第 25 卷　第 1 期
2025 年 2 月

交 通 运 输 工 程 学 报
Journal of Traffic and Transportation Engineering

Vol. 25　No. 1
Feb. 2025

文章编号:1671-1637(2025)01-0008-21

交通大模型综述

肖建力[*1],邱　雪[1],张　扬[2],苏海昇[3],李志鹏[4],张传明[5]

(1. 上海理工大学 光电信息与计算机工程学院,上海　200093; 2. 上海市城乡建设和交通发展研究院(上海市数字化城市管理中心),上海　200032; 3. 上海交通大学 计算机学院,上海　200240; 4. 同济大学电子与信息工程学院,上海　201804; 5. 北京百度网讯科技有限公司,北京　100085)

摘　要:对大语言模型(LLMs)技术在交通领域的促进作用进行了深入探讨,展现了其在改善交通管理和控制、提高交通安全以及推动自动驾驶技术发展方面的巨大潜力;系统阐述了 LLMs、视觉大模型、多模态大模型的基本概念及发展历程。针对部分交通 LLMs,总结了它们的模型架构和训练方法,探讨了 LLMs 在交通领域,如交通管理和控制、交通安全和自动驾驶方面的主要应用。研究结果表明:在交通管理和控制方面,LLMs 的应用可改善交通信号控制和交通状态预测等问题,并为城市交通管理带来新的可能性,这不仅能减少交通拥堵,还降低环境污染;在交通安全方面,相比于传统模型,LLMs 的应用显著提高交通事故分析和预测能力,通过对历史事故数据的深入学习,模型能够识别出事故高发区域和时段,从而采取预防措施,提高交通安全指数;在自动驾驶领域,传统模型向多模态自动驾驶模型的转变不仅能提高自动驾驶系统的决策和环境适应能力,还会为用户提供更加安全、舒适的驾驶体验。本文挖掘了 LLMs 在当今交通领域中的潜力和价值,同时为实现更加智能、高效的交通系统提供了有用的建议,如降低交通 LLMs 的计算成本,提升模型的实时性和可靠性等。

关键词:智能交通;通用大模型;交通大模型;交通安全;自动驾驶

中图分类号:U491　　**文献标志码**:A　　**DOI**:10.19818/j.cnki.1671-1637.2025.01.002

Review on large language models in transportation

XIAO Jian-li[*1], QIU Xue[1], ZHANG Yang[2], SU Hai-sheng[3], LI Zhi-peng[4], ZHANG Chuan-ming[5]

(1. School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China; 2. Shanghai Urban-Rural Construction and Transportation Development Research Institute (Shanghai Digital Urban Management Center), Shanghai 200032, China; 3. School of Computer Science, Shanghai Jiao Tong University, Shanghai 200240, China; 4. School of Electronic and Information Engineering, Tongji University, Shanghai 201804, China; 5. Beijing Baidu Netcom Science and Technology Co., Ltd., Beijing 100085, China)

Abstract: The promotion of large language models (LLMs) in transportation was further discussed. Their great potentials were demonstrated in improving traffic management and control, enhancing traffic safety, and advancing autonomous driving. The basic concepts and development of LLMs, large vision models, and large multimodal models were systematically expounded. Some LLMs in transportation were summarized in terms of their structures and

收稿日期:2024-07-25
基金项目:国家自然科学基金项目(92370201,61603257);中央高校基本科研业务费专项资金项目(22120230311)
* **作者简介**:肖建力(1982-),男,湖北天门人,上海理工大学副教授,工学博士,从事人工智能与大数据研究。
引用格式:肖建力,邱　雪,张　扬,等.交通大模型综述[J].交通运输工程学报,2025,25(1):8-28.
Citation: XIAO Jian-li, QIU Xue, ZHANG Yang, et al. Review on large language models in transportation[J]. Journal of Traffic and Transportation Engineering, 2025, 25(1): 8-28.

training methods. The major applications of LLMs in transportation were discussed, such as traffic management and control, traffic safety, and autonomous driving. Research results show that, when it comes to traffic management and control, issues such as traffic signal control and traffic state prediction can be significantly addressed by the application of LLMs. New possibilities are also brought for urban traffic management. Traffic congestion and environmental pollution are both reduced. As for traffic safety, compared with previous models, LLMs application significantly improve the ability to analyze and predict traffic accidents. Through deep learning of historical accident data, the models can identify those areas and time periods with a high incidence of accidents. Consequently, preventive measures can be taken to improve the traffic safety index. In the field of autonomous driving, the shift from traditional models to multimodal autonomous driving models can not only enhance the abilities of decision-making and environmental adaptation, but also provide users with a safer and more comfortable driving experience. The potential and value of LLMs in transportation are explored. Besides, practical suggestions are also offered to create a more intelligent and efficient transportation system, such as reducing the computational cost of LLMs in transportation and improving the real-time performance and reliability of models.

Keywords: intelligent transportation; general large model; LLMs in transportation; traffic safety; autonomous driving

Funding: National Natural Science Foundation of China (92370201, 61603257); Fundamental Research Funds for the Central Universities (22120230311)

* **Corresponding author**: XIAO Jian-li(1982-), male, professor, PhD, audyxiao@sjtu.edu.cn.

0 引　言

智能交通系统是现代社会的重要组成部分，其中交通管理和控制、交通安全和自动驾驶是当前交通领域的重点研究方向[1-2]。为了应对智能交通系统带来的机遇与挑战，几十年来，研究人员基于各种机器学习和深度学习的先进技术，开发了许多交通基础模型，有力促进了智能交通系统的发展。

近几年，大语言模型(Large Language Models, LLMs)发展迅速。为简洁起见，本文统一使用大模型这一简称。ChatGPT[3]作为其中的典型代表之一，具有强大的生成能力。与之相比，在 2023 年诞生的 GPT-4[4]具有更加强大的生成能力，它改善了以往模型的单模态输入方式，是一种大规模的多模态模型，可以接受图像和文本输入并且生成相应输出[5]。这也给了我们一种启示，可以根据来自不同数据源的多模态数据来对模型进行预训练，从而增强模型的能力与泛化性能。例如，在城市交通管理与控制中，将现有为特定任务设计的交通基础模型与大模型相结合，不仅摆脱了有限输入输出交互的限制，还能提高它们处理复杂交通问题的能力，并提供有效的建议[6]。

随着大模型在交通领域的认可度不断提高，相关研究人员已将其应用到自动驾驶系统和智能交通领域中。在自动驾驶系统中，通过集成数据，使车辆能够深入地感知真实世界的环境并做出相应的行为决策，从而提高驾驶的安全性和效率[7]。在智能交通领域可以使用大模型分析交通数据并预测未来交通状态，以便优化路线规划和改善交通管理；也可以使用大模型对驾驶行为进行学习，从而能够识别潜在的危险行为并提供实时驾驶建议；还可以将大模型应用于车载助手系统，使驾驶人能够通过语音与车辆进行交互，最终提高驾驶安全性和便利性[8-10]。

本文第 1 节全面介绍了大模型的发展历程，并分别详细介绍了视觉大模型和多模态大模型；第 2 节讨论了现有比较热门的交通大模型，基于相关文献总结了它们的模型架构和训练方法；第 3 节深入研究了大模型在交通领域的主要应用，其中主要包括交通管理与控制、交通安全和自动驾驶；第 4 节对交通大模型的发展和应用做出了总结与展望。

1 大模型

在本小节中，首先全面介绍大模型的发展历程，

然后分别详细介绍视觉大模型和多模态大模型。

1.1 大模型的背景

自 20 世纪 50 年代计算机学家约翰麦卡锡提出人工智能(Artificial Intelligence, AI)的概念以来,人类一直探索如何让机器掌握语言。由于复杂的语法规则,开发一个有理解能力的人工智能模型一直是一个重大的挑战[11]。近年来,大模型技术的发展受到了社会的广泛关注,这对整个 AI 界产生了重要影响。大模型的发展历程如图 1 所示,其中包含典型模型的名称及其参数大小。同时期,大模型的发展开始步入萌芽阶段。1998 年,Lecun 等[12]基于原有的神经网络结构,提出了一种特定类型的卷积神经网络 LeNet-5,形成了现代卷积神经网络的雏形,其架构包括多个卷积层、池化层和全连接层,主要用于图像识别任务,尤其是手写数字识别。从此,机器学习方法由早期的浅层机器学习转向深度学习。这对后来的深度学习和计算机视觉研究产生了深远影响,同时也为后续大模型的发展奠定了基础。

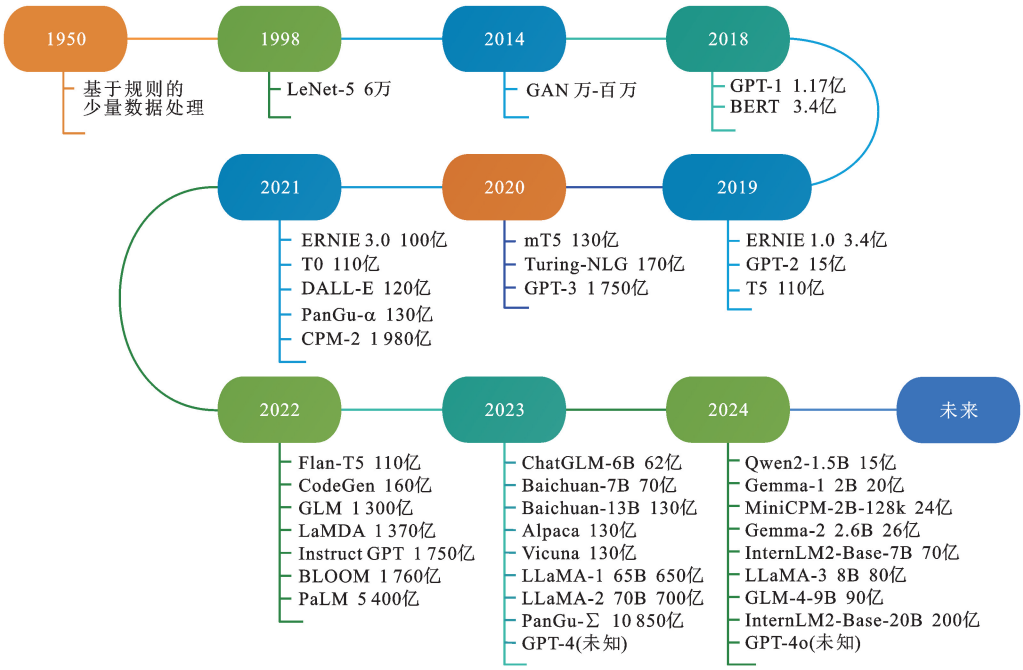


图 1 大模型的发展历程

Fig. 1 Development history of LLMs

2006 年,大模型的发展进入探索时期。2014 年,Creswell 等[13]提出了生成对抗网络(Generative Adversarial Network, GAN),当时号称“21 世纪最强大的模型”之一。它的独特之处在于其对抗性训练架构,通过其中 2 个网络(生成器和判别器)相互竞争来提升模型性能。这种架构在图像合成、艺术创作、数据增强等诸多领域展现出强大的性能,推动了人工智能领域的重大进步。2017 年,Vaswani 等[14]提出了基于自注意力机制的神经网络——Transformer,为后续大模型的主体算法架构奠定了基础。2018 年,Google 和 OpenAI 团队分别推出了 BERT[15]和 GPT-1[16]大模型,这标志着预训练大模型在自然语言处理领域的兴起。在这一时期,以 Transformer 为核心的架构为大模型的发展奠定了基础,并显著提升了大模型的性能。

大模型是人工智能领域的一项重要进展,模型通常包含数千亿参数,并且在大规模文本数据集上进行预训练,主要用于理解和生成自然语言。自 2020 年以来,多种大模型被提出,如 GPT-3[17]、GPT-4、PaLM[18]、Galactica[19]、LLaMA[20]等。大模型的一个关键特点是涌现能力,即当模型规模达到一定程度时,会出现原本小型模型中不存在的新能力,并且性能显著提升。Wei 等[21]对模型的涌现能力进行了介绍。大模型有 3 种典型的涌现能力:(1)上下文学习[17],它能够通过补全输入文本中的单词序列来生成测试实例的预期输出,而无需额外的训练或梯度更新;(2)指令遵循,通过使用自然语言描述的混合多任务数据集进行微调,大模型可以在未见过的指令形式上表现出很好的泛化能力[22];(3)逐步推理,使用思维链提示策略,大模型可以通过中间推理步骤的提示来解决一些涉及复杂推理步骤的任务[23]。大模型的另一个关键特点是扩展法则。相关研究表明,大模型通过扩展模型的规模和数据量,显著提升了模型性能[24-25]。此外,大模型还能

够对特定任务进行微调。例如，InstructGPT 采用了基于人类反馈的强化学习技术，使大模型能够按照人类期望的指令进行操作。除了自身的特点和关键技术之外，大模型还可以通过调用外部工具（如服务器或搜索引擎）来弥补其在下游任务上的不足。

尽管大模型在语言理解和生成方面取得了显著进步，但它们的发展也面临着一系列的风险和挑战[26]。例如：可解释性问题，模型的决策过程和内部工作原理不能被人类完全理解并信任；准确性问题，模型的预测结果可能因训练数据的质量、模型大小和无关特征等多种因素而出现比较高的错误率；隐私问题，模型在训练以及与用户互动过程中会接触大量隐私和个人信息，而其复杂且不透明的工作机制存在较大的隐私数据泄露风险。

1.2　视觉大模型

视觉大模型结合了视觉识别技术和大模型强大的文本处理能力，能够同时生成图像及相关的文本内容。如今视觉大模型与多模态大模型之间的界限正在逐渐模糊，这是因为多模态大模型整合了视觉和语言处理能力。

为了捕获复杂的视觉特征，科研人员使用大量数据对模型进行训练。随着计算资源和数据集的增加，视觉大模型已经在计算机视觉领域取得了显著进展。2023 年 4 月，Oquab 等[27]提出了开源的视觉大模型 DINOv2，模型采用了自监督学习，也就是从大量未标记数据中提取视觉特征。相比于以往的自监督模型（如 DINO[28]和 iBOT[29]），DINOv2 在多个基准测试上性能显著提升，不仅能对视频进行处理，还能生成高质量的分割结果。通过扩展参数量和模型规模，使得 DINOv2 具有更强的泛化能力。此外，在优化训练过程中，DINOv2 减少了内存消耗，提高了训练效率，具有良好的扩展性。然而，DINOv2 对高质量的数据依赖较大，在未经处理的数据集上的表现受限。与弱监督模型（如 CLIP[30]）相比，DINOv2 在无需文本监督的情况下表现出色，但在某些需要文本指导任务中稍显逊色。总体而言，DINOv2 在多种任务中展现了卓越性能，但在计算资源和数据集依赖方面仍具有一定的挑战。2023 年 11 月，Tuo 等[31]推出了基于扩散模型[32]的视觉大模型AnyText，它不仅能够在图像中生成文本，还可以对图像中的文本进行精确编辑，确保与周围文本风格一致。AnyText 包含 2 个核心模块——隐空间辅助模块和文本嵌入模块，在训练过程中，除了采用扩散模型的噪声预测损失，还引入了文本感知损失，对生成的文本区域精确到像素级的监督，大幅提高了文本生成的准确性。该模型还具有高度集成性，可以与现有的扩散模型结合，从而增强模型的文本生成能力。尽管支持多语言，AnyText 在处理笔画复杂的字符时，由于计算复杂，生成过程较为耗时。此外，由于训练数据主要包括中文和英文，所以在处理其他语言时会限制模型的表现。与 GlyphDraw[33]和 TextDiffuser[34]模型相比，AnyText 在多语言支持和文本生成的灵活性方面有明显优势，但是在处理复杂语言和减少对光学字符识别（Optical Character Recognition，OCR）系统依赖方面还有改进空间。2023 年 12 月，Jiang 等[35]推出了新的视觉大模型 VideoBooth，该模型的创新点是结合文本和图像的提示生成视频。模型主要采用 2 个步骤，先通过图像编码器提取图像提示的粗略视觉特征，并将特征与文本描述结合，然后再将图像提示的潜在表示注入到模型的注意力模块中，从而控制生成视频的精度。这种方法不仅使模型能够掌握图像的全局特征，还能把控图像中的细节，确保生成的视频既符合文本描述，又准确反映图像的视觉属性。与 DreamBooth[36]等模型相比，VideoBooth 在细节保留和生成视频一致性上表现更优。然而，VideoBooth 在处理视角变化较大的图像时效果有限，并且生成视频的质量依赖于图像提示的丰富度。此外，由于模型机制较为复杂，所以对计算资源的需求也比较高。为了进一步提升视频表示的性能，Wang 等[37]于 2024 年 3 月提出了新的视频基础模型 InternVideo2，在多模态视频理解方面展现了出色的性能，它通过结合视频、音频和文本的跨模态对比学习，显著提升了模型在复杂语义推理中的准确性。同时，模型能够有效捕捉视频的时空信息，尤其在长视频和复杂场景下表现优异。与其他模型相比，InternVideo2 在多模态融合和长视频理解上表现更好，并且在视频推理和复杂对话任务中超越了 VideoPrism[38]，在视频音频融合任务中超越了 UMT[39]，展现出更广泛的适用性。总体而言，InternVideo2 在多模态视频理解领域具有显著优势，但在处理高分辨率视频和降低计算复杂度方面仍有提升空间。

1.3　多模态大模型

1.3.1　基本概念

视觉大模型在视觉感知、图像识别等方面取得了显著的进步，这促使了大模型和其他模态基础模型交互融合。2023 年 3 月 GPT-4 被推出，在 GPT-3

的基础上增加了对视觉模态输入的支持，这意味着它能够同时理解文本和图像数据的输入，并生成需要的文本和图像输出。由于 GPT-4 尚未开源，为了探索其出众的多模态能力，Zhu 等[40]在 2023 年4 月提出 MiniGPT-4，这是一种具有代表性的新兴多模态大模型，期望可以模拟出类似 GPT-4 的多模态性能。研究者们讨论了多模态大模型的任务类型，大致可以分为 5 类，如图 2 所示。

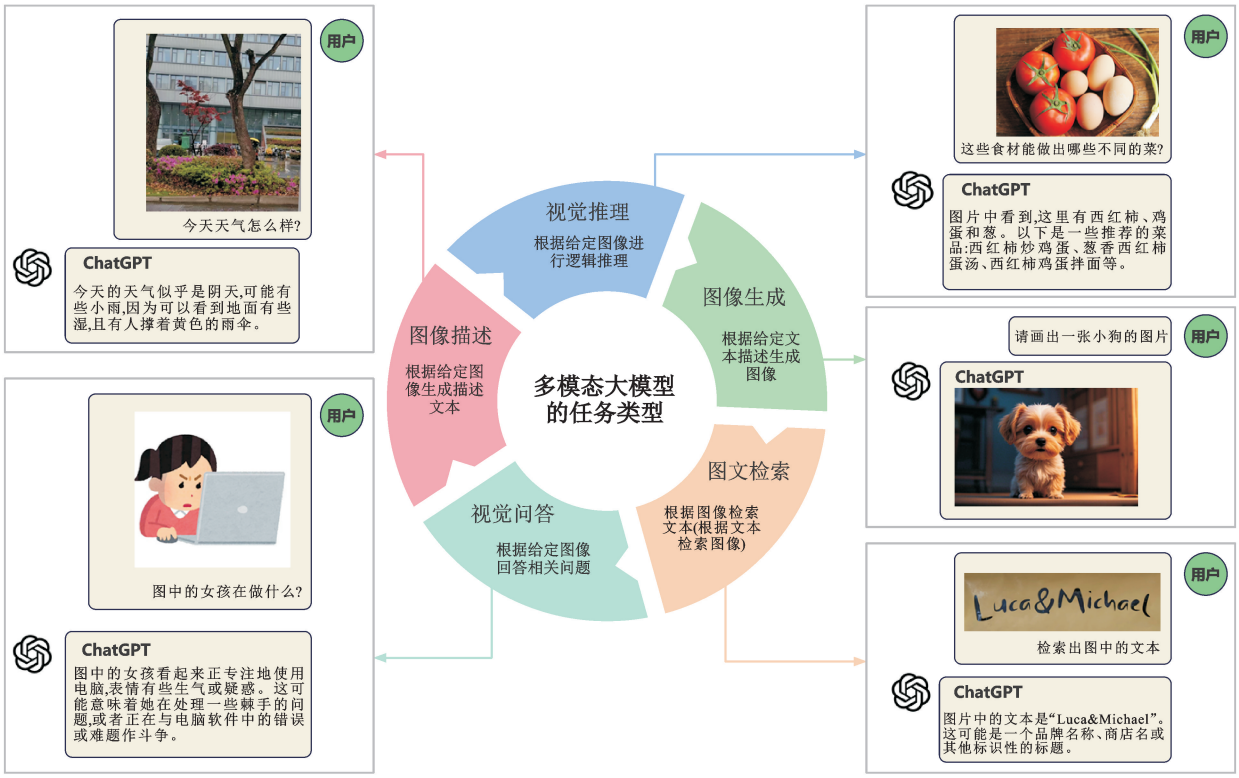


图 2 多模态大模型的任务类型

Fig. 2 Task types of large multimodal models

接下来，重点介绍以 MiniGPT-4 为代表的多模态大模型的架构和训练方法。

1.3.2 MiniGPT-4 的模型架构及训练方法

MiniGPT-4 的模型架构主要分为 3 部分：Vicuna 模型[41]、视觉编码器和线性投影层。大致模型架构如图 3 所示[40]。

框架的第 1 部分为 Vicuna 模型。为了降低模型训练的成本，MiniGPT-4 引入了 Vicuna 模型，该模型在 MiniGPT-4 中主要任务是同时理解文本和图像数据输入，并能生成符合指令的文本描述。第 2 部分为视觉编码器。它的作用是将原始的视觉输入转换成一种高级的、紧凑的特征表示(即编码)，使模型进一步完成不同的下游任务，如图像描述、视觉问答或者跨模态检索等。MiniGPT-4 的视觉编码器由两部分组成：ViT[42]和图文对齐模块 Q-Former。当图像输入之后，通过 ViT 初步编码来提取图像的特征向量，随后使用 Q-Former 模块将文本转换为嵌入向量，最终通过训练使图像嵌入与文本嵌入对齐，从而实现图文配对。由于 BERT 在处理文本方面有着优越性能，所以 MiniGPT-4 选择预训练的 BERT 模型作为 Q-Former 模块。第 3 部分为线性投影层。虽然视觉编码器模块已经在大量的图像-文本数据上面进行了预训练，但与大模型之间还存在着差距。为了弥补这一缺陷，MiniGPT-4 增加了一个可以训练的线性投影层，目的是通过训练将视觉编码器的输出特征与 Vicuna 模型对齐，方便模型进行后续计算。

MiniGPT-4 的模型训练分为 2 个阶段。(1)预训练阶段。让模型在大量的通用图像-文本数据集上进行无监督预训练，来学习基础的视觉语言知识。其中，模型使用来自 Conceptual Caption[43]、SBU[44]和 LAION[45]的组合数据集进行预训练，因为这些数据集包含丰富的图像信息和文本描述对。完成第一轮训练之后，MiniGPT-4 获得了丰富的图像知识，并且能够根据用户输入生成合理的文本描述。有时模型不能生成符合用户要求的文本输出，为了使模型输出与人类的理解保持一致，研究者们又构建了一个高质量的图像-文本数据集。该数据集的

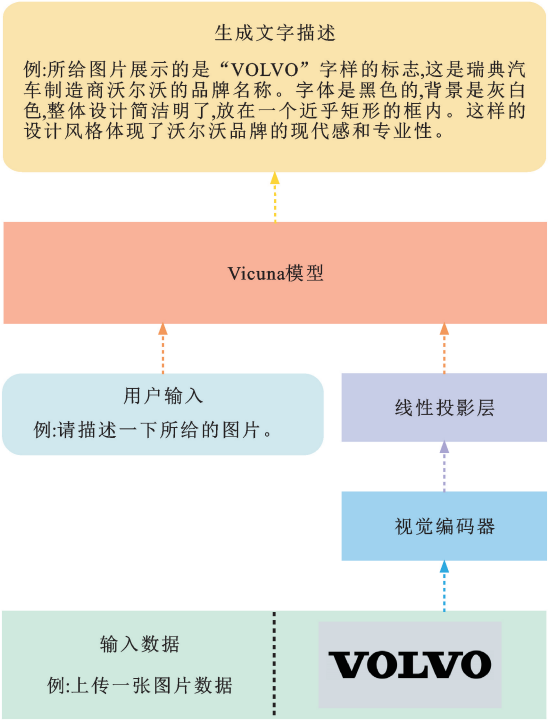


图 3　MiniGPT-4 的模型架构

Fig. 3　Model architecture of MiniGPT-4

构建考虑了 2 个基本要素，使用提示策略和自动化文本质量评估策略。使用提示策略可以使模型对给定的图像生成更全面的文本描述。研究者们使用 ChatGPT 作为自动化文本质量评估者，对预训练生成的 5 000 份图像-文本数据进行检查，并自动修正文本描述中的语法、语义和结构等错误。经过 GPT 的评估修改后，最终保留了 3 000 份符合要求的高质量数据用于第二阶段的模型训练。(2)微调阶段。使模型在少量高质量图像-文本数据集上进行有监督微调，从而进一步提高第一阶段预训练模型的生成质量和综合表现。微调的结果表明，MiniGPT-4 能够生成更加流畅、自然的视觉问答反馈，展现出强大的视觉理解能力。

1.3.3　其他多模态大模型

虽然有不少多模态大模型在图像描述和视觉问答等任务中表现优异，但有的缺乏基本的感知能力。针对这一问题，Zang 等[46]提出了一个统一的多模态大模型 ContextDET，其中包含 3 个关键子模型：视觉编码器、多模态大模型和视觉解码器。ContextDET 模型通过结合视觉编码器和多模态大模型，展现了强大的上下文理解能力，尤其在语言填空测试、视觉描述生成和问答任务中表现出色。与传统对象检测模型(如 DETR[47]和 Mask R-CNN[48])相比，ContextDET 在处理多模态上下文理解和开放词汇检测上具有显著优势，尤其在上下文对象检测任务中表现优异，但在计算效率上有待优化。Yang 等[49]对 2023 年 9 月 OpenAI 推出的多模态大模型 GPT-4V 做了深入分析，GPT-4V 能够将视觉和语言高度融合，从而处理复杂的视觉文本交互任务，如场景描述、物体定位、表格和文档推理等。此外，GPT-4V 可以处理多语言输入，能支持不同语言的视觉理解，并且在代码生成、时间推理和情感理解等领域表现出色。然而，GPT-4V 也存在一些局限性，如：推理能力依赖于上下文和提示设计；计算资源需求较高，不适合在资源受限的场景中应用。另外，模型在处理多图像输入和时序视频理解方面仍有待优化。2023 年 12 月，谷歌公司推出了系列多模态大模型 Gemini[50]，它们具备强大的多模态处理能力，能够同时处理文本、图像、音频和视频等数据，并在跨模态推理和理解任务中表现出色。Gemini Ultra 在多个基准测试中刷新了记录，特别在多模态基准测试和大规模多任务语言理解(Massive Multitask Language Understanding，MMLU[51])指标上达到了人类专家水平，这展现了模型出色的推理和多任务处理能力。与 GPT-4V 相比，Gemini Ultra 在多模态推理上表现更为出色，尤其是在跨模态任务中，同时 Gemini Nano 在资源有限的设备端应用中也具有明显优势。然而，Gemini 模型对计算资源需求较大(尤其是 Ultra 版本)，同时模型性能主要依赖于大规模高质量数据集，在处理未知领域时可能存在泛化问题。

现有的多模态大模型在处理包含图片、表格、文档和信息图表等细节的图像时表现不佳，限制了其在现实世界场景中的应用。针对这些问题，上海人工智能实验室联合商汤科技公司[52]于 2024 年 4 月推出 InternLM-XComposer2-4KHD，这是 4K 超清解析多模态大模型，具备处理从 336 像素到 4K 高清广泛分辨率的能力，在高分辨率视觉任务中表现出色，尤其在 OCR、文档理解和图表问答等任务上超越了 GPT-4V 和 Gemini Pro，展现了强大的视觉、语言理解和推理能力。模型中的动态分割和自动补丁配置策略，使其能够在超高分辨率下获得性能提升，特别是在 DocVQA[53]和 ChartQA[54]等基准测试中表现突出。然而，模型对计算资源要求较高，尤其在处理 4K 高清图像时，训练和推理的复杂度明显增加，这可能限制其在资源受限场景中的应用。相比其他模型，InternLM-XComposer2-4KHD 在高分辨率任务上表现突出，但在某些感知相关任

务中，分辨率提高带来的性能提升较为有限，需要在效率和精度间找到平衡。清华大学自然语言处理实验室和智源研究院提出了多模态大模型 MiniCPM-V 2.0，模型基于 SigLip-400M 与 MiniCPM-2.4B 进行构建，并通过感知器重采样器连接。它支持处理高达 180 万像素的图像输入，适用于中英文双语环境。在通用场景文字理解评测 TextVQA[55]基准上，实现了与 Gemini Pro 相当的性能，并在 Object HalBench[56]上的表现与 GPT-4V 相当，在抵抗幻觉上具有出色的效果。实验结果显示，模型在 OCRBench[57]测试中表现出色，并在 OpenCompass[58]评估中超越了多个大模型，如 Qwen-VL-Chat[59]、CogVLM-Chat[60] 和 Yi-VL[61]。MiniCPM-V 2.0 可高效部署在大多数个人电脑上，甚至还可以部署在手机等终端设备上，展示了卓越的适配性能和可靠性。2024 年 5 月，OpenAI 公司推出了最新的多模态大模型 GPT-4o，它可以接受文本、语音、图像、视频等任意一种类型输入，并输出文本、图像等，是一个端到端的大模型。与 GPT-4 相比，GPT-4o 的响应速度得到了显著提升，它能够更快地回应用户的查询指令，这一点在需要及时反馈的应用场景中十分重要。在对多模态理解方面，GPT-4o 能够支持对音频和视频文件的理解。在输出质量方面，GPT-4o 能够给出更详细、准确的回答，对逻辑问题和语言结构理解更加深入。

多模态大模型通过整合不同类型数据，不仅提升了信息处理精度，还扩展了应用领域，如自动驾驶、医疗诊断和个性化教育等。这预示着人工智能进入了一个新时代。在这个时代中，机器能够综合多种数据类型，实现对复杂世界的全面理解和互动。

2 交通大模型

大模型一般分为通用大模型和领域大模型，以上介绍的部分即为通用大模型。与通用大模型相比，领域大模型经过专门的训练，能够更好地理解某个特定领域的知识，它的领域专业性更高。由于在特定领域的优化，领域大模型往往比通用大模型表现的更好。本小节首先对国内首个开源的交通大模型 TransGPT 做出介绍，然后对当今社会比较热门的交通大模型做出详细介绍。

2.1 TransGPT 交通大模型

TransGPT[62]作为国内首个开源的交通大模型，在真实的交通应用场景中发挥关键作用。该模型具备多种功能，如交通状况预测、交通规划、交通安全教育、交通管理、事故报告与分析，以及自动驾驶系统支持等。TransGPT 涵盖了广泛的交通领域知识，能服务于多个相关领域，如道路、桥梁、隧道、公路和水路运输等，展现其在各种交通场景中的适用性和灵活性。下面将进行详细的介绍。

模型训练中使用到的开源数据集分为两部分：(1)通用预训练数据集，其中包括监督微调(Supervised Fine-Tuning，SFT)数据集和奖励模型数据集；(2)交通领域数据集，其中包括领域预训练数据集和领域微调数据集。这些数据分为单模态和多模态，其中单模态数据包括科技文献、科研数据、工程建设信息和管理决策信息等；多模态数据包括交通标志大全、驾考题库以及全球旅游景点等。此外，对于领域微调数据集，它的对话数据的生成方法，首先是从 pdf、docx、doc 格式文件中提取文档，然后利用大模型(如 ChatGPT 等)生成对话数据。

实验中，研究者使用 TransGPT-7B 模型对交通情况预测、交通规划、交通安全教育、事故报告和分析等任务进行测评，结果显示 TransGPT 具备多项优点：(1)模型能够处理单模态和多模态的输入，提升了在交通领域中的任务表现，特别是在驾照考试、交通标志识别和交通工程等任务中表现优异；(2)相比于其他基线模型如 VisualGLM-6B[63]，TransGPT 在生成交通场景、流量预测等多模态任务中表现优异，准确率提升了 40%；(3)模型的定制化数据集、单模态交通数据集(Single-Modal Transportation Dataset，STD)和多模态交通数据集(Multi-Modal Transportation Dataset，MTD)进一步增强了模型在交通相关任务中的精度。然而，TransGPT 也有一些局限性。由于其专注于交通领域，模型的泛化能力较弱，在处理其他领域任务时表现不如通用模型。此外，虽然模型在多模态任务中表现优异，但训练过程资源密集，特别是在处理大规模多模态数据时需要大量计算资源。总的来说，TransGPT 在交通领域具有显著优势，但在应对更广泛的任务时，其优势可能不及通用大模型(如 GPT-4)。表 1 中给出 TransGPT 系列模型的下载链接。

2.2 百度交通大模型

自百度地图上线以来，百度不断迭代基于多尺度时空速度特征的通行速度预测算法，实现了全国覆盖并且驾车通行时间预测的准确率高于 90%。这一技术支撑了精准路线规划与导航产品。2023 年 3 月，百度发布了文心一言，成为国内首个推出类似 ChatGPT 大模型的企业。基于百度文心大模型，百

表 1　TransGPT 模型的下载链接

Table 1　Download links for TransGPT models

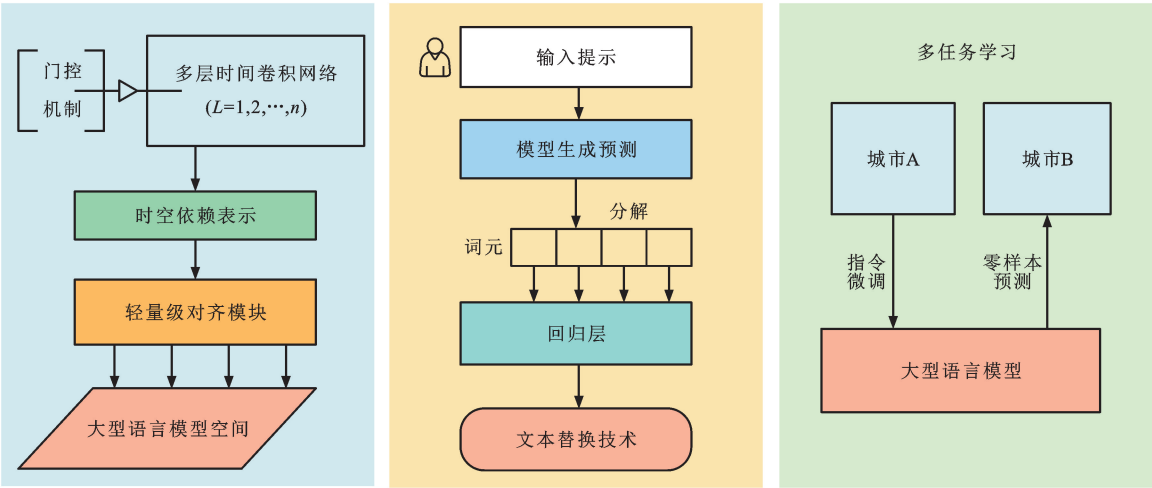
模型	下载链接
TransGPT-7B-v0	https://huggingface.co/DUOMO-Lab/TransGPT-v0
TransGPT-MM-6B-v0	https://huggingface.co/DUOMO-Lab/TransGPT-MM-v0
TransGPT-MM-6B-v1	https://huggingface.co/DUOMO-Lab/TransGPT-MM-v1

度创新性地推出了文心交通大模型，利用时空 Transformer 技术，将时间序列上的交通变化和空间上的道路拓扑结合起来，实现对交通模式的精确分析和预测。该模型综合考虑了不同地理区域、不同时间跨度及数据稀疏程度，构建了大量的交通学习任务。通过多目标、多任务预训练技术和对关键区域真实样本的模型进行微调，增强了模型的通用性和泛化能力。在百度地图 V18 版本中，文心交通大模型通过时空数据分析与交通信控优化，使信控区域的交通出行效率提升了 15%～30%，显著增强了地图的智能化并提升了用户的导航体验，开启了智慧出行时代的新篇章。

香港大学在 2024 年 3 月与百度联合推出智慧城市大模型 UrbanGPT[64]，它巧妙地结合了时空依赖编码器和指令调整范式，使模型能够理解并预测时间和空间之间复杂的相互依赖性[65]。这个模型的提出是为了解决城市感知场景中数据稀缺的挑战。

UrbanGPT 整体的模型架构如图 4 所示[64]。模型主要包含 3 个主要部分，分别是时空依赖编码器、时空指令调整和时空零样本预测。图 4(a)是时空依赖编码器[66]，其中 L 为多层时间卷积网络的层数，n 为多层时间卷积网络层数的上限，用于界定层数的范围。在多层时间卷积网络中，通过引入门控机制，生成时间依赖表示，再结合多级相关注入层生成最终的时空依赖表示。在图 4(b)中，UrbanGPT 采用创新的指令调整方法进行训练，这一过程通过指令引导语言模型理解时空上下文。例如使用历史数据和词元来指导模型生成预测词元，随后通过文本替换技术将这些词元转化为实际的数值预测结果。这种结合时空模式和指令性文本的方法，使模型能够处理包含具体数值的预测任务，最终能够从文本指令和时空信号中学习。图 4(c)阐释了模型在进行零样本预测时的工作方式，UrbanGPT 可以在一个城市(训练城市)接受训练，然后转移到另一个未见过的城市(测试城市)进行测试，在这一过程中不需要模型在测试城市上接受额外的训练。该模型在跨城市泛化中展现出卓越性能，通过多任务学习、参数训练与冻结，实现高效知识迁移，并凸显了基于文本指令进行任务分配和预测的显著优势。

UrbanGPT 模型的优化还采用了绝对误差损失和分类损失作为联合损失优化策略的一部分，使其能够有效地处理多种预测任务。研究人员通过将 UrbanGPT 与传统时空模型(如 STGCN[67] 和 AGCRN[68]等)相比，结果显示 UrbanGPT 在文本理解和零样本学习上具有明显优势，并且在跨城市



(a)　时空依赖编码器　　(b)　时空指令调整　　(c)　时空零样本预测

图 4　UrbanGPT 模型整体架构

Fig. 4　Overall architecture of UrbanGPT model

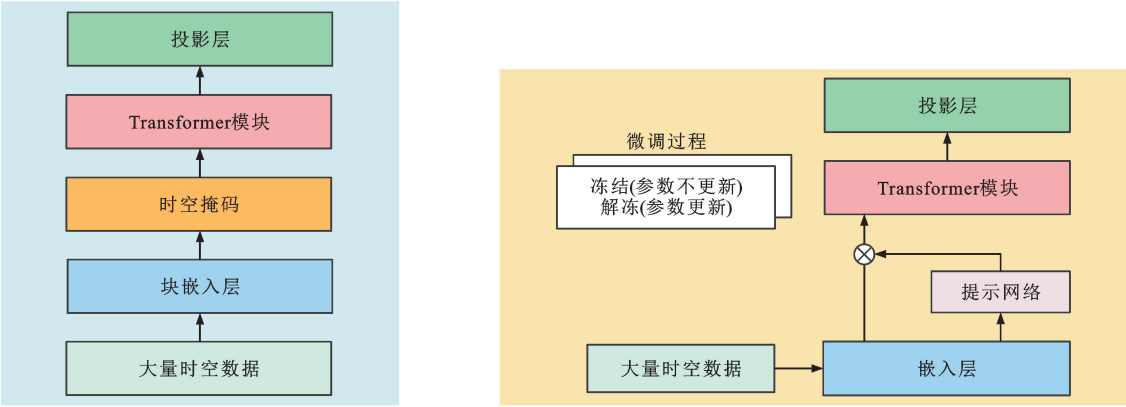
预测时泛化能力较强。将 UrbanGPT 与其他大模型(如 LLaMA-70B)相比,UrbanGPT 在处理数值型时空数据上更加具有优势。

作为一种针对城市时空数据预测的大模型,UrbanGPT 具有显著的优点和一些局限性。(1)模型具备强大的时空依赖建模能力:通过引入时空依赖编码器,可以更好地捕捉时空数据中的复杂模式与依赖关系。(2)模型的泛化能力强:在零样本学习场景下表现出色,即使在数据稀缺或缺少标记的情况下,依然能够做出准确预测。(3)结合了语言模型与时空信息:通过将语言模型与时空信息相结合,能够处理多模态数据,并从文本指令中提取有用的语义信息。(4)具有可解释性与可拓展性:通过轻量级对齐模块,可以有效整合时空依赖和语言表示,使得模型更加具有可解释性。然而,由于模型需要处理大规模城市数据,因此,对高计算资源依赖较大。虽然模型在短期预测中表现出色,但在处理长时序数据时,其表现可能会受到限制,需要进一步优化对长时间依赖的建模能力。

综上,UrbanGPT 模型的提出不仅提升了交通流量和相关事件的预测精度,还为城市交通管理和规划提供了强大的技术支持,但仍需进一步优化其计算开销和长时间依赖建模能力。

2.3 UniST 城市时空预测大模型

Yuan 等[69]在 2024 年 2 月提出了交通大模型 UniST,这是一个用于城市时空预测的通用模型。它主要用来处理并预测城市环境中的交通动态,例如车流量、人流移动、交通事故的概率等。UniST 的模型架构如图 5 所示[69],它包含 2 个模块:时空预训练模块和预训练模型微调模块。



(a) 时空预训练模块 (b) 预训练模型微调模块

图 5 UniST 的模型架构

Fig. 5 Architecture of UniST model

在时空预训练模块中,模型首先将大量城市时空数据处理成数据块,称作 ST Tokens。然后,通过块嵌入层,将 ST Tokens 编码成模型能够理解的嵌入向量,以便后续处理。在时空掩码模块中,模型引入了一种训练机制,该训练机制能使模型学习预测被遮蔽的部分,从而提高对数据的理解和预测能力。然后,嵌入的数据块将被送入 Transformer 模块进行处理,该模块是构成模型的主体部分,能够捕获数据中复杂的时空关系。最后,投影层将 Transformer 模块处理后的向量转化成最终的预测结果。

在预训练模型微调模块中,将多样的时空数据作为输入,先进入嵌入层操作,得到嵌入数据。此时,提示网络开始发挥作用,通过其中的空间记忆池和时间记忆池产生相关的提示信息,这些提示信息和嵌入数据将被一同输入到 Transformer 模块中。最后,经过投影层,可以将 Transformer 模块处理后的向量转化成最终的预测结果。此外,在整个微调过程中,模型的一些参数保持不变(即冻结参数),“冻结参数”这一步骤可以保留预训练过程中学到的通用特征,避免过拟合,还减少了计算成本。而另一些参数则会被调整(即微调),通过更新这些参数,模型能够更好地捕捉并适应特定任务的特征,从而提高预测性能。

与传统模型(如 STResNet[70] 和 ACFM[71] 等)相比,UniST 在不同城市和不同类型的时空数据上都表现出色,因此,无需为每个场景都设计特定的模型。与其他大模型(如 TrafficBERT[72] 等)相比,UniST 通过提示学习实现了对不同场景的自适应处理,在零样本和小样本学习任务中表现优异。

UniST 作为一种面向城市时空预测的大模型,具备多项优势:通过大规模时空预训练和提示学习,

模型能够在不同城市和场景中进行有效预测，尤其在零样本和小样本学习任务中表现优异；模型能够高效地处理多样化的时空数据，借助时空分块技术统一不同格式的数据，例如交通、人群流动等数据；基于时空知识的提示学习机制，使模型能够自适应不同场景中的时空模式，增强了模型的泛化能力；模型在长短期预测任务中都表现突出，能够有效地捕捉复杂的时空依赖关系，尤其在长时间序列预测中表现优于其他模型。模型的跨域学习能力也很强，在跨城市任务中，无需大量训练数据即可进行准确预测。然而，UniST 也存在一定的局限性：模型采用 Transformer 架构和大规模预训练，对计算资源的需求较高，训练成本较大；模型时空数据的多样性相较于语言或视觉数据有限，这可能限制其进一步扩展能力。

综上，UniST 在通用性和跨域学习能力上展现了明显优势，特别是在零样本或小样本场景中，其迁移学习能力比传统模型更加高效。然而，较高的计算成本和时空数据的有限多样性仍然是其主要挑战。

2.4　其他交通大模型

2023 年 4 月，商汤科技提出了“日日新 SenseNova”大模型，通过与视觉识别等先进技术相结合，显著提升了交通管理的智能化水平，推动交通行业进入 AI＋交通的 2.0 时代。为了解决绍兴市交通数据匮乏和交通管理能力有限的问题，该模型在绍兴的应用取得了显著成效。具体表现在以下几方面：首先，利用现有的摄像头收集大量交通数据并进行分析处理；其次，推动交通管理的自动化运营，减少对人力资源的依赖；然后，通过优化交通信号配时，有效管控交通流量，从而增强交通安全管理；最后，建立时空全息交通数据系统，显著提升了交通运营的智能化水平。总之，通过深入的数据分析、自动化运营、信号优化、安全管理以及全面的运营管控，“日日新 SenseNova”大模型促进了交通行业智能化的全面发展，为城市交通带来了效率和安全的双重提升。

2023 年 6 月，佳都科技发布了专为轨道交通行业设计的佳都知行交通大模型。该模型基于先进的 Transformer 技术，融合了多轮对话、复杂推理、数据分析、知识问答以及内容生成能力。通过针对性地优化交通行业的特定数据与信息，为城市交通行业提供了更加智能、高效且实时的服务。佳都知行交通大模型通过融合多种 AI 技术，其中包括深度学习和大数据分析等，致力于提升客户服务、运维管理和应急指挥的智能化水平。模型通过引入 AI 驱动的智能问答系统，显著提高了与客户互动的效率与品质。同时，模型利用 AI 进行故障模式分析，极大提高了维修工作的效率。未来，佳都科技计划加强交通大模型与城市、铁路等相关行业合作，构建一个智能交通新生态，推动传统交通系统向数字化的转型。这一转型将实现全方位感知、全面连接、多场景覆盖和全面智能化，促进智慧城市轨道交通的发展。

最后介绍交通领域内的常见数据集，如表 2 所示，其中包括智能交通和自动驾驶相关数据集。这些数据集通过数据预处理，能够处理成适合大模型输入的格式，最终应用于大模型的预训练和微调阶段。

3　大模型在交通领域的应用

在当前的交通行业中，大模型已成为一项关键技术，它在多个交通领域发挥着重要作用。本节将重点探讨大模型在交通管理和控制、交通安全和自动驾驶这 3 个领域的应用及其带来的革新。

3.1　交通管理和控制

尽管大模型在通用领域表现出色，但在应对复杂多变的交通环境时仍存在挑战。本节将对大模型在交通管理和控制领域的应用做出总结，并探讨各个大模型的优缺点。典型应用如图 6 所示。

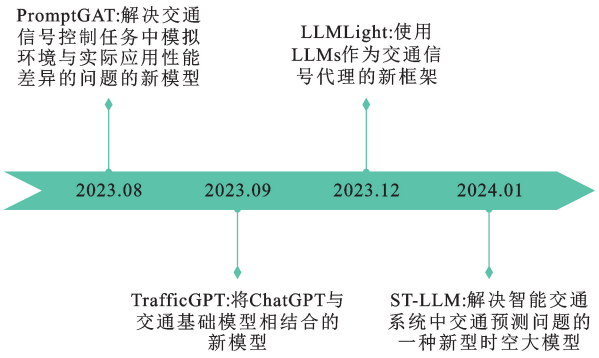


图 6　大模型在交通管理和控制领域的应用

Fig. 6　Applications of LLMs in traffic management and control

3.1.1　PromptGAT

PromptGAT 由 Da 等[83]提出，为了克服传统模型在交通管理和控制方面的局限[84]，该模型主要解决交通信号控制任务中模拟环境与实际应用性能差异的问题。模型结合了大模型的推理能力和领域知识，通过提示驱动的行为转换机制，更好地预测真实世界系统的动态。与传统GAT[85]等模型相比，

表 2 交通领域内的常见数据集

Table 2 Common datasets in transportation

数据集	简介	链接
nuScenes[73]	收集了波士顿和新加坡近 1 000 个复杂的驾驶场景。数据集由 140 万张图像、39 万次激光雷达扫描和 140 万个 3D 人工注释边界框组成	https://nuscenes.org/nuscenes
Mapillary Vistas Dataset[74]	一个大规模街道级图像数据集,包含 2.5 万个高分辨率图像,有 66 个对象类别,另有 37 个类别特定于实例的标签	https://www.mapillary.com/dataset/vistas
ApolloCar3D[75]	包含 5 277 个驾驶图像和超过 6 万的汽车实例,其中每辆汽车都配备了具有绝对模型尺寸和语义标记关键点的行业级 3D CAD 模型	https://apolloscape.auto/car_instance.html
BBD 100K	由 10 万个视频和各种注释组成,包括图像级别标记,对象边界框,可行驶区域,车道标记和全帧实例分割。该数据集具有地理、环境和天气多样性	http://bdd-data.berkeley.edu/
The SYNTHIA Dataset[76]	由 13 个类别精确的像素级语义注释:天空、建筑、道路、人行道、围栏、植被、杆、汽车、标志、行人、骑自行车的人、车道标记	https://synthia-dataset.net/
KUL Belgium Traffic Sign Dataset	包含数千个不同的交通标志,1 万多个交通标志注释。使用 8 个高分辨率摄像头录制的 4 个视频序列安装在 1 辆面包车上,录制时间总计超过 3 h	https://btsd.ethz.ch/shareddata/
Bosch Small Traffic Lights Dataset[77]	包含 13 427 个分辨率为 1 280 像素×720 像素的摄像机图像,并包含约 2.4 万个带注释的交通信号灯。其中注释包括交通信号灯的边界框以及每个交通信号灯的当前状态	https://hci.iwr.uni-heidelberg.de/content/bosch-small-traffic-lights-dataset
GTSRB[78]	德国交通标志基准测试数据集,其中有超过 40 个类别,一共超过 5 万张图像	https://benchmark.ini.rub.de/gtsrb_dataset.html
Tsinghua-Tencent 100K[79]	一个大型交通标志基准数据集,有超过 10 万张图像,包含了 3 万个交通标志,这些图像涵盖了照明度和天气变换的差异	https://cg.cs.tsinghua.edu.cn/traffic-sign/
MS COCO[80]	一个大型的物体检测、分割数据集。以场景理解为目标,通过截取复杂的日常场景,然后进行精确分割并标定位置。图像包括 91 个类别,32.8 万个影像和 250 万个标签	https://cocodataset.org/
UA-DETRAC[81]	一个多目标检测和多目标跟踪基准。其中超过 14 万个帧,标注了 8 250 个车辆和 121 万个标记的对象边界框	https://www.kaggle.com/datasets/dtrnngc/ua-detrac-dataset
BoxCars[82]	包括 11.6 万张车辆图像。这些图像由多个监控摄像头拍摄,且来自于多个观察点	https://github.com/JakubSochor/BoxCars

PromptGAT 利用大模型的推理能力,在处理未观测状态时表现更为出色,有效减少了模拟环境与真实世界之间的性能差距。模型还能动态调整交通信号策略,优化交通流量和排队长度。然而,模型的主要缺点是计算复杂度较高和对预训练模型有较高依赖。总体而言,对于解决交通信号控制中模拟到现实迁移的问题,PromptGAT 提供了一个有效的解决方案,但仍需考虑计算开销和模型依赖性的问题。

3.1.2 TrafficGPT

TrafficGPT 由 Zhang 等[5]提出,是一种将 ChatGPT 和交通基础模型相结合的新模型。首先,分析和处理交通数据,并使用 ChatGPT 为城市交通管理提供有价值的决策支持。其次,智能化地分解复杂任务,随后使用交通基础模型分别完成这些任务,从而增加模型处理复杂任务的能力。然后,通过自然语言对话,帮助人类在交通控制中进行决策。最后,通过交互式反馈和修订,提高了系统的适应性和可靠性。TrafficGPT 的优势在于能够有效处理复杂的交通数据,提供决策支持,并通过自然语言对话辅助人类进行交通管理任务。与 GPT-3 和 GPT-4 相比,TrafficGPT 在分析处理交通数据和提供决策方面表现出更高的准确性和更快的速度,尤其是在应对复杂任务和多模态数据时表现更优异。然而,其缺点在于过度依赖大模型的能力,如果遇到新颖或复杂的情况,可能会表现不佳。此外,TrafficGPT 对提示设计较为敏感,在处理模糊指令时可能需要人工干预,这在一定程度上限制了其完全自动化的能力。总体而言,TrafficGPT 通过结合 ChatGPT

和交通基础模型，在处理复杂交通任务和决策支持方面表现出色，但对提示设计的依赖性和新颖任务的适应性方面仍有待改进。

3.1.3　LLMLight

LLMLight 由 Lai 等[86]提出，是一种交通信号控制代理的新框架。交通信号控制的核心在于调整交通信号的相位变换时长，从而提高路网的运作效率。利用大模型强大的泛化能力和零样本推理能力，模型将任务描述、当前的交通状况及先验知识组合成一个提示，通过大模型的思维链推理出下一个交通信号的相位，从而优化路网的整体效率。与 GPT-4 相比，LLMLight 的专用模型（如 LightGPT）在交通信号控制任务中表现更加高效且成本更低。相较于强化学习模型，LLMLight 具有更强的泛化能力，适用于多种交通场景而无需依赖大量训练数据。在应对大规模交通流量和路网时，LLMLight 不仅展现出卓越的性能，还保持了高度的稳定性和可扩展性。然而，LLMLight 的训练过程相对复杂，计算资源需求较高，并且通用大模型在交通领域仍需进行特定的领域微调。总体而言，LLMLight 结合了大模型的泛化能力和领域专用优化，成为一种高效且具有解释性的交通信号控制解决方案，但在资源优化和领域微调方面仍有改进空间。

3.1.4　ST-LLM

ST-LLM 由 Liu 等[87]提出，是一种新型的时空大模型，专门优化智能交通系统中的交通预测任务[88]。ST-LLM 通过引入空间-时间嵌入和部分冻结注意力机制，有效捕捉了交通预测中的全局空间-时间依赖性，显著提升了预测精度。ST-LLM 还展现了强大的知识迁移能力，能够在少样本和零样本的场景中实现高效预测。然而，ST-LLM 对计算资源需求较高，尤其是在处理大规模数据时成本较大，同时模型性能高度依赖于数据质量和预训练模型的基础知识。通过在真实交通数据集上进行全面实验，对比其他大模型（如 GPT-2[89]和 LLaMA-2[90]等），ST-LLM 在交通预测任务中的表现更为优越。此外，尽管图神经网络（Graph Neural Network，GNN[91]）在捕捉局部空间依赖性方面表现出色，但在处理全局空间—时间依赖性时仍然不如 ST-LLM。总体而言，ST-LLM 在全局依赖性建模和复杂时间序列处理方面展现了强大的适应性和泛化能力。

3.2　交通安全

尽管大模型在各种通用领域的自然语言处理任务中表现出色，但在交通安全领域的任务中表现却不理想，主要是因为缺乏专门的交通安全专业数据集。本节对大模型在交通安全领域[92]的应用做出总结，并且探讨各个大模型的优缺点，典型的应用如图 7 所示。

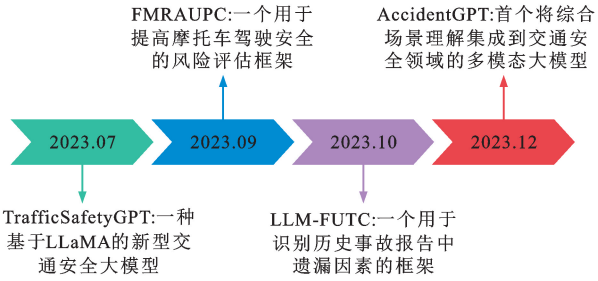


图 7　大模型在交通安全领域的应用

Fig. 7　Applications of LLMs in traffic safety

3.2.1　TrafficSafetyGPT

TrafficSafetyGPT 由 Zheng 等[93]提出，是基于 LLaMA 的新型交通安全大模型，主要解决大模型在交通安全领域任务中性能不足的问题。该模型采用有监督学习进行微调，使用 TrafficSafety-2K 数据集进行训练，该数据集包含政府指导手册中的人类标注内容和 ChatGPT 生成的指令-输出样本。通过这种方法，模型在交通安全知识问答方面表现出色，生成的文本更加专业且符合安全标准。与 LLaMA 和 ChatGPT 相比，TrafficSafetyGPT 在处理专业术语和特定场景时表现更为精准且简洁，尤其在定义、分类和指导等任务上展现了显著优势。同时，模型通过微调 LLaMA 模型的最后 2 层，大幅缩短训练时间并降低计算资源消耗，提升了效率。然而，TrafficSafetyGPT 依赖于交通安全领域的特定数据集，尽管通过生成数据进行了增强，但数据量仍有限，与通用大规模数据集相比，这可能限制其在其他领域的泛化能力。总体而言，TrafficSafetyGPT 在交通安全领域展现了较强的应用潜力。

3.2.2　FMRAUPC

基于全景相机的摩托机风险评估框架（Framework for Motorcycle Risk Assessment Using Panoramic Camera，FMRAUPC）是由 Jongwiriyanurak 等[94]提出的一个创新框架，通过摩托车头盔上的全景摄像头拍摄的视频来分析摩托车驾驶风险。与传统依赖历史事故数据和仿真模型的安全分析方法不同，该框架结合了图像到文本模型和大模型，能够从安装在摩托车驾驶员头盔上的全景摄像头采集的视频中识别潜在危险。具体而言，研究团队在泰国曼谷采集骑车人头盔上 360°全景视频数据，并通过图像

到文本模型将视频内容转化为文本描述，再利用大模型对这些描述进行分析，以识别危险并评估碰撞风险。实验结果表明，该框架能够有效识别静态和动态物体，并具备预警和事故分析的能力。进一步地，通过结合 LLaVA 模型[95]和交通事件风险相关的提示，研究团队探讨了在不同的环境条件下（如白天与夜晚）摩托车事件的风险，结果显示该模型适用于不同时间段和天气条件，具有较强的适应性。尽管这项研究结合了先进的图像处理和语言模型，但现有的图像到文本模型在处理全景图像时准确性仍然较低，尤其在识别车道数量、路面状况和交通标志方面表现不佳。总体而言，FMRAUPC 在摩托车交通风险评估中展现了明显的优势，但在全景图像处理和资源优化方面仍有改进空间。

3.2.3 LLM-FUTC

LLM-FUTC 由 Arteaga 等[96]提出，是一个结合大模型和自动化解析工具的框架，它能够识别历史事故报告中未被报告的事故因素，重点关注与酒精相关的事故案例。该框架通过对实际事故数据的分析，提高了安全分析的效率和准确性，显著减少人工总结过程中的错误和时间消耗。实验结果表明，LLM-FUTC 在识别漏报的酒精相关事故方面表现出色，特别是在采用 Flan-UL2[97]大模型的情况下，达到了 96%的准确率、88%的完全召回率和 92%的 F_1 分数，明显优于 ChatGPT 和 LLaMA-2。相比于传统的文本分类方法，LLM-FUTC 能够更高效地处理复杂的语言输入，尤其在应对隐式提示时表现突出。然而，该模型对提示设计和生成参数较为敏感，提示的隐式或显式设计可能显著影响识别结果，同时在生成参数调整时性能可能出现波动。总体而言，LLM-FUTC 展现了大模型在提升交通安全分析准确性和效率方面的潜力。

3.2.4 AccidentGPT

AccidentGPT 由 Wang 等[98]提出，是首个将综合场景理解引入交通安全领域的多模态大模型。基于车联网通信（Vehicle-to-Everything，V2X）架构，AccidentGPT 建立了一个全面的感知系统，能够全面监测道路环境，包括三维目标检测、鸟瞰图感知及车辆运动轨迹预测。通过结合自然语言推理能力与感知技术，该模型实现了对道路动态的深入理解和预测，大幅提升了复杂交通场景的响应效率。AccidentGPT 模型通过整合多传感器感知和人机交互技术，提供了对交通环境的全方位理解。其 V2X 协同感知技术能够从多个角度获取数据，精确预测潜在交通事故，不仅为自动驾驶车辆提供场景感知和碰撞规避功能，还为人工驾驶车辆提供实时安全预警和驾驶建议。此外，AccidentGPT 还引入了大模型，增强了自然语言交互能力，可以基于感知数据生成事故分析报告和安全提示，广泛适用于交通管理和执法部门。与传统交通安全模型（如 CoBEVT[99]、DiscoNet[100]、V2X-ViT[101]）和其他基于大模型的模型（如 TrafficGPT 等）相比，AccidentGPT 具备更全面的整合能力和更广泛的应用场景，尤其是在 V2X 联网环境下表现突出。尽管 AccidentGPT 在多车辆和车路协同感知上表现优异，但其表现主要依赖于传感器和基础设施的支持，在非联网或基础设施较差的环境中效果不佳。总体而言，AccidentGPT 通过结合实时感知与大模型推理技术，在智能交通管理系统中展现了强大的应用潜力。

3.3 自动驾驶

近年来，深度神经网络的快速发展极大地提升了自动驾驶技术[102]在计算机视觉和深度强化学习领域的进步。这不仅显著提升了车辆对环境的感知和解析能力，还优化了运动规划与决策制定过程。随着传感器技术的升级和计算能力的增强，更加强大的模型被应用于自动驾驶系统，进一步提升了系统的精确性和可靠性。然而，自动驾驶在极端气候、光照不佳或其他复杂环境中仍面临严峻挑战。为了应对这些问题，研究者们正探索多种方法以提高系统的安全性和鲁棒性。其中包括利用大模型提升整体性能，依托车辆间通信技术增强系统稳定性，以及通过对抗性训练强化系统在关键安全场景下的应对能力。本节将概述自动驾驶领域大模型的最新进展，典型的应用如图 8 所示。

3.3.1 DriveDreamer

DriveDreamer 由 Wang 等[103]于 2023 年 9 月提出，是一个完全基于真实驾驶场景的世界模型。该模型具备多项显著优势，包括从真实场景中生成高质量驾驶视频的能力、多样化驾驶场景中的可控性以及合理的驾驶策略生成能力。DriveDreamer 结合扩散模型和两阶段的训练方法，增强了对交通结构化信息的理解，使其适用于复杂的自动驾驶任务。在 nuScenes 基准测试上，DriveDreamer 通过广泛的实验验证，展现出精确且可控的视频生成能力，能够准确捕捉现实世界交通场景的结构约束。与 DriveGAN[104]相比，DriveDreamer 在可控性和生成质量方面表现更为优异，尤其在应对复杂的真实

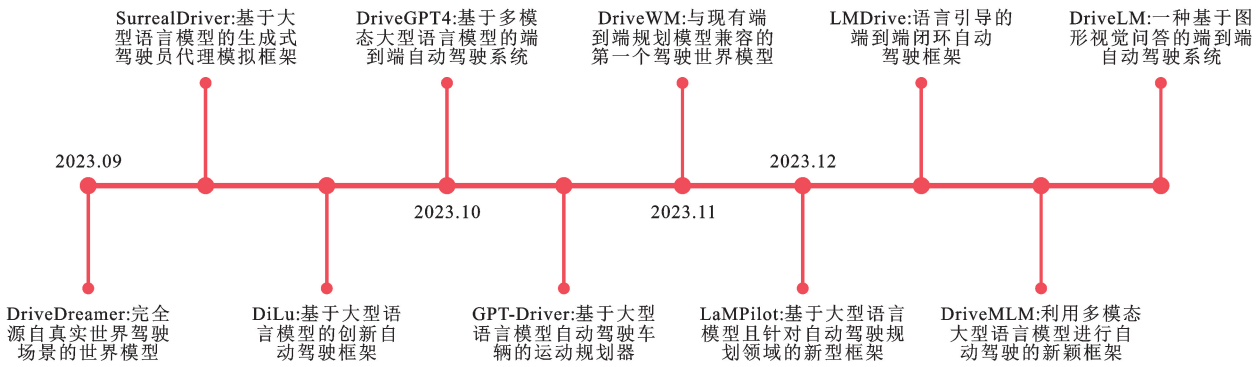


图 8　大模型在自动驾驶领域的应用

Fig. 8　Applications of LLMs in autonomous driving

驾驶场景时具有更大的优势。然而，该模型也面临一些挑战，包括计算复杂度较高、对高质量数据的强依赖性以及在处理非结构化环境时的局限性。总体而言，DriveDreamer 在复杂自动驾驶任务中的应用潜力显著，为进一步提升自动驾驶技术提供了强有力的支持。

3.3.2　SurrealDriver

SurrealDriver 由 Jin 等[105]提出，是一个基于大模型的生成式驾驶人代理模拟框架，主要用于提升交通模拟中驾驶行为的真实性和多样性。研究团队通过分析 24 名驾驶人的详细驾驶行为描述，将其作为链式思考提示，构建了一个“教练代理”模块，用于评估并指导驾驶员代理，从而培养出接近人类的驾驶风格。一系列模拟实验和用户测试验证了 SurrealDriver 的有效性，展现了其在生成真实可靠的驾驶员代理方面的卓越性能。其优势包括减少碰撞率、确保驾驶操作的连续性以及通过自然语言轻松引入驾驶规则。同时，模型具备持续学习和改进驾驶技能的能力。然而，由于依赖大模型决策，SurrealDriver 可能存在响应延迟问题，而模拟器环境的识别局限性也可能影响其在实际场景中的表现。与传统基于规则和数据的模型相比，SurrealDriver 更灵活，且更贴近人类驾驶行为，但在实时性和环境适应性方面仍有进一步提升的空间。这一框架为推进真实模拟驾驶员代理的开发提供了新思路和强大工具。

3.3.3　DiLu

DiLu 由 Wen 等[106]提出，是一个能够应对数据集偏差、过拟合和不可解释性等挑战的创新框架。该框架结合大模型的推理和反思模块，使自动驾驶系统不仅能够高效处理实时驾驶任务，还能通过自我经验的积累实现持续学习和进步。DiLu 的主要优势包括：通过内存模块的经验积累实现类人的决策能力，依靠反思模块实现自动学习与优化，具备较强的泛化能力和环境迁移能力。与传统的强化学习方法（如 GRAD[107]）相比，DiLu 拥有更强的适应性，能够更好地应对新环境和未知情况。然而，依赖大模型进行推理，DiLu 存在 5～10 s 的决策延迟，并可能出现“幻觉”问题，这些都需要进一步优化。尽管如此，相比于传统的数据驱动模型，DiLu 通过引入常识知识有效避免了过拟合和数据偏差。总体而言，DiLu 在实现自动驾驶系统的可靠性和可扩展性方面具有显著潜力，但在提升实时性和决策准确性上仍有改进空间。

3.3.4　DriveGPT4

DriveGPT4 由 Xu 等[108]提出，是一种基于多模态大模型的端到端自动驾驶系统，融合多模态输入（如视频和文本），能够预测车辆的低级控制信号并生成自然语言解释。其主要优点包括强大的多模态处理能力、端到端控制以及灵活的问答功能，特别是在车辆动态预测和决策解释方面表现出色。与传统自动驾驶方法 ADAPT[109]和其他大模型 GPT4-V 相比，DriveGPT4 能更准确地理解车辆的动态动作和回答多元化问题。此外，该系统在多任务的定性和定量分析中展现了卓越性能。为了优化模型，研究团队设计了一个专为自动驾驶调整的视觉指令数据集，使 DriveGPT4 能够解释车辆行为并提供相应的推理，而且灵活响应用户提问，显著提升了交互性。然而，DriveGPT4 在实时性和可靠性方面仍面临挑战，特别是在高计算成本和可能产生“幻觉”问题的情况下。总体而言，DriveGPT4 在多模态理解与交互能力上表现出色，为自动驾驶技术的进一步发展提供了新的方向。

3.3.5　GPT-Driver

GPT-Driver 由 Mao 等[110]提出，是一个将 GPT-3.5 模型转化为可靠自动驾驶车辆运动规划器的创新框架。运动规划作为自动驾驶的核心挑战

之一，目的是生成一个安全且合理的驾驶轨迹。GPT-Driver 通过以下流程实现这一目标：首先将自动驾驶车辆的传感器数据和环境信息转化为大模型可理解的词元；随后，模型利用其语言理解和生成能力推理出最佳的行动策略；最后，将策略转化为具体的运动指令并用于驾驶轨迹的生成。该框架采用提示-推理-微调策略，激活了大模型在数值推理方面的潜力，使其不仅能够精确描述轨迹坐标，还能通过自然语言解释决策过程。同时，GPT-Driver 具有良好的少样本学习能力，即使在数据量有限的情况下，仍然表现优异。尽管如此，GPT-Driver 在推理时间上较长，可能难以满足实时性要求。此外，目前框架对高分辨率地图等传感器数据的支持不足，且开放回路的规划评估没有完全考虑驾驶中的误差累积。与其他深度学习方法相比，GPT-Driver 在推理能力和精度方面具有显著优势，但在推理速度和复杂环境适应性上仍有改进空间。

3.3.6 DriveWM

DriveWM 由 Wang 等[111]提出，是首个与现有端到端规划模型兼容的世界驾驶模型。该模型能够根据不同的驾驶动作预测多个未来场景，并通过图像奖励模型确定最佳轨迹。DriveWM 具备卓越的多视角视频生成能力，结合了空间和时间建模，不仅能显著提升生成视频的质量和一致性，还能有效预测未来驾驶场景，增强了自动驾驶规划的合理性和鲁棒性。在分布外场景下，DriveWM 表现尤为突出，能够生成可靠的未来预测数据，并利用图像奖励函数选择最优轨迹，从而提升规划的安全性。此外，DriveWM 引入了一个统一条件接口，支持图像、文本、3D 布局和动作等多种条件输入，极大地简化了条件生成的复杂性。然而，DriveWM 生成过程计算成本较高，难以满足高实时性场景的需求，在实际应用中的表现还需要进一步的验证。相比于 VAD[112]等模型，DriveWM 在分布外场景处理上表现更优。与 DriveGAN、DriveDreamer 等单视角生成模型相比，DriveWM 首次实现了多视角一致性生成，并在生成质量上显著提升。此外，DriveWM 在生成复杂驾驶场景时也优于 MagicDrive[113]和 BEVControl[114]等多视角模型。

3.3.7 LaMPilot

LaMPilot 由 Ma 等[115]提出，是一个面向自动驾驶规划领域的新型框架。该框架将任务处理视为编程过程，通过预定义的基本动作生成可执行代码。例如，当收到“超车前方车辆”这样的自然语言指令时，LaMPilot 会启动代码生成过程，将这一复杂任务分解为一系列基本动作，并生成合理的执行序列，如“向左转”(进入超车道)、“加速”(超过前方车辆)、“向右转”(回到原先车道)。这些基本动作最终被转换为车辆控制系统可理解的具体代码指令。通过结合大模型与自动驾驶系统，LaMPilot 能够灵活理解并执行用户指令，显著提升系统的人机交互能力和任务的可解释性。其代码生成策略取代了直接控制信号，增强了任务处理的灵活性。然而，LaMPilot 在复杂场景中仍存在一定的碰撞率，且推理过程较长，难以满足实时性需求。与其他模型相比，LaMPilot 在灵活处理复杂任务方面表现更为出色。相较于 GPT-Driver 和 VAD 模型，LaMPilot 在处理复杂任务上更具优势，但实时性不如 GPT-Driver。而与传统基于规则的方法(如 IDM[116]和 MOBIL[117])相比，LaMPilot 在复杂任务完成率上有显著提升。

3.3.8 LMDrive

LMDrive 由 Shao 等[118]提出，是一种以语言为导向的端到端闭环自动驾驶框架。该框架基于 3 种输入源生成驾驶动作：(1)传感器数据(包括多视角相机和激光雷达)，用于生成与当前场景相匹配的驾驶动作；(2)导航指令(如变道、转弯等)，通常来源于人类或导航软件，确保驾驶行为符合自然语言的导航要求；(3)人类的提示指令，支持系统与人类互动，并适应建议和偏好，例如处理对抗性事件或突发状况。LMDrive 主要由视觉编码器和大模型及其相关组件这 2 个核心组件组成，其中视觉编码器负责处理多视角、多模态的传感器数据，进行场景理解并生成视觉词元。基于 LLaMA 的大模型及其相关组件(分词器、Q-Former 和适配器)，这些组件用于解析视觉编码器生成的词元，理解自然语言指令，生成控制信号并预测控制指令的完成情况。LMDrive 结合多模态传感器数据和自然语言指令，实现了端到端的闭环自动驾驶，能够实时生成控制信号并根据反馈调整输出，从而显著提升了复杂驾驶场景中的安全性和适应性。然而，LMDrive 的计算复杂度较高，推理时间较长，在应对复杂长指令时表现有所下降。与 GPT-Driver 相比，LMDrive 采用了多模态输入和闭环控制，处理复杂场景的能力更强。与 DriveGPT4 相比，LMDrive 在语言理解和闭环控制方面具有明显优势，适合实际场景的部署。

3.3.9 DriveMLM

DriveMLM 由 Wang 等[119]提出，是一个利用多

模态大模型实现自动驾驶的新框架。模型专注于模拟自动驾驶系统中的行为规划模块。其输入包括驾驶规则、用户指令和来自传感器（如摄像头、激光雷达）的多模态数据。DriveMLM 的核心特点是通过结合多模态传感器数据与大模型，进行闭环控制并提供解释说明，从而提升模型的透明性和用户信任度。它将自然语言指令与行为规划模块对齐，将复杂决策转化为具体的车辆控制信号，获得了优异的驾驶表现，尤其是在处理复杂场景时表现突出。然而，DriveMLM 的计算复杂度较高，在处理复杂长指令时存在一定的延迟。此外，模型在真实场景下的泛化能力有待优化。与 Apollo[120] 和 Interfuser[121] 等传统模型相比，DriveMLM 在处理动态场景和应对角落案例方面表现更加灵活。相比 DriveGPT4，DriveMLM 利用多模态输入显著增强了场景理解和闭环控制能力。

3.3.10　DriveLM

DriveLM 由 Sima 等[122]提出，是由上海人工智能实验室联合多方机构开发的一种基于视觉问答的端到端自动驾驶系统。该系统通过集成视觉语言模型（Visual Language Models，VLMs），实现了端到端的驾驶决策，显著提升了人机交互能力。此外，研究团队还设计了一个专门针对自动驾驶任务的 DriveLM 数据集，涵盖了感知、预测和规划等多个方面的问题—回答对注释，使 DriveLM 能够模拟人类驾驶员的推理过程。借助该数据集，DriveLM 在开放循环规划和端到端驾驶任务中表现卓越，尤其在处理未知场景和新对象时表现出强大的泛化能力。然而，DriveLM 的推理速度较慢，难以满足实时性要求，目前缺乏闭环规划支持，并且对传感器数据的处理能力也有限。与 GPT-Driver 相比，DriveLM 采用了更复杂的多步推理过程，而与 UniAD[123] 相比，DriveLM 在处理复杂场景和泛化能力上更有优势。相较于 BLIP-RT2，DriveLM 通过图结构问答能够在复杂场景下增强推理和决策能力。

4　结　语

（1）本文首先对大模型、视觉大模型和多模态大模型进行了概述和介绍，并分析了这些模型在处理自然语言、视觉信息以及多模态数据方面不同的特点和优势。然后，文章对现有的交通大模型做出总结。最后，文章展示了大模型如何为交通领域提供有力的技术支撑，具体到交通管理和控制、交通安全和自动驾驶等方面的主要应用。大模型的应用极大推动了交通领域的进步和发展，为智能交通系统带来了前所未有的可能性。可以预见，大模型在交通的众多领域将有广泛的应用前景，例如交通基础设施性能建模与分析[124]、自动驾驶安全提升等[125]。

（2）本文系统性地总结了交通大模型的现有应用，并通过对不同模型的实验比较，提出了大模型在交通流量预测、事故预防和自动驾驶等方面的应用策略。此外，本文发现了现有研究中的若干关键问题，例如：在稀疏数据条件下大模型的表现明显下降；大模型的训练和运行需要大量的数据和计算资源；在计算成本高和可能产生“幻觉”问题的情况下，模型的实时性和可靠性较低；模型在处理实时动态交通场景时也面临响应延迟等问题。在未来的研究中，针对这些问题，可以对交通大模型进一步优化。

（3）未来研究应当更加注重提升大模型的可解释性、准确性和安全性。通过引入基于可解释 AI 的透明模型结构，能够提高用户对模型输出结果的信任度。此外，通过针对大规模交通数据的优化训练和参数调整，模型的预测误差有望进一步降低。现有研究（如 UniAD、TransFuser 等）已经验证了此类改进在特定场景下的有效性，未来可以进一步拓展至更多的交通应用场景。最后，强化隐私保护机制，如通过差分隐私技术，确保用户数据在模型训练中的安全性。可以预见，随着 DeepSeek 这样开源、高性能大模型的出现，大模型的构建成本将迅速下降，大模型的应用将迅速普及，交通大模型必然会成为交通领域的研究热门方向。

（4）大模型在交通领域的研究将侧重于几个关键点：首先，深度结合交通领域知识的大模型，已被证明在交通信号控制和事故预防中具有显著效果（如 LLMLight、AccidentGPT 等），未来应进一步优化其任务处理能力；其次，探索多模态学习和智能体协作的潜力，以增强模型在复杂交通场景中的适应性和应变能力。通过集成外部 API，如天气和实时交通数据，模型能够更高效地执行交通管理任务，这一思路已经在部分交通系统中成功应用（如 LMDrive 等），未来可进一步推广至更广泛的智能交通管理领域。

参 考 文 献：

References：

[1] DIMITRAKOPOULOS G, DEMESTICHAS P. Intelligent transportation systems[J]. IEEE Vehicular Technology Magazine, 2010, 5(1): 77-84.

[2] LIN Yang-xin, WANG Ping, MA Meng. Intelligent transportation

system (ITS): concept, challenge and opportunity[C]// IEEE. 2017 IEEE 3rd International Conference on Big Data Security on Cloud (Bigdatasecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS). New York: IEEE, 2017: 167-172.

[3] OUYANG L, WU J, JIANG X, et al. Training language models to follow instructions with human feedback[C]// ACM. Advances in Neural Information Processing Systems. New York: ACM, 2022: 27730-27744.

[4] ACHIAM J, ADLER S, AGARWAL S, et al. GPT-4 technical report[R]. San Francisco: OpenAI, 2023.

[5] ZHANG Si-yao, FU Dao-cheng, LIANG Wen-zhe, et al. TrafficGPT: viewing, processing and interacting with traffic foundation models[J]. Transport Policy, 2024, 150: 95-105.

[6] ZHOU Xing-cheng, LIU Ming-yu, YURTSEVER E, et al. Vision language models in autonomous driving and intelligent transportation systems[J]. arXiv, 2023, DOI: 10.48550/arXiv.2310.14414.

[7] SHOAIB M R, EMARA H M, ZHAO Jun. A survey on the applications of frontier AI, foundation models, and large language models to intelligent transportation systems[C]// IEEE. 2023 International Conference on Computer and Applications (ICCA). New York: IEEE, 2023: 1-7.

[8] CUI Can, MA Yun-sheng, CAO Xue, et al. A survey on multimodal large language models for autonomous driving[C]// IEEE. 2024 IEEE/CVF Winter Conference on Applications of Computer Vision. New York: IEEE, 2024: 958-979.

[9] ZHENG Ou, ABDEL-ATY M, WANG Dong-dong, et al. ChatGPT is on the horizon: could a large language model be suitable for intelligent traffic safety research and applications?[J]. arXiv, 2023, DOI: 10.48550/arXiv.2303.05382.

[10] CUI Can, MA Yun-sheng, CAO Xu, et al. Receive, reason, and react: drive as you say, with large language models in autonomous vehicles[J]. IEEE Intelligent Transportation Systems Magazine, 2024, 16(4): 81-94.

[11] MCCORDUCK P, CFE C. Machines Who Think: a Personal Inquiry into the History and Prospects of Artificial Intelligence[M]. Natick: A. K. Peters, 2004.

[12] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[13] CRESWELL A, WHITE T, DUMOULIN V, et al. Generative adversarial networks: an overview[J]. IEEE Signal Processing Magazine, 2018, 35(1): 53-65.

[14] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// MIT Press. Proceedings of the 31st International Conference on Neural Information Processing Systems. Massachusetts: MIT Press, 2017: 6000-6010.

[15] DEVLIN J, CHANG Ming-wei, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[J]. arXiv, 2018, DOI: 10.48550/arXiv.1810.04805.

[16] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training[R]. San Francisco: OpenAI, 2018.

[17] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[C]// ACM. Advances in Neural Information Processing Systems. New York: ACM, 2020: 1877-1901.

[18] CHOWDHERY A, NARANG S, DEVLIN J, et al. PaLM: scaling language modeling with pathways[J]. Journal of Machine Learning Research, 2023, 24(240): 11342-11436.

[19] TAYLOR R, KARDAS M, CUCURULL G, et al. Galactica: a large language model for science[J]. arXiv, 2022, DOI: 10.48550/arXiv.2211.09085.

[20] TOUVRON H, LAVRIL T, IZACARD G, et al. LLaMA: open and efficient foundation language models[J]. arXiv, 2023, DOI: 10.48550/arXiv.2302.13971.

[21] WEI J, TAY Y, BOMMASANI R, et al. Emergent abilities of large language models[J]. arXiv, 2022, DOI: 10.48550/arXiv.2206.07682.

[22] SANH V, WEBSON A, RAFFEL C, et al. Multitask prompted training enables zero-shot task generalization[J]. arXiv, 2021, DOI: 10.48550/arXiv.2110.08207.

[23] WEI J, WANG Xue-zhi, SCHUURMANS D, et al. Chain-of-thought prompting elicits reasoning in large language models[J]. arXiv, 2022, DOI: 10.48550/arXiv.2201.11903.

[24] TAO Chao-fan, LIU Qian, DOU Long-xu. Scaling laws with vocabulary: larger models deserve larger vocabularies[C]// MIT Press. Proceedings of the 38th International Conference on Neural Information Processing Systems. Massachusetts: MIT Press, 2024: 1-33.

[25] HOFFMANN J, BORGEAUD S, MENSCH A, et al. Training compute-optimal large language models[J]. arXiv, 2022, DOI: 10.48550/arXiv.2203.15556.

[26] ZHAO W X, ZHOU Kun, LI Jun-yi, et al. A survey of large language models[J]. arXiv, 2023, DOI: 10.48550/arXiv.2303.18223.

[27] OQUAB M, DARCET T, MOUTAKANNI T, et al. DINOv2: learning robust visual features without supervision[J]. arXiv, 2023, DOI: 10.48550/arXiv.2304.07193.

[28] CARON M, TOUVRON H, MISRA I, et al. Emerging properties in self-supervised vision transformers[C]// IEEE. 2021 IEEE/CVF International Conference on Computer Vision(ICCV). New York: IEEE, 2021: 9630-9640.

[29] ZHOU Jing-hao, WEI Chen, WANG Hui-yu, et al. iBOT: image BERT pre-training with online tokenizer[J]. arXiv, 2021, DOI: 10.48550/arXiv.2111.07832.

[30] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]// PMLR. International Conference on Machine Learning. New York: PMLR, 2021: 8748-8763.

[31] TUO Yu-xiang, XIANG Wang-meng, HE Jun-yan, et al.

AnyText: multilingual visual text generation and editing[J]. arXiv, 2023, DOI: 10.48550/arXiv.2311.03054.

[32] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models[C]// MIT Press. Proceedings of the 33rd International Conference on Neural Information Processing Systems. Massachusetts: MIT Press, 2020: 6840-6851.

[33] MA Jian, ZHAO Ming-jun, CHEN Chen, et al. GlyphDraw: seamlessly rendering text with intricate spatial structures in text-to-image Generation[J]. arXiv , 2023, DOI: 10.48550/arXiv.2303.17870.

[34] CHEN Jing-ye, HUANG Yu-pan, LYU Teng-chao, et al. Textdiffuser: diffusion models as text painters[C]// MIT Press. Proceedings of the 37th International Conference on Neural Information Processing Systems. Massachusetts: MIT Press, 2023: 1-35.

[35] JIANG Yu-ming, WU Tian-xing, YANG Shuai, et al. Videobooth: diffusion-based video generation with image prompts[C]// IEEE. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2024: 6689-6700.

[36] RUIZ N, LI Yuan-zhen, JAMPANI V, et al. DreamBooth: fine tuning text-to-image diffusion models for subject-driven generation[C]// IEEE. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2023: 22500-22510.

[37] WANG Yi, LI Kun-chang, LI Xin-hao, et al. Computer Vision-ECCV 2024 [M]. Berlin: Springer International Publishing, 2024.

[38] ZHAO Long, GUNDAVARAPU N B, YUAN Liang-zhe, et al. Videoprism: a foundational visual encoder for video understanding[J]. arXiv, 2024, DOI: 10.48550/arXiv.2402.13217.

[39] LIU Ye, LI Si-yuan, WU Yang, et al. Umt: unified multi-modal transformers for joint video moment retrieval and highlight detection[C]// IEEE. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2022: 3032-3041.

[40] ZHU De-yao, CHEN Jun, SHEN Xiao-qian, et al. MiniGPT-4: enhancing vision-language understanding with advanced large language models[J]. arXiv, 2023, DOI: 10.48550/arXiv.2304.10592.

[41] CHIANG W L, LI Z, LIN Z, et al. Vicuna: an open-source chatbot impressing GPT-4 with 90% * ChatGPT quality[R/OL]. 2023, https: //vicuna. lmsys. org.

[42] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale[J]. arXiv, 2020, DOI: 10.48550/arXiv.2010.11929.

[43] SHARMA P, DING Nan, GOODMAN S, et al. Conceptual captions: a cleaned, hypernymed, image alt-text dataset for automatic image captioning[C]// USAACL. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Stroudsburg: USAACL, 2018: 2556-2565.

[44] ORDONEZ V, KULKARNI G, BERG T. Im2text: describing images using 1 million captioned photographs [C] // ACM. Proceedings of the 25th International Conference on Neural Information Processing Systems. New York: ACM, 2011: 1143-1151.

[45] SCHUHMANN C, VENCU R, BEAUMONT R, et al. Laion-400m: open dataset of CLIP-filtered 400 million image-text pairs[J]. arXiv, 2021, DOI: 10.48550/arXiv.2111.02114.

[46] ZANG Yu-hang, LI Wei, HAN Jun, et al. Contextual object detection with multimodal large language models[J]. arXiv, 2023, DOI: 10.48550/arXiv.2305.18279.

[47] CARION N, MASSA F, SYNNAEVE G, et al. Computer Vision-ECCV 2020 [M]. Berlin: Springer International Publishing, 2020.

[48] HE K M, GKIOXARI G, DOLLáR P, et al. Mask R-CNN[C]// IEEE. 2017 IEEE International Conference on Computer Vision(ICCV). New York: IEEE, 2017: 2980-2988.

[49] YANG Zheng-yuan, LI Lin-jie, LIN K, et al. The dawn of LMMs: preliminary explorations with GPT-4V(ision)[J]. arXiv, 2023, DOI: 10.48550/arXiv.2309.17421.

[50] ANIL R, BORGEAUD S, ALAYRAC J B, et al. Gemini: a family of highly capable multimodal models[J]. arXiv, 2023, DOI: 10.48550/arXiv.2312.11805.

[51] HENDRYCKS D, BURNS C, BASART S, et al. Measuring massive multitask language understanding[J]. arXiv preprint, 2020, DOI: 10.48550/arXiv.2009.03300.

[52] DONG Xiao-yi, ZHANG Pan, ZANG Yu-hang, et al. InternLM-XComposer2-4KHD: a pioneering large vision-language model handling resolutions from 336 pixels to 4KHD[J]. arXiv, 2024, DOI: 10.48550/arXiv.2404.06512.

[53] MATHEW M, KARATZAS D, JAWAHAR C V. DocVQA: a dataset for VQA on document images[C]// IEEE. 2021 IEEE Winter Conference on Applications of Computer Vision (WACV). New York: IEEE, 2021: 2200-2209.

[54] MASRY A, LONG Do X, TAN J Q, et al. ChartQA: a benchmark for question answering about charts with visual and logical reasoning [C] // USAACL. Findings of the Association for Computational Linguistics: ACL 2022. Stroudsburg: USAACL, 2022: 2263-2279.

[55] SINGH A, NATARAJAN V, SHAH M, et al. Towards VQA models that can read[C]// IEEE. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2019: 8317-8326.

[56] ROHRBACH A, HENDRICKS L A, BURNS K, et al. Object hallucination in image captioning[J]. arXiv, 2018, DOI: 10.48550/arXiv.1809.02156.

[57] LIU Yu-liang, LI Zhang, HUANG Ming-xin, et al. OCRBench: on the hidden mystery of OCR in large multimodal models[J]. arXiv, 2023, DOI: 10.48550/arXiv.2305.07895.

[58] CONTRIBUTORS O C. Opencompass: a universal evaluation

platform for foundation models[R]. GitHub Repository, 2023.

[59] BAI Jin-ze, BAI Shuai, YANG Shu-sheng, et al. Qwen-VL: a versatile vision-language model with versatile abilities[J]. arXiv, 2023, DOI: 10.48550/arXiv.2308.12966.

[60] WANG Wei-han, LYU Qing-song, YU Wen-meng, et al. CogVLM: visual expert for pretrained language models[J]. arXiv, 2023, DOI: 10.48550/arXiv.2311.03079.

[61] YOUNG A, CHEN Bei, LI Chao, et al. Yi: open foundation models by 01. ai[J]. arXiv, 2024, DOI: 10.48550/arXiv.2403.04652.

[62] WANG Peng, WEI Xiang, HU Fang-xu, et al. TransGPT: multi-modal generative pre-trained transformer for transportation[C]//IEEE. 2024 International Conference on Computational Linguistics and Natural Language Processing (CLNLP). New York: IEEE, 2024: 96-100.

[63] DU Zheng-xiao, QIAN Yu-jie, LIU Xiao, et al. GLM: general language model pretraining with autoregressive blank infilling[C]//ACL. Proceedings of the 60th Annual Meeting of the Association for Computational linguistics. Stroudsburg: ACL, 2022: 320-335.

[64] LI Zhong-hang, XIA Liang-hao, TANG Jia-bin, et al. UrbanGPT: spatio-temporal large language models[J]. arXiv, 2024, DOI: 10.48550/arXiv.2403.00813.

[65] 关为生,肖建力.联合时空特征的交通流参数预测综述[J].上海理工大学学报,2022,44(6):592-602.
GUAN Wei-sheng, XIAO Jian-li. A review on parameters prediction of traffic flow by combining spatio-temporal features[J]. Journal of University of Shanghai for Science and Technology, 2022, 44(6): 592-602.

[66] 龙佰超,关为生,肖建力.基于数据编解码的时空交通流预测方法[J].上海理工大学学报,2023,45(2):120-127.
LONG Bai-chao, GUAN Wei-sheng, XIAO Jian-li. Spatio-temporal traffic flow prediction method based on data encoding and decoding[J]. Journal of University of Shanghai for Science and Technology, 2023, 45(2): 120-127.

[67] YU Bing, YIN Hao-teng, ZHU Zhan-xing. Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting[J]. arXiv, 2017, DOI: 10.48550/arXiv.1709.04875.

[68] BAI Lei, YAO Li-na, LI Can, et al. Adaptive graph convolutional recurrent network for traffic forecasting[C]// ACM. Proceedings of the 34th International Conference on Neural Information Processing Systems. New York: ACM, 2020: 17804-17815.

[69] YUAN Yuan, DING Jing-tao, FENG Jie, et al. UniST: a prompt-empowered universal model for urban spatio-temporal prediction[C]//ACM. Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York: ACM, 2024: 4095-4106.

[70] ZHANG Jun-bo, ZHENG Yu, QI De-kang. Deep spatio-temporal residual networks for citywide crowd flows prediction[C]// ACM. Proceedings of the AAAI Conference on Artificial Intelligence. New York: ACM, 2017: 1655-1661.

[71] LIU Ling-bo, ZHANG Rui-mao, PENG Jie-feng, et al. Attentive crowd flow machines[C]//ACM. Proceedings of the 26th ACM International Conference on Multimedia. New York: ACM, 2018: 1553-1561.

[72] JIN K H, WI J A, LEE E J, et al. TrafficBERT: pre-trained model with large-scale data for long-range traffic flow forecasting[J]. Expert Systems with Applications, 2021, 186: 115738.

[73] CAESAR H, BANKITI V, LANG A H, et al. NuScenes: a multimodal dataset for autonomous driving[C]// IEEE. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2020: 11621-11631.

[74] NEUHOLD G, OLLMANN T, BULÒ S R, et al. The mapillary vistas dataset for semantic understanding of street scenes[C]// IEEE. 2017 IEEE International Conference on Computer Vision(ICCV). New York: IEEE, 2017: 5000-5009.

[75] SONG Xi-bin, WANG Peng, ZHOU Ding-fu, et al. ApolloCar3D: a large 3D car instance understanding benchmark for autonomous driving[C]//IEEE. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). New York: IEEE, 2019: 5447-5457.

[76] ROS G, SELLART L, MATERZYNSKA J, et al. The SYNTHIA dataset: a large collection of synthetic images for semantic segmentation of urban scenes[C]// IEEE. 2016 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2016: 3234-3243.

[77] BEHRENDT K, NOVAK L, BOTROS R. A deep learning approach to traffic lights: detection, tracking, and classification[C]//IEEE. 2017 IEEE International Conference on Robotics and Automation (ICRA). New York: IEEE, 2017: 1370-1377.

[78] STALLKAMP J, SCHLIPSING M, SALMEN J, et al. The German traffic sign recognition benchmark: a multi-class classification competition[C]//IEEE. The 2011 International Joint Conference on Neural Networks. New York: IEEE, 2011: 1453-1460.

[79] ZHU Zhe, LIANG Dun, ZHANG Song-hai, et al. Traffic-sign detection and classification in the wild[C]//IEEE. 2016 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2016: 2110-2118.

[80] LIN T Y, MAIRE M, BELONGIE S, et al. Computer Vision-ECCV 2014[M]. Berlin: Springer International Publishing, 2014.

[81] WEN Long-yin, DU Da-wei, CAI Zhao-wei, et al. UA-DETRAC: a new benchmark and protocol for multi-object detection and tracking[J]. Computer Vision and Image Understanding, 2020, 193: 102907.

[82] SOCHOR J, HEROUT A, HAVEL J. BoxCars: 3D boxes as CNN input for improved fine-grained vehicle recognition[C]//

IEEE. 2016 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2016: 3006-3015.

[83] DA Long-chao, GAO Min-quan, MEI Hao, et al. Prompt to transfer: sim-to-real transfer for traffic signal control with prompt learning[J]. arXiv, 2023, DOI: 10.48550/arXiv.2308.14284.

[84] 秦严严，罗钦中，贺正冰. 网联自动驾驶车辆混合交通流专用道管控方法[J]. 交通运输工程学报，2023，23(3)：221-231.
QIN Yan-yan, LUO Qin-zhong, HE Zheng-bing. Management and control method of dedicated lanes for mixed traffic flows with connected and automated vehicles[J]. Journal of Traffic and Transportation Engineering, 2023, 23(3): 221-231

[85] HANNA J, STONE P. Grounded action transformation for robot learning in simulation[C]// ACM. Proceedings of the AAAI Conference on Artificial Intelligence. New York: ACM, 2017: 4931-4932.

[86] LAI Si-qi, XU Zhao, ZHANG Wei-jia, et al. Large language models as traffic signal control agents: capacity and opportunity[J]. arXiv, 2023, DOI: 10.48550/arXiv.2312.16044.

[87] LIU Chen-xi, YANG Sun, XU Qian-xiong, et al. Spatial-temporal large language model for traffic prediction[J]. arXiv, 2024, DOI: 10.48550/arXiv.2401.10134.

[88] 户佐安，邓锦程，韩金丽，等. 图神经网络在交通预测中的应用综述[J]. 交通运输工程学报，2023(5)：39-61.
HU Zuo-an, DENG Jin-cheng, HAN Jin-li, et al. Review on application of graph neural network in traffic prediction[J]. Journal of Traffic and Transportation Engineering, 2023(5): 39-61.

[89] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners[EB/OL]. (2020-09-18)[2024-12-01], https://insightcivic.s3.us-east-1.amazonaws.com/language-models.pdf.

[90] TOUVRON H, MARTIN L, STONE K, et al. LLaMA 2: open foundation and fine-tuned chat models[J]. arXiv, 2023, DOI: 10.48550/arXiv.2307.09288.

[91] ZHOU Jie, CUI Gan-qu, HU Sheng-ding, et al. Graph neural networks: a review of methods and applications[J]. AI Open, 2020, 1: 57-81.

[92] 吴　兵，王文璇，李林波，等. 多前车影响的智能网联车辆纵向控制模型[J]. 交通运输工程学报，2020，20(2)：184-194.
WU Bing, WANG Wen-xuan, LI Lin-bo, et al. Longitudinal control model for connected autonomous vehicles influenced by multiple preceding vehicles[J]. Journal of Traffic and Transportation Engineering, 2020, 20(2): 184-194.

[93] ZHENG O, ABDEL-ATY M, WANG D D, et al. TrafficSafetyGPT: tuning a pre-trained large language model to a domain-specific expert in transportation safety[J]. arXiv, 2023, DOI: 10.48550/arXiv.2307.15311.

[94] JONGWIRIYANURAK N, ZENG Z C, WANG M H, et al. Framework for motorcycle risk assessment using onboard panoramic camera (short paper)[C]// Roger B, Dianna S, Sarah W, et al. Leibniz International Proceedings in Informatics (LIPIcs). Leeds: Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2023: 44: 1-44: 7.

[95] LIU Hao-tian, LI Chun-yuan, WU Qing-yang, et al. Visual instruction tuning[C]// MIT Press. Proceedings of the 37th International Conference on Neural Information Processing Systems. Massachusetts: MIT Press, 2023: 1-25.

[96] ARTEAGA C, PARK J W. A large language model framework to uncover underreporting in traffic crashes[J]. Journal of Safety Research, 2023, 92: 1-13.

[97] TAY Y, DEHGHANI M, TRAN V Q, et al. UL2: unifying language learning paradigms[J]. arXiv, 2022, DOI: 10.48550/arXiv.2205.05131.

[98] WANG Le-ning, REN Yi-long, JIANG Han, et al. AccidentGPT: accident analysis and prevention from V2X environmental perception with multi-modal large model[J]. arXiv, 2023, DOI: 10.48550/arXiv.2312.13156.

[99] XU Run-sheng, TU Zheng-zhong, XIANG Hao, et al. CoBEVT: cooperative bird's eye view semantic segmentation with sparse transformers[J]. arXiv, 2022, DOI: 10.48550/arXiv.2207.02202.

[100] MEHR E, JOURDAN A, THOME N, et al. DiscoNet: shapes learning on disconnected manifolds for 3D editing[C]// IEEE. Proceedings of the IEEE/CVF International Conference on Computer Vision. New York: IEEE, 2019: 3473-3482.

[101] XU Run-sheng, XIANG Hao, TU Zheng-zhong, et al. Computer vision - ECCV 2022[M]. Berlin: Springer International Publishing, 2022.

[102] 王润民，朱　宇，赵祥模，等. 自动驾驶测试场景研究进展[J]. 交通运输工程学报，2021，21(2)：21-37.
WANG Run-min, ZHU Yu, ZHAO Xiang-mo, et al. Research progress on test scenario of autonomous driving[J]. Journal of Traffic and Transportation Engineering, 2021, 21(2): 21-37.

[103] WANG Xiao-feng, ZHU Zheng, HUANG Guan, et al. DriveDreamer: towards real-world-driven world models for autonomous driving[J]. arXiv, 2023, DOI: 10.48550/arXiv.2309.09777.

[104] KIM S W, PHILION J, TORRALBA A, et al. DriveGAN: towards a controllable high-quality neural simulation[C]// IEEE. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2021: 5816-5825.

[105] JIN Y, SHEN X, PENG H, et al. SurrealDriver: designing generative driver agent simulation framework in urban contexts based on large language model[J]. arXiv, 2023, DOI: 10.48550/arXiv.2309.13193.

[106] WEN Li-cheng, FU Dao-cheng, LI Xin, et al. DiLu: a knowledge-driven approach to autonomous driving with large language models[J]. arXiv, 2023, DOI: 10.48550/arXiv.2309.16292.

[107] XI Z, SUKTHANKAR G. A graph representation for autonomous driving[C]// MIT Press. Proceedings of the 36th

International Conference on Neural Information Processing Systems. Massachusetts: MIT Press, 2022: 1-11.

[108] XU Zhen-hua, ZHANG Yu-jia, XIE En-ze, et al. DriveGPT4: interpretable end-to-end autonomous driving via large language model[J]. IEEE Robotics and Automation Letters, 2004, 9: 8186-8193.

[109] JIN Bu, LIU Xin-yu, ZHENG Yu-peng, et al. ADAPT: action-aware driving caption transformer[C]// IEEE. 2023 IEEE International Conference on Robotics and Automation (ICRA). New York: IEEE, 2023: 7554-7561.

[110] MAO Jia-geng, QIAN Yu-xi, YE Jun-jie, et al. GPT-Driver: learning to drive with GPT[J]. arXiv, 2023, DOI: 10.48550/arXiv.2310.01415.

[111] WANG Yu-qi, HE Jia-wei, FAN Lue, et al. Driving into the future: multiview visual forecasting and planning with world model for autonomous driving[J]. arXiv, 2023, DOI: 10.48550/arXiv.2311.17918.

[112] JIANG Bo, CHEN Shao-yu, XU Qing, et al. VAD: vectorized scene representation for efficient autonomous driving[C]// IEEE. 2023 IEEE/CVF International Conference on Computer Vision. New York: IEEE, 2023: 8340-8350.

[113] GAO Rui-yuan, CHEN Kai, XIE En-ze, et al. Magicdrive: street view generation with diverse 3D geometry control[J]. arXiv, 2023, DOI: 10.48550/arXiv.2310.02601.

[114] YANG Kai-rui, MA En-hui, PENG Ji-bin, et al. BEVControl: accurately controlling street-view elements with multi-perspective consistency via bev sketch layout[J]. arXiv, 2023, DOI: 10.48550/arXiv.2308.01661.

[115] MA Yun-sheng, CUI Can, CAO Xu, et al. LaMPilot: an open benchmark dataset for autonomous driving with language model programs[C]// IEEE. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2024: 15141-15151.

[116] TREIBER M, HENNECKE A, HELBING D. Congested traffic states in empirical observations and microscopic simulations[J]. Scientific Reports, 2000, 62(2): 1805-1824.

[117] KESTING A, TREIBER M, HELBING D. General lane-changing model MOBIL for car-following models [J]. Transportation Research Record, 2007, 1999(1): 86-94.

[118] SHAO Hao, HU Yu-xuan, WANG Le-tian, et al. LMDrive: closed-loop end-to-end driving with large language models[C]// IEEE. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2024: 15120-15130.

[119] WANG Wen-hai, XIE Jiang-wei, HU Chuan-yang, et al. DriveMLM: aligning multi-modal large language models with behavioral planning states for autonomous driving[J]. arXiv, 2023, DOI: 10.48550/arXiv.2312.09245.

[120] FAN Hao-yang, ZHU Fan, LIU Chang-chun, et al. Baidu apollo em motion planner[J]. arXiv, 2018, DOI: 10.48550/arXiv.1807.08048.

[121] SHAO Hao, WANG Le-tian, CHEN Ruo-bing, et al. Safety-enhanced autonomous driving using interpretable sensor fusion transformer[C]// PMLR. Conference on Robot Learning. New York: PMLR, 2023: 726-737.

[122] SIMA Chong-hao, RENZ K, CHITTA K, et al. DriveLM: driving with graph visual question answering[J]. arXiv, 2023, DOI: 10.48550/arXiv.2312.14150.

[123] HU Yi-han, YANG Jia-zhi, CHEN Li, et al. Planning-oriented autonomous driving[C]// IEEE. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2023: 17853-17862.

[124] HAN Bing-ye, DU Zeng-ming, DAI Lei, et al. Modeling the dynamic performance of transportation infrastructure using panel data model in state-space specifications[J]. Journal of Traffic and Transportation Engineering (English Edition), 2023, 10(3): 441-453.

[125] OLAYODE O I, DU B, SEVERINO A, et al. Systematic literature review on the applications, impacts, and public perceptions of autonomous vehicles in road transportation system[J]. Journal of Traffic and Transportation Engineering (English Edition), 2023, 10(6): 1037-1060.