中图法分类号:TP39　　**文献标识码**: A　　**文章编号**: 1006-8961(XXXX)XX-0001-18

论文引用格式:Hu Jianfang, Huang Linjiang, Zhai Wei, Yan Ruisong, Li Chenglin, Zheng Weishi, He Ran, Zha Zhengjun, Xiong Hongkai. Intelligent Driving Foundation Model[J/OL]. Journal of Image and Graphics, XXXX:1-18. DOI: 10.11834/jig.260085. (胡建芳, 黄林江, 翟伟, 闫瑞松, 李成林, 郑伟诗, 赫然, 查正军, 熊红凯. 智能驾驶大模型[J/OL]. 中国图象图形学报,XXXX:1-18. DOI: 10.11834/jig.260085.)[DOI: 10.11834/jig.260085]

智能驾驶大模型

胡建芳[1*],黄林江[2*],翟伟[3*],闫瑞松[4],李成林[4],郑伟诗[1],赫然[5],查正军[3],熊红凯[4]

1. 中山大学计算机学院,广东广州　510275; 2. 北京航空航天大学人工智能学院,北京　100191; 3. 中国科学技术大学信息科学技术学院,安徽合肥　230026; 4. 上海交通大学信息与电子工程学院,上海　200240; 5. 中国科学院自动化研究所,北京　100190

摘　要: 智能驾驶大模型融合了视觉、语言与动作多模态学习,正引领自动驾驶从传统"感知—规划—控制"架构向端到端一体化演进。其统一表征、生成式推理及少样本泛化的能力,显著提升了智能驾驶系统的鲁棒性与决策智能。报告首先系统梳理了国际国内智能驾驶大模型领域的最新进展,包括决策规划、环境感知、视觉问答、数据生成等方面。其中,决策规划部分讨论了端到端可解释决策模型的兴起、多模态与序列化决策模型的融合以及世界模型与认知智能体的引入;环境感知部分从多模态感知与语义解释的融合、语言提示驱动的运动轨迹预测与行为理解两条主线出发进行探讨;视觉问答部分讨论了国内外研究者针对推理可解释性与决策验证提出的系列方法;数据生成部分则以数据来源为区分,探讨自动标注、生成式数据合成、世界模型、虚实一体仿真等手段如何解决自动驾驶数据收集成本高、长尾场景覆盖率不足的问题。在此基础上进行横向对比,分析了我国在数据资源、算力生态、算法创新与标准体系方面的优势与短板。面向未来,提出应强化基础研究与公共底座、完善可信AI评测体系、推进个性化驾驶与人机对齐、构建自主可控生态等建议。智能驾驶大模型已成为我国汽车产业高质量发展的关键突破口与人工智能应用的新高地。本文提及的算法及相关开源代码已汇总至:https://github.com/Ruisong-Yan/Intelligent-Driving-Foundation-Model,亦可通过https://www.scidb.cn/detail?dataSetId=3921ce7e24e44cf98428e3bc1494c410获取。

关键词: 智能驾驶;大模型;多模态学习;世界模型;端到端;可解释性

Intelligent Driving Foundation Model

Hu Jianfang[1*], Huang Linjiang[2*], Zhai Wei[3*], Yan Ruisong[4], Li Chenglin[4], Zheng Weishi[1], He Ran[5], Zha Zhengjun[3], Xiong Hongkai[4]

1. School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510275, China; 2. School of Artificial Intelligence, Beihang University, Beijing 100191, China; 3. School of Information Science and Technology, University of Science and Technology of China, Hefei 230026, China; 4. School of Information Science and Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China; 5. Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

Abstract: The intelligent driving foundation model integrates vision, language, and action through multimodal learning, driving the evolution of autonomous systems from the traditional "perception-planning-control" architecture towards an end-to-end unified paradigm. By leveraging the capabilities of unified representation, generative reasoning, and few-shot generalization, it significantly enhances the system robustness and decision-making intelligence. From the perspective of

收稿日期:2026-02-06;**修回日期**:2026-03-04

*通信作者:熊红凯,上海交通大学特聘教授,主要研究方向为信号处理、多媒体通信、机器学习等。E-mail: xionghongkai@sjtu.edu.cn;共同一作　†通信作者:熊红凯 xionghongkai@sjtu.edu.cn

基金项目:国家自然科学基金(项目编号:62431017;U24A20251;62320106003)

Supported by:National Natural Science Foundation of China(Grant No. 62431017;U24A20251;62320106003)

research, the intelligent driving foundation model incorporates achievements from multiple disciplines: the latest advances in visual computing, natural language processing, reinforcement learning, cognitive science, computer graphics, and virtual simulation are comprehensively applied within this system. At the industry level, global leading automotive and technology companies have also regarded large models as the technological cornerstone of the next generation of intelligent driving systems. This report systematically reviews the latest progress in the intelligent driving foundation model at both the international and domestic levels, including: the decision planning, environmental perception, visual question answering, and data generation. Specifically, the decision planning section is dedicated to achieving a direct mapping from perception inputs to planning outputs through a unified large model architecture, while maintaining explainability and generalization capabilities. On one hand, to address the shortcomings of traditional end-to-end driving models in terms of interpretability, generalization ability, and safety verification, researchers have proposed a series of approaches that incorporate tegrate language models with visual models, achieving "model thinking readability" through natural languages. On the other hand, large language models also provide a unified language-level interface for the collaborative optimization of multi-vehicle planning, enabling different vehicle agents to engage in semantic communication, share intentions and polices, and form a group intelligence similar to the "implicit collaboration" among human drivers. The task of perception and motion prediction section includes not only includes detecting surrounding objects, but also understanding environmental semantics, inferring the behavioral intentions of other traffic participants, and performing multi-target trajectory prediction in dynamic scenarios. Traditional perception systems rely on the dense labeling and geometric reconstruction models, which often experience performance degradation in long-tail scenarios (such as extreme weather, emergencies). To address these issues, the academic community has recently introduced large language models and multimodal fusion mechanisms, incorporating semantic reasoning into visual perception to achieve "semantic-enhanced visual understanding". This perception-semantic integration design significantly enhances the depth of understanding of complex environments by autonomous driving systems. Predictive capabilities can also be enhanced by introducing language reasoning. Some methods use language prompts as the semantic guidance, combining visual and motion features for future trajectory prediction, which are referred to as "language prompt guided prediction". In order to enable the model to explain and communicate with human drivers in natural language, people further introduced visual question answering into intelligent driving foundation models. With this approach, driving models can not only answer questions, such as "why is the vehicle slowing down" and "can we change lanes", but also adjust polices based on semantic questions, achieving an explainable and intervenable driving intelligence. Retrieval-augmented-generation and chain-of-thought techniques are applied as an effective means to enhance the question-answering capabilities into autonomous driving systems, this report discusses related methods in the visual question answering section. Data is a key driving factor for enhancing the capabilities of autonomous driving systems. High-quality, diverse, and semantically consistent driving data thud directly determines the generalization and safety performance of the model. Traditional data collection and annotation methods are extremely costly: for example, the annotation cost for urban-level autonomous driving can reach $3-5 per frame, with less than 5% coverage of long-tail scenarios. To address this issue, research focus has shifted to the automatic annotation, self-supervised learning, and generative data synthesis. The target of these methods is to reduce dependence on manual annotation, synthesize rare samples that are difficult to capture in the real world in the virtual space, and form a closed-loop data engine, enabling the co-evolution of models and data. The data generation section, distinguished by the data sources, explores how automatic annotation, generative data synthesis, world models, and integrated virtual and real simulation methods solve the problems of high cost and insufficient coverage of long-tail scenarios in autonomous driving data. On these bases, this report conducts a horizontal comparison and further analyzes China's strengths and limitations in terms of the data resources, computational infrastructure, algorithmic innovation, and standardization. International research is leading in the theoretical depth and integration of multimodal fusion, especially showing a great innovative potential in unified architecture, generative world models, and collective intelligence. While China has significant advantages in the engineering applications, real-time optimization, and scenario adaptation, particularly with a unique practical experience in data closed-loop, automatic annotation, and computational optimization. Looking forward to the future development trends, intelligent driving technology faces challenges in real-time, safety, and personalization. This report recommends strengthening the fundamental research and public infra-

structure, establishing a unified open-source data platform to promote the sharing and collaboration of multimodal data, building trustworthy AI evaluation systems, advancing personalized driving and human-AI alignment, and fostering an autonomous and controllable innovation ecosystem. Intelligent driving foundation models have become a crucial enabler for the high-quality development of China's automotive industry and a new frontier in applied artificial intelligence. The algorithms and open-source codes mentioned have been summarized at: https://github. com/Ruisong-Yan/Intelligent-Driving-Foundation-Model, and can also be accessed via: https://www. scidb. cn/detail? dataSetId=3921ce7e24e44cf98428e3bc1494c410.

Key words: intelligent driving; foundation model; multimodal learning; world model; end-to-end; interpretability

0 引 言

智能驾驶大模型(intelligent driving foundation model, IDFM)是近年来人工智能、计算机视觉与交通工程交叉融合的前沿方向。它代表着从算法范式到产业体系的系统性变革——即以统一的大规模模型取代分散的模块化组件,通过大数据驱动的跨模态表征与推理能力,实现从感知、决策到控制的端到端(end to end, E2E)闭环优化。该方向的快速发展不仅标志着自动驾驶技术进入"智能化2.0"阶段,也预示着人工智能在复杂物理世界中实现通用认知与决策的可行路径。

传统自动驾驶系统通常采用"感知—预测—规划—控制"的模块化架构。各模块之间通过中间接口传递有限语义特征,虽然便于调试与验证,但也引入了信息割裂、冗余计算、误差累积与跨模块优化困难等问题。尤其在城市驾驶场景中,交通参与者行为多样、道路拓扑复杂、环境变化剧烈,模块化系统难以在全局一致的目标下实现最优决策。近年来,大模型技术的兴起为这一困境提供了新的解决方案。借助 Transformer 架构的统一表征能力、世界模型的可解释推理机制以及视觉语言动作(vision - language - action, VLA)模型的跨模态理解能力,智能驾驶正从"任务驱动"转向"场景驱动",从"规则约束"走向"语义推理",由封闭域向开放世界演化。

智能驾驶大模型的核心特征在于"统一性"、"涌现性"和"泛化性"。所谓统一性,是指大模型以单一网络同时处理多源输入(摄像头、激光雷达、毫米波雷达、高精地图、语音指令等),在共享潜空间中进行感知、预测与决策的一体化学习;所谓涌现性,是指随着模型规模和数据量的提升,系统自发展现出人类未显式编程的能力,如自监督语义理解、场景问答、风险预测等;而泛化性,则体现为模型在未见过的场景中依然能够做出合理的驾驶决策,具备跨域迁移与零样本学习能力。这些特征使大模型成为突破复杂交通环境"长尾问题"的关键技术路径。

从研究视角看,智能驾驶大模型融合了多学科成果:视觉计算、自然语言处理、强化学习、认知科学、图像图形学与虚拟仿真等领域的最新进展在该体系中实现了综合应用。例如,世界模型(world model)以生成式方式重建交通环境,支持闭环仿真与反事实推理;多模态 Transformer 通过跨域对齐机制实现语义一致性学习;而强化学习与模仿学习的结合,使模型能够在虚拟世界中自我优化驾驶策略。这种"虚实共生"的研究框架,也推动了仿真图形学、三维重建和生成式人工智能(artificial intelligence, AI)在交通场景中的大规模应用,形成了从视觉生成到运动控制的全链条创新生态。

在产业层面,全球主流汽车与科技企业均已将大模型视为新一代智能驾驶系统的技术基石。例如,特斯拉的 FSD V12 采用端到端 Transformer 架构,将感知、规划与控制统一到单一网络中;Waymo 与 NVIDIA 通过世界模型实现虚拟环境的自监督训练;国内的华为乾崑 ADS 3.0、小鹏 XNet、比亚迪"天神之眼"系统等,也已在商用车辆中应用端到端学习框架。与此同时,百度 Apollo、地平线、商汤科技等企业在大模型语义理解、场景生成和安全评测方面持续布局,推动形成了涵盖数据采集、训练框架、车规芯片与安全法规的完整生态体系。

政策层面的推动亦为大模型的发展提供了强劲支撑。自《智能汽车创新发展战略》发布以来,国家层面明确提出"到2025年实现有条件自动驾驶规模化生产"的目标。北京市、深圳市等地陆续出台自动驾驶条例,首次允许 L3 级别私家车上路试点运行,为高风险、高复杂度的大模型测试提供了制度保障。

与此同时,国家智能网联汽车创新中心、开放测试示范区等基础设施建设,为海量数据采集与仿真验证提供了独特场景优势。这种“政策+场景+技术”的协同机制,使我国在智能驾驶大模型领域具备了快速迭代与先行落地的现实土壤。

总体而言,智能驾驶大模型不仅是技术路线的革新,更是产业范式的重构。它以统一的大模型为枢纽,连接算法、算力、数据与场景四大要素,实现从“单车智能”向“车路云一体化”的体系跃迁。其发展将深刻影响未来交通安全、能源效率与社会治理方式,对构建智慧城市、数字中国和交通强国具有里程碑意义。

本文系统梳理了国际国内智能驾驶大模型的最新进展,包括决策规划、环境感知、视觉问答、数据生成等方面。在此基础上,第3节和第4节横向比较了国内外研究进展,并展望了智能驾驶大模型未来的发展趋势与应用前景。章节结构如图1所示。

1 国际研究现状

国际上,智能驾驶大模型的研究已成为人工智能与自动驾驶技术融合的核心前沿。与传统模块化

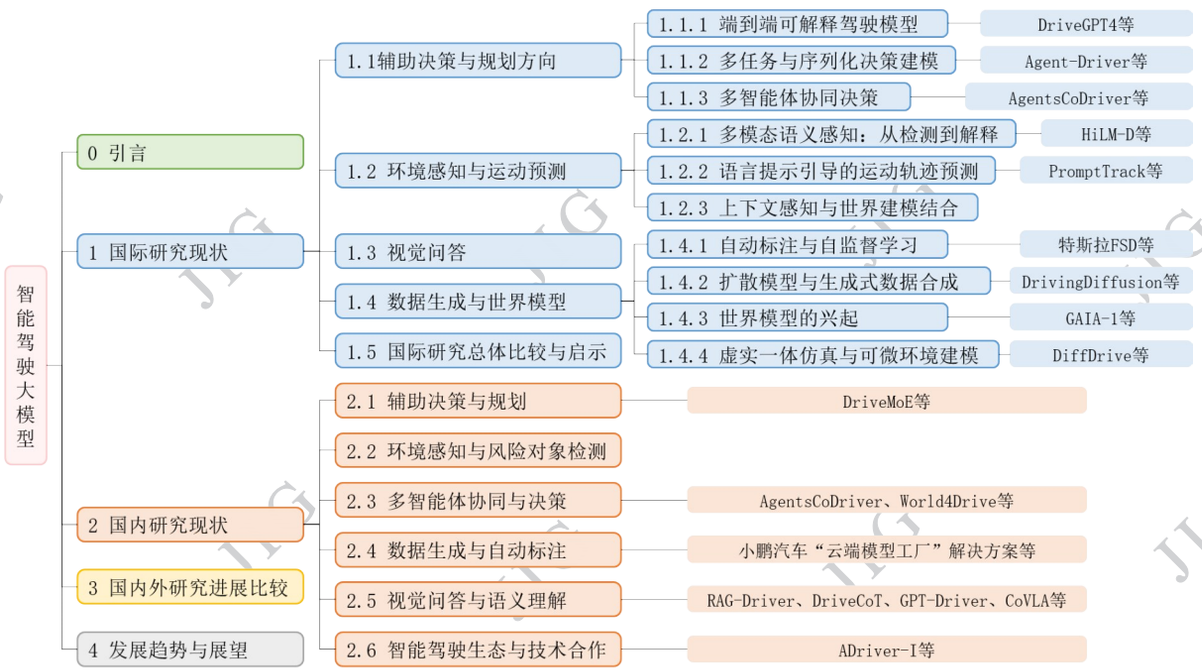

图1 本文章节结构

Fig. 1 Outline diagram of this paper

体系(perception - prediction - planning - control)相比,大模型驱动的端到端范式更加强调语义统一、任务共享与推理一致性。2022年起,欧美学术界与产业界纷纷将通用人工智能(artificial general intelligence,AGI)的思想引入自动驾驶领域,探索视觉-语言-动作多模态融合的统一模型结构。

整体上,国际研究发展趋势可概括为以下四个特征:

1)从模块化到统一模型(unified model)转变:以Transformer架构为核心的模型开始替代独立的感知、预测、规划模块,通过统一参数化与共享语义空间提升全局决策一致性。

2)从视觉特征到语义理解(semantic-level reasoning)转变:大模型具备语言推理与知识调用能力,能在复杂驾驶场景中进行解释性推理与多目标权衡。

3)从监督学习到自监督与世界建模(self-supervised world modeling)转变:生成式世界模型能够在虚拟空间中自我训练,减少对昂贵标注数据的依赖。

4)从单车智能到群体智能(collaborative multi-agent intelligence)转变:多智能体强化学习框架开始在交通仿真与策略协调中落地,形成“云-车-车”智能体交互格局。

在研究路径上，国际前沿主要集中于四大方向：辅助决策与规划（decision and planning）、环境感知与运动预测（perception and motion prediction）、视觉问答与语义解释（visual question answer and explainable interaction）、数据生成与世界模型（data generation and world model）。以下各节将基于典型国际研究工作进行系统梳理，并结合代表性论文分析主要成果与发展趋势。

1.1 辅助决策与规划方向（decision and planning）

辅助决策与规划是智能驾驶大模型的核心研究领域。该方向的总体目标是：以统一大模型架构实现从感知输入到行为输出的直接映射，同时保留可解释性与泛化能力。国际上，该方向的发展大致经历了三个阶段：端到端可解释决策模型的兴起（2022－2023）、多模态与序列化决策模型的融合（2023－2024）、世界模型与认知智能体的引入（2024－至今）。整体发展脉络与代表性工作如图2所示。

1.1.1 端到端可解释驾驶模型

传统的端到端驾驶模型虽然能够直接从感知输入预测控制信号，但在可解释性、泛化能力及安全验证方面存在不足。为解决这一问题，国际学界提出了一系列结合语言模型与视觉模型的方案，以增强可解释决策。

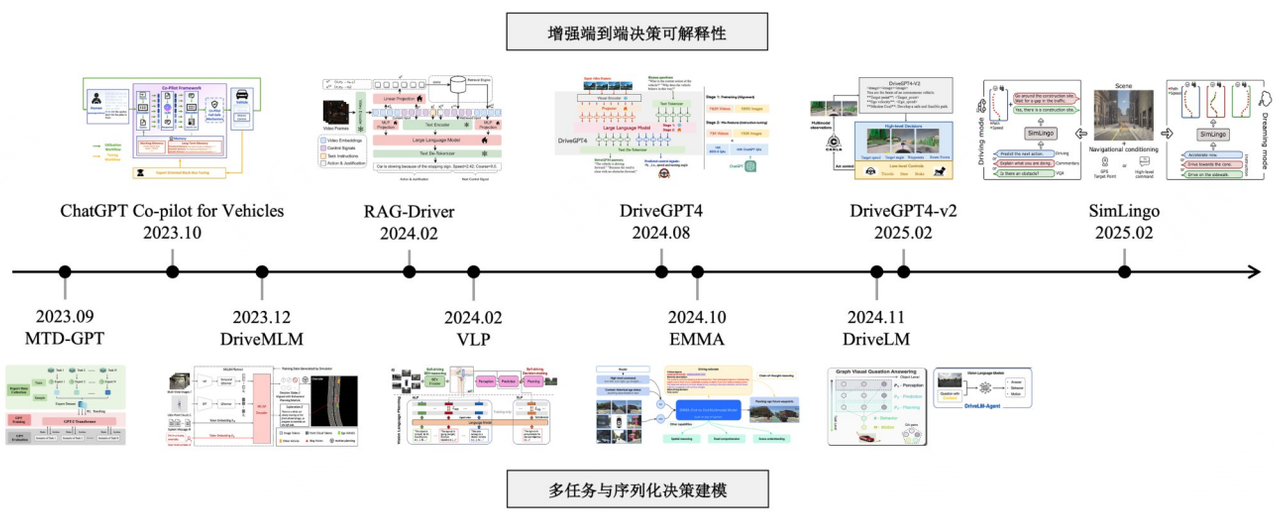

图2　智能驾驶大模型辅助决策与规划发展脉络

Fig. 2　The history of IDFM for Decision and Planning

DriveGPT4系列（Xu等，2024；Xu等，2025）是最具代表性的工作之一，由香港中文大学与清华大学合作团队提出（国际期刊IEEE Transactions on Intelligent Vehicles收录）。该模型在多模态大模型Valley（Luo等，2023）的基础上构建，结合视觉输入与自然语言描述，实现了可解释的端到端驾驶决策。DriveGPT4的核心创新是引入“视觉指令微调（visual instruction tuning）”机制：通过含语义标签和驾驶指令的数据集微调模型，使其能够在预测转向、加减速等控制信号的同时生成行动解释文本。例如，在测试场景中，模型可输出“减速以避让右侧行人”，显著提升了透明度与信任度。RAG-Driver框架（Yuan等，2024）由牛津大学Mobile Robotics Group开发，是另一项具有代表性的国际研究。该模型采用检索增强生成（retrieval-augmented generation，RAG）机制，从历史专家决策库中检索与当前场景相似的样本作为上下文提示，使大模型在生成控制信号时具备记忆引用能力。实验结果显示，RAG-Driver在未见过的新城市场景中实现了零样本驾驶能力，同时能生成自然语言解释，显著改善了可解释性问题。ChatGPT Co-pilot for Vehicles（Wang等，2023）则来自美国的Wang等人团队，是早期将通用大语言模型（large language model，LLM）嵌入驾驶系统的尝试。研究者将ChatGPT用作辅助驾驶对话引擎，能在驾驶过程中提供解释与建议，例如回答“前方为什么减速”或“是否可以变道”等问题。虽然该系统仍处于实验阶段，但标志着自然语言模型开始承担“认知副驾驶”的功能，为后续人机共驾系

统奠定基础。SimLingo(Renz 等,2025)通过语言-动作对齐机制,将视觉输入转为隐式的“语言式语义”,进而进行规划。模型采用视觉表示经过语言编码器转化为潜在的类文本标记,再与动作规划模块对齐,实现闭环自主驾驶,减少对真实语言输入的依赖。

这些研究表明,国际学界正在形成一种共识:自动驾驶大模型的核心竞争力,不仅在于预测精度,更在于可解释性与语义一致性。通过自然语言接口实现“模型思维可读化”,成为国际可解释驾驶的重要趋势。

1.1.2 多任务与序列化决策建模

随着 Transformer 架构在时序任务中的成功应用,国际学界开始探索将驾驶决策问题转化为语言或序列建模问题,从而实现多任务统一学习。

MTD-GPT(multi-task decision GPT)模型(Liu 等,2023)首次将无信号交叉口等复杂交通场景的多种决策任务转化为序列预测问题。研究团队通过设计统一的文本模板,将不同驾驶任务(如左转、超车、避让)编码为语言序列,使模型能够在共享参数下执行多类动作。这一思路将“多任务学习”与“大语言模型”结合,突破了传统强化学习模型在任务特化上的瓶颈。Agent-Driver 框架(Mao 等,2023)则由斯坦福大学与 NVIDIA 合作提出。该模型将 LLM 的推理能力与强化学习的策略优化结合,引入认知记忆体(cognitive memory unit)与推理引擎(reasoning engine),支持在不同场景下生成逻辑一致的动作计划。实验结果表明,Agent-Driver 在 nuScenes 数据集(Caesar 等,2020)上的决策准确率较传统基准提升 15%,且能输出自然语言解释说明决策依据。DriveLM 模型(Sima 等,2023)(由德国图宾根大学与香港大学联合提出)进一步推进了“语言化驾驶”理念。研究团队将车辆轨迹离散化为文本 token,通过“轨迹分词器(trajectory tokenizer)”和图结构问答模块实现多轮逻辑推理。DriveLM 能在多轮问答后得出最终控制决策,体现了语言模型在复杂决策树推理中的潜力。而 DriveMLM(Wang 等,2023)(由加州大学圣地亚哥分校与清华联合团队提出)则强调语言输出与控制动作之间的对齐关系。研究者将 Apollo 系统中定义的驾驶动作(如 accelerate、brake、turn_left)编码为行为状态集合,并将 LLM 输出与这些状态映射,从而实现真正意义上的“语言到控制”桥接。DriveMLM 被认为是“语义驾驶”范式的重要突破之一。最新的视觉语言规划(vision-language-planning,VLP)模型(Pan 等,2024)进一步提出了两种新范式:智能体学习范式(agent-wise learning paradigm,ALP):通过比较模型生成的鸟瞰图(bird's eye view,BEV)与真实地图,提高模型对环境推理的几何一致性;自车为中心学习范式(self-driving-car-centric learning paradigm,SLP):以自车为中心对齐视觉与文本规划特征,强化策略表示。该工作发表于 CVPR 2024,被视为多模态规划领域的重要里程碑。此外,DME-Driver 框架(Han 等,2024)则来自澳门大学与美国密歇根大学合作团队。研究者将人类决策逻辑与三维场景感知模型结合,提出“决策与感知双流融合(decision-perception co-learning)”机制,实现可解释且可执行的规划输出。该方法显著提升了复杂交互场景下的决策稳定性。EMMA(Hwang 等,2024)通过多模态 Transformer 将视觉、地图、运动历史、车辆状态等多源信息统一到一个时空特征空间中,实现从 检测 → 跟踪 → 预测 的统一表达。其感知模块不仅执行传统的检测与轨迹估计,还强调高维语义解释能力(意图、交互关系),属于典型的“多模态语义理解驱动的感知预测系统”。

综上可见,国际上在辅助决策与规划方向的研究呈现以下趋势:语言化与序列化成为核心思路,将驾驶行为映射为文本序列以便统一学习。记忆增强与多模态融合提高模型推理一致性与可解释性。从可解释到可审计的决策链逐步建立,为监管和责任界定提供基础。

1.1.3 多智能体协同决策

在多车协作与交通博弈场景下,单车最优决策并不等同于全局最优。为应对群体智能场景,国际学界提出了多智能体协同驱动架构。

AgentsCoDriver(Hu 等,2024)是由美国佛罗里达大学与华盛顿大学合作开发的典型系统。该框架利用 LLM 驱动的多智能体结构,使不同车辆代理(agents)能够进行语义通信、共享意图与策略,形成类似人类驾驶员之间“隐性协同”的群体智能。模型包括推理引擎、认知记忆、强化反射(reinforced reflex)以及通信模块,支持实时博弈决策。实验显示,该系统在密集交通与交互场景下表现优于单智能体模型。LangCoop(Gao 等,2025)强调多车辆协作,通过语言作为中间表征来传递协作意图(如让道、合流)。模型使用 VLM 将视觉场景转换为自然

语言式驾车意图，再利用语言通信实现多车规划的协作优化。

这一方向的意义在于：为车—车—云协同控制提供了语言层面的统一接口；拓展了大模型从单智能体推理到社会智能推理的能力。国际学界普遍认为，多智能体 LLM 将成为未来车路协同（vehicle to everything，V2X）与车联网场景的核心技术支撑。

1.2 环境感知与运动预测（perception and motion prediction）

感知与预测是自动驾驶系统的基础环节，其任务不仅包括检测周围物体，还需理解环境语义、推测其他交通参与者的行为意图，并在动态场景下进行多目标轨迹预测。传统的感知体系依赖稠密标注与几何重建模型，往往在长尾场景（如极端天气、突发事件）中性能下降，例如基于 BEV 的表征学习方法如 BEVFormer（Li 等，2022）和 DETR3D（Wang 等，2022）。为解决这些问题，国际学界近年引入大语言模型（LLM）与多模态融合机制，将语义推理引入视觉感知，从而实现“语义增强的视觉理解”。国际研究可分为两条主线：多模态感知与语义解释的融合；语言提示驱动的运动轨迹预测与行为理解。

1.2.1 多模态语义感知：从检测到解释

HiLM-D 模型（Ding 等，2023）是该方向的重要代表。该模型由美国华盛顿大学与香港科技大学合作提出，旨在让感知模块具备“解释性输出”能力。研究者将高分辨率图像输入与语言描述相结合，通过跨模态 Transformer 结构实现同时检测、定位与语义解释。例如，当模型检测到“前方有障碍物”时，能够进一步输出语言解释：“车辆在前方 15 米处检测到静止自行车，应减速通过”。这种感知—语义一体化设计显著提高了自动驾驶系统对复杂环境的理解深度。

与传统基于卷积的 BEV 感知不同，HiLM-D 的多模态特征在语义空间中共享表示，使得视觉模型能够借助语言知识库补全信息缺失。在单风险对象数据集上，其检测准确率接近完全监督方法，同时在语义解释维度上具备更高鲁棒性。尽管论文指出其数据集（每段视频仅含单一风险目标）仍有局限，但该研究首次验证了“语言驱动的视觉理解”在自动驾驶感知任务中的可行性，为国际后续工作提供了范式基础。SafeAuto（Zhang 等，2025）通过引入安全知识库（交通规则、意外情况）增强自动驾驶决策。利用多模态基础模型提取视觉与语言语义，再通过知识增强模块对策略做安全约束，使动作输出具备法规一致性和安全优先性。

1.2.2 语言提示引导的运动轨迹预测

在复杂交通场景中，预测周围车辆与行人未来的运动轨迹是实现安全决策的关键。传统方法依赖动力学模型或纯视觉 Transformer，在应对突发事件时容易缺乏上下文推理。国际学界因此提出了“语言提示（prompt）引导的预测模型”，通过引入自然语言上下文增强模型的语义推理能力。

PromptTrack 模型（Wu 等，2023）由英国帝国理工学院与清华大学联合团队开发，是首个将自然语言提示引入 3D 运动预测任务的国际工作。该模型构建了基于 Transformer 的时序预测网络，以语言提示作为语义引导，结合视觉与运动特征进行未来轨迹预测。其训练集“NuPrompt”提供了多场景、多语义提示的匹配样本，如“前方车辆即将左转”“行人加速横穿道路”等，使模型能够以文本方式理解并预测行为。实验表明，在 nuScenes 数据集上，PromptTrack 的 ADE（平均误差）比传统 Transformer 模型降低约 12%，显著提升了预测精度。

在更复杂的驾驶行为预测任务中，LC-LLM 框架（Peng 等，2024）由美国华盛顿大学与加州大学戴维斯分校团队提出，针对车辆变道行为进行可解释预测。模型在理解语义提示后可同时输出轨迹预测结果与行为解释，如“右前方有慢车，计划变道超越”。LC-LLM 的贡献在于引入语言描述与行为信号的对齐机制，使预测结果兼具数值准确性与语义可解释性，被国际同行视为“语言化运动预测”的典型范式。DiffVLA（Jiang 等，2025）将扩散模型用于规划决策，通过视觉-语言提示（如“保持车道行驶”）指导扩散过程生成连续的未来轨迹或控制序列。算法利用 VLM 的语义理解能力 + 扩散模型的轨迹生成能力，实现高鲁棒性的规划。

1.2.3 上下文感知与世界建模结合

2024 年，Zheng 等人（2024）在《arXiv preprint arXiv：2403.11057》发表了开创性工作，提出上下文感知运动预测模型（context-aware motion prediction）。研究者利用 GPT-4V 的语义理解能力解析复杂交通语境，将其输出作为“高层语义上下文”输入至运动预测模型 MTR（Shi 等，2023）。该设计有效提升了模型在交叉路口、多车交互等长尾场景下的

泛化性能。这项研究的意义在于首次证明:通用多模态大模型(如 GPT-4V)具备理解驾驶场景语境的能力,可为下游预测模块提供可解释先验。这标志着世界模型(world model)与感知预测的耦合成为国际研究新方向。

从整体上看,国际在感知与预测领域的研究展现出三大趋势:多模态语义融合成为主流:从纯视觉检测向语言增强的语义感知转变(如 HiLM-D)(Ding 等,2023)。语言提示化的预测模型兴起:利用 Prompt 与思维链(chain of thought,CoT)推理弥补数据稀疏性(如 PromptTrack)(Wu 等,2023;Peng 等,2024)。通用模型与领域模型耦合:将 GPT-4V 等通用模型输出引入专用 Transformer 模块,实现开放语义对接(Zheng 等,2024)。这些研究共同推动了"理解型感知(comprehension-based perception)"的形成,即模型不仅"看见物体",更能"理解场景语义",体现了国际自动驾驶系统由感知驱动向推理驱动的演化趋势。

1.3 视觉问答(visual question answering,VQA)与可解释交互

在传统自动驾驶系统中,模型的行为决策往往是"黑箱式"的,即系统无法对外解释其动作依据。为此,国际学界提出了"视觉问答(VQA)+大模型"框架,让模型能以自然语言与人类进行交流与解释。VQA 模型不仅能回答诸如"为什么车辆减速""能否变道"等问题,还能根据语义提问调整策略,实现可解释与可干预的驾驶智能,如图 3 所示。

视觉语言模型(visual language model,VLM)在驾驶可解释性中的早期探索来自南加州大学的

GPT-Driver(Mao 等,2023),该模型首次将 GPT 类大语言模型直接用于自动驾驶,通过'视觉指令微调'机制,在输出转向、加减速控制信号的同时,生成'减速以避让右侧行人'这类自然语言解释,开创了'控制 + 解释'双输出的端到端范式,也为后续可解释驾驶模型提供了设计思路。最具代表性的研究是牛津大学提出的 RAG-Driver 框架(Yuan 等,2024)。该模型将检索增强生成机制引入自动驾驶 VQA 系统。其工作原理为:从历史驾驶数据库中检索相似场景的专家示范;将这些示范的语义嵌入作为提示输入多模态 LLM;模型在生成驾驶信号的同时输出自然语言解释。这一机制让系统在面对未见场景时仍可借鉴相似案例进行合理推理,实现了零样本驾驶与自然语言解释双统一。实验结果显示,RAG-Driver 在多城市测试集上相较端到端基线模型提升 9% 的成功率,同时输出的语言解释被 85% 的受试者评为"可信"或"合理"。该框架被视为国际上"可解释自动驾驶"最具代表性的工作之一。CoVLA(Arai 等,2025)提供面向 VLA 任务的大规模视觉-语言-动作数据集。包括场景图像、语言描述(如意图、规则)、对应驾驶控制。其目的在于提升模型在语言引导规划、解释问答、场景生成等任务的训练能力。

图 3 智能驾驶大模型 VQA 任务示意图(Li 等,2025)

Fig. 3 Illustration of IDFM for VQA task (Li et al., 2025)

在 VQA 的推理机制中,思维链(CoT)技术的引入极大地增强了模型的逻辑透明性。国际研究普遍认为,CoT 能将复杂常识推理分解为逐步推断过程,有助于在动态交通情境下实现合理判断。例如,当模型接收到语义提示"雨天路滑"时,CoT 机制可依次推断:①路面摩擦力下降 → ②刹车距离增加 → ③应提前减速并增大车距,从而将常识性认知转化为控制信号。这种推理路径在欧美学术界被称为

“解释可视化路径(reasoning visualization path)”。目前，多家国际研究机构(如 MIT CSAIL、斯坦福 AI Lab)正在探索将 CoT 嵌入 BEV+Transformer 架构中，以实现逻辑推理可视化与决策验证。

视觉问答模型的另一个关键挑战是实时性。通用大模型在处理视觉与语言联合输入时计算量极大，难以直接部署于车载平台。为此，特斯拉在其 2023 - 2024 年的 FSD V12 系统中采用了 BEV+Transformer 架构结合模型蒸馏技术(Shi 等，2023)。研究团队将云端大模型压缩为轻量级版本(参数减少约 10 倍)，在不显著损失精度的情况下，将推理时延控制在 100 毫秒级别，从而实现了实时 VQA 响应。这一实践表明：蒸馏与边缘优化是推动大模型“上车”的关键工程方向。

为了使自动驾驶模型具备持续适应新场景的能力，国际学界提出将持续学习(continual learning)与世界模型结合。代表性工作包括 R-UniAD 框架，该方法由香港大学与新加坡国立大学研究团队提出，借助强化学习策略让世界模型持续探索与更新知识结构。模型在每次任务执行后将新经验融入潜在语义空间，实现“终身学习(lifelong learning)”。这种能力使模型能在未见场景下通过类比推理做出合理决策，从而具备零样本泛化性能。国际学界认为，这类持续学习范式为未来“开放世界自动驾驶”提供了关键思路。

1.4 数据生成与世界模型(data generation and world model)

在自动驾驶大模型体系中，数据是推动能力提升的关键驱动因素。高质量、多样化、语义一致的驾驶数据直接决定了模型的泛化与安全性能。传统数据采集与标注方式成本极高：例如城市级自动驾驶标注成本可达每帧 3 - 5 美元，长尾场景覆盖率不足 5%。为应对这一问题，国际研究重点转向自动标注、自监督学习与生成式数据合成。这一方向的核心目标是：减少人工标注依赖；在虚拟空间合成真实世界难以捕获的稀有样本；形成闭环数据引擎，使模型—数据共同演化。

1.4.1 自动标注与自监督学习

自动标注(auto-labeling)与自监督学习(self-supervised learning)是国际研究的两大支撑机制。例如，特斯拉自 2023 年起在其全自动驾驶系统 FSD Beta 中部署了自监督标注引擎：利用数百万辆车的行驶数据，通过冗余视角匹配与时序一致性算法自动生成语义标签。该体系大幅降低了人工标注成本，同时持续扩充了稀有场景样本库。

学术界也提出了一系列自监督框架，用于在无标注条件下学习视觉 - 语言对齐表示。例如 NVIDIA 的 UniAD(unified autonomous driving)系列研究证明，通过 BEV Transformer 融合时间一致性损失函数，可显著提升模型在弱标注环境下的目标识别与轨迹预测能力。S4-Driver(Xie 等，2025)强调利用语言任务(叙述、推理、问答)作为自监督目标来增强时序视觉特征学习，使模型能够在复杂交通场景中进行上下文相关的运动预测与风险推断。语言任务在训练中扮演“提示(prompt)”的角色，提升模型对动态场景的理解与未来轨迹预测能力，因此更适合归类为语言提示增强的时序预测范式。虽然这些系统属于工业闭源，但其核心思想与前期研究报告中的分析一致：通过自监督学习提升数据利用效率，构建端到端的可扩展数据生态体系。

1.4.2 扩散模型与生成式数据合成

近年来，生成式模型(generative model)在计算机视觉领域的突破，为自动驾驶提供了全新的数据生成范式。DrivingDiffusion 模型(Li 等，2024)是该方向的代表工作之一。研究团队提出一种“基于场景布局引导的多视角扩散生成框架”，可根据预设的交通场景描述生成逼真的驾驶视频。其核心创新在于：使用潜空间扩散模型(latent diffusion model，LDM)生成图像序列，确保视觉连续性；将车辆位置、行人分布、道路结构等语义要素嵌入条件编码器，形成“语义驱动的视频生成”；支持自由视角渲染与天气、光照控制，覆盖现实中难以采集的极端情境。DrivingDiffusion 在生成的多视角视频中实现了结构一致性与物理合理性，并在 Waymo Open Dataset 的仿真测试中显著扩展了数据多样性。该研究首次将扩散模型从图像生成成功迁移至驾驶视频合成领域，为世界模型训练提供了新的数据来源。

1.4.3 世界模型(world model)的兴起

世界模型(world model)是国际智能驾驶研究的另一大前沿。其核心思想是让模型在潜空间中模拟物理世界的状态转移，从而在虚拟环境中实现推理、规划与仿真。

国际上最具影响力的工作是 GAIA-1(generative artificial intelligence for autonomous driving)(Hu

等,2023),由英国初创公司 Wayve 在 2024 年推出。GAIA-1 是首个结合视频、语言与动作信号的生成式世界模型,能在端到端系统中执行感知、预测与控制的闭环任务。GAIA-1 的结构包括三部分:多模态感知编码器:接收视频帧、文本描述与驾驶指令输入;潜空间世界模型(latent world model):在隐变量空间内学习状态转移方程,用于预测未来环境演化;生成式控制器:基于潜变量生成未来帧与控制信号,实现视觉—决策同步预测。该模型的一个重要突破在于:无需显式高精地图或先验规则,便能在仿真环境中生成真实可控的驾驶视频和动作序列。

在国际评测中,GAIA-1 实现了对多车交互、雨雪夜间场景的高保真模拟,被视为"世界模型在自动驾驶领域的 GPT 时刻"。Wayve 进一步提出"Data Engine 2.0"体系,强调模型与数据的双向进化:模型生成新数据 → 数据再训练模型 → 不断逼近真实驾驶分布。

1.4.4 虚实一体仿真与可微环境建模

除生成模型外,世界模型还被用于构建可微仿真平台,使端到端系统在虚拟环境中实现闭环优化。美国 MIT、斯坦福及 ETH Zürich 的研究团队正在联合开发"Differentiable Simulation for Driving (DiffDrive)"平台,该系统将物理引擎与神经网络环境生成器相结合,使模型能在仿真世界中进行梯度可导的策略学习。这种"可微世界"框架使自动驾驶策略优化从试错式采样转向连续可优化空间,显著提升了数据效率。与 GAIA-1 的生成式世界模型不同,DiffDrive 更注重物理一致性与强化学习结合,被认为是通向"具身智能(embodied intelligence)"的重要桥梁。OpenDriveVLA(Zhou 等,2025)提出将视觉-语言-动作(VLA)模型集成到端到端自动驾驶任务中。模型利用多模态编码器融合视觉输入与语言指令,然后通过动作解码器输出低层控制信号。其核心算法结构基于 VLM + 动作规划的统一框架,最终实现端到端的控制预测。

1.5 国际研究总体比较与启示

综合分析可知,国际智能驾驶大模型研究已形成从感知到决策、从世界建模到解释推理的完整体系,具有以下四大技术特征:

1)端到端统一架构(unified E2E framework)

以 Transformer 为核心,融合视觉、语言与动作三模态,通过单一参数化体系实现从传感器到控制的直接映射(Xu 等,2024;Pan 等,2024)。与传统的层级结构相比,E2E 模型显著降低了误差传递,提高了系统整体可优化性。

2)语言驱动的可解释决策(language-driven explainability)

通过自然语言接口实现模型决策的语义透明化,典型代表包括 RAG-Driver(Yuan 等,2024)、DriveLM(Sima 等,2023)等。此外,思维链(CoT)推理技术正逐步嵌入规划层,为安全监管提供可视化路径。

3)生成式世界模型与仿真验证(generative world modeling)

Wayve GAIA-1(Hu 等,2023)等工作表明,生成式世界模型可在虚拟空间中实现大规模训练与风险评估,为减少实车测试依赖提供新途径。这种虚实一体结构正成为国际研究的主流方向。

4)多智能体协同与群体智能(collaborative multi-agent intelligence)

AgentsCoDriver(Hu 等,2024)的提出标志着自动驾驶大模型开始具备"社会智能"特征,即多车体之间的语义通信与协同决策。

尽管进展迅速,国际上智能驾驶大模型的研究仍存在若干核心挑战:

可解释性与安全验证:黑箱化 E2E 模型的安全性评估仍是监管难题,国际标准化组织(ISO/PAS 8800:2024)正讨论引入模型级日志与语义审计机制。**分布漂移与数据偏差:**生成式数据存在"自循环"风险(A→B→A 模型评估闭环),需建立可信数据认证体系。**计算效率与车载部署:**如何在算力受限的边缘平台上运行超大模型仍是瓶颈,特斯拉与 NVIDIA 正探索张量剪枝与结构蒸馏方案。**伦理与责任划分:**国际法规尚未明确模型错误决策的责任归属问题。欧盟、美国交通安全局(NHTSA)均提出"AI 责任透明化"原则,但缺乏统一评测指标。这些共识表明,大模型驱动的智能驾驶不仅是技术革命,也涉及数据治理与伦理监管的新挑战。

2 国内研究进展

随着智能驾驶技术的快速发展,国内在这一领域的研究取得了显著进展。尤其是在大模型技术的引入后,自动驾驶系统逐渐从传统的模块化架构向更加智能化、一体化的决策体系转型。国内的研究

主要集中在以下几个领域：

2.1 辅助决策与规划

智能驾驶的决策与规划是保证驾驶安全和效率的关键领域。国内在这一领域的研究取得了显著进展，尤其是在如何利用大模型提升决策精度与可解释性方面。传统的自动驾驶系统通常采用模块化架构，各模块之间的分工明确，但这种架构也暴露了许多问题，如跨模块信息割裂、冗余计算和误差积累等。因此，越来越多的研究者开始尝试将大模型应用于决策与规划中，以实现更高效的端到端学习和决策推理。陈妍妍等人(2024)系统梳理了端到端自动驾驶系统的结构和发展。

国内的研究者探索了如何通过大语言模型(LLM)与自动驾驶决策过程结合，提升决策过程的透明度和可解释性。例如，某些研究提出了利用大语言模型的推理能力来增强决策系统的智能化，使得驾驶决策不仅能够提高精度，还能够提供明确的决策依据，增加决策过程的透明度和可解释性(Xu等，2024)。这些研究旨在通过结合大语言模型的推理能力，使得自动驾驶系统在面对复杂交通环境时，能够做出更加合理的决策，并通过可解释性增强驾驶员对系统的信任。

在决策系统的设计中，国内还出现了一些针对历史决策经验的研究。通过引入记忆模块，系统能够“记住”过去的决策经验，从而在遇到类似情况时能够进行更合理的推理与决策。这一技术的应用，不仅提升了自动驾驶系统的适应能力，也增加了其在复杂情境下的鲁棒性(Luo等，2023)。例如，通过记忆模块，系统能够在应对城市道路复杂情况时，结合历史经验做出更加合理的调整，从而提升决策的精度。再例如，上海交通大学提出的DriveMoE(Yang等，2025)提出将混合专家模型(MoE)引入VLA模型，以提升复杂驾驶场景下的泛化能力。不同专家网络处理不同驾驶模式(如跟车、超车、避障)，通过门控网络选择合适专家。使模型在不同场景下自动切换最优策略。

2.2 环境感知与风险对象检测

环境感知是智能驾驶系统的核心模块之一，国内的研究者尝试将大模型引入到环境感知中，以提升系统对复杂驾驶环境的适应能力。传统的环境感知系统通常依赖单一的传感器或信息源，但随着大模型技术的引入，多模态数据的融合成为可能。视觉相机数据更具语义丰富性，而激光雷达数据更具深度准确性，多模态数据的融合可以提升感知的鲁棒性和准确性(李熙莹等，2023)。国内的研究工作集中在如何通过大模型将视觉、激光雷达、高清地图等多模态数据进行深度融合，从而提高感知系统对动态风险对象(如行人、其他车辆等)的检测精度和实时反应能力。

例如，国内的小鹏汽车在其“云端模型工厂”项目中，利用自动标注技术大幅降低了数据标注的成本，为环境感知提供了更为高效的支持。这一项目不仅提高了数据处理的效率，也加速了模型的迭代更新，从而提升了智能驾驶系统在复杂环境中的感知能力(Wang等，2023)。在目标检测与分类方面，国内的研究也取得了许多创新性进展。研究表明，通过将大语言模型与视觉感知系统结合，可以有效提升自动驾驶系统对复杂场景中风险对象的实时识别与定位能力(Yuan等，2024)。

在风险对象检测方面，国内的研究尤其注重如何通过大模型提高对复杂驾驶环境中潜在风险对象的检测精度。例如，在高速公路、城市路段等复杂场景中，如何及时识别并定位行人、突然出现的障碍物及其他交通参与者是当前的研究热点。多模态大语言模型的引入，不仅提升了系统的感知精度，也增强了系统在高风险场景中的应对能力(Hu等，2024)。

2.3 多智能体协同与决策

随着智能驾驶技术的不断发展，交通环境逐渐从单一智能体的驾驶转变为多个智能体(如多辆自动驾驶汽车)的协同决策。国内的研究者也开始探索如何在复杂交通场景中实现多个智能体的协同与决策，这不仅能提高系统的安全性，也能提升交通流的整体效率。多智能体协同决策在城市拥堵、交叉口调度等复杂场景中具有重要的应用价值。

国内提出的多智能体协同框架(如AgentsCo-Driver)采用了大语言模型作为核心推理引擎，通过智能体之间的通信与协作，优化决策过程。这些研究通过强化学习和多智能体博弈理论，使得智能体能够在多变的环境中进行信息共享与联合决策，从而避免了单一智能体在复杂场景下决策失误的问题(Hu等，2024)。此外，通过引入大语言模型的推理能力，系统能够更好地处理多个智能体之间的相互影响与冲突，提升决策的合理性和协调性。例如中科院自动化所与理想汽车合作的World4Drive

(Zheng 等,2025),融合 VLM 空间语义先验构建意图感知世界模型,使多车在无通信协议的情况下,通过场景语义推理实现‘隐性协同’,碰撞率较单智能体方案下降 46.7%,尤其适用于城市无保护路口的多车交互场景。香港大学团队则聚焦 V2X 协同场景(Yu 等,2025),通过端到端架构融合车路协同数据,将无保护左转场景的通过率提升至 96%,解决了单车感知的盲区问题,为‘云 - 车 - 路’一体化决策提供了端到端实现路径。

多智能体协同与决策的研究还包括如何通过强化学习、模仿学习等方法,让系统能够在新的交通环境中进行实时学习和自我优化。这一领域的研究为未来智能驾驶系统的普及和智能交通的实现提供了新的解决方案。

2.4 数据生成与自动标注

高质量的数据是智能驾驶系统成功的关键,然而传统的数据采集和标注方式成本高昂且效率低下。为此,国内的研究者开始探索如何通过生成模型和自动标注技术降低数据采集和标注的成本,并加速模型的训练与迭代。基于生成对抗网络(generative adversarial networks,GANs)和扩散模型的生成技术,研究者能够合成多样化的驾驶场景数据,特别是在长尾场景和极端条件下的数据生成,以提升系统在极端场景下的可靠性(刘江帆等,2025)。

例如,国内的小鹏汽车的“云端模型工厂”便是一个成功的案例。该项目通过自动标注技术显著降低了每张图像的标注成本,从而加速了模型训练的迭代。这一技术不仅提高了数据处理的效率,也为大规模的自动驾驶数据集的构建提供了可行的解决方案(Wen 等,2023)。此外,国内一些研究也关注如何利用生成模型合成多视角、多场景的驾驶数据,丰富模型的训练数据集。这些合成的数据能够在现实中难以获得的场景中,帮助训练出更加鲁棒的自动驾驶系统。

2.5 视觉问答与语义理解

视觉问答(VQA)技术在自动驾驶系统中的应用,显著提升了系统与驾驶员的自然语言交互能力,使得自动驾驶能够在复杂的驾驶环境中理解并应对驾驶员的需求。国内研究者在这一领域的探索,集中在如何利用大模型提升系统在复杂驾驶场景中的推理能力、决策能力及其可解释性。其中,检索增强技术(RAG)被认为是提升自动驾驶视觉问答可解释性和决策可靠性的有效途径。RAG 技术通过引入检索增强的上下文学习,使得模型能够从预先构建的专家示范库中检索出与当前场景相似的案例,进而参考这些案例生成驾驶控制信号,并同时给出可理解的决策解释。例如,牛津大学提出的 RAG-Driver 框架通过这种方式提高了决策过程的透明度,使得复杂驾驶问答的可解释性得到了显著增强。

除了检索增强技术,思维链推理(CoT)也是提升自动驾驶系统问答能力的有效手段。思维链推理通过模拟人类的思维过程,能够将复杂的常识性判断转化为具体的行动决策。国内的小鹏汽车研发的驾驶认知模型便应用了思维链推理机制,使得模型能够在面对如“雨天路滑”这种常识性知识时,自动推理得出“增大跟车距离”的决策,从而将抽象的常识转化为实际的驾驶行为调整。这种推理方式使得模型在面对复杂情境时,能够做出符合人类逻辑的决策,提高了自动驾驶系统的理解能力与应对效率(Zeng 等,2025;Liao 等,2025;Wang 等,2024;Nie 等,2024)。表 1 对代表性工作采用的 CoT 类型及主要贡献进行了对比分析。

然而,大模型在执行视觉问答时通常需要较长的计算时间,自动驾驶系统对时效性的要求非常严格。为了满足这一需求,模型蒸馏技术被提出用于降低推理时延。模型蒸馏通过对大模型进行压缩和简化,大幅度减少计算开销,进而提高系统的实时性。例如,特斯拉采用了 BEV+Transformer 架构的感知与决策模型,并通过蒸馏技术将云端的大模型压缩为轻量级版本,使其能够在车载硬件上以接近实时的速度运行。精简后的模型能够在保证决策准确性的同时,满足自动驾驶对毫秒级响应的要求。

随着自动驾驶技术逐步接入更加复杂的驾驶环境,如何应对海量长尾场景和不断涌现的新情况成为一大挑战。为此,持续学习(增量学习)技术的引入,使得模型在不遗忘已有知识的前提下,能够不断扩展其知识边界。这使得模型能够在面对训练阶段未见过的驾驶情境时,利用已积累的类似经验进行类比推理,从而做到零样本下的合理决策。例如,商汤科技提出的 R-UniAD 框架,采用强化学习策略让模型自主探索复杂场景,并通过持续学习逐步提升对罕见场景的适应能力。这种方法能够显著提升系统在面对新的驾驶情境时的应对能力,使其具备更加广泛的适应性。华中科学技术大学提出的

ReCogDrive（Li 等，2025）采用“认知推理 + 强化学习”结合的思路。模型具备“CoT 解释”功能，通过语言模块解释驾驶决策，并将解释的逻辑用于调节策略学习。该框架强调解释性与决策一致性。

综上所述，国内在视觉问答技术应用方面，结合了检索增强、思维链推理、模型蒸馏和持续学习等多种方法，极大地提升了自动驾驶系统的智能化水平、决策精度和实时性。随着这些技术的不断发展与完善，未来智能驾驶系统将能够更好地应对复杂驾驶场景，提高系统的可解释性、实时性和适应性，进一步推动自动驾驶技术的普及和发展。

2.6 智能驾驶生态与技术合作

在构建智能驾驶生态方面，国内不仅注重技术的突破，也重视产业链的协同。国内如华为、百度等企业正在加大对智能驾驶技术的投资，并推动大模型技术在自动驾驶领域的应用。同时，国家层面的政策也为智能驾驶的发展提供了支持。随着基础设施的逐步完善，智能驾驶系统的商用化进程正在加速。比如，旷视科技提出的 ADriver-I（Jia 等，2023）构建大规模自动驾驶世界模型（world model），学习从视频中提取动态过程，并能够在潜空间中进行场景“rollout”。可用于闭环模拟、数据补全、未来预测与策略训练。

在智能驾驶大模型的应用中，国内还面临着数据孤岛问题、算力不足等挑战。尽管如此，随着产业链的完善和技术的不断进步，国内的智能驾驶大模型技术有望在未来几年迎来快速增长，并在全球智能驾驶领域占据重要地位。

3 国内外研究进展比较

在国际智能驾驶大模型研究的推进中，端到端统一模型逐渐替代了传统的模块化体系（如感知、预测、规划和控制）。这种模式强调语义统一、任务共享与推理一致性，特别是通过大规模预训练模型（如 Transformer 架构）将感知、决策和控制的任务结合为一个统一的系统，从而提高了全局决策的一致性与可解释性。与此同时，多模态融合（视觉、语言、动作）与世界模型的引入使得大模型能够进行更加复杂的推理与规划，推动了自动驾驶技术从传统的“模块化”向“多任务共享”的范式转变。相比之下，国内在智能驾驶大模型研究中，更

表 1 智能驾驶大模型 VQA 任务的代表性工作

Table. 1 Representative works of IDFM for VQA task

方法名称	主要研究单位	COT	具体描述	年份
GPT-Driver	南加州大学	\	首次将 GPT 类大语言模型直接用于自动驾驶，开创了“控制 + 解释”双输出的端到端范式。	2023
RAG-Driver	牛津大学	\	将“检索增强生成机制”引入自动驾驶 VQA 系统，从历史驾驶数据库中检索相似场景的专家示范，作为提示输入模型。	2024
DriveCoT	香港大学	文本	提出了包含 CoT 推理过程的自动驾驶数据集，可用于评估 CoT 的准确性和最终决策。	2024
Reason2Drive	复旦大学	文本	提出了包含超过 60 万视频-文本对的 VQA 数据集，并针对 CoT 的推理性能提出新的评估指标。	2024
CoVLA	图灵公司	文本	提供面向 VLA 任务的大规模视觉-语言-动作数据集，可用于自动驾驶 VLA 系统的训练和评估。	2025
FutureSightDrive	西安交通大学	视觉	将“视觉时空 CoT”引入自动驾驶，使得 VLA 系统能够通过预测未来帧进行视觉思考。	2025
Cot-Drive	澳门大学	文本	使用大模型为驾驶场景生成推理描述，并通过知识蒸馏训练一个轻量级模型，方便端侧部署。	2025
ReCogDrive	华中科技大学	文本	开源自动驾驶 VLA 模型，VLM 部分用于编码和推理，以指导扩散规划器部分生成轨迹。	2025

注：表中“\”表示该方法未使用思维链（CoT）推理。

多聚焦于技术的落地与工程化。尽管国内在可解释性、数据闭环和量产部署方面已取得重要进展，但在模型的统一性和推理的深度整合方面，依然面临挑战。国内的研究多为模块化突破，虽然在自动标注、数据平台建设和多场景适配方面的进展快速，但体系化融合与推理一致性的深度整合相较国际略显薄弱，整体上还没有形成国际那样高度一体化的框架。

模型与学习范式：国际研究在智能驾驶领域的技术发展，尤其是在大模型的推理能力与学习范式上，呈现出较为系统的进展。以生成式世界模型为代表的技术，正在推动智能驾驶向更深层次的多任务和多模态融合发展。例如，DriveGPT4 和 VLP 等模型，通过将驾驶任务转化为语言或序列化的决策任务，实现了大模型的多任务共享学习，从而突破了传统强化学习模型在任务特化上的局限。此外，世界模型的引入不仅使得模型能够在虚拟环境中进行仿真训练，还通过生成控制信号和物理推演增强了系统的推理与规划能力，为闭环决策提供了新的思路。而国内的研究则强调可解释性与实时性，尤其是在大模型的推理速度与计算成本方面，国内注重通过模型蒸馏、轻量化与边缘优化来降低推理时延，确保大模型能够在车载硬件上运行。RAG（检索增强生成）、CoT（思维链推理）和持续学习等技术已经在国内的研究中初步应用，并取得了积极进展，但在模型的深度整合与推理一致性上仍存在一定差距。国内的研究更多关注工程化与实用性，尤其是对复杂场景下的应对能力和实时决策的优化。

数据与评测生态：国际上的智能驾驶大模型研究普遍拥有较为完善的数据与评测体系，尤其是高质量、跨城市和多场景的数据采集与评估框架。随着大语言模型（LLM）和生成模型（如生成式世界模型）的引入，国际学界已经实现了基于虚拟环境的自我训练与自监督学习，有效减少了对昂贵标注数据的依赖。同时，国际研究还通过标准化评测和开放基准来确保模型的可靠性与可验证性，例如 RAG-Driver 和 DriveLM 等模型在多城市环境下的广泛测试，能够反映出自动驾驶系统在复杂场景中的表现。国内在这一领域的研究重点则集中在数据闭环和自动标注技术的创新应用，尤其是通过自动标注引擎和生成模型来降低数据采集和标注的成本，加速大模型的迭代。尽管国内在数据处理和实时反馈优化方面取得了较快的进展，但在高质量、多样化的开放数据集建设和统一的评测框架方面，与国际领先水平相比仍有差距。国内需要加强数据管理体系的建设，提升开放数据集的质量和覆盖面。

可解释性与人机共驾：国际上的智能驾驶大模型研究特别强调可解释性，尤其是在端到端决策模型中，通过自然语言接口和检索增强生成（RAG）技术来实现模型决策的语义透明化。例如，RAG-Driver 和 DriveGPT4 都在视觉问答（VQA）系统中加入了自然语言解释和决策过程的可追溯性，这使得自动驾驶系统不仅能够执行任务，还能够向驾驶员提供详细的决策理由，增强了系统的可审计性。国内的研究则主要集中在提升驾驶认知与交互能力方面，尤其是在 VQA 技术与可解释驾驶之间的结合，国内研究者通过引入思维链推理（CoT）与 RAG 来增强系统的推理能力，并通过实时反馈优化来提高决策的可执行性。然而，尽管国内在可解释性方面取得了进展，但与国际在推理路径可视化和解释的深度方面，尚存在一定差距。未来，国内可以借鉴国际的研究经验，通过多模态融合和思维链推理的进一步发展，提升自动驾驶系统的推理透明度和可验证性。

实时部署与工程化：在自动驾驶的工程化应用上，国内的优势十分明显，尤其是在实时部署与计算效率方面。国内的研究者将大模型与边缘计算相结合，利用模型蒸馏和轻量化网络，成功将大规模模型压缩为适用于车载平台的轻量版，从而满足了自动驾驶对实时性的严苛要求。此外，国内的智能驾驶系统在复杂道路环境适配、实时数据处理和场景定制化方面展现了强大的工程能力，能够迅速应用于实际的车载平台。相比之下，国际上的智能驾驶大模型更加注重多智能体协同与群体智能的研究，在 V2X 通信和交通仿真中实现了智能体之间的语义协作和行为协调。例如，AgentsCoDriver 提出了多车协作的语义通信和决策共享机制，推动了群体智能的发展。但在实时性和算力约束方面，国际研究仍需进一步优化和整合。

多智能体与协同智能：随着智能驾驶技术的进步，多智能体协同决策成为新的研究热点，尤其是在城市复杂交通环境中，单车决策往往难以实现全局最优。国际上，基于多智能体强化学习的协同框架，如 AgentsCoDriver，使得不同车辆之间能够通过语义

通信共享决策策略，提高了整体的交通效率与安全性。这一方向的研究在 V2X 通信和交通博弈中得到广泛应用，形成了云－车－车的智能体协作网络。国内也开始关注这一方向，提出了基于大语言模型的多智能体协同框架，通过智能体之间的协作和博弈来优化交通决策。这些研究不仅提升了自动驾驶系统的智能化水平，也为未来智能交通系统的建设奠定了基础。

综合来看，国际与国内在智能驾驶大模型研究方面各有优势和特点。国际在理论深度和多模态融合的整合上走在前列，尤其在统一架构、生成式世界模型和群体智能方面展现了巨大的创新潜力。国内则在工程化应用、实时性优化和场景适配上具有显著优势，尤其是在数据闭环、自动标注与计算优化方面具有独到的实践经验。未来，国内研究可以借鉴国际在可解释性和推理深度方面的经验，进一步推动技术的整合与创新，并在多智能体协同和世界模型的构建方面加强跨学科合作。同时，国际研究可以从国内的实践经验中汲取灵感，在更高效的工程化与量产部署方面进一步发展。

4　发展趋势与展望

智能驾驶大模型作为当前汽车行业与人工智能技术交汇的前沿领域，正快速成为全球竞争的重要战略制高点。随着自动驾驶技术的不断进步，大模型技术的崛起为解决自动驾驶中的核心问题提供了新机遇，也促使智能驾驶系统向更高层次的融合发展。未来，智能驾驶大模型将在多个领域展现出重要的研究进展和广泛的应用前景。

随着自动驾驶技术的进步，感知与决策过程逐渐从传统的模块化架构向感知决策一体化演进。传统的分模块架构在复杂环境下容易暴露出误差积累、长尾场景难以覆盖以及可解释性差等问题，而大模型凭借强大的推理能力和统一的表征能力，能够在多个领域实现端到端一体化。这种一体化的方案通过视觉、语言和动作的融合网络，在减少信息损耗的同时，提高了系统的整体性能和适应性。这一趋势的推动将使智能驾驶能够更加高效地应对复杂的道路环境，提高安全性和决策的可靠性。

此外，生成式世界模型的引入，为智能驾驶大模型的训练提供了重要的创新解决方案。生成式模型能够在虚拟空间中生产高保真、长尾化的训练样本，降低数据采集的成本并缩短算法的迭代周期。随着虚拟仿真技术的发展，智能驾驶系统能够在模拟环境中进行广泛的场景训练，包括应对稀有碰撞、极端天气等情境，这显著减少了对实际路况数据的依赖，从而加速了模型的迭代进程并提升了系统的适应能力。

然而，随着大模型的应用，智能驾驶技术面临的挑战也日益显著。首先是实时性和安全性问题。为了满足大规模智能驾驶的需求，如何将大模型推理时延压缩至百毫秒以内，成为技术突破的关键。此外，智能驾驶系统在处理复杂交通环境时，必须具备冗余安全设计和可验证的安全边界，确保在遇到突发情况时能够稳定运行。因此，如何确保大模型在确保安全的前提下具备实时性，是未来技术发展的重要方向。

随着智能驾驶技术的不断成熟，个性化驾驶的需求也日益增长。未来的大模型将能够根据驾驶者的偏好进行在线微调，结合个性化的风格模板支持驾驶过程中的决策调整。通过对驾驶者行为的建模，大模型将能够精准对接驾驶者的需求，在提升驾驶体验的同时，保持系统的安全性和稳定性。长期记忆和短期记忆的协同工作将显著提升跨区域的适应能力，帮助系统快速适应不同地区的交通规则和道路文化差异，从而进一步推动智能驾驶的普及。

为了更好地推动智能驾驶大模型的进步，行业需要建设统一的开源数据平台，推动多模态数据的共享与协同。这不仅有助于解决数据孤岛问题，也能为大规模训练和模型优化提供必要的支持。此外，智能驾驶技术的标准化建设也变得尤为重要，尤其是在场景标签、评测标准和安全性评估方面的统一，将有助于行业快速发展并提升全球技术竞争力。通过加强国内外标准的协调和一致性建设，中国将在全球智能驾驶技术的竞争中占据有利位置。

未来，智能驾驶大模型的研究和应用将不仅限于技术创新，更多的社会、政策和经济层面的考虑将推动这一领域的快速发展。随着国家政策的支持和行业的快速发展，智能驾驶大模型将成为推动汽车产业升级和实现智能交通系统的关键技术。同时，智能驾驶大模型所产生的海量数据也将成为国家重要的战略资源，在保障数字主权和信息安全方面发挥重要作用。

综上所述，智能驾驶大模型的未来充满潜力。随着技术的不断突破，智能驾驶系统将更加智能化、个性化且安全性更高。面对这一发展趋势，国家和企业应加强基础研究与产业协同，推动技术创新与标准化建设，从而抢占智能驾驶大模型的未来发展制高点。

致谢：本文由中国图象图形学学会视觉大数据专业委员会组织撰写，该专业委员会链接为https://www.csig.org.cn/16/201704/49323.html，在此一并致谢。

参考文献

Arai H, Miwa K, Sasaki K, Watanabe K, Yamaguchi Y, Aoki S and Yamamoto I. 2025. Covla: comprehensive vision-language-action dataset for autonomous driving//2025 IEEE/CVF Winter Conference on Applications of Computer Vision. Tucson, USA: IEEE: 1933 - 1943 [DOI: 10.1109/WACV61041.2025.00195]

Caesar H, Bankiti V, Lang A H, Vora S, Liong V E, Xu Q, Krishnan A, Pan Y, Baldan G and Beijbom O. 2020. Nuscenes: a multimodal dataset for autonomous driving//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 11621-11631 [DOI: 10.1109/CVPR42600.2020.01164]

Chen Y Y, Tian D X, Lin C M and Yin H B. 2024. Survey of end-to-end autonomous driving systems. Journal of Image and Graphics, 29(11): 3216-3237 (陈妍妍, 田大新, 林椿眄, 殷鸿博. 2024. 端到端自动驾驶系统研究综述. 中国图象图形学报, 29(11): 3216-3237) [DOI: 10.11834/jig.230787]

Ding X P, Han J H, Xu H, Zhang W and Li X M. 2023. Hilm-d: towards high-resolution understanding in multimodal large language models for autonomous driving [EB/OL]. [2025-04-15]. https://arxiv.org/pdf/2309.05186v1

Gao X B, Wu Y H, Wang R J, Liu C X, Zhou Y and Tu Z Z. 2025. Langcoop: collaborative driving with language//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Tennessee, USA: IEEE: 4226 - 4237 [DOI: 10.1109/CVPRW67362.2025.00406]

Han W C, Guo D Q, Xu C Z, and Shen J B. 2024. Dme-driver: integrating human decision logic and 3d scene perception in autonomous driving [EB/OL]. [2024-01-08]. https://arxiv.org/pdf/2401.03641

Hu A, Russell L, Yeo H, Murez Z, Fedoseev G, Kendall A, Shotton J and Corrado G. 2023. Gaia-1: a generative world model for autonomous driving [EB/OL]. [2023-09-29]. https://arxiv.org/pdf/2309.17080

Hu S K, Fang Z R, Fang Z H, Chen X H and Fang Y G. 2024. Agentscodriver: large language model empowered collaborative driving with lifelong learning [EB/OL]. [2024-04-21]. https://arxiv.org/pdf/2404.06345

Hwang J J, Xu R S, Lin H, Hung W C, Ji J W, Choi K, Huang D, He T, Covington P, Sapp B, Zhou Y, Guo J, Anguelov D and Tan M. 2024. Emma: end-to-end multimodal model for autonomous driving. [EB/OL]. [2024-10-30]. https://arxiv.org/pdf/2410.23262

Jia F, Mao W X, Liu Y F, Zhao Y C, Wen Y Q, Zhang C, Zhang X Y and Wang T C. 2023. Adriver-i: a general world model for autonomous driving. [EB/OL]. [2023-11-22]. https://arxiv.org/pdf/2311.13549

Jiang A Q, Gao Y, Sun Z G, Wang Y R, Wang J J, Chai J H, Cao Q, Heng Y W, Jiang H, Dong Y D, Zhang Z Z, Guo X D, Sun H and Zhao H. 2025. Diffvla: vision-language guided diffusion planning for autonomous driving. [EB/OL]. [2025-06-03]. https://arxiv.org/pdf/2505.19381?

Jin Y, Shen X X, Peng H L, Liu X A, Qin J L, Li J Y, Xie J T, Gao P Z, Zhou G Y and Gong J T. 2023. Surrealdriver: designing generative driver agent simulation framework in urban contexts based on large language model. [EB/OL]. [2023-09-22]. https://arxiv.org/pdf/2309.13193v1

Kuang, J Y, Shen Y, Xie J Y, Luo H H, Xu Z, Li R H, Li Y H, Cheng X F, Lin X K and Han Y. 2025 Natural language understanding and inference with mllm in visual question answering: a survey. ACM Computing Surveys, 57(8): 1-36 [DOI: 10.1145/3711680]

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

Li X Y, Ye Z H, Wei S K, Chen Z, Chen X T, Tian Y H, Dang J W, Fu S J and Zhao Y. 2023. 3D object detection for autonomous driving from image: a survey——benchmarks, constraints and error analysis. Journal of Image and Graphics, 28(6): 1709-1740 (李熙莹, 叶芝桧, 韦世奎, 陈泽, 陈小彤, 田永鸿, 党建武, 付树军, 赵耀. 2023. 基于图像的自动驾驶3D目标检测综述——基准、制约因素和误差分析. 中国图象图形学报, 28(6): 1709-1740) [DOI: 10.11834/jig.230036]

Li Y K, Xiong K X, Guo X Y, Li F, Yan S X, Xu G W, Zhou L J, Chen L, Sun H Y, Wang B, Ma K, Chen G, Ye H J, Liu W Y, Wang X G. 2025. Recogdrive: a reinforced cognitive framework for end-to-end autonomous driving [EB/OL]. [2025-09-29]. https://arxiv.org/pdf/2506.08052

Li Z Q, Wang W H, Li H Y, Xie E Z, Sima C H, Lu T, Qiao Y and Dai J F. 2022. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. [EB/OL]. [2022-07-13]. https://arxiv.org/pdf/2203.17270

Liao，H C，Kong H L，Wang B N，Wang C Y，Ye W，He Z B，Xu C Z and Li Z N. 2025. Cot-drive：Efficient motion forecasting for autonomous driving with llms and chain-of-thought prompting. IEEE Transactions on Artificial Intelligence：1-15［DOI：10.1109/TAI.2025.3564594］

Liu J F，Zhang T Y，Zhong F Z，Yue P，Liu A and Liu X L. 2023. A survey of safety evaluation data generation techniques for autonomous driving. Journal of Image and Graphics，30(11)：3413-3437（刘江帆，张天缘，钟芳桔，岳鹏，刘艾杉，刘祥龙. 2025. 面向自动驾驶的安全评测数据生成技术综述. 中国图象图形学报，30(11)：3413-3437）［DOI：10.11834/jig.250181］

Liu J Q，Hang P，Qi X，Wang J Q and Sun J. 2023. Mtd-gpt：A multi-task decisionmaking gpt model for autonomous driving at unsignalized intersections//2023 IEEE 26th International Conference on Intelligent Transportation Systems. Bilbao，Spain：IEEE：5154-5161［DOI：10.1109/ITSC57777.2023.10421993］

Luo R P，Zhao Z W，Yang M，Dong J W，Da L，Lu P C，Wang T，Hu L M，Qiu M H and Wei Z Y. 2023. Valley：Video assistant with large language model enhanced ability［EB/OL］.［2023-10-08］. https://arxiv.org/pdf/2306.07207v2

Mao J G，Qian Y X，Ye J J，Zhao H and Wang Y. 2023. Gpt-driver：Learning to drive with gpt［EB/OL］.［2023-12-05］. https://arxiv.org/pdf/2310.01415v3

Mao J G，Ye J J，Qian Y X，Pavone M and Wang Y. 2023. A language agent for autonomous driving［EB/OL］.［2024-07-28］. https://arxiv.org/pdf/2311.10813

Nie M，Peng R Y，Wang C W，Cai X Y，Han J H，Xu H and Zhang L. 2024. Reason2drive：Towards interpretable and chain-based reasoning for autonomous driving//European Conference on Computer Vision. Milan，Italy：Springer：292-308［DOI：10.1007/978-3-031-73347-5_17］

Pan C B，Yaman B，Nesti T，Mallik A，Allievi A G，Velipasalar M and Ren L. 2024. Vlp：Vision language planning for autonomous driving//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle，USA：IEEE：14760 - 14769［DOI：10.1109/CVPR52733.2024.01398］

Peng M X，Guo X S，Chen X D，Zhu M X，Chen K H，Yang H，Wang X S and Wang Y H. 2024. Lc-llm：Explainable lane-change intention and trajectory predictions with large language models.［EB/OL］.［2024-03-27］. https://arxiv.org/pdf/2403.18344v1

Renz K，Chen L，Arani E and Sinavski O. 2025. Simlingo：vision-only closed-loop autonomous driving with language-action alignment//Proceedings of the Computer Vision and Pattern Recognition Conference. Nashville，USA：IEEE：11993 - 12003［DOI：10.1109/CVPR52734.2025.01120］

Shi S S，Jiang L，Dai D X and Schiele B. 2023. Motion transformer with global intention localization and local movement refinement//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans，USA：Curran Associates Inc.：6531-6543［DOI：10.5555/3600270.3600743］

Sima C H，Renz K，Chitta K，Chen L，Zhang H X，Xie C G，Beißwenger J，Luo P，Geiger A and Li H Y. 2023. Drivelm：driving with graph visual question answering.［EB/OL］.［2023-12-21］. https://arxiv.org/pdf/2312.14150

Wang，T Q，Xie E Z，Chu R H，Li Z G and Luo P. 2024. Drivecot：integrating chain-of-thought reasoning with end-to-end driving［EB/OL］.［2024-03-25］. https://arxiv.org/pdf/2403.16996

Wang S Y，Zhu Y X，Li Z H，Wang Y T，Li L and He Z B. 2023. Chatgpt as your vehicle co-pilot：an initial attempt. IEEE Transactions on Intelligent Vehicles，8(12)：4706-4721［DOI：10.1109/TIV.2023.3325300］

Wang W H，Xie J W，Hu C Y，Zou H M，Fan J N，Tong W W，Wen Y，Wu S L，Deng H M，Li Z Q，Tian H，Lu L W，Zhu X Z，Wang X G，Qiao Y and Dai J F. 2023. Drivemlm：aligning multi-modal large language models with behavioral planning states for autonomous driving［EB/OL］.［2023-12-14］. https://arxiv.org/pdf/2312.09245v1

Wang Y，Guizilini V C，Zhang T Y，Zhao H and Solomon J. 2022. Detr3d：3d object detection from multi-view images via 3d-to-2d queries//Proceedings of the 5th Conference on Robot Learning. London，UK：PMLR：180-191［DOI：10.48550/arXiv.2110.06922］

Wen L C，Fu D C，Li X，Cai X Y，Ma T，Cai P L，Dou M，Shi B T，He L and Qiao Y. 2023. Dilu：a knowledge-driven approach to autonomous driving with large language models［EB/OL］.［2023-09-28］. https://arxiv.org/pdf/2309.16292

Wu D M，Han W C，Wang T C，Liu Y F，Zhang X Y and Shen J B. 2023. Language prompt for autonomous driving［EB/OL］.［2023-09-08］. https://arxiv.org/pdf/2309.04379v1

Xie Y C，Xu R S，He T，Hwang J，Luo K T，Ji J W，Lin H，Chen L T，Lu Y R，Leng Z Q，Anguelov D and Tan M X. 2025. S4-driver：scalable self-supervised driving multimodal large language model with spatio-temporal visual representation//Proceedings of the Computer Vision and Pattern Recognition Conference. Nashville，USA：IEEE：1622 - 1632［DOI：10.1109/CVPR52734.2025.00159］

Xu Z H，Zhang Y J，Xie E Z，Zhao Z，Guo Y，Wong K K，Li Z G，and Zhao H S. 2024. Drivegpt4：interpretable end-to-end autonomous driving via large language model. IEEE Robotics and Automation Letters，9(10)：8186-8193［DOI：10.1109/LRA.2024.3440097］

Xu，Z H，Bai Y，Zhang Y J，Li Z L，Xia F，Wong K K，Wang J Q，Zhao H S. 2025. Drivegpt4-v2：harnessing large language model capabilities for enhanced closed-loop autonomous driving//Proceedings of the Computer Vision and Pattern Recognition Conference. Nashville，USA：IEEE：17261-17270［DOI：10.1109/CVPR52734.2025.01609］

Yang Z J, Chai Y L, Jia X S, Li Q F, Shao Y Q, Zhu X K, Su H S and Yan J C. 2025. Drivemoe: mixture-of-experts for vision-language-action model in end-to-end autonomous driving [EB/OL]. [2025-05-22]. https://arxiv.org/pdf/2505.16278

Yu H B, Yang W X, Zhong J R, Yang Z W, Fan S Q, Luo P and Nie Z Q. 2025. End-to-end autonomous driving through v2x cooperation// Proceedings of the AAAI Conference on Artificial Intelligence. Philadelphia, USA: AAAI Press: 9598-9606 [DOI: 10.1609/aaai.v39i9.33040]

Yuan J H, Sun S Y, Omeiza D, Zhao B, Newman P, Kunze L and Gadd M. 2024. Rag-driver: generalisable driving explanations with retrieval-augmented in-context learning in multi-modal large language model [EB/OL]. [2024-05-29]. https://arxiv.org/pdf/2402.10828

Zeng S, Chang X Y, Xie M W, Liu X R, Bai Y F, Pan Z, Xu M and Wei X. 2025. Futuresightdrive: thinking visually with spatio-temporal cot for autonomous driving [EB/OL]. [2025-05-23]. https://arxiv.org/pdf/2505.17685v1

Zhang J W, Xuan Y, Wang T Q, Yao Y, Petiushko A and Li B. 2025. Safeauto: knowledge-enhanced safe autonomous driving with multi-modal foundation models [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2503.00211v1

Zheng X J, Wu L X, Yan Z J, Tang Y R, Zhao H, Zhong C, Chen B K and Gong J T. 2024. Large language models powered context-aware motion prediction [EB/OL]. [2024-03-17]. https://arxiv.org/pdf/2403.11057v1

Zheng Y P, Yang P X, Xing Z B, Zhang Q C, Zheng Y H, Gao Y F, Li P F, Zhang T, Xia Z P, Jia P, Lang X P and Zhao D B. 2025. World4drive: end-to-end autonomous driving via intention-aware physical latent world model//Proceedings of the IEEE/CVF International Conference on Computer Vision. Hawaii, USA: IEEE: 28632-28642 [DOI: 10.48550/arXiv.2507.00603]

Zhou X C, Han X Y, Yang F, Ma Y P and Knoll A C. 2025. Opendrivevla: towards end-to-end autonomous driving with large vision language action model [EB/OL]. [2025-03-30]. https://arxiv.org/pdf/2503.23463v1

作者简介

闫瑞松，上海交通大学在读博士生，主要研究方向为强化学习、视觉语言动作模型、智能决策等。E-mail：yanruisong@sjtu.edu.cn

李成林，上海交通大学教授，主要研究方向为多媒体信号处理及通信、强化学习、联邦学习等。E-mail：lcl1985@sjtu.edu.cn

郑伟诗，中山大学教授，主要研究方向为视频图像处理、计算机三维建模与生成式人工智能。机器人学习等。E-mail：wszheng@ieee.org

赫然，中国科学院自动化研究所研究员，主要研究方向为人工智能、模式识别、计算机视觉等。E-mail：ran.he@ia.ac.cn

查正军，中国科学技术大学教授，主要研究方向为图像视频处理与分析、计算机视觉、多模态大模型等。E-mail：zhazj@ustc.edu.cn