提出的基于深度信念网络 (deep belief network, DBN) 的开创性模型重启了深度学习热潮, 随后 ImageNet 数据集[91] 的发布确立了"大数据+大算力+深度网络"的工程范式. 而真正为大模型时代奠定基石的是 2017 年 Google 提出的 Transformer 架构[10], 该架构彻底摒弃了循环与卷积结构的归纳偏置, 利用自注意力机制实现了对长距离依赖的高效并行建模, 成为后续大部分主流模型的骨干网络.

在此基础上, LLM 率先实现了突破. LLM 通常是指包含数千亿或更多参数的基于 Transformer 架构的语言模型. OpenAI 于 2018 年发布的 GPT-1[92] 模型与 Google 发布的 BERT[93] 模型分别探索了自回归与自编码两种预训练路径, 使用海量的大规模无标注文本数据进行训练, 学习通用的语言、逻辑和知识, 验证了"无监督预训练+有监督微调"范式在自然语言处理 (natural language processing, NLP) 任务上的普适性. 随后, OpenAI 坚持规模定律[94], 相继发布 GPT-2 与 GPT-3[11], 将参数规模推至千亿量级, 首次展现了模型在零样本和少样本场景下的学习与涌现能力. 除此之外, LLaMA[95]、Galaactica[96] 与 PaLM[97] 等也是大规模预训练模型的

代表作. 而 2022 年底 ChatGPT 的问世更是引起了全球对“对话式 AI”的关注, 随后 GPT-4[98] 则进一步引入基于人类反馈的强化学习 (reinforcement learning from human feedback, RLHF)[99], 使模型具备了类人的逻辑推理与指令遵循能力, 标志着 LLM 已初步具备通用人工智能的雏形.

然而, 面临高质量文本数据有限和单模态无法完全理解现实世界的挑战, 研究重心迅速向多模态大模型转移, 旨在打破自然语言文本与物理世界的感知壁垒. 2021 年以 CLIP[100] 为代表的视觉语言预训练模型的发布首次打通了文本与图像的语义空间, 实现了图像与文本语义空间的统一. 在此基础上, 为赋予模型基于视觉上下文生成文本与少样本学习的能力, DeepMind 推出的 Flamingo[101] 引入交叉注意力 (cross-attention) 机制, 成为视觉和语言融合架构设计的里程碑. 随后, BLIP-2[102] 设计的 Q-Former 结构进一步优化了视觉特征与语言模型的对齐策略, 极大降低了训练成本, 推动了多模态大模型从专用向通用的跃迁. 此后, 以 GPT-4V[103]、Gemini[104] 以及开源的 LLaVA[105] 为代表的 VLM, 通过基于视觉编码器与 LLM 的深度融合, 赋予了模型强大的图像理解与视觉推理的能力. 然而, VLM 在迈向实际应用时仍面临诸多挑战, 主要包括由幻觉引发的可靠性风险、缺乏对物理世界常识的深度对齐以及高质量多模态数据稀缺的瓶颈.

随后, 智能体技术由早期面向封闭环境的强化学习智能体, 向具备开放世界认知能力的具身智能体演进. 具身智能是人工智能与机器人学的深度融合, 旨在构建具备可感知、可思考、可行动能力的物理智能体, 其应用场景非常广泛, 在商业服务、农业、自动驾驶等各个领域均有所涉及, 如图 6 所示. 具身智能所强调的“感知–决策–执行”的闭环联动, 要求智能体在多模态环境中准确理解语义并输出物理动作, 这一核心逻辑与 VLA 模型的设计初衷高度同构. 正是为了满足这种从“多模态认知”到“物理执行”的深度统一需求, VLA 范式应运而生, 该范式将连续的物理动作空间离散化, 从而把复杂的机器人控制转化为序列建模问题, 成为具身智能在大模型时代的最佳载体. 如 RT-2 模型[106] 是一个典型的 VLA 模型, 它将机器人的物理动作编码为语言 Token, 直接输出控制信号; 而 Gato[107] 更是将图像块、文本字符与离散动作统一编码为同一语义空间下的 Token 序列, 实现了多任务之间的可迁移, 但同时也面临着可解释性较弱、数据需求庞大等挑战. 针对上述模型闭源或难以在消费级硬件上部署的局限, OpenVLA[108] 进一步推动了 VLA 的大众化. 该模型构建于 LLaMA 2[109] 强大的视觉–语言底座之上, 验证了参数高效微调, 如低秩适配 (low-rank adaptation, LoRA)[110] 及其优化版本 QLoRA[111] 在具身控制中的有效性, 显著降低了 VLA 模型的定制与运行门槛, 使得在消费级 GPU 上定制高性能 VLA 策略成为可能, 标志着具身智能进入了开源生态协同发展的新阶段. 可以说, 多模态大模型通过具身智能实现具身化, 而具身智能借助多模态大模型提升智能化, 二者是相辅相成的促进关系.



(a) Google AutoRT 大规模机器人集群编排系统
(a) Google AutoRT large-scale robotic cluster orchestration system



(b) Keenon Robotics 具身服务机器人 XMAN-R1
(b) Keenon Robotics embodied service robot XMAN-R1



(c) WeRide 无人出租
(c) WeRide robotaxi



(d) JITRI 农业采摘机器人
(d) JITRI agricultural harvesting robot

图 6　具身智能部分应用场景

Fig. 6　Case studies on application scenarios of embodied intelligence

纵观大模型的发展历程, 其技术演进呈现出一条由语言智能到多模态感知再到具身控制的清晰脉络. 具体而言, LLM 率先赋予了机器卓越的常识推理与泛化能力, 奠定了认知层面的基座; 随后, VLM 的涌现打破了语义与视觉表征之间的感知壁垒, 实现了语言与视觉的跨模态对齐; 而 VLA 则进一步打通了从高层认知到底层物理执行的闭环, 使智能体具备了理解并干预现实世界的能力. 这种跨模态的通用认知能力, 正在为自动驾驶等垂直领域的智能化转型注入核心驱动力[112], 推动其从感知理解向决策行动的闭环智能加速演进.

3.2　大模型在自动驾驶领域的应用

受益于深度学习技术的突破, 自动驾驶技术已取得长足进展, 但面对复杂的开放动态场景, 传统的基于模块化或端到端的系统仍难以有效应对长尾分布的挑战. 近年来, 大模型凭借其卓越的语义理解、逻辑推理与泛化能力, 为突破上述瓶颈提供了全新的范式. 大模型正在重塑自动驾驶从感知理解到决策规划的流程[113], 尤其是深度融合了视觉、语

言等多模态信息的大模型. 本节将分别从 LLM 的逻辑推理、VLM 的多模态感知、VLA 模型的端到端具身控制 3 个维度, 系统阐述大模型在自动驾驶领域的应用现状与演进趋势.

3.2.1 大语言模型

大语言模型的爆发式增长正在推动自动驾驶技术从传统的数据驱动向知识驱动演进. 传统的深度学习模型虽然在感知任务上表现优异, 但在面对训练集之外的长尾场景时往往束手无策, 且缺乏可解释性, 无法给出符合人类认知的决策解释. 相比之下, LLM 凭借其在海量文本数据中习得的世界知识与常识推理能力, 能够理解复杂的交通交互逻辑与未曾见过的长尾场景. 这种能力的引入不仅打破了传统自动驾驶模型的"黑盒"壁垒, 增加了可解释的思维链过程, 更使得自动驾驶车辆具备了类人的泛化推理能力.

在多模态大模型普及之前, 研究者首先探索了基于纯文本的 LLM 在自动驾驶决策与规划中的潜力. 这类方法的核心思路是"环境文本化", 即利用上游的感知模块将复杂的视觉场景抽象为结构化的文本描述, 如目标坐标、速度、车道拓扑等, 进而利用 LLM 强大的逻辑推理能力生成驾驶决策或代码等具体的控制内容.

在初期探索阶段, Fu 等[114] 提出的"Drive like a human"框架将 LLM 定义为自动驾驶的认知中枢, 利用 CoT 推理使系统能够在面对长尾场景时像人类司机一样运用常识进行解释与决策. 随后, Mao 等[115] 提出的 GPT-Driver 在进一步将运动规划重构为语言建模问题的基础上, 引入一种不同于普通提示方法[80, 99, 116] 的"提示–推理–微调"策略, 成功激发了 GPT-3.5 强大的逻辑推理与上下文理解能力. 而 Wen 等[117] 提出的 DiLu 则是首个以 LLM 为"认知核心"与"决策大脑"的模仿人类学习范式的自动驾驶框架, 其结构如图 7 所示. 它创新性地设计一个经验存储模块来存储历史驾驶经验, 使智能体能够通过检索过往类似场景的决策逻辑来优化当前行为, 从而在不进行参数微调的情况下实现类似人类驾驶员的持续学习与反思的能力. 与此同时, 为弥补 LLM 在连续数值控制上的不足, Sha 等[118] 提出的 LanguageMPC 采取一种新颖的分层决策架构, 将高层语义推理与底层运动控制相分离. 该方法利用 LLM 作为认知中枢, 深入分析复杂路况并进行宏观决策, 如判断超车时机或是否让行, 进而动态调节底层模型预测控制的权重参数与偏置矩阵, 使控制器能灵活适应不同驾驶场景. 这种设计有效地结合了 LLM 的逻辑推理优势与传统控制算法的动力学精准性, 从根本上解决了直接利用 LLM 输出底层控制指令所导致的不稳定等问题, 显著提升了系统在真实环境中的鲁棒性与安全性.

除了决策规划, LLM 在人机交互与指令理解方面也展现出巨大潜力. 现有的自驾系统往往仅能执行预定义的僵化指令, 难以满足用户的个性化需求. 针对这一问题, Liao 等[119] 提出一种基于 GPT-4 增强的多模态定位框架. 该研究利用 GPT-4 作为语义分析引擎, 精准捕捉自然语言命令中细腻的语义特征与乘客意图, 并结合跨模态注意力机制将抽象

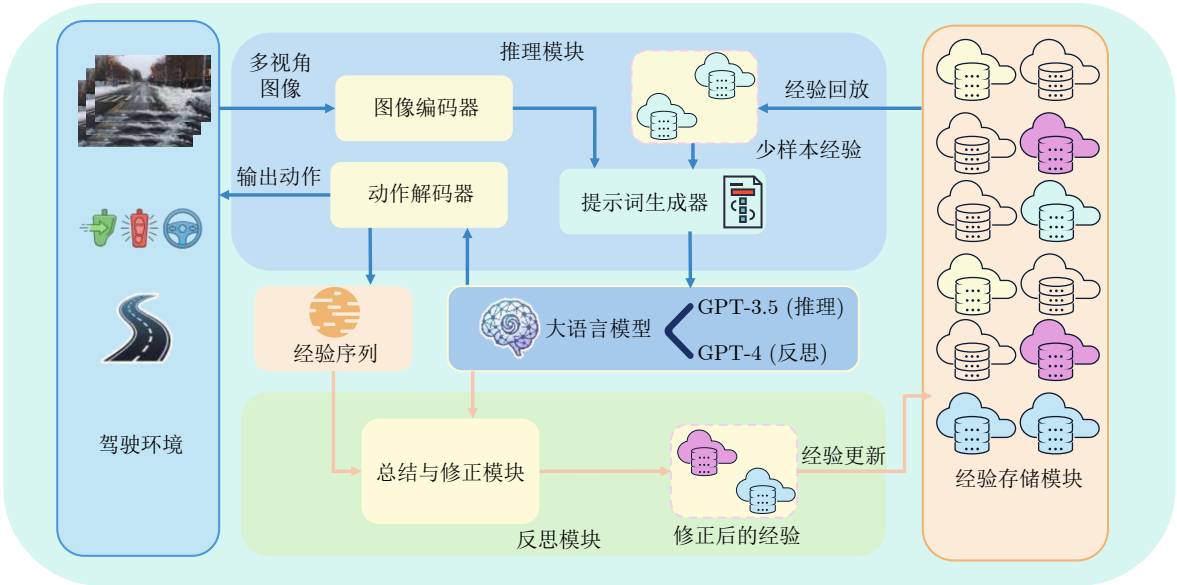


图 7 基于记忆检索与持续反思机制的 DiLu 框架

Fig. 7 DiLu framework based on memory retrieval and continuous reflection mechanism

的语言指令映射至具体的物理道路区域. 配合其设计的区域特定动态层注意力解码器, 该方法能够帮助车辆在复杂的交通场景中高效锁定符合指令的关键区域, 从而实现了一种自然流畅的“与车对话”的人机交互新范式. 在此基础上, 为进一步提升交互的深度与个性化体验, Cui 等[120] 提出的 Talk2Drive 框架将关注点从单纯的指令理解拓展到了对驾驶风格的个性化适配. 该系统利用 LLM 强大的语义推理能力, 不仅能解析诸如“我赶时间”这类隐含的高层语义指令并自动调整为相对激进一些的跟车策略, 还引入了基于数据库的记忆模块. 这使得车辆能够记录并检索用户的历史反馈与偏好, 从而在长期的交互中不断自我进化, 成为真正懂用户驾驶风格的系统.

3.2.2　视觉–语言模型

然而, 单纯基于文本的 LLM 高度依赖上游感知模块将视觉信息转化为文本, 这一过程不可避免地导致了环境细节的丢失. 自动驾驶本质上是一个高度依赖视觉感知的任务, 即便纯文本大模型拥有强大的逻辑推理能力, 也难以直接理解复杂的交通场景. 鉴于此, 融合视觉感知能力的 VLM 打破了模态间的壁垒, 旨在实现对物理世界的端到端感知与理解, 并成为连接底层视觉特征与高层语义认知的关键桥梁. 通过多模态融合的输入, 它能够更全面地获取环境信息, 增强自动驾驶系统的环境理解能力. 例如, 将文本描述作为提示输入, 可以帮助模型更准确地解读复杂的交通标识、模糊的地面标线以及临时的交通管制信息, 从而显著提升系统对复杂场景的环境理解能力. 同时, 多模态融合也有助于提升人机交互能力, 使系统能够与驾驶员和乘客进行更自然、流畅的交流. 此外, 这种融合还能为决策过程提供更丰富的依据, 从而提高决策的准确性和可解释性. Zhou 等[121] 指出, VLM 在自动驾驶中的应用已不再局限于传统的开放词汇检测, 而是正在向场景理解、长尾数据生成以及语义规划等深层任务渗透, 为实现从感知到规划的认知闭环提供了新的技术路径.

VLM 的核心在于通过视觉编码器将图像对齐到语言空间. 2021 年, 作为多模态大模型发展的基石, CLIP[100] 模型利用 4 亿对图文数据进行对比学习, 打破了视觉与语言的语义壁垒, 成为后续主流多模态模型中不可或缺的视觉编码器. 在此基础上, Liu 等[105] 提出的 LLaVA 确立了“视觉指令微调”的标准范式, 即通过投影层将 CLIP 提取的视觉特征映射到 LLM 的输入空间, 赋予了模型极强的图像理解与对话能力. 这一高效架构迅速被研究领域采纳, 成为构建驾驶专用多模态模型, 如 DriveGPT4[84] 等的首选网络.

在感知层面, 通用的多模态基础模型展现了极高的零样本迁移能力. GPT-4V 能够精准识别交通标志、信号灯, 并能推断其他交通参与者的意图. 为实现更精细的目标检测, Kirillov 等[122] 提出的 SAM (segment anything model) 与 Grounding-DINO[123] 相结合, 构建了统一的分割与跟踪框架, 利用文本提示即可在视频流中实现对车辆与行人的开放集检测与分割.

然而, 尽管通用 VLM 在自然图像理解上表现优异, 但直接将其迁移至对安全要求极高的自动驾驶场景时, 仍面临着细粒度感知缺失与空间推理匮乏这两大核心瓶颈. 首先是分辨率限制导致的小目标丢失问题. 针对此, Ding 等[124] 提出一种融合多尺度高分辨率细节的增强框架 HiLM-D, 并指出传统 VLM 往往将图像压缩至低分辨率, 导致远处的红绿灯或车辆信息丢失, 而通过引入高分辨率视觉分支, 在风险物体定位与意图预测任务中显著优于低分辨率基线, 证明了高保真视觉信息对驾驶安全的重要性. 同时, 除二维平面的像素级感知外, 自动驾驶更需要理解复杂的三维空间结构. 为赋予 VLM 空间感知与推理能力, Talk2BEV[125] 设计一种基于鸟瞰图的场景表示法. 它先生成包含语义信息的 BEV 图, 再利用 VLM 执行视觉问答 (visual question answering, VQA) 任务, 从而实现了对车辆空间关系的深度推理. 类似地, LiDAR-LLM[126] 通过独特的视角感知变换器将 3D 点云数据与文本嵌入对齐, 使模型能够直接理解 3D 环境. 针对 3D 增强的成本问题, Lübberstedt 等[127] 提出的 V3LMA 展示了一种免训练策略, 通过将 3D 目标检测器提取的结构化文本描述注入大模型, 在减少成本的情况下显著提升了空间问答的准确率.

VLM 的另一大优势在于其通过自然语言解释驾驶决策的能力, 这对于提升自动驾驶系统的透明度至关重要. Atakishiyev 等[128] 的研究证实了 VLM 能够基于道路几何与障碍物清晰度解释驾驶行为. 在此基础上, Xu 等[84] 提出的 DriveGPT4 是利用指令微调后的 VLM 统一实现可解释决策与端到端控制的自动驾驶框架之一, 为该范式的发展奠定了基础. 该工作构建了一个专门的视觉指令微调数据集, 利用 ChatGPT 生成大量基于 BDD-X 数据集[129] 的问答对, 旨在描述驾驶场景中的车辆动作、解释其背后的决策逻辑, 并最终预测所需的控制信号. 其中包含车辆动作描述、决策解释及控制信号预测. 通过处理多帧视频输入, DriveGPT4 不仅能

够以端到端的方式输出车辆的底层控制信号, 还能通过自然语言与用户交互, 解释其决策背后的逻辑, 实现了感知、决策与解释的统一. 而在随后, 针对 VLM 在空间感知精度与推理延迟上的局限性, Tian 等[86] 提出了 DriveVLM. 不同于直接输出控制信号, DriveVLM 引入了受人类认知启发的 CoT 机制, 将驾驶任务解耦为场景描述、场景分析与分层规划 3 个阶段, 特别强化了对长尾场景的理解能力. 更重要的是, 为解决大模型推理速度慢且缺乏精确 3D 空间推理的问题, 该文献提出了 DriveVLM-Dual 混合系统. 该系统采用“快慢系统”设计理念, 将 DriveVLM 的高层语义规划 (慢系统) 与传统自动驾驶的高频轨迹规划 (快系统) 相结合, 通过融合 3D 感知结果与高频轨迹优化, 成功实现了 VLM 在实车上的实时部署与高效运行. 进一步地, 针对紧急避险场景, 如前车急刹等, 该类混合架构须设计严格的动态仲裁与安全兜底机制. 由于大模型的单次推理通常存在数百毫秒的不可控延迟, 系统在底层逻辑上需要采用异步时钟运行机制, 即“快系统”保持高频运转, 绝不阻塞等待“慢系统”的最新决策. 当瞬时危险发生而慢系统推理尚未完成时, 仲裁机制会立即触发越权接管, 由快系统依托极低延迟的 3D 几何感知与硬性物理碰撞约束, 如为自动驾驶系统定义了可数学验证的纵向与横向安全距离公式的责任敏感安全 (responsibility sensitive safety, RSS) 模型[130], 直接执行紧急制动或局部避让策略. 这种“慢系统决定认知上限、快系统守住安全下限”的非对称仲裁机制, 能够有效防止因大模型推理滞后而导致关键决策窗口关闭的风险. 为进一步弥合大模型语言空间与车辆底层控制空间之间的鸿沟, Cui 等[131] 提出了 DriveMLM. 不同于 DriveGPT4 的端到端预测或 DriveVLM 的轨迹生成, DriveMLM 将驾驶任务建模为对行为规划状态 (如借道、减速、跟随等) 的预测, 通过将 VLM 的语义决策与成熟的工业级规划控制模块 (如 Apollo) 相对接, 既保留了大模型处理复杂长尾指令的优势, 又确保了闭环控制的安全性与稳定性, 证明了语言接口的兼容性价值. 这一系列工作共同推动了 VLM 从单纯的感知问答向具备决策规划能力的驾驶认知大模型演进.

在实现场景理解的基础上, 如何利用自然语言主动指导车辆航行成为新的研究方向. 进一步地, Shao 等[132] 提出的闭环端到端自动驾驶框架 LMDrive 将 VLM 的语义推理能力从被动的问答拓展到了主动的运动规划领域. 该框架能够根据用户的自然语言指令, 如“在前方路口左转”, 并结合视觉输入, 生成高层的导航路点与具体的运动轨迹, 实现了从场景理解到轨迹生成的过程, 并在 CARLA[71] 仿真环境中完成了闭环测试, 验证了语言指令直接指导端到端驾驶的可行性, 也为后续实现感知到控制闭环的 VLA 模型的发展提供了宝贵的经验.

尽管 VLM 在场景理解与高层规划上表现卓越, 但其输出通常仍停留在离散的语言描述或元动作层面, 在执行高频、连续的底层控制时仍面临精度与稳定性的挑战. 这也促使学术界进一步探索将视觉感知、语言推理与物理动作直接统一的 VLA 模型.

3.2.3 视觉–语言–动作模型

VLA 模型在自动驾驶领域的兴起, 很大程度上受益于通用具身智能的发展, 主要体现在“语义理解–轨迹生成–行为执行”的一体化上. 作为首个通才智能体, DeepMind 提出的 Gato[107] 具有奠基性意义. 它采用“序列化一切”的策略, 将图像、文本及机械臂动作统一编码为 Token 序列, 证明了单一 Transformer 架构处理跨模态控制任务的可行性. 随后, VIMA[133] 引入“多模态提示”机制, 支持图文交织的输入形式, 显著提升了模型对复杂指令的泛化能力. 这些工作为自动驾驶实现“感知–决策–控制”一体化提供了核心方法论参考.

在 VLA 发展的初期, 研究者倾向于采用模块化架构, 将语言作为连接感知与控制的中间件. 在这一阶段, VLM 不再是被动的解释器, 而是主动的规划者. OpenDriveVLA[134] 是这一阶段的代表, 它将视觉与文本指令融合, 生成中间路径点, 如“20 米后右转”供下游控制器执行. 为处理更复杂的场景, DriveMoE[135] 采用混合专家 (mixture-of-experts, MoE) 架构, 根据语言线索动态激活“超车”或“启停”等子模块. 针对安全性问题, SafeAuto[136] 引入符号化交通规则来验证语言驱动的规划, 而 RAG-Driver[137] 则通过检索历史驾驶案例来增强长尾场景下的决策能力. 此外, CoVLA-Agent[138] 和 Lang-Coop[139] 也分别在动作映射和多车协作方面进行了探索. 尽管模块化设计缩小了语义鸿沟, 但其多阶段流程不可避免地引入了误差累积与推理延迟.

受益于多模态基础模型的发展以及自动驾驶端到端范式的启发, 构建完全统一的自动驾驶端到端 VLA 模型成为主流趋势. 这类模型在一个单一的前向传播中, 直接将多模态输入映射为轨迹或控制信号, 并通过反向传播使得模型能够在同一表征空间中进行联合优化. DriveGPT4-V2[140] 是从解释向控制跨越的典型, 它通过统一的架构将视觉输入、历史控制与语言指令映射到语言模型空间, 实现了在

同一模型中既能回答关于驾驶决策的问题, 又能直接输出底层控制信号, 完成了从“驾驶解说”到“驾驶操控”的进化. OpenEMMA[141] 作为开源先驱, 在 nuScenes 数据集上证明了基于思维链的端到端规划在复杂真实场景中的有效性. 为解决泛化与对齐难题, SimLingo[142] 提出“动作梦境 (action dreaming)”技术. 该技术通过在相同视觉状态下模拟多种可能的未来轨迹并与不同语言指令进行配对, 强制模型将语言指令作为决策的核心依据, 从而实现语言与动作的强对齐. Drive anywhere[143] 则利用基于文本的隐空间模拟, 如将晴天替换为雪天, 实现了仿真到现实的零样本迁移.

而近期的 VLA 研究浪潮开始将重心转向推理增强能力, 试图解决大模型“虽然会开, 但不能解释为何这么开”的问题. ORION[144] 指出了语义空间与数值动作空间之间的鸿沟, 设计一个混合架构: 利用 QT-Former 融合长期的上下文信息, 生成显式指令, 实现了复杂指令与多视角场景的跨模态语义理解, 并能够据此输出高精度轨迹. 而 Chi 等[145] 提出的 Impromptu VLA 构建了一个专注于挑战性场景的大规模、高质量的 VLA 数据集. 该数据集从 8 个开源数据集提炼出约 8 万个视频片段, 专门覆盖道路边界不清等四类长尾场景, 并通过大模型与人工相结合的方式, 提供了丰富的规划问答与动作轨迹标注. 该工作通过实验证明, 基于此数据集微调的模型会“先思考后行动”, 在感知、推理与规划能力上均获得显著提升. AutoVLA[146] 和 Drive-R1[147] 则进一步融合了自适应推理与强化学习, 使模型在闭环测试中展现出顶尖的规划能力. 为直观阐述此类推理增强型 VLA 模型的工作机理, 图 8 展示了当前自动驾驶 VLA 模型经典的两阶段训练范式. 整个流程分为“监督微调 (supervised fine-tuning, SFT)”与“后训练”两个阶段. 在第一阶段, 模型通过多视角图像、专家轨迹与包含推理逻辑的问答对的监督微调, 建立起视觉感知与动作输出的基础映射, 目的是实现模态对齐并赋予模型基础的驾驶与推理能力. 在更关键的第二阶段, 为突破模仿学习的上限, 模型冻结了视觉编码器与动作头的参数, 将 LLM 置于强化学习闭环中. 通过与在线环境的交互以及奖励模型的反馈, LLM 能够不断试错并优化其决策逻辑, 从而实现向具备深层推理能力的“具身智能体”跃迁. 而 FutureSightDrive[148] 则通过时空思维链增强了视觉推理, 使 VLA 模型能够以视觉方式思考未来场景. 这一阶段的模型标志着 VLA 从单纯的条件反射向具备深层认知的决策核心演进.

VLA 模型的引入标志着具身智能理念在自动驾驶领域的深度渗透, 它打破了高层语义认知与底层物理控制之间的异构壁垒, 推动自动驾驶系统向具备“感知–决策–执行”一体化闭环能力的具身智能体跃迁. 然而, 尽管 VLA 取得了显著进展, 但将庞大的多模态大模型部署在资源受限的车端仍面临巨大挑战, 主要体现在实时性与计算开销上. 为此, LightVLA[149] 设计一种可微 Token 剪枝机制, 利用动态自适应评估视觉 Token 的重要性, 在不损失性能的前提下显著降低了推理延迟. 此外, 早期的探索工作如 DriveGPT4[84]、TS-VLM[150] 和 DynRsl-VLM[151] 也在不同方向上推动了 VLA 在自动驾驶

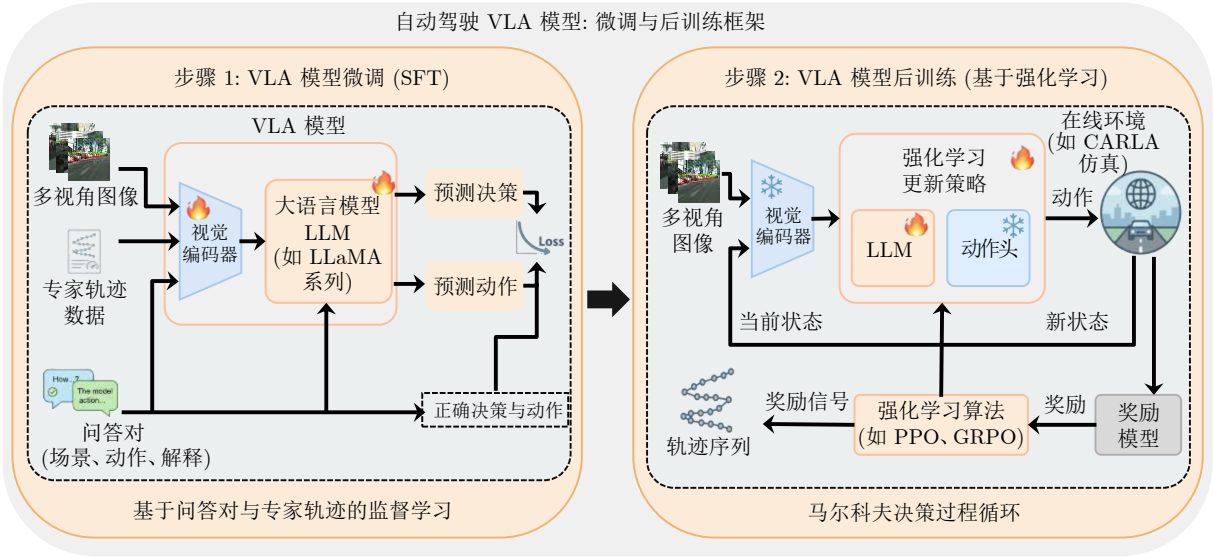


图 8　融合监督微调与强化学习的自动驾驶 VLA 训练框架

Fig. 8　An autonomous driving VLA training framework integrating SFT and RL

中的应用. 虽然它们主要侧重于感知增强与系统解释性, 但也为后续研究提供了重要启示. 例如, TS-VLM 通过文本引导的软排序池化优化了多视图注意力融合, DynRsl-VLM 则通过动态分辨率处理提升了细粒度特征对齐能力, 这些都为 VLA 在复杂驾驶场景中的感知与推理能力奠定了基础. 未来, 如何进一步解决鲁棒性验证、多模态时空对齐以及社会化交互等挑战, 将是 VLA 迈向大规模落地的关键.

4 世界模型的兴起与应用

近年来伴随着生成式人工智能 (generative AI) 的蓬勃发展[152], 研究视野逐渐从静态的图文生成向动态的时空推演延伸. 在 LLM、VLM、VLA 等大模型高速发展的现在, 为赋予模型在生成任务中对物理世界因果规律的深刻理解与精准模拟能力, 世界模型[153] 正成为通往通用人工智能的新技术高地. 为进一步疏通这些模型与自动驾驶结合的技术脉络, 表 1 梳理了 LLM、VLM、VLA 模型及 WM 这四类模型的核心优势、主要局限及其在自动驾驶领域的代表性工作.

4.1 时空表征与场景生成

世界模型的本质是让机器具备对物理世界的内部表征和模拟能力, 主要包括理解世界和预测未来两大视角, 即构建内部表示以理解世界机制, 并预测未来状态以模拟和指导决策[160]. 这类模型不再局限于预测下一个 Token, 而是致力于在潜空间中构建物理世界的高保真时空表征, 具备生成连贯视频与推演未来的能力. 例如, Google DeepMind 提出的 Genie[161] 证明了一个模型可以仅从互联网视频中无监督地学习出潜在的动作空间和环境动态, 并能根据一张图片和动作指令生成全新的、可交互的虚拟环境. 这项研究为实现一个通用的、能够模拟多样化物理和视觉规则的世界模型迈出了关键一步. 与此同时, Meta 推出的 V-JEPA[162] 则摒弃了像素级的生成预测, 转而在抽象的特征空间中进行掩码预测, 证明了模型能够在不重建每一帧细节的情况下, 高效习得物理世界的时空结构与因果关系.

区别于通用的视频生成模型, 如 Sora[163] 等, 自动驾驶领域的世界模型更侧重于对道路场景的结构化理解与物理约束遵循, 通过扩散模型生成仿真感知数据, 提升系统对极端场景的处理能力. 它将多视角相机图像、激光雷达点云及高精地图等异构感知数据映射至紧凑的潜在场景空间, 不仅实现了感知与预测的一体化融合, 更在潜空间中学习了可微分的动态状态转移机制. 这种机制能够捕捉交通场景的物理演化规律, 在严格遵循几何与动力学约束的前提下, 对多智能体的未来状态进行长时程推演, 从而为端到端自动驾驶的可解释性、闭环规划及长尾场景泛化提供了全新的理论范式. 世界模型的基础在于如何高效表征并生成复杂的驾驶场景. 随着

表 1 自动驾驶领域大模型与世界模型应用的对比与分析

Table 1 Comparison and analysis of large models and world model for autonomous driving applications

模型类型	核心优势	主要局限	代表性工作
LLM	核心优势在于其通过海量文本训练获得的、高度抽象的世界知识与常识推理能力, 这在早期为自动驾驶系统注入了“通用智能”的雏形	模型过于依赖文本描述作为环境输入, 对视觉信息理解较弱, 存在“语言世界”与“物理世界”的割裂, 无法直接、精准、安全地操控物理实体	Drive like a human[114] GPT-Driver[115] DiLu[117] Talk2Drive[120] LanguageMPC[118]
VLM	实现了视觉与文本信息的对齐, 使自动驾驶系统拥有了强大的“感知–理解”能力, 实现了对环境的直观、多模态、语义级理解, 是连接感知与认知的关键桥梁	虽然它实现了视觉与语义的跨模态对齐, 但是它主要是一个观察者和评论者, 而非一个决策者与执行者. 它缺乏直接与物理环境进行具身交互和闭环控制的能力	DriveGPT4[84] DriveVLM[86] DriveMLM[131] LMDrive[132] Talk2BEV[125]
VLA	它是迈向通用具身智能的关键一步, 通过构建从多模态感知到车辆控制的端到端映射, 打破了高层语义认知与底层物理控制之间的异构壁垒, 推动自动驾驶系统向具备感知–决策–执行一体化闭环能力的具身智能体跃迁	在数据、安全、可解释性、算力等方面依旧面临着巨大的挑战, 距离大规模、高可靠性的商用部署仍有距离, 需在系统架构、验证方法与基础设施上取得关键突破	LangCoop[139] OpenEMMA[141] AutoVLA[146] OpenDriveVLA[134] OmniDrive[154] CoVLA-Agent[138]
WM	被认为是实现 L4 级及以上全自动驾驶的潜在关键技术路径之一, 其重要性不仅源于技术架构的革新, 更在于它直接指向并试图复现人类驾驶行为的深层认知本质——对复杂交通环境未来演变的动态预判以及在此基础上的多步推理与想象性规划	目前主要处于前沿研究和大规模数据训练阶段, 仍存在高质量多模态数据稀缺与鲁棒性较低等问题, 并在可靠性、实时响应能力以及安全等方面尚未通过严格的车规级验证, 仍需在理论框架、算法设计、系统集成与标准化评价体系等方面完成关键突破	GAIA-2[155] SimGen[156] DrivingDiffusion[157] DriveWM[158] DriveDreamer[159] DriveDreamer4D[87]

技术演进, 相关研究正从二维视觉向三维空间及多模态融合的方向发展.

早期的生成式世界模型主要基于二维图像平面, 利用自回归模型或扩散模型预测未来的视觉观测. 例如, GAIA-1[164] 能够整合视频、文本和动作信息, 并将异构输入映射为统一的离散令牌序列, 通过自回归预测生成高度逼真的未来驾驶场景视频, 不仅精准捕捉了场景的几何结构与物理动态, 更在未见过的复杂长尾场景中展现了卓越的泛化能力. 类似地, DriveDreamer[159] 利用扩散模型, 通过文本提示和结构化交通信息引导场景生成, 不仅能生成高质量的驾驶视频, 还能作为神经模拟器训练感知模型. DrivingDiffusion[157] 则进一步引入多视角一致性约束, 通过跨视角注意力来解决自动驾驶领域多视角视频数据生成的难题.

尽管上述二维生成模型能够产生逼真的视觉效果, 但它们往往忽视了场景内在的几何一致性, 导致生成的长时序视频容易出现物体形变或结构崩塌. 为弥补这一缺陷, Liang 等[165] 提出的 UniFuture 创新性地引入了显式的深度几何约束. 该模型构建了一个图像与深度共享的统一世界模型, 通过“双潜空间共享”与“多尺度潜交互”机制, 实现了 RGB 外观特征与深度几何特征的双向修正. 这使得模型在生成高保真未来画面的同时, 能够保持严格的深度结构时空连贯性, 为从纯二维视频生成向具备几何感知的世界模型演进构建了重要桥梁.

在此基础上, 为实现对三维空间的精细化表征, 学术界开始进一步探索基于俯视视角或三维体素的生成模型. 作为先驱工作, MILE[166] 率先采用基于模型的模仿学习思路, 在低维的 BEV 潜在空间中联合建模环境的动态演化与自车的动作执行, 使得智能体能够摆脱物理环境的束缚, 在隐式空间中进行长时程的未来场景想象与轨迹规划, 显著提升了应对未见工况的泛化能力. 然而, BEV 表征在高维度上仍存在信息压缩与丢失. 为进一步提升表征精度, OccWorld[167] 不再依赖 2D 表征, 而是采用生成式架构直接预测三维占用栅格的时序演化, 从而在 3D 空间中对场景动态进行建模, 显著提升了对动态障碍物的理解能力. 随后, 为弥补传统 LLM 缺乏 3D 空间感知能力的短板, Wei 等[85] 进一步提出 OccLLaMA. 该研究设计了一种高效的 3D 场景 Tokenizer, 将稀疏且类别不平衡的语义占据数据离散化为类语言的 Token 序列. 通过构建统一的多模态词表, OccLLaMA 利用预训练的 LLaMA[95] 模型在一个统一的 Transformer 架构中实现了 4D 场景演变预测、运动规划及视觉问答. 这种对物理世界的高保真推演能力, 不仅提升了感知的鲁棒性, 更促使端到端自动驾驶从单纯的轨迹回归向一种全新的范式演变——引入世界模型作为动态环境模拟器, 构建了“感知–模拟–规划”的闭环体系.

4.2　预测性训练与决策

在这一新范式下, “在想象中规划”成为核心研究点. 在端到端开创性工作 UniAD[8] 的基础上, GenAD[168] 首次在该统一框架内采用生成式方法, 利用变分自编码器对轨迹分布进行建模, 以生成的方式同步处理预测与规划任务, 图 9 为该端到端生成式框架的示意图.

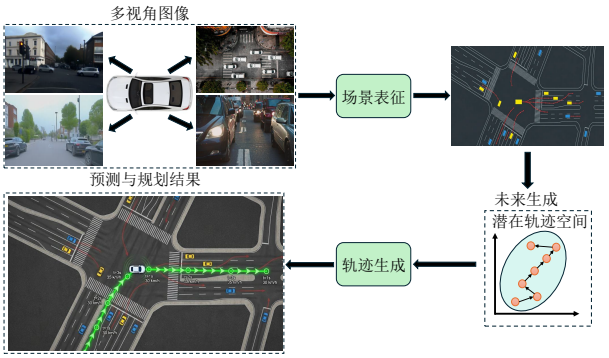


图 9　基于生成式端到端范式的 GenAD 自动驾驶框架

Fig. 9　GenAD autonomous driving framework based on generative end-to-end paradigm

进一步地, 为实现更主动的风险规避, Wang 等[158] 提出了 DriveWM, 一个兼容现有端到端规划器的多视角世界模型. 它能够为规划器采样的不同候选轨迹生成对应的高保真多视角未来视频, 并通过基于图像的奖励函数评估每条轨迹的未来场景优劣, 辅助规划器选择最优行驶路径. 而作为将具身智能与环境交互深度融合的典型代表, MindDrive[169] 则针对传统模型缺乏闭环纠错能力的痛点, 构建了基于在线强化学习 (online RL) 的训练范式. 该研究利用仿真器充当外部世界模型提供即时奖励反馈, 指导 LLM 在持续的“试错”中优化策略, 有力证明了与世界模型的高频交互是生成式智能体从开环预测走向闭环控制的关键一环.

4.3　高保真仿真的应用

然而在测试验证环节, 实车路测的高昂成本、人力投入与安全风险始终是制约自动驾驶规模化落地的瓶颈. 针对这一痛点, 世界模型在离线数据生成与闭环仿真中的应用, 旨在打破传统仿真平台的“Sim-to-Real”鸿沟, 构建高保真的数字孪生世界[170]. 作为该领域的里程碑式工作, Yang 等[171] 提出的

UniSim 摒弃了传统游戏引擎的建模路径, 转而利用神经辐射场技术从真实驾驶日志中重建可交互的动态场景. 本质上, UniSim 构建了一个具备物理一致性的神经模拟器, 它不仅支持对历史场景进行“反事实”编辑 (如动态插入变道车辆、移除或调整交通参与者行为等) 以生成边缘测试用例, 还能让自动驾驶系统在其中执行动作、感知反馈并预测环境演变, 实现真正意义上的闭环交互与状态推演. UniSim 的成功落地标志着工业界研发范式的重要转型——从单一的模型结构优化转向基于生成式 AI 与世界模型的高保真数据闭环, 为实现可扩展、低风险的系统安全边界验证与持续进化提供了关键基础设施. 在这之后, 结合最新的三维高斯泼溅 (3D Gaussian splatting, 3DGS) 技术, DrivingGaussian[172] 实现了对动态驾驶场景的实时、高保真重建与渲染, 有效模拟了传感器的物理特性, 为跨模态感知验证提供了可靠基础.

然而, 仅依靠重建难以应对极端工况, 为系统性地生成此类稀缺且高价值的长尾场景, 研究范式正逐渐从“重建已有数据”转向“生成未知场景”. 其中, 融合强生成先验的世界模型成为实现这一跨越的关键路径, DriveDreamer4D[87] 正是这一思路的典型代表. 它提出一种将生成式世界模型作为数据引擎的创新范式. 针对 3DGS 难以渲染未见过的复杂驾驶动作这一痛点, 该研究利用视频生成模型强大的幻觉能力, 生成同一场景下不同运动轨迹的合成视频. 通过生成式训练策略, DriveDreamer4D 将生成模型的丰富先验知识成功蒸馏进 4DGS 表示中, 不仅保持了 3D 场景的几何一致性, 更显著提升了模型在生成变道、加减速等长尾交互场景下的时空连贯性. 在此基础上, RAD[173] 进一步挖掘了高保真环境在策略学习中的潜力. 针对传统模仿学习的因果混淆与误差累积问题, 该框架利用 3DGS 构建几何一致的数字孪生世界, 并开创性地将其应用于大规模闭环强化学习训练. 通过引入安全关键奖励与人类驾驶风格正则化, RAD 证明了基于高保真世界模型的“低成本试错”是提升端到端策略鲁棒性的有效路径. 然而, 真实的仿真不仅需要静态背景, 更需要逼真的交通流. TrafficBots[174] 是一种基于世界模型的多智能体策略, 它结合条件变分自编码器与运动预测架构, 通过目的地与个性化潜变量生成多样化且符合物理约束的交通行为. 该策略支持闭环交互, 使背景车辆能对自车行为做出符合交通规则的动态响应, 从而构建了一个可微分、数据驱动的仿真环境. 这种基于学习的交互仿真显著降低了真实路测的成本与风险, 为自动驾驶规划算法的训练与验证提供了高效平台.

5 自动驾驶系统的训练范式与验证体系

随着大模型与自动驾驶系统的深度融合, 传统的基于几何感知的训练体系正逐步向端到端范式演进. 这一变革不仅重塑了数据的标注形态, 也推动了验证体系从单一地采用离线数据集进行轨迹误差评估等任务向语义理解与闭环控制能力的综合测试转型. 本节将从开环与闭环 2 个维度, 探讨这一新范式下的关键数据集、训练策略与验证平台.

5.1 开环训练与验证

传统自动驾驶模型的训练与验证主要采用基于数据集的开环范式, 即通过模仿学习最小化预测轨迹与专家轨迹之间的误差以及根据已记录的专家驾驶行为评估系统的性能. 该方法需要利用包含车辆在真实行驶过程中的多模态传感器观测 (如图像、点云)、环境上下文信息 (如高精地图、导航) 以及标注的专家驾驶行为真值 (如轨迹、控制指令) 等内容的离线数据集进行评估, 侧重于评估模型对静态场景的感知理解、逻辑推理以及基于历史数据的轨迹预测能力, 是模型构建世界知识与驾驶常识的基石. 总体而言, 自动驾驶数据集的演进正呈现出从基础感知向预测与规划拓展的趋势. 一方面, 数据是自动驾驶的基础, 海量训练数据能够赋予模型精准的环境理解与决策规划能力, 从而保障驾驶系统的安全与稳定; 另一方面, 近年来多模态大模型的迅速发展, 更加印证了海量高质量数据的重要性, 给予自动驾驶数据集的建设以启发.

5.1.1 从感知到理解

感知类数据集是开发高可靠性自动驾驶系统的基石, 其演进历程体现了学术界与工业界对环境理解深度与广度的不断追求. KITTI 数据集[175] 自 2012 年发布以来, 对自动驾驶领域产生了深远影响. 作为早期的综合性基准, 它包含了由摄像头、激光雷达和 GPS 等记录的真实驾驶场景, 凭借高分辨率数据和丰富的注释, 极大地促进了目标检测、跟踪、深度估计及视觉里程计等基础算法的开发与评测, 成为该领域的开山之作. 随着对城市场景理解需求的提升, 数据集开始向更精细的语义分割和更大规模的样本量发展. Cityscapes[176] 专注于复杂的城市环境, 提供了包括车辆、行人及交通设施在内的 30 类物体的像素级精细标注, 确立了城市场景语义分割的标准基准. ApolloScapes[177] 则进一步提升了数据维度, 提供了超过 14 万帧的高分辨率图像及 2D/3D 语义标签, 尤其在车道分割和多目标跟踪任

务上, 为高精度的感知与导航系统验证提供了关键支持. 在数据规模与多样性方面, 伯克利大学发布的 BDD100K 数据集[178] 则是一座重要的里程碑, 它包含 10 万个涵盖多种驾驶条件的视频序列, 不仅规模庞大, 而且标签种类丰富, 成为研究人员验证算法在复杂动态环境中的泛化能力的重要平台.

针对真实数据采集成本高昂及长尾场景稀缺的挑战, 合成数据集提供了新的解决方案. 如 VIPER[179] 利用虚拟游戏引擎生成并提取了超过 25 万帧涵盖多种天气条件的高分辨率视频数据, 有效缓解了真实场景数据获取困难的问题; SYNTHIA[180] 则通过提供 13 400 张带有像素级注释的合成图像, 致力于弥合虚拟与现实世界之间的差距, 为跨域鲁棒性算法的研究提供了经济高效的工具.

而近年来, 随着多模态融合技术的兴起, 涌现出大量融合多模态的自动驾驶数据集, 为自动驾驶的发展作出了重要的贡献. 如 nuScenes[27] 通过集成六个摄像头、激光雷达和雷达数据, 实现了对波士顿和新加坡复杂城市场景的 360° 全方位覆盖, 不仅设立了多视角目标检测的新基准, 更推动了多传感器融合技术的发展. 随后推出的 Waymo 开放数据集[181] 则以其工业级的高质量多模态数据和广泛的地理覆盖范围, 进一步提升了感知算法在检测、跟踪和分割任务上的稳健性上限. 而为了解决恶劣天气下的感知难题, RADIATE[182] 作为首个公开的雷达数据集填补了空白. 该数据集专注于雨、雾、雪等恶劣天气条件, 结合 LiDAR 与摄像头数据, 重点验证了雷达在视觉失效场景下的感知能力, 对于提升自动驾驶系统在全天候环境下的安全性和鲁棒性均具有重要意义.

在理解层面, 为评估模型对驾驶场景中自然语言指令的理解能力, Deruyttere 等[183] 发布了首个自动驾驶对象指称数据集 Talk2Car. 该数据集构建于 nuScenes 之上, 包含 11 959 条用自然语言编写的驾驶命令. 不同于传统的物体检测, Talk2Car 要求模型必须结合文本语义与视觉特征, 精准定位出指令所指涉的特定地标或交通参与者. 作为早期视觉定位任务的基准, 它为后续研究如何将 LLM 与自动驾驶感知系统对齐提供了重要的数据支撑与评测标准. 前文提到的 GPT-4 增强框架[119] 便是基于 Talk2Car 数据集进行算法验证的典型案例, 形成了“数据集–解决方案”的闭环.

5.1.2　从预测到规划控制

继感知任务数据集的蓬勃发展之后, 预测与规划控制类数据集已成为推动自动驾驶系统向决策智能进阶的必备要素. 此类数据集蕴含了丰富的交通动态与交互行为信息, 对于提升模型在复杂场景下的预测与决策能力至关重要. 根据任务的覆盖范围与应用侧重, 相关数据集可分为特定任务数据集与多任务数据集两类[184].

特定任务数据集主要聚焦于交通流的具体行为特征、行人意图以及短时运动预测, 为理解道路参与者的微观动态提供了基础. 例如, 为突破传统车载视角在捕捉全局交互行为上的局限, highD[185] 开创性地采用了基于无人机的高速公路采集方案. 该数据集包含 11 万条汽车和卡车的处理后轨迹, 有效弥补了传统测量技术在场景安全验证中的数据缺陷. 然而, 其仅覆盖理想环境, 缺乏复杂环境下的数据. 作为其演进版本, 2022 年发布的 exiD[186] 同样利用无人机采集以避免遮挡, 但在场景上更专注于高速公路出入口的复杂变道交互, 在保证数据质量与效率的同时, 显著提升了交互场景的多样性. 而针对城市环境中最难以预测的弱势道路使用者, PIE 数据集[187] 致力于解决行人行为理解难题. 该数据集在多伦多市中心多种光照条件下记录了超过 6 h 的数据, 提供了带遮挡标记的边框、过街意图置信度及动作文本标签, 为研究细粒度的行人视觉推理提供了丰富支撑. 而在 3D 目标跟踪与运动预测方面, Argoverse[188] 是运动预测领域的经典基准. 它提供了 360° 图像、激光雷达点云及高清语义地图, 覆盖 290 km 已标注车道和 30 万条轨迹. 得益于丰富的传感器配置, Argoverse 在预测算法研发中发挥了关键作用, 但其在不同地理区域的泛化能力仍存在一定局限.

前文所提到的 Talk2Car 侧重于静态的视觉定位, 而为了赋予模型更深层的因果推理能力, 学术界进一步提出基于视觉问答的认知数据集. nuScenes-QA[189] 提供了大规模的多模态视觉问答基准, 涵盖了从物体计数到空间关系判断的各类问题. 不过, 在可解释性方面, 简单的视觉问答往往难以捕捉驾驶决策背后复杂的因果逻辑链条. 针对这一局限, Sima 等[190] 的 DriveLM 工作引入了图视觉问答 (graph visual question answering, GVQA) 的新范式, 构建了首个包含这种逻辑依赖标注的语言–驾驶数据集 DriveLM-data, 还提出了基于 VLM 的基线方法 DriveLM-Agent. 不同于孤立的问答对, DriveLM 数据集将驾驶过程建模为由感知、预测和规划问题组成的逻辑图结构, 如“那辆车为什么减速?”到“因为前方有红灯”再到“那么我应该怎么做?”. 这种 CoT 式的标注使得模型能够在开环环境下学习类人的决策逻辑, 证明了通过显式建模推理过程的逻辑结构, 能够显著提升模型在复杂场景下的规划能力与可解释性.

在具备感知与推理能力后, 开环训练的最终目标是生成准确的控制信号. 然而, 长期以来, 训练一直面临一个关键瓶颈, 即缺乏大规模、高质量且“感知–规划–控制”全链路对齐的数据集. 传统数据集往往只提供感知标注或孤立的行为片段, 难以支撑端到端驾驶策略的连贯学习与可靠评估. 针对传统数据集缺乏动作对齐的瓶颈, Arai 等[138] 发布了 CoVLA 数据集, 一个大规模的“视觉–语言–动作”全对齐数据集. 该数据集通过采样构建了 10 000 个场景, 每帧均提供驾驶场景描述与未来的 3 s 轨迹, 并同步记录车辆控制信号. 这种“感知–认知–控制”全链路对齐的数据结构, 率先在真实场景下支持多模态大模型基于语言指令进行轨迹规划的离线端到端训练与评估, 有效弥补了以往数据集在规模与动作对齐上的不足. 同时, 为弥补从感知到认知的鸿沟, 伯克利大学进一步推出了 BDD-X 数据集[129]. 它基于 BDD100K 的视觉数据, 引入人类驾驶员的动作描述与决策逻辑标注, 建立了感知输入与决策输出之间的因果链路, 形成了“感知–决策–解释”的数据对. 这种将驾驶行为文本化的处理方式不仅提升了系统的可解释性, 更在当前大模型时代成为训练 VLM 学习驾驶思维链的关键方案.

在此类高质量数据的支撑下, 涌现出了一批优秀的端到端模型. 例如, DriveGPT4[84] 基于 BDD-X 数据集的真实行车记录进行视觉指令微调, 使模型在生成控制指令的同时, 同步输出“减速以避让右侧行人”等自然语言解释. 在 BDD-X 上的实验表明, 其控制信号预测误差显著低于基线方法, 验证了其在提升规划可解释性与控制精度方面的有效性. 同样, 理想汽车联合清华大学提出的 Drive-VLM-Dual[86] 在 nuScenes 验证集上将平均规划轨迹误差 (average L2 error) 降低至 0.31 m, 且碰撞率控制在 0.10%, 性能超越了 UniAD[8]、VAD-Base[74] 等 SOTA 方法, 在开环条件下达到了新的高度.

5.2 闭环训练与验证

尽管基于真实数据的开环训练能赋予模型基础的驾驶能力, 但其无法模拟决策对环境状态的反馈影响, 容易在实际部署中产生累积误差. 因此, 基于高保真模拟器甚至基于物理世界实车的闭环验证, 成为评估模型鲁棒性与真正能够落地的核心手段. 闭环验证需要构建一个模拟现实世界驾驶环境的仿真环境, 并将系统部署在仿真环境中, 随着时间或设定任务测试其性能. 该系统需要在模拟交通流中实现安全驾驶, 并能够在遇到一些突发情况时作出合理的决策. 为直观阐述这两种范式的差异, 图 10 对比了开环与闭环的运行机理: 左侧的开环范式依赖离线数据集, 其模型输出与环境状态解耦, 通过计算与专家轨迹的误差对模型进行优化或测试, 数据保真度高且收敛较稳定, 但应对数据集以外场景的能力较差; 而右侧的闭环范式构建了“感知–决策–更新”的完整反馈回路, 模型输出的控制动作会实时改变仿真环境的状态, 进而影响下一帧的传感器输入, 增强了模型的泛化性和鲁棒性.

5.2.1 仿真环境中的闭环测试

随着端到端模型复杂度的提升, 闭环测试已

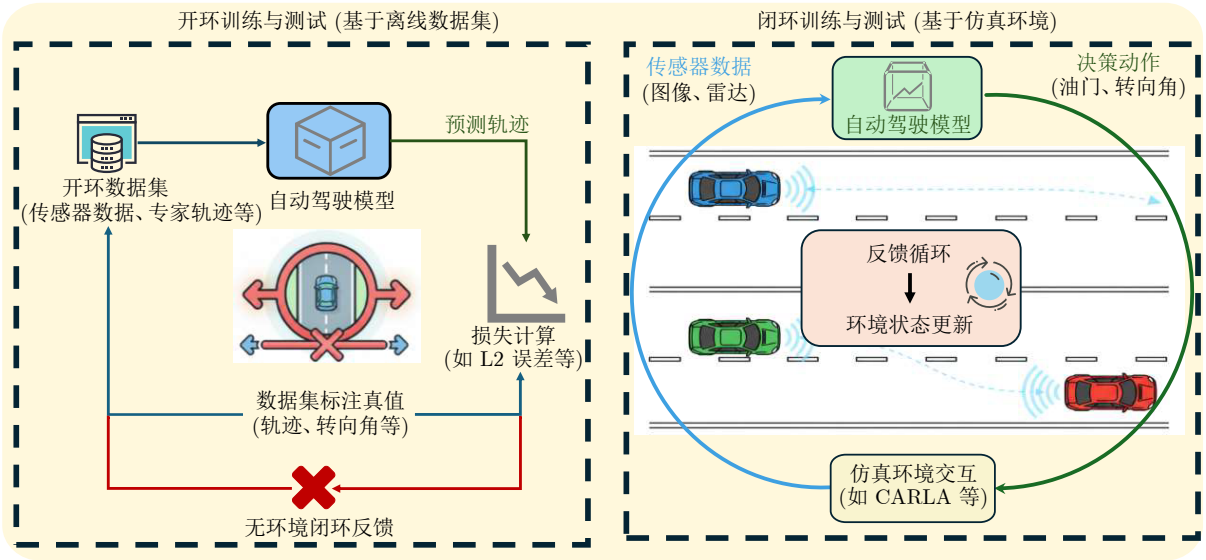


图 10 自动驾驶训练与验证——开环范式与闭环范式

Fig. 10 Autonomous driving training and validation——Open-loop and closed-loop paradigms

成为评估模型鲁棒性的核心手段. 在研究中, 可用于自动驾驶的仿真平台有很多, 如 CARLA[71]、CarSim[191]、CarMaker[192]、AirSim[193] 等. 以开源的 CARLA 仿真平台为例, 其提供的动态交通流与复杂长尾场景为大模型上车提供了低成本的试错环境. 在此背景下, 一系列结合大模型的闭环驾驶工作相继涌现. DriveLM[190] 是一个基于图结构视觉问答的自动驾驶推理框架, 通过构建感知、预测与规划之间的问答依赖图, 模拟人类驾驶的逐步推理过程, 显著增强了模型在复杂场景中的可解释性与零样本泛化能力. 与此同时, LMDrive[132] 则推进了语言模型在自动驾驶中的闭环控制能力, 构建了一个端到端的语言引导驾驶系统, 能够实时解析多模态自然语言指令并执行闭环驾驶任务, 实现了语言模型在仿真环境中的指令跟随与决策控制.

随着 VLM 技术的发展, Renz 等[78] 提出的 CarLLaVA 框架脱颖而出, 成为 CARLA 挑战赛 2.0 中的典型代表. 该研究证明了在纯视觉输入下, 无需复杂的标签, 依靠预训练的 VLM 与大规模自动驾驶数据集, 即可训练出具备专家级驾驶能力的闭环自动驾驶模型. 而 AutoVLA[146] 则设计了一种新型的 VLA 架构, 通过设计特定的动作词表, 直接将驾驶动作作为语言的一部分进行自回归生成, 有效增强了规划策略的一致性. 该工作在仿真环境中的开环和闭环基准都进行了全面评估, 其中闭环评估在 CARLA 的 Bench2Drive 基准中对各种驾驶指标进行衡量, 结果表明在闭环测试中, AutoVLA 在综合驾驶评分和成功率方面均优于 UniAD[8]、VAD[74]、ORION[144] 等经典端到端驾驶模型. 随后的 SimLingo[142] 进一步在 Bench2Drive 等多个高难度仿真基准中刷新了 SOTA 记录, 并特别强调了语言与动作对齐的重要性, 验证了通过动作梦境等机制进行训练, 可以让模型在闭环控制中不仅开得稳健, 还能够听得懂复杂的指令.

为直观展示大模型时代自动驾驶数据集从单纯的几何感知向语义理解与逻辑推理演进的特征, 表 2 系统梳理了涵盖开环与闭环的经典感知、认知与推理自动驾驶数据集. 其中, “模型输入”区分单模态 (即纯视觉或纯文本) 与多模态数据, 反映了模型对异构信息的整合能力; “数据来源”涵盖基于现有公开数据集 (如 nuScenes) 的二次语义增强、高保真仿真环境 (如 CARLA) 生成以及真实物理世界采集等多种途径; “数据集任务定位”是核心指标, 重点展示了数据集对视觉问答、思维链推理、驾驶指令跟随及场景描述等高阶认知任务的支持情况, 而非仅限于传统的物体检测框; 而“规模”则统计了包含的场景片段数量或问答对数量, 体现了数据体量对大模型训练泛化能力的支撑水平, 其中“K”代表单位“千”, “M”代表单位“百万”.

5.2.2　物理世界中的闭环实测

然而, 仿真终究不能对标现实世界, 真实的物理世界才是检验自驾模型的最终平台, 但实车验证也面临着成本高昂、不安全等难题. 早期有许多奠基性工作为后续的实车验证做出了重要的贡献. 例如, 作为端到端自动驾驶的开山之作, 前文所提到的 Bojarski 等[9] 的工作不仅提出了 DAVE-2, 更开创性地完成了实车环境下的闭环验证. 该研究将训练好的 CNN 直接部署于物理车辆, 在包括高速公路、乡村土路及复杂施工区域在内的真实场景中实现了对方向盘的实时闭环控制. 通过统计人工接管前的自主行驶里程, 该工作有力地证明了端到端模型在无需人工特征工程的情况下, 也可具备在非结构化真实环境中保持横向控制稳定性的泛化潜力, 是一个非常经典的前期工作. 而后为验证端到端模型在真实物理世界的闭环能力, Hawke 等[208] 提出一种基于条件模仿学习的城市驾驶方案. 该工作并未局限于仿真环境, 而是将由 30 h 人类数据训练出的策略部署于自动驾驶实车中, 在复杂的欧洲城市街道完成了 35 km 的闭环路测. 这些工作都为端到端算法的实车部署落地提供了早期的重要实证.

进一步地, 针对自动驾驶长尾场景中端到端模型鲁棒性不足的难题, NVIDIA 提出了 Alpamayo-R1[209], 一个将显式因果推理与动作预测紧密耦合的通用架构. 该工作通过构建可追溯的因果链与多阶段强化学习对齐, 有效解决了传统 VLA 模型中推理与控制脱节的问题. 尤为关键的是, 该研究成功完成了高难度的实车闭环验证, 在涵盖红绿灯路口、施工区域及行人交互等复杂的城市公开道路测试中, 实现了零接管驾驶. 此外, 系统在 RTX Pro 6000 上达到了 99 ms 的端到端推理延迟, 满足百毫秒级的实时控制标准, 从而有力证明了具备深层因果推理能力的 VLA 模型在物理世界部署的实时性与可行性. 同时, 自然语言接口的兴起也推动了人车协同决策的发展, 如 Talk2Drive[120] 构建一套通过自然语言指令控制实车的闭环框架. 区别于大多数仅停留在仿真环境中的研究, 该工作设计了云端 LLM 与车载电子控制单元 (ECU) 的各种通信接口, 将人类的口头驾驶指令 (如“我想开快点”) 实时转化为可执行的底层控制代码. 真实道路实验表明, 该系统不仅实现了 100% 的代码执行成功率, 更通过个性化的长程推理显著降低了驾驶员的接管频率, 为解决 LLM 在真实交通场景中的即时性与安全性

表 2 大模型时代面向感知、认知与推理任务的自动驾驶数据集汇总

Table 2 Summary of autonomous driving datasets for perception, cognition, and reasoning tasks in the era of large models

模型输入	数据集名称	数据来源	数据集任务定位	规模
单模态	Cityscapes[176]	真实世界	像素级语义与实例分割	5 K 个细粒度与 20 K 个粗粒度图像
	MetaAD	真实世界	闭环驾驶评测	110 K 个训练片段和 10 K 个验证片段
	SimLingo[142]	CARLA	驾驶动作指令跟随与视觉问答	基于 1 M 帧片段的 28 M 个问答对
多模态	nuScenes-QA[189]	nuScenes[27]	多模态场景问答	34 K 个场景下的 460 K 个问答对
	NuPrompt[194]	nuScenes	目标场景描述与轨迹预测	35.4 K 个目标–提示词对
	DriveGPT4[84]	BDD-X[129]	感知与推理问答	56 K 个视频–文本指令样本
	DriveLM-nuScenes[190]	nuScenes	驾驶任务图结构问答	4.8 K 帧片段下的 440 K 个问答对
	DriveLM-CARLA	CARLA[71]	驾驶任务图结构问答	180 K 帧片段下的 3.75 M 个问答对
	Reason2Drive[195]	nuScenes、Waymo[181] 等	感知与推理问答	600 K 个问答对
	LingoQA[196]	真实世界	驾驶行为问答与解释推理	419 K 个问答对
	OmniReason-nuScenes[197]	nuScenes	时空因果推理与问答	1 K 个约 20 s 的驾驶场景
	OmniReason-Bench2Drive	Bench2Drive[76]	闭环驾驶因果推理	1 K 个约 150 m 的驾驶片段
	WOMD[198]	Waymo	交互式轨迹预测	约 104 K 个真实城市和郊区驾驶场景
	LMDrive[132]	CARLA	驾驶指令跟随	64 K 个指令跟随数据片段
	Talk2Car[183]	nuScenes	自然语言指令的视觉定位与指称	12 K 条指令与 9.2 K 张图像
	CoVLA-Dataset[138]	真实世界	轨迹预测与帧级场景推理描述	10 K 个视频片段/6 M 帧片段对应字幕
	DATAD[199]	CARLA	驾驶员接管注意力与视线预测	在 12 种接管场景下的 600 K 帧数据
	W^{3}DA[200]	BDD-A[201]、DADA-2000[202] 等	可解释性注意力预测	3.5 K 个场景中的 70 K 个关键样本
	OmniDrive[154]	nuScenes	规划、推理与反事实问答	通过对抗性推理生成大规模问答对
	Rank2Tell[203]	真实世界	重要性分级与因果推理描述	116 个密集推理片段
	DriveBench[204]	DriveLM	VLM 驾驶可靠性评测	20.5 K 个问答对与 19.2 K 帧片段
	ImpromptuVLA[145]	nuScenes、Waymo 等	驾驶 VLA 任务	80 K 余个长尾场景片段
	Chat-B2D[144]	CARLA	闭环驾驶 VQA 与因果推理	1.1 M 个问答对
	NuGrounding[205]	nuScenes	多视角 3D 视觉定位	2.2 M 个文本提示与 34 K 个关键帧
	RefAV[206]	Argoverse 2[207]	基于语言的时空场景挖掘	10 K 条自然语言查询

挑战提供了宝贵的实测数据.

自动驾驶的训练与验证体系正经历着一场深刻的范式转移. 从数据层面看, 数据集完成了从单一感知的像素级标注向蕴含因果逻辑的认知级标注的跨越, 为构建具备认知能力的驾驶智能体提供了知识底座; 从验证维度看, 单一的开环轨迹误差评估正在向融合高保真仿真与实车部署的闭环测试演进. 这些转变共同指向一个根本目标: 推动自动驾驶系统从“感知–反应”向“认知–决策”演进, 以构建更安全、可靠且真正智能的驾驶智能体.

6 当前研究存在的挑战与展望

尽管大模型与端到端范式的结合为自动驾驶开辟了从“感知智能”向“认知智能”跃迁的渠道, 但要真正实现全场景、L4 级及以上无人驾驶的商业化落地, 当前的系统在实时性、安全性、可解释性以及架构设计上仍面临着严峻的挑战. 本节将对这些核心问题进行剖析, 并对未来的技术演进方向进行展望.

6.1 当前研究存在的挑战

尽管大模型技术为自动驾驶注入了强大的认知潜力, 但要将其从实验室推向物理世界, 首先需要解决的便是车端有限算力与模型庞大参数量之间的矛盾. 自动驾驶作为对安全要求极高的实时应用, 要求系统必须在毫秒级时间内完成对环境的感知与决策. 然而, 当前的多模态大模型动辄拥有数百亿参数, 其巨大的计算开销与推理延迟在现有车载硬件上难以忽视. 虽然学术界尝试通过如 DriveVLM-Dual[86] 设计的“快慢系统”解耦架构来缓解这一压力, 试图用小模型保障实时性、大模型提供高层指导, 但在高动态的复杂交通流中, 如何界定“快”与“慢”的动态切换边界, 防止因大模型推理滞后而导致关键决策窗口关闭, 仍是制约其工程化落地的首要瓶颈.

然而, 即便跨越了算力门槛, 相比于传统模块

化系统基于规则的“白盒”逻辑, 端到端大模型的“黑盒”特性也会使得故障边界模糊, 难以进行归因分析与解释. 因此, 如何将大模型的泛化创造力严格约束在控制理论的稳定域内, 构建有效的“安全护栏”, 是当前亟待解决的安全性难题. 除安全隐患外, 大模型在空间认知上的“模态错位”也是导致系统不稳定的关键因素. 现有的主流 VLM 大多基于互联网 2D 图文数据预训练, 虽然在自然图像理解上表现卓越, 但是缺乏对驾驶场景中深度、遮挡关系及三维几何结构的固有认知. 尽管 Talk2BEV[125] 等工作试图引入 BEV 表征来弥补这一缺陷, 但大模型离散符号化的语义空间与车辆控制所需的连续精确的物理度量空间之间始终横亘着巨大的鸿沟. 简单的 Tokenizer 往往难以承载过于精细的微操意图, 这种映射的精度损失经常导致模型呈现出控制指令振荡的现象.

归根结底, 上述所有挑战的根源在于高质量认知数据的稀缺. 数据集的演进虽然解决了量的问题, 但包含高质量思维链标注的认知数据 (如 BDD-X[129]) 依然匮乏, 导致模型更多是在模仿表层的驾驶行为, 而非学习底层的因果逻辑. 尽管 DriveGPT4[84] 等模型能通过自然语言输出看似合理的驾驶理由, 但由于主流架构多采用解耦的输出头, 使得文本与轨迹虽共享感知特征, 却缺乏强逻辑约束的因果对齐. 这种语义解释与物理执行之间的因果错位, 使得可解释性在功能安全层面失去了作为事故定责或调试依据的可信度, 在关键时刻甚至可能造成对人类驾驶员的误导. 为克服这一痛点, 如何使模型在生成思维链时与物理动作保持强一致性已成为当前的关键研究方向. 目前的潜在解决路径主要集中在三个维度: 一是构建结构化的因果推理链与对齐评估体系, 如 Reason2Drive[195] 构建了“感知–预测–推理–动作”的序列化数据集与评估指标, 为大模型在自回归解码阶段实现“先思考、后行动”提供了关键的高质量数据与优化基准; 二是基于行为状态空间的接口对齐, 如 DriveMLM[131] 放弃了直接输出底层连续控制信号的模式, 而是将大模型的多模态输出与标准化的行为规划状态严格对齐, 从而构建起语义决策与下游控制模块之间的可靠桥梁; 三是引入多模态渐进式推理与世界模型推演机制, 例如 Zhang 等[210] 的最新工作 MindDriver 提出了从“文本语义理解”到“视觉物理空间想象”, 再到“轨迹规划”的渐进式推理架构, 该方案通过强制模型在输出控制动作前, 先将抽象的文本思维链具象化为符合物理规律的未来场景演变, 从而在架构层面构建了从语义到物理度量的强因果桥梁, 赋予自动驾驶系统“先想后看再行动”的类人认知能力, 为该挑战提供了一条可行的解决路径.

因此, 虽然大模型的爆发为自动驾驶带来了无限可能, 但客观存在的挑战使得其发展与大规模落地应用依旧任重道远.

6.2 未来展望

展望未来, 自动驾驶系统的架构演进将不再局限于单一模型的参数扩张, 而是向着类人认知的协同方向深化. 面对车端算力约束与开放世界无限长尾场景之间的根本矛盾, 未来的理想形态应当是构建一种动态的分层认知架构: 底层依托高效、轻量级的端到端网络, 凭借类人的记忆认知负责处理绝大多数高频、确定性的常规驾驶任务, 从而保障毫秒级的响应速度与控制稳定性; 而拥有深层推理能力的多模态大模型则在遭遇复杂博弈、罕见长尾或需要接管的时刻被激活, 发挥其逻辑推演与全局策略指导优势. 进一步地, 针对上述分层认知架构, 如何精确界定“快”与“慢”的动态切换边界, 是未来突破大模型上车瓶颈的关键技术挑战. 与本文第 3.2.2 节中所探讨的极端情况下的安全兜底机制相呼应, 未来的系统架构还可将重点放在按需唤醒机制上, 从而形成完整的双向动态仲裁. 现有的架构多依赖人工规则进行生硬的系统切换, 而未来的理想范式将深度融合认知不确定性估计理论[211–213]. 在常态化巡航中, 大模型以低频率在后台提供全局路由与语义导航; 而在微观规控层面, 底层“快网络”在输出高频规控指令的同时, 必须具备可信的自我审视能力, 即实时量化自身的场景认知置信度, 如基于证据深度学习[214] (evidential deep learning, EDL) 或模型集成的认知不确定性建模. 当常规轻量级网络在遭遇罕见长尾障碍物或复杂博弈场景导致置信度低于安全阈值, 且系统基于运动学预判当前距离物理危险仍有足够的时间冗余时, 才会精准地“向上唤醒”大模型进行高层语义推演与策略指导. 这种动态的边界双向划分, 能够在一定程度上应对分层架构中的时延与安全悖论.

然而, 要让动辄数百亿参数的大模型真正在车端实现百毫秒级的实时闭环运行, 单纯依赖架构解耦仍显不足, 极致的端侧部署与加速技术将成为未来不可或缺的工程基石. 在模型轻量化方面, 模型压缩技术, 如基于 OPTQ[215] 等经典算法的低比特量化、结构化剪枝与知识蒸馏将走向深度融合. 未来的重要趋势是通过云端超大模型提取高质量的常识推理与思维链数据, 将其高效蒸馏至数十亿参数级别的车端小模型中, 并在低数据位宽 (如 INT4

或 INT8 量化) 下确保认知性能的极低损失. 这一技术路径已在产业界得到初步验证, 如小鹏汽车通过云端 720 亿参数基座模型蒸馏至车端 70 亿模型, 结合自研芯片异构加速, 使 VLA 模型具备了实时闭环运行能力. 同时, 理想汽车的软硬协同设计定律表明, 在 INT8 量化与硬件原生支持的协同下, 模型可在相同延迟预算下实现精度跃升[216]. 而在底层算力支撑方面, 异构计算平台的软硬协同优化将发挥决定性作用. 未来的车载计算单元将针对大模型的核心算子 (如注意力机制) 进行专属定制, 通过动态调度 CPU、GPU 与专用神经网络处理器的异构资源, 实现吞吐量的最大化与推理延迟的最小化. 这种宏观层面的“直觉行动”与“逻辑监督”的有机融合, 辅以微观层面的端侧极致优化, 不仅能有效打破大模型上车的算力瓶颈, 更能通过大模型的语义输出来实时监督小模型的行为边界, 实现系统安全性与智能性的高度统一.

伴随着架构的升级, 自动驾驶的迭代范式也将发生根本性变革, 从依赖海量实车数据的被动灌输, 转向基于多模态大模型与世界模型的终身学习与内生仿真. 关键的是, 目前世界模型还正处于发展阶段, 随着未来生成式世界模型的成熟与广泛应用, 意味着智能体将不再仅仅依赖昂贵的物理路测来积累经验, 而是拥有一个能够理解物理规律的“数字大脑”, 使其能够在潜空间中进行类似于人类“思维预演”的低成本试错与反事实推演. 这种内生仿真能力将彻底打破传统仿真与真实世界之间的鸿沟, 赋予自动驾驶系统在车辆出厂后持续进化的能力, 即通过不断的虚拟演练与真实环境反馈的交互, 实现对新环境、新交通规则乃至新突发状况的自适应迁移, 真正迈向具有高度类人认知的具身智能体.

从更宏观的视角来看, 自动驾驶与具身智能的联结将更加紧密. 随着 VLA 范式的成熟, 轮式车辆与人形机器人、无人机等形态的底层感知–决策–控制逻辑将趋于统一, 通用机器人基础模型的跨形态知识共享将极大地加深车辆对物理世界的理解, 实现从“专用司机”向“通用具身智能体”的跨越. 在此基础上, 未来的智能驾驶系统将由被动的执行者角色, 演变为可交流、可解释的合作伙伴, 能够解释决策动机并精准理解复杂的个性化意图. 这种基于深层语义理解的人机交互[217] 不仅是消除公众对“算法黑箱”恐惧的关键, 更是构建和谐人机共生关系、推动高阶自动驾驶社会化落地的必由之路.

综上所述, 大模型能赋予车辆强大的场景理解与交互能力, 使其能解读复杂的路况语义, 甚至通过自然交互与乘客、其他道路使用者沟通, 达到人机协作的效果. 世界模型能让车辆在虚拟环境中进行海量的社会实践和推理试错, 快速学习应对各类长尾场景和突发事件, 从而大幅提升其应对未知环境的泛化与快速学习能力. 而具身智能则强调“感知–决策–控制”的一体化闭环, 确保这种高级认知能力能转化为安全、流畅、拟人化的实际驾驶行为. 未来, 随着端到端大模型对长尾场景理解能力的深化, 构建“数据驱动、认知增强、虚实闭环”的一体化迭代体系, 将是实现自动驾驶系统向全场景落地突破的关键路径. 为将上述“数据驱动与认知增强相结合”的理念转化为可实施的工程方案, 本文提出如下包含三个核心环节的一体化技术落地路线.

1) 认知数据的自动化挖掘与对齐 (数据底座)

摒弃传统高度依赖人工的几何框标注, 转而利用云端大规模视觉–语言基础模型, 对海量无监督路测数据进行自动化的高维语义挖掘. 重点在于提取人类驾驶员的接管片段与极低频的长尾危险场景, 并利用大模型自动生成包含因果逻辑的思维链标注, 从而构建高质量的认知级自动驾驶数据集[84, 189, 195].

2) 规则先验与深层推理的注入 (认知增强)

在车端部署阶段, 不应让端到端模型在完全无约束的黑盒中自由试错. 实施路线应依托前文所述的“快慢系统”架构[86], 将人类的交通法规常识与博弈逻辑, 通过大模型转化为结构化的安全条例或惩罚函数; 随后, 将这些高阶认知特征作为软性约束或引导提示深度融合到端到端规控网络的特征空间中, 实现数据拟合与逻辑推理的优势互补[118].

3) 基于世界模型的反事实闭环演练 (虚实闭环)

针对物理路测成本高昂且危险场景难以穷尽的痛点, 该路线的最终闭环在于内生仿真, 即利用高保真生成式世界模型[87, 171, 173] 对真实采样的边缘场景进行“反事实推演”. 智能体能够在合成的平行数字空间中, 针对同一危险场景尝试成千上万种不同的避险动作序列, 并利用强化学习进行低成本试错, 再将优化后的策略迁移至物理世界并部署, 最终形成“数据采集–认知理解–虚拟试错–策略进化”的终身学习闭环.

7 结束语

本文针对大模型时代端到端自动驾驶的演进趋势, 回顾了自动驾驶技术从传统模块化架构向端到端范式的转型历程, 深入剖析了包括 LLM、VLM、VLA 模型在内的大模型及世界模型赋能自动驾驶感知、推理与控制的核心机制. 在此基础上, 本文梳理了伴随范式转移而来的数据演进与闭环验证体系

变革, 分析了当前研究在实时性、安全性及可解释性等方面面临的挑战, 并展望了基于多模态大模型强大理解能力与世界模型内生仿真能力的未来发展方向, 旨在为推动自动驾驶从感知智能向具备类人决策能力的具身认知智能体跃迁提供参考.

参考文献

1 National Development and Reform Commission, Office of the Central Cyberspace Affairs Commission, Ministry of Science and Technology, Ministry of Industry and Information Technology, Ministry of Public Security, Ministry of Finance, et al. Strategy for innovation and development of intelligent vehicles [Online], available: https://www.ndrc.gov.cn/xxgk/zcfb/tz/202002/t20200224_1221077.html, September 9, 2026
(国家发展改革委, 中央网信办, 科技部, 工业和信息化部, 公安部, 财政部, 等. 智能汽车创新发展战略 [Online], available: https://www.ndrc.gov.cn/xxgk/zcfb/tz/202002/t20200224_1221077.html, 2026-09-09)

2 Ministry of Industry and Information Technology, National Standardization Administration. Guidelines for the construction of the national internet of vehicles industry standard system (intelligent connected vehicles) (2023 Version) [Online], available: https://wap.miit.gov.cn/jgsj/kjs/wjfb/art/2023/art_28a7501f51ae4b408f32f3fb2c49e271.html, September 9, 2026
(工业和信息化部, 国家标准化管理委员会. 国家车联网产业标准体系建设指南 (智能网联汽车) (2023 版) [Online], available: https://wap.miit.gov.cn/jgsj/kjs/wjfb/art/2023/art_28a7501f51ae4b408f32f3fb2c49e271.html, 2026-09-09)

3 Ministry of Industry and Information Technology, Ministry of Public Security, Ministry of Housing and Urban-Rural Development, Ministry of Transport. Notice on carrying out pilot work for the admission and road traffic of intelligent connected vehicles [Online], available: https://www.gov.cn/zhengce/zhengceku/202311/content_6915788.htm, September 9, 2026
(工业和信息化部, 公安部, 住房和城乡建设部, 交通运输部. 关于开展智能网联汽车准入和上路通行试点工作的通知 [Online], available: https://www.gov.cn/zhengce/zhengceku/202311/content_6915788.htm, 2026-09-09)

4 Ministry of Industry and Information Technology, Ministry of Public Security, Ministry of Natural Resources, Ministry of Housing and Urban-Rural Development, Ministry of Transport. Notice on announcing the list of pilot cities for "vehicle-road-cloud integration" application of intelligent connected vehicles [Online], available: https://www.gov.cn/zhengce/zhengceku/202407/content_6965771.htm, September 9, 2026
(工业和信息化部, 公安部, 自然资源部, 住房和城乡建设部, 交通运输部. 关于公布智能网联汽车"车路云一体化"应用试点城市名单的通知 [Online], available: https://www.gov.cn/zhengce/zhengceku/202407/content_6965771.htm, 2026-09-09)

5 Li X S, Guo Z Z, Dai X Y, Lin Y L, Jin J C, Zhu F H, et al. Deep imitation learning for traffic signal control and operations based on graph convolutional neural networks. In: Proceedings of the 23rd International Conference on Intelligent Transportation Systems (ITSC). Rhodes, Greece: IEEE, 2020. 1−6

6 Han S S, Wang X, Zhang J J, Cao D P, Wang F Y. Parallel vehicular networks: A CPSS-based approach via multimodal big data in IoV. *IEEE Internet of Things Journal*, 2019, **6**(1): 1079−1089

7 Outline of the 14th five-year plan for national economic and social development of the People's Republic of China and the long-range objectives through the year 2035 [Online], available: http://www.gov.cn/xinwen/2021-03/13/content_5592681.htm, September 9, 2026
(中华人民共和国国民经济和社会发展第十四个五年规划和 2035 年远景目标纲要 [Online], available: http://www.gov.cn/xinwen/2021-03/13/content_5592681.htm, 2026-09-09)

8 Hu Y H, Yang J Z, Chen L, Li K Y, Sima C H, Zhu X Z, et al. Planning-oriented autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 17853−17862

9 Bojarski M, del Testa D, Dworakowski D, Firner B, Flepp B, Goyal P, et al. End to end learning for self-driving cars. arXiv preprint arXiv: 1604.07316, 2016.

10 Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS). Long Beach, USA: Curran Associates Inc., 2017. 6000−6010

11 Brown T B, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. In: Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS). Vancouver, Canada: Curran Associates Inc., 2020. Article No. 159

12 Yang Y X, Han C R, Mao R H, Wang H S, Chen Z W, Yang Y T, et al. Survey of general end-to-end autonomous driving: A unified perspective. DOI: 10.36227/techrxiv.176523315.56439138/v1

13 Chib P S, Singh P. Recent advancements in end-to-end autonomous driving using deep learning: A survey. *IEEE Transactions on Intelligent Vehicles*, 2024, **9**(1): 103−118

14 Chen Yan-Yan, Tian Da-Xin, Lin Chun-Mian, Yin Hong-Bo. Survey of end-to-end autonomous driving systems. *Journal of Image and Graphics*, 2024, **29**(11): 3216−3237
(陈妍妍, 田大新, 林椿眄, 殷鸿博. 端到端自动驾驶系统研究综述. 中国图象图形学报, 2024, **29**(11): 3216−3237)

15 Zhu Y X, Wang S Y, Zhong W Q, Shen N C, Li Y Q, Wang S Q, et al. A survey on large language model-powered autonomous driving. *Engineering*, 2026, **63**: 190−203

16 Jiang S C, Kong M L, Tang Y H, Sun L J, Qian K A, Luo Z A, et al. A survey on vision-language-action models for autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). Honolulu, USA: IEEE, 2025. 4583−4595

17 Janai J, Güney F, Behl A, Geiger A. Computer vision for autonomous vehicles: Problems, datasets and state of the art. *Foundations and Trends in Computer Graphics and Vision*, 2020, **12**(1−3): 1−308

18 Li L, Wang F Y. *Advanced Motion Control and Sensing for Intelligent Vehicles*. New York: Springer, 2007.

19 Yurtsever E, Lambert J, Carballo A, Takeda K. A survey of autonomous driving: Common practices and emerging technologies. *IEEE Access*, 2020, **8**: 58443−58469

20 Yeong D J, Velasco-Hernandez G, Barry J, Walsh J. Sensor and sensor fusion technology in autonomous vehicles: A review. *Sensors*, 2021, **21**(6): Article No. 2140

21 Dai X R. HybridNet: A fast vehicle detection system for autonomous driving. *Signal Processing: Image Communication*, 2019, **70**: 79−88

22 Huang K L, Shi B T, Li X, Li X, Huang S Y, Li Y K. Multi-modal sensor fusion for auto driving perception: A survey. arXiv preprint arXiv: 2202.02703, 2022.

23 Philion J, Fidler S. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D. In: Proceedings of the 16th European Conference on Computer Vision (ECCV). Glasgow, UK: Springer, 2020. 194−210

24 Li H Y, Sima C H, Dai J F, Wang W H, Lu L W, Wang H J,

et al. Delving into the devils of bird′s-eye-view perception: A review, evaluation and recipe. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, **46**(4): 2151−2170

25 Li Z Q, Wang W H, Li H Y, Xie E Z, Sima C H, Lu T, et al. BEVFormer: Learning bird′s-eye-view representation from LiDAR-camera via spatiotemporal Transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025, **47**(3): 2020−2036

26 Liu Z J, Tang H T, Amini A, Yang X Y, Mao H Z, Rus D L, et al. BEVFusion: Multi-task multi-sensor fusion with unified bird′s-eye view representation. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). London, UK: IEEE, 2023. 2774−2781

27 Caesar H, Bankiti V, Lang A H, Vora S, Liong V E, Xu Q, et al. nuScenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 11618−11628

28 Zhu Bing, Zhang Pei-Xing, Zhao Jian, Chen Hong, Xu Zhi-Gang, Zhao Xiang-Mo, et al. Review of scenario-based virtual validation methods for automated vehicles. *China Journal of Highway and Transport*, 2019, **32**(6): 1−19
(朱冰, 张培兴, 赵健, 陈虹, 徐志刚, 赵祥模, 等. 基于场景的自动驾驶汽车虚拟测试研究进展. 中国公路学报, 2019, **32**(6): 1−19)

29 Mozaffari S, Al-Jarrah O Y, Dianati M, Jennings P, Mouzakitis A. Deep learning-based vehicle behavior prediction for autonomous driving applications: A review. *IEEE Transactions on Intelligent Transportation Systems*, 2022, **23**(1): 33−47

30 Huang Y J, Du J T, Yang Z R, Zhou Z W, Zhang L, Chen H. A survey on trajectory-prediction methods for autonomous driving. *IEEE Transactions on Intelligent Vehicles*, 2022, **7**(3): 652−674

31 Chi Rong-Hu, Hou Zhong-Sheng, Huang Biao. Optimal iterative learning control of batch processes: From model-based to data-driven. *Acta Automatica Sinica*, 2017, **43**(6): 917−932
(池荣虎, 侯忠生, 黄彪. 间歇过程最优迭代学习控制的发展: 从基于模型到数据驱动. 自动化学报, 2017, **43**(6): 917−932)

32 Min K, Kim D, Park J, Huh K. RNN-based path prediction of obstacle vehicles with deep ensemble. *IEEE Transactions on Vehicular Technology*, 2019, **68**(10): 10252−10256

33 Nikhil N, Morris B T. Convolutional neural network for trajectory prediction. In: Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018. 186−196

34 Zhang K P, Feng X L, Wu L, He Z B. Trajectory prediction for autonomous driving using spatial-temporal graph attention Transformer. *IEEE Transactions on Intelligent Transportation Systems*, 2022, **23**(11): 22343−22353

35 Gao J Y, Sun C, Zhao H, Shen Y, Anguelov D, Li C C, et al. VectorNet: Encoding HD maps and agent dynamics from vectorized representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 11522−11530

36 Zhao H, Gao J Y, Lan T, Sun C, Sapp B, Varadarajan B, et al. TNT: Target-driven trajectory prediction. In: Proceedings of the 4th Conference on Robot Learning (CoRL). Cambridge, USA: PMLR, 2020. 895−904

37 Zu C Y, Yang C, Wang J, Gao W B, Cao D P, Wang F Y. Simulation and field testing of multiple vehicles collision avoidance algorithms. *IEEE/CAA Journal of Automatica Sinica*, 2020, **7**(4): 1045−1063

38 Guo Y Q, Yao D Y, Li B, He Z M, Gao H C, Li L. Trajectory planning for an autonomous vehicle in spatially constrained environments. *IEEE Transactions on Intelligent Transportation Systems*, 2022, **23**(10): 18326−18336

39 Guo Y Q, Guo Z L, Wang Y Z, Yao D Y, Li B, Li L. A survey of trajectory planning methods for autonomous driving-part I: Unstructured scenarios. *IEEE Transactions on Intelligent Vehicles*, 2024, **9**(9): 5407−5434

40 Paden B, Čáp M, Yong S Z, Yershov D, Frazzoli E. A survey of motion planning and control techniques for self-driving urban vehicles. *IEEE Transactions on Intelligent Vehicles*, 2016, **1**(1): 33−55

41 Soltani A R, Tawfik H, Goulermas J Y, Fernando T. Path planning in construction sites: Performance evaluation of the Dijkstra, A*, and GA search algorithms. *Advanced Engineering Informatics*, 2002, **16**(4): 291−303

42 Wang X, Tang K, Dai X Y, Xu J T, Xi J H, Ai R, et al. Safety-balanced driving-style aware trajectory planning in intersection scenarios with uncertain environment. *IEEE Transactions on Intelligent Vehicles*, 2023, **8**(4): 2888−2898

43 LaValle S M. Rapidly-exploring Random Trees: A New Tool for Path Planning, Technical Report 98-11, Iowa State University, USA, 1998.

44 Teng S Y, Hu X M, Deng P, Li B, Li Y C, Ai Y F, et al. Motion planning for autonomous driving: The state of the art and future perspectives. *IEEE Transactions on Intelligent Vehicles*, 2023, **8**(6): 3692−3711

45 Bernhard J, Gieselmann R, Esterle K, Knoll A. Experience-based heuristic search: Robust motion planning with deep Q-learning. In: Proceedings of the 21st International Conference on Intelligent Transportation Systems (ITSC). Maui, USA: IEEE, 2018. 3175−3182

46 Puterman M L. Markov decision processes. *Handbooks in Operations Research and Management Science*, 1990, **2**: 331−434

47 Codevilla F, Müller M, López A, Koltun V, Dosovitskiy A. End-to-end driving via conditional imitation learning. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Brisbane, Australia: IEEE, 2018. 4693−4700

48 Bai Tian-Xiang, Wang Shuai, Shen Zhen, Cao Dong-Pu, Zheng Nan-Ning, Wang Fei-Yue. Parallel robotics and parallel unmanned systems: Framework, structure, process, platform and applications. *Acta Automatica Sinica*, 2017, **43**(2): 161−175
(白天翔, 王帅, 沈震, 曹东璞, 郑南宁, 王飞跃. 平行机器人与平行无人系统: 框架、结构、过程、平台及其应用. 自动化学报, 2017, **43**(2): 161−175)

49 Wang J K, Chi W Z, Li C M, Wang C Q, Meng M Q H. Neural RRT*: Learning-based optimal path planning. *IEEE Transactions on Automation Science and Engineering*, 2020, **17**(4): 1748−1758

50 Ma N C, Wang J K, Liu J B, Meng M Q H. Conditional generative adversarial networks for optimal path planning. *IEEE Transactions on Cognitive and Developmental Systems*, 2022, **14**(2): 662−671

51 Qureshi A H, Miao Y L, Simeonov A, Yip M C. Motion planning networks: Bridging the gap between learning-based and classical motion planners. *IEEE Transactions on Robotics*, 2021, **37**(1): 48−66

52 Chen Hong, Gong Xun, Hu Yun-Feng, Liu Qi-Fang, Gao Bing-Zhao, Guo Hong-Yan. Automotive control: The state of the art and perspective. *Acta Automatica Sinica*, 2013, **39**(4): 322−346
(陈虹, 宫洵, 胡云峰, 刘奇芳, 高炳钊, 郭洪艳. 汽车控制的研究现状与展望. 自动化学报, 2013, **39**(4): 322−346)

53 Duan Yan-Jie, Lv Yi-Sheng, Zhang Jie, Zhao Xue-Liang, Wang Fei-Yue. Deep learning for control: The state of the art and prospects. *Acta Automatica Sinica*, 2016, **42**(5): 643−654
(段艳杰, 吕宜生, 张杰, 赵学亮, 王飞跃. 深度学习在控制领域的

研究现状与展望. 自动化学报, 2016, **42**(5): 643−654)

54 Liu T, Tian B, Ai Y F, Wang F Y. Parallel reinforcement learning-based energy efficiency improvement for a cyber-physical system. *IEEE/CAA Journal of Automatica Sinica*, 2020, **7**(2): 617−626

55 Chen L, Li Y C, Huang C, Xing Y, Tian D X, Li L, et al. Milestones in autonomous driving and intelligent vehicles-part I: Control, computing system design, communication, HD map, testing, and human behaviors. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2023, **53**(9): 5831−5847

56 Ang K H, Chong G, Li Y. PID control system analysis, design, and technology. *IEEE Transactions on Control Systems Technology*, 2005, **13**(4): 559−576

57 Thrun S, Montemerlo M, Dahlkamp H, Stavens D, Aron A, Diebel J, et al. Stanley: The robot that won the DARPA grand challenge. *Journal of Field Robotics*, 2006, **23**(9): 661−692

58 Falcone P, Borrelli F, Asgari J, Tseng H E, Hrovat D. Predictive active steering control for autonomous vehicle systems. *IEEE Transactions on Control Systems Technology*, 2007, **15**(3): 566−580

59 Yu X K, Wang H, Teng C L, Sun X Q, Chen L, Cai Y F. DG-PR-MPC: Learning-based model predictive controller for autonomous vehicle path following. *IET Intelligent Transport Systems*, 2023, **17**(10): 1992−2003

60 Kabzan J, Hewing L, Liniger A, Zeilinger M N. Learning-based model predictive control for autonomous racing. *IEEE Robotics and Automation Letters*, 2019, **4**(4): 3363−3370

61 Zhu Y X, Li Z H, Wang F Y, Li L. Control sequences generation for testing vehicle extreme operating conditions based on latent feature space sampling. *IEEE Transactions on Intelligent Vehicles*, 2023, **8**(4): 2712−2722

62 Tampuu A, Matiisen T, Semikin M, Fishman D, Muhammad N. A survey of end-to-end driving: Architectures and training methods. *IEEE Transactions on Neural Networks and Learning Systems*, 2022, **33**(4): 1364−1384

63 Singh A. End-to-end autonomous driving using deep learning: A systematic review. arXiv preprint arXiv: 2311.18636, 2023.

64 Chen L, Wu P H, Chitta K, Jaeger B, Geiger A, Li H Y. End-to-end autonomous driving: Challenges and frontiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, **46**(12): 10164−10183

65 Xiao Y, Codevilla F, Gurram A, Urfalioglu O, López A M. Multimodal end-to-end autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 2022, **23**(1): 537−547

66 Bansal M, Krizhevsky A, Ogale A S. ChauffeurNet: Learning to drive by imitating the best and synthesizing the worst. In: Proceedings of the Robotics: Science and Systems 2019. Freiburg im Breisgau, Germany: RSS Foundation, 2019.

67 Wu P H, Jia X S, Chen L, Yan J C, Li H Y, Qiao Y. Trajectory-guided control prediction for end-to-end autonomous driving: A simple yet strong baseline. In: Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS). New Orleans, USA: Curran Associates Inc., 2022. Article No. 443

68 Pomerleau D A. ALVINN: An autonomous land vehicle in a neural network. In: Proceedings of the 2nd International Conference on Neural Information Processing Systems (NeurIPS). Denver, USA: Morgan Kaufmann, 1988. 305−313

69 Chitta K, Prakash A, Geiger A. NEAT: Neural attention fields for end-to-end autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, 2021. 15773−15783

70 Chitta K, Prakash A, Jaeger B, Yu Z H, Renz K, Geiger A. TransFuser: Imitation with Transformer-based sensor fusion for autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, **45**(11): 12878−12895

71 Dosovitskiy A, Ros G, Codevilla F, López A M, Koltun V. CARLA: An open urban driving simulator. In: Proceedings of the 1st Annual Conference on Robot Learning (CoRL). Mountain View, USA: PMLR, 2017. 1−16

72 Ye T J, Jing W, Hu C Y, Huang S K, Gao L P, Li F Z, et al. FusionAD: Multi-modality fusion for prediction and planning tasks of autonomous driving. arXiv preprint arXiv: 2308.01006, 2023.

73 Hu S C, Chen L, Wu P H, Li H Y, Yan J C, Tao D C. ST-P3: End-to-end vision-based autonomous driving via spatial-temporal feature learning. In: Proceedings of the 17th European Conference on Computer Vision (ECCV). Tel Aviv, Israel: Springer, 2022. 533−549

74 Jiang B, Chen S Y, Xu Q, Liao B C, Chen J J, Zhou H L, et al. VAD: Vectorized scene representation for efficient autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 8306−8316

75 Sun W C, Lin X W, Shi Y N, Zhang C, Wu H R, Zheng S F. SparseDrive: End-to-end autonomous driving via sparse scene representation. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Atlanta, USA: IEEE, 2025. 8795−8801

76 Jia X S, Yang Z J, Li Q F, Zhang Z Y, Yan J C. Bench2Drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. In: Proceedings of the 38th International Conference on Neural Information Processing Systems (NeurIPS). Vancouver, Canada: Curran Associates Inc., 2024. Article No. 25

77 Zhang B Z, Song N, Jin X, Zhang L. Bridging past and future: End-to-end autonomous driving with historical prediction and planning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2025. 6854−6863

78 Renz K, Chen L, Marcu A M, Hünermann J, Hanotte B, Karnsund A, et al. CarLLaVA: Vision language models for camera-only closed-loop driving. arXiv preprint arXiv: 2406.10165, 2024.

79 Lu Y H, Tu J D, Ma Y X, Zhu X G. ReAL-AD: Towards human-like reasoning in end-to-end autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Honolulu, USA: IEEE, 2025. 27783−27793

80 Wei J, Wang X Z, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-thought prompting elicits reasoning in large language models. In: Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS). New Orleans, USA: Curran Associates Inc., 2022. Article No. 1800

81 Jiang B, Chen S Y, Zhang Q, Liu W Y, Wang X G. AlphaDrive: Unleashing the power of VLMs in autonomous driving via reinforcement learning and reasoning. arXiv preprint arXiv: 2503.07608, 2025.

82 Kahneman D. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux, 2011.

83 Jia X S, Wu P H, Chen L, Xie J W, He C H, Yan J C, et al. Think twice before driving: Towards scalable decoders for end-to-end autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 21983−21994

84 Xu Z H, Zhang Y J, Xie E Z, Zhao Z, Guo Y, Wong K Y K, et al. DriveGPT4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*, 2024, **9**(10): 8186−8193

85 Wei J L, Yuan S S, Li P F, Hu Q D, Gan Z X, Ding W C. OccLLaMA: An occupancy-language-action generative world model for autonomous driving. arXiv preprint arXiv: 2409.03272, 2024.

86 Tian X Y, Gu J R, Li B L, Liu Y C, Wang Y, Zhao Z Y, et al. DriveVLM: The convergence of autonomous driving and large vision-language models. In: Proceedings of the 8th Conference on Robot Learning (CoRL). Munich, Germany: PMLR, 2024. 4698−4726

87 Zhao G S, Ni C J, Wang X F, Zhu Z, Zhang X Y, Wang Y D, et al. DriveDreamer4D: World models are effective data machines for 4D driving scene representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2025. 12015−12026

88 Graves A, Wayne G, Reynolds M, Harley T, Danihelka I, Grabska-Barwińska A, et al. Hybrid computing using a neural network with dynamic external memory. *Nature*, 2016, **538**(7626): 471−476

89 Wang Xiao, Zhang Xiang-Yu, Zhou Rui, Tian Yong-Lin, Wang Jian-Gong, Chen Long, et al. An intelligent architecture for cognitive autonomous driving based on parallel testing. *Acta Automatica Sinica*, 2024, **50**(2): 356−371
(王晓, 张翔宇, 周锐, 田永林, 王建功, 陈龙, 等. 基于平行测试的认知自动驾驶智能架构研究. 自动化学报, 2024, **50**(2): 356−371)

90 Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks. *Science*, 2006, **313**(5786): 504−507

91 Deng J, Dong W, Socher R, Li L J, Li K, Li F F. ImageNet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Miami, USA: IEEE, 2009. 248−255

92 Radford A, Narasimhan K, Salimans T, Sutskever I. Improving Language Understanding by Generative Pre-training, Technical Report, OpenAI, USA, 2018.

93 Devlin J, Chang M W, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional Transformers for language understanding. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, USA: Association for Computational Linguistics, 2019. 4171−4186

94 Kaplan J, McCandlish S, Henighan T, Brown T B, Chess B, Child R, et al. Scaling laws for neural language models. arXiv preprint arXiv: 2001.08361, 2020.

95 Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M A, Lacroix T, et al. LLaMA: Open and efficient foundation language models. arXiv preprint arXiv: 2302.13971, 2023.

96 Taylor R, Kardas M, Cucurull G, Scialom T, Hartshorn A, Saravia E, et al. Galactica: A large language model for science. arXiv preprint arXiv: 2211.09085, 2022.

97 Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, et al. PaLM: Scaling language modeling with pathways. *The Journal of Machine Learning Research*, 2023, **24**(1): Article No. 240

98 Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman F L, et al. GPT-4 technical report. arXiv preprint arXiv: 2303.08774, 2023.

99 Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C L, Mishkin P, et al. Training language models to follow instructions with human feedback. In: Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS). New Orleans, USA: Curran Associates Inc., 2022. Article No. 2011

100 Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning (ICML). Virtual Event: PMLR, 2021. 8748−8763

101 Alayrac J B, Donahue J, Luc P, Miech A, Barr I, Hasson Y, et al. Flamingo: A visual language model for few-shot learning. In: Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS). New Orleans, USA: Curran Associates Inc., 2022. Article No. 1723

102 Li J N, Li D X, Savarese S, Hoi S C H. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proceedings of the 40th International Conference on Machine Learning (ICML). Honolulu, USA: PMLR, 2023. 19730−19742

103 Yang Z Y, Li L J, Lin K, Wang J F, Lin C C, Liu Z C, et al. The dawn of LMMs: Preliminary explorations with GPT-4V(ision). arXiv preprint arXiv: 2309.17421, 2023.

104 Anil R, Borgeaud S, Alayrac J B, Yu J H, Soricut R, Schalkwyk J, et al. Gemini: A family of highly capable multimodal models. arXiv preprint arXiv: 2312.11805, 2023.

105 Liu H T, Li C Y, Wu Q Y, Lee Y J. Visual instruction tuning. In: Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS). New Orleans, USA: Curran Associates Inc., 2023. Article No. 1516

106 Zitkovich B, Yu T H, Xu S C, Xu P, Xiao T, Xia F, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In: Proceedings of the 7th Conference on Robot Learning (CoRL). Atlanta, USA: PMLR, 2023. 2165−2183

107 Reed S E, Zolna K, Parisotto E, Colmenarejo S G, Novikov A, Barth-Maron G, et al. A generalist agent. *Transactions on Machine Learning Research*, **2022**: 1−42

108 Kim M J, Pertsch K, Karamcheti S, Xiao T, Balakrishna A, Nair S, et al. OpenVLA: An open-source vision-language-action model. In: Proceedings of the 8th Conference on Robot Learning (CoRL). Munich, Germany: PMLR, 2024. 2679−2713

109 Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. LLaMA 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv: 2307.09288, 2023.

110 Hu E J, Shen Y L, Wallis P, Allen-Zhu Z, Li Y Z, Wang S A, et al. LoRA: Low-rank adaptation of large language models. In: Proceedings of the 10th International Conference on Learning Representations (ICLR). Virtual Event: OpenReview.net, 2022.

111 Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L. QLoRA: Efficient finetuning of quantized LLMs. In: Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS). New Orleans, USA: Curran Associates Inc., 2023. Article No. 441

112 Huang D W, Yan C, Li Q, Peng X J. From large language models to large multimodal models: A literature review. *Applied Sciences*, 2024, **14**(12): Article No. 5068

113 Gao H X, Wang Z R, Li Y Q, Long K W, Yang M, Shen Y Q. A survey for foundation models in autonomous driving. arXiv preprint arXiv: 2402.01105, 2024.

114 Fu D C, Li X, Wen L C, Dou M, Cai P L, Shi B T, et al. Drive like a human: Rethinking autonomous driving with large language models. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW). Waikoloa, USA: IEEE, 2024. 910−919

115 Mao J G, Qian Y X, Ye J J, Zhao H, Wang Y. GPT-driver: Learning to drive with GPT. arXiv preprint arXiv: 2310.01415, 2023.

116 Yao S Y, Zhao J, Yu D, Du N, Shafran I, Narasimhan K R, et al. ReAct: Synergizing reasoning and acting in language models. In: Proceedings of the 11th International Conference on Learning Representations (ICLR). Kigali, Rwanda: OpenReview.net, 2023.

117 Wen L C, Fu D C, Li X, Cai X Y, Ma T, Cai P L, et al. DiLu: A knowledge-driven approach to autonomous driving with large language models. In: Proceedings of the 12th International Conference on Learning Representations (ICLR). Vienna, Austria: OpenReview.net, 2024.

118 Sha H, Mu Y, Jiang Y X, Zhan G J, Chen L, Xu C F, et al. LanguageMPC: Large language models as decision makers for autonomous driving. arXiv preprint arXiv: 2310.03026, 2023.

119 Liao H C, Shen H M, Li Z N, Wang C Y, Li G F, Bie Y M, et al. GPT-4 enhanced multimodal grounding for autonomous driving: Leveraging cross-modal attention with large language models. *Communications in Transportation Research*, 2024, **4**: Article No. 100116

120 Cui C, Yang Z C, Zhou Y P, Ma Y S, Lu J W, Li L X, et al. Personalized autonomous driving with large language models: Field experiments. In: Proceedings of the 27th International Conference on Intelligent Transportation Systems (ITSC). Edmonton, Canada: IEEE, 2024. 20−27

121 Zhou X C, Liu M Y, Yurtsever E, Zagar B L, Zimmer W, Cao H, et al. Vision language models in autonomous driving: A survey and outlook. *IEEE Transactions on Intelligent Vehicles*, DOI: 10.1109/TIV.2024.3402136

122 Kirillov A, Mintun E, Ravi N, Mao H Z, Rolland C, Gustafson L, et al. Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 3992−4003

123 Liu S L, Zeng Z Y, Ren T H, Li F, Zhang H, Yang J, et al. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In: Proceedings of the 18th European Conference on Computer Vision (ECCV). Milan, Italy: Springer, 2024. 38−55

124 Ding X P, Han J H, Xu H, Zhang W, Li X M. HiLM-D: Enhancing MLLMs with multi-scale high-resolution details for autonomous driving. *International Journal of Computer Vision*, 2025, **133**(8): 5379−5395

125 Choudhary T, Dewangan V, Chandhok S, Priyadarshan S, Jain A, Singh A K, et al. Talk2BEV: Language-enhanced bird′s-eye view maps for autonomous driving. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Yokohama, Japan: IEEE, 2024. 16345−16352

126 Yang S Q, Liu J M, Zhang R R, Pan M J, Guo Z Y, Li X Q, et al. LiDAR-LLM: Exploring the potential of large language models for 3D LiDAR understanding. In: Proceedings of the 39th AAAI Conference on Artificial Intelligence (AAAI). Philadelphia, USA: AAAI Press, 2025. 9247−9255

127 Lübberstedt J, Rivera E, Uhlemann N, Lienkamp M. V3LMA: Visual 3D-enhanced language model for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Nashville, USA: IEEE, 2025. 4769−4778

128 Atakishiyev S, Salameh M, Yao H S, Goebel R. Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions. *IEEE Access*, 2024, **12**: 101603−101625

129 Kim J, Rohrbach A, Darrell T, Canny J, Akata Z. Textual explanations for self-driving vehicles. In: Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018. 577−593

130 Shalev-Shwartz S, Shammah S, Shashua A. On a formal model of safe and scalable self-driving cars. arXiv preprint arXiv: 1708.06374, 2017.

131 Cui E F, Wang W H, Li Z Q, Xie J W, Zou H M, Deng H M, et al. DriveMLM: Aligning multi-modal large language models with behavioral planning states for autonomous driving. *Visual Intelligence*, 2025, **3**(1): Article No. 22

132 Shao H, Hu Y X, Wang L T, Song G L, Waslander S L, Liu Y, et al. LMDrive: Closed-loop end-to-end driving with large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 15120−15130

133 Jiang Y F, Gupta A, Zhang Z C, Wang G Z, Dou Y Q, Chen Y J, et al. VIMA: Robot manipulation with multimodal prompts. In: Proceedings of the 40th International Conference on Machine Learning (ICML). Honolulu, USA: PMLR, 2023. 14975−15022

134 Zhou X C, Han X Y, Yang F, Ma Y P, Tresp V, Knoll A. OpenDriveVLA: Towards end-to-end autonomous driving with large vision language action model. In: Proceedings of the 40th AAAI Conference on Artificial Intelligence (AAAI). Singapore: AAAI Press, 2026. 13782−13790

135 Yang Z J, Chai Y L, Jia X S, Li Q F, Shao Y Q, Zhu X K, et al. DriveMoE: Mixture-of-experts for vision-language-action model in end-to-end autonomous driving. arXiv preprint arXiv: 2505.16278, 2025.

136 Zhang J W, Yang X, Wang T Q, Yao Y, Petiushko A, Li B. SafeAuto: Knowledge-enhanced safe autonomous driving with multimodal foundation models. In: Proceedings of the 42nd International Conference on Machine Learning (ICML). Vancouver, Canada: PMLR, 2025. 76497−76517

137 Yuan J H, Sun S Y, Omeiza D, Zhao B, Newman P, Kunze L, et al. RAG-driver: Generalisable driving explanations with retrieval-augmented in-context multi-modal large language model learning. In: Proceedings of the Robotics: Science and Systems XX (RSS). Delft, The Netherlands: RSS Foundation, 2024.

138 Arai H, Miwa K, Sasaki K, Watanabe K, Yamaguchi Y, Aoki S, et al. CoVLA: Comprehensive vision-language-action dataset for autonomous driving. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Tucson, USA: IEEE, 2025. 1933−1943

139 Gao X B, Wu Y H, Wang R J, Liu C X, Zhou Y, Tu Z Z. LangCoop: Collaborative driving with language. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Nashville, USA: IEEE, 2025. 4226−4237

140 Xu Z H, Bai Y, Zhang Y J, Li Z L, Xia F, Wong K Y K, et al. DriveGPT4-V2: Harnessing large language model capabilities for enhanced closed-loop autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2025. 17261−17270

141 Xing S, Qian C Y, Wang Y P, Hua H Y, Tian K X, Zhou Y, et al. OpenEMMA: Open-source multimodal model for end-to-end autonomous driving. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW). Tucson, USA: IEEE, 2025. 911−919

142 Renz K, Chen L, Arani E, Sinavski O. SimLingo: Vision-only closed-loop autonomous driving with language-action alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville,

USA: IEEE, 2025. 11993−12003

143 Wang T H, Maalouf A, Xiao W, Ban Y T, Amini A, Rosman G, et al. Drive anywhere: Generalizable end-to-end autonomous driving with multi-modal foundation models. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Yokohama, Japan: IEEE, 2024. 6687−6694

144 Fu H Y, Zhang D K, Zhao Z C, Cui J F, Liang D K, Zhang C, et al. ORION: A holistic end-to-end autonomous driving framework by vision-language instructed action generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Honolulu, USA: IEEE, 2025. 24823−24834

145 Chi H H, Gao H A, Liu Z M, Liu J N, Liu C Y, Li J W, et al. Impromptu VLA: Open weights and open data for driving vision-language-action models. In: Proceedings of the 39th Conference on Neural Information Processing Systems (NeurIPS). San Diego, USA: Curran Associates Inc., 2025.

146 Zhou Z W, Cai T H, Zhao S Z, Zhang Y, Huang Z Y, Zhou B L, et al. AutoVLA: A vision-language-action model for end-to-end autonomous driving with adaptive reasoning and reinforcement fine-tuning. In: Proceedings of the 39th Conference on Neural Information Processing Systems (NeurIPS). San Diego, USA: Curran Associates Inc., 2025.

147 Li Y, Tian M, Zhu D C, Zhu J T, Lin Z Y, Xiong Z W, et al. Drive-R1: Bridging reasoning and planning in VLMs for autonomous driving with reinforcement learning. In: Proceedings of the 40th AAAI Conference on Artificial Intelligence (AAAI). Singapore: AAAI Press, 2026. 6708−6716

148 Zeng S, Chang X Y, Xie M W, Liu X R, Bai Y F, Pan Z, et al. FutureSightDrive: Thinking visually with spatio-temporal CoT for autonomous driving. In: Proceedings of the 39th Conference on Neural Information Processing Systems (NeurIPS). San Diego, USA: Curran Associates Inc., 2025.

149 Jiang T T, Jiang X F, Ma Y, Wen X, Li B L, Zhan K, et al. The better you learn, the smarter you prune: Towards efficient vision-language-action models via differentiable token pruning. arXiv preprint arXiv: 2509.12594, 2025.

150 Chen L H, Hassani H, Nikan S. TS-VLM: Text-guided soft-sort pooling for vision-language models in multi-view driving reasoning. arXiv preprint arXiv: 2505.12670, 2025.

151 Zhou X R, Shan L L, Gui X L. DynRsl-VLM: Enhancing autonomous driving perception with dynamic resolution vision-language models. arXiv preprint arXiv: 2503.11265, 2025.

152 Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. In: Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS). Vancouver, Canada: Curran Associates Inc., 2020. Article No. 574

153 Ha D, Schmidhuber J. Recurrent world models facilitate policy evolution. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS). Montreal, Canada: Curran Associates Inc., 2018. 2455−2467

154 Wang S H, Yu Z D, Jiang X H, Lan S Y, Shi M, Chang N, et al. OmniDrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2025. 22442−22452

155 Russell L, Hu A, Bertoni L, Fedoseev G, Shotton J, Arani E, et al. GAIA-2: A controllable multi-view generative world model for autonomous driving. arXiv preprint arXiv: 2503.20523, 2025.

156 Zhou Y S, Simon M, Peng Z H, Mo S C, Zhu H Z, Guo M Y, et al. SimGen: Simulator-conditioned driving scene generation. In: Proceedings of the 38th International Conference on Neural Information Processing Systems (NeurIPS). Vancouver, Canada: Curran Associates Inc., 2024. Article No. 1548

157 Li X F, Zhang Y F, Ye X Q. DrivingDiffusion: Layout-guided multi-view driving scenarios video generation with latent diffusion model. In: Proceedings of the 18th European Conference on Computer Vision (ECCV). Milan, Italy: Springer, 2024. 469−485

158 Wang Y Q, He J W, Fan L, Li H X, Chen Y T, Zhang Z X. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 14749−14759

159 Wang X F, Zhu Z, Huang G, Chen X Z, Zhu J G, Lu J W. DriveDreamer: Towards real-world-drive world models for autonomous driving. In: Proceedings of the 18th European Conference on Computer Vision (ECCV). Milan, Italy: Springer, 2024. 55−72

160 Ding J T, Zhang Y K, Shang Y, Zhang Y H, Zong Z F, Feng J, et al. Understanding world or predicting future? A comprehensive survey of world models. *ACM Computing Surveys*, 2026, **58**(3): Article No. 57

161 Bruce J, Dennis M D, Edwards A, Parker-Holder J, Shi Y G, Hughes E, et al. Genie: Generative interactive environments. In: Proceedings of the 41st International Conference on Machine Learning (ICML). Vienna, Austria: PMLR, 2024. 4603−4623

162 Bardes A, Garrido Q, Ponce J, Chen X L, Rabbat M, LeCun Y, et al. Revisiting feature prediction for learning visual representations from video. arXiv preprint arXiv: 2404.08471, 2024.

163 Liu Y X, Zhang K, Li Y, Yan Z L, Gao C J, Chen R X, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. arXiv preprint arXiv: 2402.17177, 2024.

164 Hu A, Russell L, Yeo H, Murez Z, Fedoseev G, Kendall A, et al. GAIA-1: A generative world model for autonomous driving. arXiv preprint arXiv: 2309.17080, 2023.

165 Liang D K, Zhang D Y, Zhou X, Tu S F, Feng T R, Li X F, et al. Seeing the future, perceiving the future: A unified driving world model for future generation and perception. arXiv preprint arXiv: 2503.13587, 2025.

166 Hu A, Corrado G, Griffiths N, Murez Z, Gurau C, Yeo H, et al. Model-based imitation learning for urban driving. In: Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS). New Orleans, USA: Curran Associates Inc., 2022. Article No. 1505

167 Zheng W Z, Chen W L, Huang Y H, Zhang B R, Duan Y Q, Lu J W. OccWorld: Learning a 3D occupancy world model for autonomous driving. In: Proceedings of the 18th European Conference on Computer Vision (ECCV). Milan, Italy: Springer, 2024. 55−72

168 Zheng W Z, Song R Q, Guo X D, Zhang C M, Chen L. GenAD: Generative end-to-end autonomous driving. In: Proceedings of the 18th European Conference on Computer Vision (ECCV). Milan, Italy: Springer, 2024. 87−104

169 Fu H Y, Zhang D K, Zhao Z C, Cui J F, Xie H W, Wang B, et al. MindDrive: A vision-language-action model for autonomous driving via online reinforcement learning. arXiv preprint arXiv: 2512.13636, 2025.

170 Yang Lin-Yao, Chen Si-Yuan, Wang Xiao, Zhang Jun, Wang Cheng-Hong. Digital twins and parallel systems: State of the art, comparisons and prospect. *Acta Automatica Sinica*, 2019, **45**(11): 2001−2031

(杨林瑶, 陈思远, 王晓, 张俊, 王成红. 数字孪生与平行系统: 发展现状、对比及展望. 自动化学报, 2019, **45**(11): 2001−2031)

171 Yang Z, Chen Y, Wang J K, Manivasagam S, Ma W C, Yang A J, et al. UniSim: A neural closed-loop sensor simulator. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 1389−1399

172 Zhou X Y, Lin Z W, Shan X J, Wang Y T, Sun D Q, Yang M H. DrivingGaussian: Composite Gaussian splatting for surrounding dynamic autonomous driving scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 21634−21643

173 Gao H, Chen S Y, Jiang B, Liao B C, Shi Y, Guo X Y, et al. RAD: Training an end-to-end driving policy via large-scale 3DGS-based reinforcement learning. In: Proceedings of the 39th Conference on Neural Information Processing Systems (NeurIPS). San Diego, USA: Curran Associates Inc., 2025.

174 Zhang Z J, Liniger A, Dai D X, Yu F, van Gool L. TrafficBots: Towards world models for autonomous driving simulation and motion prediction. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). London, UK: IEEE, 2023. 1522−1529

175 Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Providence, USA: IEEE, 2012. 3354−3361

176 Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, et al. The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 3213−3223

177 Huang X Y, Cheng X J, Geng Q C, Cao B B, Zhou D F, Wang P, et al. The ApolloScape dataset for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Salt Lake City, USA: IEEE, 2018. 954−960

178 Yu F, Chen H F, Wang X, Xian W Q, Chen Y Y, Liu F C, et al. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 2633−2642

179 Richter S R, Hayder Z, Koltun V. Playing for benchmarks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 2232−2241

180 Ros G, Sellart L, Materzynska J, Vazquez D, López A M. The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 3234−3243

181 Sun P, Kretzschmar H, Dotiwalla X, Chouard A, Patnaik V, Tsui P, et al. Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 2443−2451

182 Sheeny M, de Pellegrin E, Mukherjee S, Ahrabian A, Wang S, Wallace A. RADIATE: A radar dataset for automotive perception in bad weather. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Xi′an, China: IEEE, 2021. 1−7

183 Deruyttere T, Vandenhende S, Grujicic D, van Gool L, Moens M F. Talk2Car: Taking control of your self-driving car. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, 2019. 2088−2098

184 Liu M Y, Yurtsever E, Fossaert J, Zhou X C, Zimmer W, Cui Y N, et al. A survey on autonomous driving datasets: Statistics, annotation quality, and a future outlook. *IEEE Transactions on Intelligent Vehicles*, 2024, **9**(11): 7138−7164

185 Krajewski R, Bock J, Kloeker L, Eckstein L. The highD dataset: A drone dataset of naturalistic vehicle trajectories on German highways for validation of highly automated driving systems. In: Proceedings of the 21st International Conference on Intelligent Transportation Systems (ITSC). Maui, USA: IEEE, 2018. 2118−2125

186 Moers T, Vater L, Krajewski R, Bock J, Zlocki A, Eckstein L. The exiD dataset: A real-world trajectory dataset of highly interactive highway scenarios in Germany. In: Proceedings of the IEEE Intelligent Vehicles Symposium (IV). Aachen, Germany: IEEE, 2022. 958−964

187 Rasouli A, Kotseruba I, Kunic T, Tsotsos J. PIE: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, South Korea: IEEE, 2019. 6261−6270

188 Chang M F, Lambert J, Sangkloy P, Singh J, Bak S, Hartnett A, et al. Argoverse: 3D tracking and forecasting with rich maps. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019. 8740−8749

189 Qian T W, Chen J J, Zhuo L H, Jiao Y, Jiang Y G. nuScenes-QA: A multi-modal visual question answering benchmark for autonomous driving scenario. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence (AAAI). Vancouver, Canada: AAAI Press, 2024. 4542−4550

190 Sima C H, Renz K, Chitta K, Chen L, Zhang H X, Xie C E, et al. DriveLM: Driving with graph visual question answering. In: Proceedings of the 18th European Conference on Computer Vision (ECCV). Milan, Italy: Springer, 2024. 256−274

191 Benekohal R F, Treiterer J. CARSIM: Car-following model for simulation of traffic in normal and stop-and-go conditions. *Transportation Research Record*, 1988, **1194**: 99−111

192 Hong C J, Aparow V R. System configuration of human-in-the-loop simulation for level 3 autonomous vehicle using IPG CarMaker. In: Proceedings of the IEEE International Conference on Internet of Things and Intelligence Systems (IoTaIS). Bandung, Indonesia: IEEE, 2021. 215−221

193 Shah S, Dey D, Lovett C, Kapoor A. AirSim: High-fidelity visual and physical simulation for autonomous vehicles. In: Proceedings of the 11th International Conference on Field and Service Robotics (FSR). Zurich, Switzerland: Springer, 2018. 621−635

194 Wu D M, Han W C, Liu Y F, Wang T C, Xu C Z, Zhang X Y, et al. Language prompt for autonomous driving. In: Proceedings of the 39th AAAI Conference on Artificial Intelligence (AAAI). Philadelphia, USA: AAAI Press, 2025. 8359−8367

195 Nie M, Peng R Y, Wang C W, Cai X Y, Han J H, Xu H, et al. Reason2Drive: Towards interpretable and chain-based reasoning for autonomous driving. In: Proceedings of the 18th European Conference on Computer Vision (ECCV). Milan, Italy: Springer, 2024. 292−308

196 Marcu A M, Chen L, Hünermann J, Karnsund A, Hanotte B, Chidananda P, et al. LingoQA: Visual question answering for autonomous driving. In: Proceedings of the 18th European Conference on Computer Vision (ECCV). Milan, Italy: Springer, 2024. 252−269

197 Liu P, Ning Q T, Lu X Y, Liu H P, Ma W L, She D G, et al. OmniReason: A temporal-guided vision-language-action framework for autonomous driving. arXiv preprint arXiv: 2509.00789, 2025.

198 Ettinger S, Cheng S Y, Caine B, Liu C X, Zhao H, Pradhan S, et al. Large scale interactive motion forecasting for autonomous driving: The Waymo open motion dataset. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, 2021. 9690−9699

199 Huang Y X, Lin Y B, Yue L S S, Yao Z H, Wang J. From gaze to movement: Predicting visual attention for autonomous driving human-machine interaction based on programmatic imitation learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Honolulu, USA: IEEE, 2025. 26146−26155

200 Zhou Y C, Tang J Y, Xiao X Y, Lin Y Y, Liu L K, Guo Z P, et al. Where, what, why: Towards explainable driver attention prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Honolulu, USA: IEEE, 2025. 2675−2685

201 Xia Y, Zhang D Q, Kim J, Nakayama K, Zipser K, Whitney D. Predicting driver attention in critical situations. In: Proceedings of the 14th Asian Conference on Computer Vision (ACCV). Perth, Australia: Springer, 2018. 658−674

202 Fang J W, Yan D X, Qiao J H, Xue J R, Wang H, Li S. DADA-2000: Can driving accident be predicted by driver attention? Analyzed by a benchmark. In: Proceedings of the IEEE Intelligent Transportation Systems Conference (ITSC). Auckland, New Zealand: IEEE, 2019. 4303−4309

203 Sachdeva E, Agarwal N, Chundi S, Roelofs S, Li J C, Kochenderfer M, et al. Rank2Tell: A multimodal driving dataset for joint importance ranking and reasoning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Waikoloa, USA: IEEE, 2024. 7498−7507

204 Xie S Y, Kong L D, Dong Y H, Sima C H, Zhang W W, Chen Q A, et al. Are VLMs ready for autonomous driving? An empirical study from the reliability, data, and metric perspectives. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Honolulu, USA: IEEE, 2025. 6285−6297

205 Li F H, Jin H, Gao B, Fan L Y, Jiang L H, Zeng L. NuGrounding: A multi-view 3D visual grounding framework in autonomous driving. arXiv preprint arXiv: 2503.22436, 2025.

206 Davidson C, Ramanan D, Peri N. RefAV: Towards planning-centric scenario mining. arXiv preprint arXiv: 2505.20981, 2025.

207 Wilson B, Qi W, Agarwal T, Lambert J, Singh J, Khandelwal S, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. arXiv preprint arXiv: 2301.00493, 2023.

208 Hawke J, Shen R, Gurau C, Sharma S, Reda D, Nikolov N, et al. Urban driving with conditional imitation learning. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Paris, France: IEEE, 2020. 251−257

209 Wang Y, Luo W J, Bai J J, Cao Y L, Che T, Chen K, et al. Alpamayo-R1: Bridging reasoning and action prediction for generalizable autonomous driving in the long tail. arXiv preprint arXiv: 2511.00088, 2025.

210 Zhang L J, Yuan Y J, Wu C J, Chang X Y, Cai X, Zeng S, et al. MindDriver: Introducing progressive multimodal reasoning for autonomous driving. arXiv preprint arXiv: 2602.21952, 2026.

211 Wang Y F, Wang T X, Yue T. Uncertainty propagation from sensor data to deep learning models in autonomous driving. *Information and Software Technology*, 2025, **183**: Article No. 107735

212 Feng D, Harakeh A, Waslander S L, Dietmayer K. A review and comparative study on probabilistic object detection in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 2022, **23**(8): 9961−9980

213 Lu D Y, Du H Y, Wu Z F, Yang S. Risk assessment in autonomous driving: A comprehensive survey of risk sources, methodologies, and system architectures. *Autonomous Intelligent Systems*, 2025, **5**(1): Article No. 24

214 Sensoy M, Kaplan L, Kandemir M. Evidential deep learning to quantify classification uncertainty. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS). Montreal, Canada: Curran Associates Inc., 2018. 3183−3193

215 Frantar E, Ashkboos S, Hoefler T, Alistarh D. OPTQ: Accurate quantization for generative pre-trained Transformers. In: Proceedings of the 11th International Conference on Learning Representations (ICLR). Kigali, Rwanda: OpenReview.net, 2023.

216 Sun L Y, Jiang J W, Ding Y F, Li F F, Song Y, Zhang H F, et al. Hardware co-design scaling laws via roofline modelling for on-device LLMs. arXiv preprint arXiv: 2602.10377, 2026.

217 Hu Yun-Feng, Qu Ting, Liu Jun, Shi Zhu-Qing, Zhu Bing, Cao Dong-Pu, et al. Human-machine cooperative control of intelligent vehicle: Recent developments and future perspectives. *Acta Automatica Sinica*, 2019, **45**(7): 1261−1280
(胡云峰, 曲婷, 刘俊, 施竹清, 朱冰, 曹东璞, 等. 智能汽车人机协同控制的研究现状与展望. 自动化学报, 2019, **45**(7): 1261−1280)



刘瑞琪 安徽大学人工智能学院硕士研究生. 2024 年获得安徽大学计算机科学与技术专业学士学位. 主要研究方向为社会计算, 认知自动驾驶.
E-mail: wa24301152@stu.ahu.edu.cn
(**LIU Rui-Qi** Master student at the School of Artificial Intelligence, Anhui University. He received his bachelor degree in computer science and technology from Anhui University in 2024. His research interests include social computing and cognitive autonomous driving.)



方 可 安徽大学人工智能学院硕士研究生. 2025 年获得兰州大学计算机科学与技术专业学士学位. 主要研究方向为社会计算, 认知自动驾驶.
E-mail: W125211033@stu.ahu.edu.cn
(**FANG Ke** Master student at the School of Artificial Intelligence, Anhui University. He received his bachelor degree in computer science and technology from Lanzhou University in 2025. His research interests include social computing and cognitive autonomous driving.)



王远大 安徽大学机器人学院副研究员. 2020 年获得东南大学控制科学与工程专业博士学位. 主要研究方向为深度强化学习, 具身智能和机器人系统控制. E-mail: wangyd@ahu.edu.cn
(**WANG Yuan-Da** Associate researcher at the School of Robotics, Anhui University. He received his Ph.D. degree in control science and engineering from Southeast University in 2020. His research interests include deep reinforcement learning, embodied intelligence, and robotic system control.)



王 晓 安徽大学机器人学院教授. 2016 年获得中国科学院大学社会计算专业博士学位. 主要研究方向为具身智能, 认知自动驾驶. 本文通信作者. E-mail: xiao.wang@ahu.edu.cn
(**WANG Xiao** Professor at the School of Robotics, Anhui University. She received her Ph.D. degree in social computing from the University of Chinese Academy of Sciences in 2016. Her research interests include embodied intelligence and cognitive autonomous driving. Corresponding author of this paper.)



孙长银 安徽大学人工智能学院教授. 2003 年获得东南大学控制科学与工程专业博士学位. 主要研究方向为具身智能, 模式识别和优化理论.
E-mail: cysun@seu.edu.cn
(**SUN Chang-Yin** Professor at the School of Artificial Intelligence, Anhui University. He received his Ph.D. degree in control science and engineering from Southeast University in 2003. His research interests include embodied intelligence, pattern recognition, and optimization theory.)