车用人工智能标准体系研究报告

汽标委智能网联汽车分委会

中国汽车标准化技术委员会
智能网联汽车分技术委员会
车用人工智能标准专项组

2025 年 7 月

前言

近年来，人工智能（AI）技术的飞速发展为智能汽车领域带来了前所未有的机遇。得益于 AI 技术的支持，自动驾驶不仅能够高效地完成行驶任务，还能显著提升交通安全、优化出行效率、减少交通事故，并有效降低能源消耗。AI 赋能自动驾驶技术，使车辆能够实时感知、决策并执行复杂的驾驶任务，逐步实现从辅助驾驶到完全自主驾驶的转变。作为智能交通的重要组成部分，自动驾驶汽车正逐步走向商业化，并成为全球汽车产业技术创新和竞争的核心。

在全球智能网联汽车发展的竞争中，中国正处于关键的技术突破期，面临着从传统汽车制造到智能化、网联化转型的巨大挑战。AI 技术在自动驾驶领域的突破为中国汽车产业提供了巨大的发展潜力。为了在激烈的国际竞争中占据一席之地，中国必须紧抓智能驾驶、车联网与共享出行等新兴技术的发展机遇，推动汽车产业的快速转型升级，实现 AI 驱动的“智能化”成为中国汽车产业的核心竞争力。

因此，为了推动自动驾驶技术的全面应用与发展，促进智能交通体系的构建，提高交通安全与出行效率，推动产业升级与技术创新，中国汽车技术研究中心有限公司联合中国科学技术大学、理想汽车、小鹏汽车、吉利汽车、科大讯飞、赛迪、上检、东风汽车、中科南京软件院、上海交通大学、广汽、中国汽研、长安汽车、清华大学、比亚迪、阿里、百度、中国政法大学等 66 家单位，共同编写完成《车用人工智能技术及标准体系研究报告》。本报告旨在全面探讨智能网联汽车领域中的 AI 技术，从基础数据到智能系统的各个环节，详细分析了 AI 在智能网联领域各个功能任务中的应用与发展。

报告首先描述了智能网联汽车中的基础要素——数据，作为 AI 系统运行的基石，数据的高效收集与处理是实现智能网联汽车的关键。接着，分析并研究了车用 AI 相关术语及安全风险等级分类准则，确定了不同安全等级车用 AI 系统的管理思路。然后，通过深入分析车用 AI 的基础共性和平台建设，探讨了支撑智能网联汽车技术的技术平台与架构。随后，分析了 AI 在感知、预测、决策等驾驶任务中的应用，揭示了这些核心技术通过 AI 算法，提升环境感知的精度、预测交通行为的能力，并优化车辆的自主决策过程。在后续章节中，本报告还对车

汽标委智能网联汽车分委会

用 AI 的系统测试与评估进行了详细讨论，提出了智能网联汽车在数据安全、信息安全及功能安全等方面的测试评估标准需求，确保 AI 系统在实际应用中的安全性和可靠性。此外，本报告还涉及了智能座舱 AI 技术及其应用场景的标准化问题，以及伦理安全在智能网联汽车 AI 中的应用。

本研究报告集合行业头部车企、主流技术供应商、顶尖科研院所及第三方检测机构，通过分析行业现状、技术发展、法律法规及国内外标准，对车用人工智能标准体系构建及其可行性进行梳理与分析，为下一步国家标准制定提供参考。

在本研究报告编制过程中，各起草单位参阅了大量材料，并借鉴了行业的部分素材，鉴于篇幅有限，这里不一一列举，仅作诚挚的感谢！

在此，再次衷心感谢参与研究报告编写的各个单位和组织：

中国汽车技术研究中心有限公司、中汽研汽车检验中心（天津）有限公司、中国科学技术大学、中汽智联技术有限公司、北京车和家汽车科技有限公司、广州汽车集团股份有限公司汽车工程研究院、上海交通大学、东风汽车集团有限公司、科大讯飞股份有限公司、上海机动车检测认证技术研究中心有限公司、中国软件评测中心（工业和信息化部软件与集成电路促进中心）、广州小鹏汽车科技有限公司、吉利汽车研究院（宁波）有限公司、中科南京软件技术研究院、中国汽车工程研究院股份有限公司、阿里云计算有限公司、重庆长安汽车股份有限公司、中国政法大学、清华大学、清华大学苏州汽车研究院（相城）、大众汽车（中国）投资有限公司、比亚迪汽车工业有限公司、北京赛目科技股份有限公司、北京航空航天大学、重庆大学、同济大学、吉林大学、上海临港绝影智能科技有限公司、上海人工智能实验室、地平线征程（上海）科技有限公司、东软集团股份有限公司、高通无线通信技术（中国）有限公司、北京汽车研究总院有限公司、浙江吉利远程新能源商用车集团有限公司、通用汽车（中国）投资有限公司、一汽解放汽车有限公司、中国重型汽车集团有限公司、北汽福田汽车股份有限公司、上海汽车集团股份有限公司创新研究开发总院、上汽大众汽车有限公司、上汽大通有限公司、宁波路特斯机器人有限公司、长城汽车股份有限公司、上汽通用五菱汽车股份有限公司、广州文远知行科技有限公司、北京智能车联产业创新中心有限公司、江铃汽车股份有限公司、北京航迹科技有限公司、斑马网络技术有限

汽标委智能网联汽车分委会

公司、博世汽车部件（苏州）有限公司、浙江万安科技股份有限公司、惠州市德赛西威汽车电子有限公司、西华大学、北京百度智行科技有限公司、蔚来汽车科技（安徽）有限公司、中国移动上海产业研究院、中兴通讯股份有限公司、厦门金龙旅行车有限公司、岚图汽车科技有限公司、中汽研软件测评（天津）有限公司、北京与之科技有限公司、卡尔动力（北京）科技有限公司、中国汽车战略与政策研究中心、北京星也律师事务所、西南政法大学、重庆中科汽车软件创新中心。

主要编写人：

孙航、华一丁、吴含冰、秦一鉉、贺可勋、薛吟兴、刘子毅、赵帅、王琪瑞、王超、李涛、王文博、陈学文、王亚飞、汪博文、吴梁广、周文峰、李红林、雷琴辉、童鹏、王雪初、李淑玲、卢俊蓉、沈金华、于峰、张帆、王云英、王荣、邹博松、王越优、王钰、赵雄、常君璧、王新元、王潼、霍国君、许辉雨、曹川、陈勇、尹杨、蒋睿、黄承超、房科、李庆庆、吴天舒、潘登、梁桥、王琳、周超、邓兵、朱胜贤、杜相龙、罗薇、李昌、陈玥蓉、张凌寒、苏航、杨啸、王红、彭亮、陈海、李升波、陈晨、李金华、翟洋、李浩然、蔡孟池、丁延超、孙川、姜彦吉、杨俊儒、夏曦、孙琬、王玮、黄健超、郭光智、袁甜甜、高文、谢国樑、黄富荣、张伟豪、翁鹏、白智敏、薛晓卿、杨世春、曹耀光、王朋成、任秉韬、刘凯、李楚照、陈虹、张琳、孟强、杜严勇、胡云峰、麻斌、范亦卿、吴思雨、郭茜茜、周泽杨、程思颖、沈叶沁、许亮、陶莹、蔡品隆、石博天、窦民、翟宇麟、朱俊、汪鹏、李春林、何子豪、贾龙、李天然、陈书平、王江胜、于腾飞、苏冬静、林大洋、黄殿辉、王金奎、梁海强、黄伟伟、韦健林、刘庆勃、林元则、叶森、鲍凤卿、毛威、叶楠、韩睿、石俊杰、薛剑鸣、卢晶、谢良辉、田磊、王鹏程、尹俊、秦涛、刘阳、孙心洁、李军、田俊涛、金大鹏、厉承龙、侯件件、陈艳梅、朱丽敏、李莉、张力凡、王奇源、丁先山、刘翔、董旭、孙占朋、赵云霄、唐陆奇、甘鑫、林茂青、李绍鹏、唐顺成、周继祥、韩钧宇、闫晗、苏鹏飞、常婉渲、汪向阳、梁振宝、薛峰、任现、徐国荣、许瑞琛、全代勇、罗咏刚、李萱、陈军、李子兮、周博林、李若尧、杨东方、肖友、赵梓健、都大卫、李冲、孔炤、燕翔江、梁长乐、时乐泉、江会华、刘卫东、王艳华、李澜、徐季莹、余

汽标委智能网联汽车分委会

东瑾、韩勇、焦明顺、曹旭、龚悦、刘丁、赵洁、钱鹏程、朱奇章、姚玉丽、刘怡初、孙靖虎、张裁会、钱汉华、刘旭磊、白礼婷、程周、彭伟、陈思云、王凯、牟文珺、丁俊勇、王小亮、陈晓、黄程、俞海山、刘彦革、司华超、王文涛、夏显召、李予佳、翟瑞卿、闫明凯、刁兆拓、赵瑞、陈韵巧、沈萌、卫璁、赵佳、郑玮、谷珊琳子、郑志峰、袁梦、张立军、李仞钰。

汽标委智能网联汽车分委会

目录

一、车用 AI 基础技术及标准化 7
第一章 车用人工智能 基础要素 数据 7
1.1 车用人工智能数据现状 7
1.2 车用人工智能数据获取 17
1.3 车用人工智能数据管理和质量要求 22
1.4 总结与展望 29
第二章 车用 AI 基础共性 30
2.1 车用 AI 风险等级分类标准化研究 30
2.2 车用 AI 应用场景风险等级分类方法 41
2.3 车用 AI 治理等级的分类方法 53
2.4 总结与展望 59
第三章 车用 AI 基础系统平台标准化研究 60
3.1 车用 AI 基础系统平台研究概述 60
3.2 车用 AI 基础系统平台标准化研究 61
3.3 车用 AI 平台架构 77
3.4 总结与展望 90
参考文献 91
二、车用 AI 技术应用及标准化 93
第四章 车用 AI 智能网联汽车感知标准化研究 93
4.1 车用 AI 感知标准化研究概述 93
4.2 车用 AI 感知系统架构、关键任务与主要实现方法 95
4.3 车用 AI 感知测试 106
4.4 车用 AI 感知标准化需求与挑战 119
4.5 总结与展望 120
第五章 车用 AI 自动驾驶预测与决策标准化研究 121
5.1 车用 AI 预测与决策标准化研究概述 121
5.2 车用 AI 预测与决策系统简介 125
5.3 车用 AI 预测与决策测试及其评价方法 139
5.4 车用 AI 预测与决策评价指标 149
5.5 总结与展望 169
第六章 端到端系统标准化研究 171
6.1 研究背景与范围 171
6.2 端到端系统技术概述 187
6.3 端到端系统测试要求 206
6.4 总结与展望 214
第七章 智能座舱 AI 技术应用场景与标准化研究 215
7.1 智能座舱 AI 技术应用现状 215
7.2 智能座舱 AI 技术应用场景与技术路线探索 220
7.3 智能座舱 AI 应用的关键技术 245
7.4 智能座舱 AI 技术应用测试与评价的流程和要求 256
7.5 智能座舱 AI 技术标准化需求与关键挑战 266
7.6 总结与展望 268
参考文献 269

汽标委智能网联汽车分委会

三、车用 AI 安全及标准化 284
第八章 车用 AI 测试与评估标准化研究 284
8.1 车用 AI 测试评价技术概述 284
8.2 车用 AI 测试评价方法 284
8.3 车用 AI 测试评价流程 305
8.4 车用 AI 测试评价工具 307
8.5 总结与展望 314
第九章 车用 AI 数据安全测试与评估标准化研究 319
9.1 车用 AI 数据安全测试与评估现状 319
9.2 车用 AI 数据安全总体要求 327
9.3 车用 AI 数据安全测试技术研究 343
9.4 车用 AI 数据安全评估技术研究 361
9.5 总结与展望 367
第十章 车用 AI 信息安全测试与评估标准化研究 368
10.1 车用 AI 信息安全测试与评估现状 368
10.2 车用 AI 信息安全测试与评估通用要求标准化研究 374
10.3 车用 AI 信息安全测试工具标准化研究 379
10.4 车用 AI 信息安全测试流程标准化研究 388
10.5 车用 AI 信息安全评估流程标准化研究 410
10.6 总结与展望 421
第十一章 车用 AI 功能安全及标准化研究 423
11.1 车用 AI 功能安全标准化研究 423
11.2 车用 AI 功能安全要求与验证 426
11.3 总结与展望 451
第十二章 车用 AI 伦理安全标准化研究 452
12.1 车用 AI 伦理安全概述 452
12.2 车用 AI 伦理安全风险识别、评估与应对策略 456
12.3 车用 AI 伦理安全标准化研究 473
12.4 总结与展望 481
第十三章 车用 AI 安全管理体系标准化研究 482
13.1 研究背景与范围 482
13.2 车用 AI 安全管理体系标准化需求研究及挑战 483
13.3 车用 AI 安全管理体系搭建思路研究 492
13.4 总结与展望 499
参考文献 500
附录 507
附录 A：国内车用 AI 相关标准明细表 507
附录 B：国际车用 AI 相关标准明细表 510
附录 C：车用 AI 领域预计标准研制规划明细表 524
附录 D：车用 AI 领域标准研制路线图 528

汽车智能网联汽车分委会

一、车用 AI 基础技术及标准化

第一章 车用人工智能 基础要素 数据

1.1 车用人工智能数据现状

随着人工智能、大数据等新一代信息技术的蓬勃发展，极大拓展了智能网联汽车的功能边界。数据是人工智能赖以发展的核心资源。车用人工智能数据是指在汽车领域中，为支持人工智能算法的训练、测试、验证和应用所采集、处理和使用的数据。本章通过对国内外人工智能数据政策、标准分析，阐释车用人工智能数据研究的意义和目的，探讨车用人工智能数据的发展态势，识别车用人工智能数据的问题和难点，了解车用人工智能数据的应用现状，明确车用人工智能数据现状，结合智能网联汽车技术深度融合与跨领域协同的发展特点，构建车用人工智能数据生态，确保智能网联汽车产业高质量发展。

1.1.1 国内外人工智能数据法规标准情况调研分析

面对人工智能技术发展带来的机遇与挑战，国内外的人工智能数据法规正在逐步制定和完善，如下表 1-1 所示。

表 1-1 国内外人工智能数据法规汇总表

序号	国家/机构	法规英文名称	主要内容	发布时间
1	美国	Data Accountability and Transparency Act of 2020	将人工智能算法的自动化决策纳入监管，提出消费者应当有权对收集数据的理由提出质疑。	2020年06月
2	美国	Artificial Intelligence Act	致力于联邦政府更好、更安全地利用人工智能技术。	2021年05月
3	中国	中华人民共和国数据安全法	旨在加强和规范数据安全的管理和保护。除此之外，应明确的车用人工智能数据安全的法规标准，为各个组织和企业提供清晰的指导。	2021年6月
4	中国	生成式人工智能服务管理暂行办法	旨在促进生成式人工智能健康发展和规范应用，维护国家安全和社会公共利益，保护公民、法人和其他组织的合法权益。	2023年7月

汽标委智能网联汽车分委会

在人工智能领域，国际标准组织（ISO）等机构已经开展了大量的标准化工作，并且专门设立了 ISO/IEC JTC1 SC42 人工智能分技术委员会。目前，与人工智能安全相关的国际标准主要集中在基础概念和技术框架等通用标准方面，重点关注人工智能管理、可信性、安全以及隐私保护等领域。

表 1-2 ISO 数据相关标准

序号	标准号	标准英文名称	标准中文名称	主要内容
1	ISO/IEC 8183:2023	Information technology-Artificial intelligence-Data life cycle framework	信息技术人工智能数据生命周期框架	描述了一个全面的框架，涵盖数据管理的各个方面，从数据采集和创建到开发、部署、维护和最终退役的整个人工智能(AI)系统生命周期的数据处理操作要求，但并未定义特定的服务、平台或工具，该标准适用于在人工智能系统的开发和使用过程中的数据。
2	ISO/IEC 5259-1：2024	Artificial intelligence-Data quality for analytics and machine learning(ML） Part 1: Overview，terminology, and examples	人工智能分析和机器学习（ML）的数据质量 第1部分：概述、术语和示例	是 ISO/IEC 5259 系列的基础部分，专注于分析和机器学习(ML)的数据质量。建立了跨数据生命周期不同阶段评估和增强数据质量的框架，这对于可靠的分析和机器学习的结果至关重要。
5	ISO/IEC 5259-2	Artificial intelligence-Data quality for analytics and machine learning(ML） Part 2: Data quality measures	人工智能分析和机器学习（ML）的数据质量第2部分：数据质量措施	该标准主要针对人工智能领域中数据分析与机器学习过程中的数据质量要求，详细规定了数据质量评估的相关措施。
6	ISO/IEC 5259-3：2024	Artificial intelligence-Data quality for analytics and machine learning(ML） Part 3:Data quality management requirements and guidelines	人工智能分析和机器学习（ML）的数据质量第3部分：数据质量管理要求和指南	规定了在分析和机器学习(ML)领域中建立、实施、维护和持续改进数据质量的要求并提供了指导。 标准定义了质量管理过程的需求和指导，以及可定制的参考过程和方法，以满足文档中的需求。
7	ISO/IEC 5259-4	Artificial intelligence-Data quality for analytics and machine	人工智能分析和机器学	建立了通用的组织方法，无论应用组织的类

	：2024	learning(ML） Part 4: Data quality process framework	习（ML）的数据质量第 4 部分：数据质量过程框架	型、规模或性质如何，以确保分析和机器学习(ML)中培训和评估的数据质量。
8	ISO/IEC 5259-5	Artificial intelligence-Data quality for analytics and machine learning(ML） Part 5: Data quality governance framework	人工智能分析和机器学习（ML）的数据质量第 5 部分：数据质量治理框架	主要针对人工智能领域中数据分析和机器学习应用的数据质量治理问题。内容涵盖数据质量治理的总体框架，包括治理结构、流程设计、责任分配机制以及数据评估方法。
9	ISO/IEC 5259-6	Artificial intelligence-Data quality for analytics and machine learning(ML） Part 6: Visualization framework for data quality	人工智能分析和机器学习（ML）的数据质量 第 6 部分：数据质量的可视化框架	主要概述了在人工智能和机器学习应用背景下的数据可视化。提供一些示例，展示在 AI 生命周期的不同阶段，各利益相关方如何运用数据可视化。
10	ISO PAS 8800	Road Vehicles — Safety and artificial intelligence	道路车辆人工智能安全	定义在道路车辆环境中影响人工智能（AI）性能不足和故障行为的安全相关属性及风险因素，并描述涵盖全生命周期的框架，包括推导与数据质量及完整性相关的功能考虑的适当安全要求等。

此外，欧洲电信标准化协会（ETSI）关注的重点议题包括人工智能数据安全、完整性和隐私性、透明性、可解释性、伦理与滥用、偏见缓解等，已经发布多份人工智能安全研究报告。

国内车用人工智能数据相关标准尚在起步阶段，已发布标准主要集中在智能网联汽车数据、数据安全等领域。

1.1.2 车用人工智能数据研究的意义与目的

人工智能模型依赖于海量的数据训练，数据是人工智能系统的基石，也是人工智能发展最重要的要素。高质量的数据是训练准确模型的前提，人工智能做出的任何判断，都是对数据进行整合。人工智能模型训练所需的数据不仅需要多，而且还需要多样化、高质量化。目前在汽车行业，人工智能训练及测试使用的数据技术存在许多挑战，第一，目前车用人工智能训练、测试数据来源于第三方公

汽标委智能网联汽车分委会

开数据集、企业自建数据库、实车道路采集数据、道路 V2X 设备采集数据等，这些数据需要经过诸多步骤进行预处理才能真正用于人工智能模型训练，其流程复杂且无明确的指导规范，导致数据质量不一，没有规范化评判标准。第二，车厂或者供应商通过车辆自采的或者用户回传的数据存在较多敏感信息，数据安全问题是不可忽略的技术难题[1]。第三，各家数据采集方式、数据格式不统一，存在数据闭塞问题，数据利用率极低。因此通过对人工智能全过程数据处理标准化的制定，具备如下实际意义：

(1) 提高数据质量：通过标准化规范，降低数据错误、重复和不一致的现象，提高数据的准确性、完整性和可靠性。

(2) 促进数据资源共享，降低数据采集成本：建立统一的数据标准，提高数据在不同系统、部门之间的互操作性，方便数据的传递、共享和交换。

(3) 降低数据管理成本。通过标准化数据管理要求，降低数据维护、整合和处理的成本。

(4) 提升数据价值。数据规范化能够将不同格式、不同单位、不同标准的数据转换为统一的格式、统一的单位、统一的标准。数据标准化能够确保这些数据在处理和分析时具有一致性和可比性，从而提高数据的利用价值。

(5) 提高车用人工智能模型开发效率：标准化后的数据可以减少数据处理过程，减少人工智能模型开发的外在影响条件，提高研发效率。

车用人工智能数据标准化在提高数据质量、降低数据采集、管理成本、提升数据价值等方面具有重要意义。随着汽车产业的智能化进程不断加速，数据标准化的重要性将愈发凸显。

1.1.3 车用人工智能数据的问题与难点

从自动驾驶到智能导航，从语音交互到情感识别，人工智能正在深刻改变着人们的出行方式和驾驶体验。然而，在这个变革的浪潮中，车用人工智能数据的问题与难点也逐渐浮出水面，成为制约技术发展的关键因素。

汽标委智能网联汽车分委会

1.1.3.1 数据质量管理问题

车用人工智能数据的质量管理是一个复杂、繁琐而重要的任务，涉及多个方面，包括数据的一致性、完整性以及可追溯性等。

(1) 一致性问题：在智能化管理系统中，通常会使用多个数据源，并且这些数据源可能包含不同的数据格式和结构。这导致了数据在输入和输出过程中需要进行有效的验证和校准，以确保数据的一致性。不一致的数据可能会影响到模型及系统的泛化能力和稳定性，从而对驾驶安全产生负面影响。

(2) 完整性问题：缺失、不完整或不准确的数据会对模型及系统产生严重的影响。为了确保数据的完整性，需要建立一套完善的数据采集和录入机制，确保所采集的数据是完整和准确的。同时，还需要建立严格的数据审核和审计机制，对数据的完整性进行有效监控和管理。

(3) 可追溯性问题：在车辆智能化过程中，需要对数据进行有效的追踪和溯源，以便在出现问题时能够迅速定位并解决。这要求建立一个完善的数据追踪系统，记录数据的来源、处理过程和结果等信息，以便在需要时进行追溯。

1.1.3.2 法律遵从性问题

车用人工智能数据的使用和处理涉及到数据保护、数据跨境传输、个人信息保护、数据权益、知识产权等多个方面，涉及一系列复杂的法律遵从性问题，包括：

(1) 法律法规滞后：随着车用人工智能技术的快速发展，现有的法律法规体系往往跟不上技术进步的步伐。这意味着在收集、处理和使用车辆数据的过程中，可能没有明确的法律规定可以依循，从而导致无法律法规可遵从的问题。

(2) 跨境数据传输：数据跨境也是车用人工智能数据面临的重要法规问题。车用人工智能的发展，涉及到国内外企业的合作、跨国企业等。相关企

汽标委智能网联汽车分委会

业在开发过程中如需将数据跨境传输至其他国家的服务器进行处理和分析，数据跨境需要遵守国家相关法律法规要求，如：《网信办第 7 号令 汽车数据安全管理若干规定（试行）》、《网信办第 11 号令 数据出境安全评估办法》等。此外，不同国家有不同的数据保护法律，如欧盟的通用数据保护条例，均增加了法律遵从性的复杂性。

(3) 个人信息保护：个人信息保护是车用人工智能数据面临的重要法规问题。车辆在运行过程中会收集到大量的个人信息，如驾驶员的行驶轨迹、习惯偏好等。这些数据如果被滥用或泄露，将可能侵犯个人隐私权。因此，在使用和处理这些数据时，必须遵守相关的隐私保护法规，确保数据的合法性和安全性。

(4) 地理信息：车用人工智能在模型训练和实际使用等环节，都需要采集大量的地理位置和环境数据，相关活动已被自然资源部《自然资规[2022]1号 关于促进智能网联汽车发展维护测绘地理信息安全的通知》明确定义为测绘行为，因此相关活动需要满足《中华人民共和国测绘法》相关要求，当前的法律要求也对车用人工智能的发展造成一定的困扰。

(5) 数据确权和所有权：数据的所有权、使用权和收益权等都需要明确界定和分配。在数据共享和交换的过程中，法律需要明确各方的权利和责任，需要确保各方权益的公平性和合法性，避免产生纠纷和冲突，确保数据的合法使用。

(6) 透明度和用户同意：用户应该被充分告知他们的数据如何被收集、存储和使用，并且有权决定是否同意这些处理活动。这要求汽车制造商提供透明的隐私政策和用户同意机制。

(7) 事故责任归属：在自动驾驶模式下发生事故时，确定责任归属变得复杂。法律需要明确制造商、软件开发商、车主和其他可能涉及方的责任。

(8) 公众意识和教育：公众对于智能汽车的数据收集和使用可能缺乏足够的认识。因此，提高公众意识和进行相关教育是确保法律遵从性的重要组成部分。

汽标委智能网联汽车分委会

(9) 知识产权问题：知识产权也是车用人工智能数据不可忽视的一个方面。在开发和使用人工智能模型的过程中，可能会涉及到算法、模型结构、数据集等多个方面的知识产权。因此，需要遵守相关的知识产权法规，尊重他人的知识产权，避免侵权行为的发生。

除以上问题和难点外，车用人工智能数据还面临着伦理层面的挑战。例如，如何确保人工智能系统的决策公正性和透明性，避免因为数据偏见或算法歧视而导致的不公平现象；如何平衡人工智能系统的安全性和效率性，确保在保障安全的前提下提高系统的运行效率等。

1.1.4 车用人工智能数据的应用现状

近年来，汽车行业已与人工智能深度融合，为解决交通安全、道路拥堵、能源消耗和环境污染等问题提供了新的解决方案[2]。人工智能算法需要大量不同的数据进行训练和测试，以确保其在实际应用中的性能。目前，车用人工智能数据的应用主要是在智能驾驶与智能座舱领域，对应的数据分类有感知、预测与决策、端到端、驾驶大模型、AI 状态监测数据、舱驾融合、大模型人机交互数据等，此外，还有预测性维护与远程诊断等其他领域的应用。

1.1.4.1 在智能驾驶中的应用现状

人工智能通过深度学习和计算机视觉使汽车能够识别道路标志、行人、车辆及其他障碍物等，并实时规划出合理路径及控制指令，控制车辆完成驾驶任务。目前，主流的智能驾驶系统包括环境感知、定位与建图、预测、决策、规划和控制六大核心功能[3]。智能驾驶常用的数据可以分为三类：数字地图、车载传感器数据和车联网（Vehicle to Everything，V2X）数据。

数字地图为智能驾驶车辆提供了大量准确且丰富的先验知识，能够辅助车辆的环境感知、定位建图、决策规划和控制功能。数字地图主要包括导航地图、ADAS 地图和高精地图[4]。其中，导航地图主要通过二维经纬度来表达空间位置，能够与空间定位系统相结合，精准引导智能驾驶车辆从出发地到达目的地。它还支持位置查询、目的地检索、路线规划等功能，并能反映动态交通信息，完成实时路径计算与引导。在部分场景下，导航地图还可以支持 2.5D 伪三维展示，用

汽标委智能网联汽车分委会

于呈现重要地标或复杂路况。导航地图的定位精度一般在 10~20 m。ADAS 地图是在传统导航地图的基础上，增加了坡度、曲率、航向、车道数等信息而形成的电子地图。它不仅支持路线引导功能，还能为车辆的其他辅助驾驶功能提供支持，例如自适应速度建议、车道偏离预警和车道保持等功能。通过地图数据，ADAS 地图能够弥补车辆内置传感器探测范围有限的不足，利用数字地图数据沿车辆行驶路径“查看”更远距离的情况。ADAS 地图的定位精度一般可以达到 0.5~1 米。

高精地图则更为详细，它不仅包含了道路车道布局以及车道间的拓扑连接关系，还涵盖了人行道、不同类型隔离设施、车道线施画类型等信息，并且提供了丰富的环境信息，例如道路交通标志标牌、限速信息、交通灯位置、道路坡度、弯道曲率等。这些信息能够辅助车辆实现高精度定位、车道级路线规划、驾驶行为决策与控制等功能。高精地图的定位精度能够达到 0.1~0.5 米。

车载传感器数据包括视觉数据、雷达数据、GPS 数据和运动传感器数据等。视觉数据最接近于人类感知，包含有丰富的环境信息，可以用于目标检测、目标跟踪、构建语义地图等，但是，在晴、雨、雪、雾等恶劣天气下，相机的性能会迅速下降，而且，相机的观测距离有限，仅靠相机无法准确了解远距离的环境信息。同时，相机捕捉到的图像是 2D 图像，虽然可以在 2D 图像的基础上建立 3D 模型，但效果难以保证[5]。用于采集雷达数据的传感器有毫米波雷达、激光雷达、超声波雷达等。毫米波雷达是智能驾驶系统中最为成熟的传感器之一。它具有强大的环境适应性，能够直接获取目标的速度、距离和角度等准确信息，可用于目标检测、跟踪以及运动预测等任务，从而弥补了视觉数据的不足。然而，毫米波雷达输出的障碍物列表存在噪声较大的问题，漏检率和误检率也相对较高，并且无法提供跟踪目标的类别和尺寸信息。因此，毫米波雷达的数据主要用于补充其他传感器的数据。激光雷达与毫米波雷达在工作原理上有所不同，它采用更高频率的激光束进行探测，因此具备高分辨率和强隐蔽性的优势。不过，激光雷达在恶劣天气条件下的表现相对较弱。与晴朗天气相比，其在雨、浓烟、浓雾等异常天气中的感知距离会显著缩短，对异常天气的适应性较差。超声波雷达则根据探测距离分为近程和远程两种类型，通常安装在汽车的前后保险杠和侧面，用于纵向和侧向测距。尽管超声波雷达的测距精度有限且探测距离较短，但它具有频率

汽车零部件网联汽车分会

高、穿透能力强和成本低的特点。因此，它主要适用于倒车泊车等近距离感知场景，并且通常通过多个超声波传感器组成阵列来完成近距离感知任务。此外，还有 GPS 数据和 IMU、转角及轮速编码器等运动传感器数据，用于位置信息和运动状态信息的获取。不同类型的数据以不同的形式呈现周边的环境信息，在不同场景中发挥各自优势。因此，基于单一类型数据的感知方案往往难以实现全工况下的精准感知。而基于多源数据融合的感知技术可有效利用跨模态互补信息，能够克服单一类型数据的局限性，提升车载感知系统的性能表现。

V2X 包括车辆与交通系统中的任何其他组件之间的数据共享，包括车辆对车辆（V2V）、车辆对基础设施（V2I）和车辆对网络（V2N）[6]。V2V 使附近的车辆能够共享车载传感器和数字地图等数据，有助于更全面地了解驾驶场景。V2I 促进了智能驾驶车辆与基础设施组件之间的通信，例如嵌入在道路基础设施中的交通灯、标志或路边传感器等，协同工作以扩展感知范围并提高智能驾驶车辆的情境感知。V2N 指的是在车辆和更广泛的网络基础设施之间交换数据，通常利用蜂窝网络为车辆提供对云数据的访问。V2N 通过共享跨区域数据或提供有关交通拥堵或道路封闭的实时更新，帮助 V2V 和 V2I 的合作感知。智能驾驶车辆提供了丰富的与外界数据交流的方式，但同时众多的对外接口也带来了安全风险。物理层、网络层、传输层和应用层都有可能面临信息安全攻击[7]。

1.1.4.2 在智能座舱中的应用现状

智能座舱可以看作一个结合了“人、车、路、云”的智能化的服务空间，与人机交互有直接关系[8]。车用人工智能数据能够助力车企打造更聪明、更舒适、更好用的第三生活空间，实现对复杂车控、车书问答、闲聊陪伴、生活服务、多模态交互等智能化、场景化的应用。智能座舱常用的交互数据主要有视觉数据、听觉数据、触觉数据、嗅觉数据等。

智能座舱内的视觉数据主要通过仪表盘、中控、HUD 和氛围灯等实现交互，另外，手势识别技术还可以通过检测和识别驾驶员或乘客的手势来控制座舱环境和娱乐系统等，减少对座舱内物理按键的依赖。

汽标委智能网联汽车分委会

听觉数据涵盖音乐、语音提醒、警示音以及语音识别等方面。在与驾驶任务相关的场景中，听觉在感知事态方面发挥着至关重要的作用，有时甚至无法被其他模态的交互所替代。目前，智能座舱内听觉数据的研究重点集中在自然语言处理上，这包括处理连续对话的能力以及语音情绪识别的能力。随着语音助手的不断发展和成熟，智能座舱已经能够通过车载语音助手实现车机控制和娱乐信息控制等功能。

触觉数据是视觉数据的有效补充，能够为用户提供物体的纹理、温度和振动幅度等信息。在智能座舱中，触觉数据的交互主要包括指纹识别和振动反馈。指纹识别在安全领域具有独特优势，应用于汽车时，可以实现指纹启动、身份识别等功能，并且能够根据不同指纹对应的用户偏好自动调整车辆设置，从而极大提升操作的安全性和便捷性。振动反馈则主要用于驾驶功能，例如通过座椅或方向盘的振动向驾驶员提供触觉警报，提醒驾驶分心或疲劳驾驶，其在驾驶安全领域具有广阔的应用前景。

智能座舱中的嗅觉数据主要用于提升虚拟环境的沉浸感。一方面，通过检测驾驶员的气味特征可以获取其心理状态，从而执行相应的辅助或提醒功能；另一方面，嗅觉数据还可以用于检测疲劳驾驶，并结合智能香氛系统自动调节香氛种类或浓度，以舒缓情绪或提神醒脑。此外，嗅觉数据还可以根据具体场景营造不同的氛围感，例如将环境气味与音乐风格和节奏相匹配，或者在雨后天晴时提供泥土气息，根据不同天气和季节提供不同的气味等。未来，嗅觉交互有望实现更智能和数字化的交互方式。

1.1.5 车用人工智能数据研究框架

基于车用人工智能数据的获取、生成和应用的全生命周期脉络，重点研究车用人工智能数据的获取、车用人工智能数据的管理及质量要求，分别在第二节和第三节体现。

其中，第二节主要研究车用人工智能数据的获取，包括人工智能数据获取目标、真实环境数据要求、AI 合成方法研究和 AI 合成场景覆盖度研究。第三节主

汽标委智能网联汽车分委会

要研究车用人工智能数据管理及质量要求，包括人工智能数据一般要求、数据质量要求、数据质量评估、数据声明周期、管理相关产物等。

1.2 车用人工智能数据获取

1.2.1 概述

车用人工智能（Automotive AI）在现代智能交通系统中扮演着至关重要的角色。其应用范围广泛，包括自动驾驶、智能导航、车载娱乐系统、驾驶员辅助系统等。这些系统依赖于大量高质量的数据来训练和优化 AI 模型，以确保其在各种复杂的驾驶环境中表现出色。获取这些数据的过程和方法对于提升车用 AI 的性能和安全性至关重要。

数据获取可以分为两大类：真实环境数据和 AI 合成数据。真实环境数据是通过车辆传感器（如摄像头、雷达、激光雷达等）直接从实际驾驶环境中采集的。除这种视觉数据外，语音数据也是车用人工智能不可或缺的一部分，特别是在车载娱乐系统、智能导航中。这些数据反映了真实世界中的各种场景和情况，具有高度的现实性和准确性。然而，获取和处理这些数据往往成本高昂且时间耗费巨大。

为了补充和增强真实环境数据，研究人员还开发了多种 AI 合成方法，生成虚拟数据。这些合成数据可以模拟真实驾驶环境中的各种场景，包括不同的天气条件、道路类型、交通状况、各种环境下的语音指令和对话场景等。通过 AI 合成方法，能够快速生成大量多样化的数据，弥补真实数据的不足，特别是在一些危险或罕见场景中。

在数据获取过程中，不仅需要关注数据的数量，还需要重视数据的质量和多样性。有效的数据获取策略应该确保覆盖各种驾驶场景，以全面训练和测试车用 AI 系统。通过科学的评估方法，可以验证 AI 合成数据对真实场景的覆盖度，确保其在实际应用中的可靠性和有效性。

汽标委智能网联汽车分委会

本部分将详细探讨车用人工智能数据获取的目标、真实环境数据的获取方法、AI 合成方法介绍、AI 合成数据的作用以及 AI 合成场景覆盖度的评估，旨在为开发高性能、安全可靠的车用 AI 系统提供全面的指导。

1.2.2 目标

车用人工智能数据获取的目标是确保 AI 系统在各种驾驶环境中能够安全、可靠地运行。具体目标包括以下几个方面：

(1) 高质量数据采集：

真实世界数据：从实际驾驶环境中获取高分辨率、多传感器融合的数据，以确保数据的真实性和多样性。这些数据应包括不同天气条件、昼夜变化、各种道路类型和交通状况下的驾驶情况。AI 合成数据：利用先进的合成技术生成高质量的虚拟数据，模拟复杂和危险的场景，补充真实数据的不足。这些数据应具备高保真度，以便于与真实世界数据结合使用。

(2) 数据多样性与覆盖度：

确保数据覆盖各种驾驶场景，包括城市道路、高速公路、乡村道路、不同的交通密度和行为模式（如行人、骑行者、其他车辆的行为）。包含不同天气条件（晴天、雨天、雪天、雾天等）和不同光照条件（白天、夜晚、黄昏等），以增强 AI 系统在不同环境下的适应能力。

(3) 数据标注和质量控制：

确保所有数据都经过准确、详细的标注，包括物体检测、路径跟踪、语义分割等信息，以便于 AI 模型的训练和验证。实施严格的数据质量控制措施，确保数据的准确性、一致性和完整性，减少噪声和误差的影响。

(4) 高效的数据处理和存储：

采用高效的数据处理技术，快速处理和转换大量数据，以满足 AI 模型的训练需求。确保数据存储的可靠性和安全性，便于随时调用和更新，同时保护数据隐私和安全。

汽标委智能网联汽车分委会

(5) 提升 AI 系统的鲁棒性和安全性：

通过多样化的数据训练 AI 系统，使其具备在复杂和动态环境中做出安全决策的能力。评估并改进 AI 系统在罕见或极端场景下的表现，确保其在各种情况下的安全性和稳定性。

(6) 成本效益优化：

平衡数据获取和处理的成本与效益，最大化资源利用效率。采用 AI 合成数据的方法，降低高成本、危险场景的数据获取难度。利用自动化和半自动化的标注工具，减少人工标注的工作量和成本。

通过实现这些目标，车用人工智能数据获取策略将有助于开发出性能优越、安全可靠的 AI 驾驶系统，推动智能交通系统的进步和应用。

1.2.3 真实数据获取方法

自然采集数据是指在自然环境和正常使用场景下，通过各种设备或手段，在没有刻意设置实验条件或进行人工干预的情况下采集得到的数据，是智能网联汽车开发、测试过程中的重要数据来源，现阶段，基于自动驾驶技术的发展和进步，基于多传感器单车采集的驾驶场景数据采集技术已经相对成熟，数据采集包括以下三个维度，分别为环境信息采集数据、道路信息采集数据、交通参与者信息采集数据。

(1) 环境信息采集数据

在数据采集过程中，需要采集得到的环境信息包括采集日期、采集时间、天气、温度、光照条件、经纬度等信息。其中，采集时间和经纬度信息均由 GPS 获得，光照条件信息数据由光照计获得。

(2) 道路信息采集数据

道路信息包括道路类型、道路交通标识、车道数量、车道属性、车道之间位置关系以及车道线属性等信息。

汽标委智能网联汽车分委会

道路类型包括高速公路、城市道路，其中，高速公路按照其不同区间的功能进行细分，可分为高速收费站、高速匝道、高速公路、高速服务区、高速隧道以及高架桥等；城市道路可分为主干道、路口、环岛、立交桥等。道路交通标识信息包括指示牌和地面指示线，地面指示线属性包括地面指示线颜色、地面指示线线型、地面指示线完整度、地面指示线清晰度等。车道属性包括车道数量、车道宽度、车道坡度、车道材质、车道减速带数量等。车道线信息包括车道线颜色、车道线线型、车道线完整度、车道线清晰度等。

(3) 交通参与者信息采集数据

交通参与者信息包括交通参与者类型、数量、ID、速度、加速度、尺寸、坐标以及和本车的距离等。

1.2.4 AI 合成数据方法

人工智能（AI）合成数据方法是基于人工智能模型直接进行编辑或生成数据的方法，这些合成数据可用于训练和测试人工智能模型，特别是在真实数据难以获取或存在隐私问题的情况下。自动驾驶 AI 合成数据中可以主要分为两类，一类是视觉数据合成，一类是语音数据合成。通常 AI 合成数据中针对视觉数据合成的方法分为两大类，即数据编辑方法和数据生成方法，其中数据编辑方法在原有的数据上进行编辑生成与原始数据分布之外的数据，如自然损坏、对抗攻击。而数据生成方法包括但不限于生成对抗网络（GANs）、变分自编码器（VAEs）和扩散方法(Diffusion)等。这些方法能够生成与真实数据相似的数据分布，从而提供有价值的训练数据。视觉合成数据可解决 3D 数据的标注问题，例如自动生成仿真场景，包括道路、车辆、行人、建筑物等，可结合真实世界的地图数据和 3D 模型，以及通过人工智能技术生成的动态元素，来创建逼真的驾驶环境，快速生成大量不同的场景数据，用于测试自动驾驶系统在各种条件下的性能。另外，在自动驾驶中，AI 合成数据中语音合成技术主要用于车载语音助手、导航提示、警告信息等场景。通常该技术主要是为自动驾驶的智能座舱服务，旨在提供更加舒适、便捷和安全的驾驶体验，同时提高驾驶员和乘客的满意度。语音合成技术可以生成语音与驾驶员语音交互，实现对车辆的控制和信息获取，也可以告诉乘

汽标委智能网联汽车分委会

客相关乘坐信息，提高乘客的使用体验。语音合成技术可以分为基于规则的语音合成，基于统计参数的语音合成两类。基于规则的语音合成是最传统的方法，通过预先设计的语言学规则来生成语音。基于统计参数的语音合成利用统计学习方法建立文本与语音之间的映射关系。

1.2.5 AI 合成数据的作用

合成数据在自动驾驶中不可或缺，它弥补了数据采集昂贵的费用和人力投入，并且能够生成与真实数据几乎一致的高质量数据，它的主要作用可以从基准测试和提高下游任务性能两方面进行探讨。

1.2.5.1 基准测试

合成数据在基准测试中的应用提供了一个控制严格且可重复的测试环境。无论是数据增强方法还是数据生成方法，都可以生成各种复杂场景和条件下的数据，用于评估算法和模型的性能。这些合成数据能够覆盖真实世界数据中难以获得或极端稀有的情况，帮助系统地测试和验证自动驾驶模型的鲁棒性和适应性。例如，通过合成极端天气、复杂交通环境等场景，可以全面测试自动驾驶系统在不同条件下的表现，确保其安全性和可靠性。此外，合成数据还可以用于不同模型和算法的公平比较，提供统一的测试标准，推动领域内的技术进步。

1.2.5.2 提高下游任务性能

将合成数据整合到训练集中可以显著提高下游任务的性能。由于实际数据集通常存在数据不足、标注昂贵或涉及隐私等问题，合成数据提供了一种有效的补充手段。通过生成多样化的合成数据，可以扩展训练数据集的覆盖范围，增强模型的泛化能力。例如，在自动驾驶的对象检测和分割任务中，通过生成不同姿势和环境下的车辆、行人等合成图像，训练集的数据多样性得以提升，模型能够更好地适应各种复杂场景。这不仅提高了模型在真实世界中的表现，还减少了对昂贵数据采集和标注的依赖。此外，合成数据还可以用于数据增强技术，通过生成与真实数据相似的样本，进一步提升模型的鲁棒性和性能。

汽标委智能网联汽车分委会

1.3 车用人工智能数据管理和质量要求

1.3.1 概述

随着人工智能技术的快速发展，其在汽车行业的应用愈发广泛。车用人工智能数据作为智能网联汽车的重要组成部分，对于提升汽车的安全性、智能化水平以及用户体验起着至关重要的作用。因此，对车用人工智能数据的合理应用与性能要求的研究显得尤为重要。

车用人工智能数据是指智能网联汽车在行驶过程中产生、收集、处理的各种数据，包括车辆状态数据、环境感知数据、用户行为数据等。这些数据是智能网联汽车实现自动驾驶、智能交通管理、智能服务等功能的基础。然而，由于车用人工智能数据的复杂性和多样性，如何科学、系统地管理和应用这些数据，确保数据的准确性和安全性，成为了当前汽车行业亟待解决的问题。

1.3.2 一般要求

(1) 科学性：依据智能网联汽车数据的多维特征及其内在逻辑关系，进行科学且系统化的分类和分级。

(2) 实用性：确保智能网联汽车数据的分类分级中，每个类目都有实际对应的数据，避免设置无意义的分类。

(3) 扩展性：分类分级方案应具有概括性和包容性，能够涵盖现有各类数据，并适应未来可能出现的新数据类型。

(4) 合规性：数据分类分级需严格遵循国家法律法规及行业主管部门的相关规定。

(5) 可执行性：分类分级规则应简洁明了，避免过于复杂，以确保实际操作的可行性。

(6) 时效性：数据分级应设定有效期限，超过期限后，数据级别需根据既定的变更策略及时调整。

汽标委智能网联汽车分委会

(7) 稳定性：分类分级应基于智能网联汽车数据最稳定的特征和属性，以保证分类分级结果的稳定性，并便于对同一类别或级别的数据统一适用相同的安全要求。

(8) 显著性：根据数据在产生、采集、使用等方面的重要特征或内容，确定智能网联汽车数据的分类方案。

1.3.3 数据通用质量要求及评估

数据的准确性、完整性、一致性和时效性等数据质量指标，是构建高效、可靠的人工智能系统的基础。评估机制用于持续监测和确保数据质量，及时识别并解决数据质量问题。结合车用人工智能数据的特点，以下从数据准确性、数据完整性、数据一致性、数据可信度、数据时效性、数据均衡性、数据多样性等维度阐述车用人工智能数据质量定义及评估要求。

1.3.3.1 数据准确性

数据集的准确性是指数据集中每个数据项具有正确数据值的程度。语法准确性指数据值与相关域中的一组语法正确的数据值的接近度，语义准确性指数据值与相关域中的一组语义正确的数据值的接近度。

如果数据项的数据值具有与其显式数据类型相同的类型，则该数据项在语法上是正确的；如果其数据值具有与机器学习任务相对应的期望值，则该项在语义上是正确。机器学习模型是数学结构的，这意味着训练、验证、测试或生产数据集中数据值的句法或语义准确性低，可能导致模型本身不正确或模型做出的推断不正确。

对于监督学习分类系统、图像目标检测系统等，样本标签、标签序列内容的正确性会影响训练模型从数据中推理的准确性。

1.3.3.2 数据完整性

数据完整性根据具有所有预期属性和实体实例的值的数据来描述。在某些情况下，当机器学习算法在训练、验证或测试数据集中遇到一个或多个空数据项或标注项时，它们可能会失败。此外，当生产数据包含空数据值时，经过训练的机

汽标委智能网联汽车分委会

器学习模型也可能失败。数据集中标记数据的完整性特征是相对的。在不同的场景中，完整性的含义可能不同，应根据具体的使用上下文进行考虑。

1.3.3.3 数据可信度

数据的可信度可以通过数据在特定使用环境中具有用户认为可信的属性的程度来定义。可信度适用于单个数据项、数据记录中的相关数据项以及整个数据集。数据使用的环境可能会影响其感知的真实性和可信度。数据在处理（例如传输、存储、计算）过程中可能会受到授权方和未授权方的干扰。机器学习的一个新问题是，未经授权的各方干扰训练、验证、测试和生产数据，故意使训练后的模型不可用，或操纵训练后的建模所做的推断。

数据可信度可以从数值的可信度、来源可信度、数据字典可信度、数据模型可信度等多方面评估。

1.3.3.4 数据时效性

数据时效性致力于判断数据集或数据库中的数据哪些是当前的值，哪些是过期的值。如果数据都带时间戳并且准确的话，这就不算个问题了，但存在各种各样的原因导致这个条件在现实中很难满足。ISO/IEC 25012 以相对于数据使用的正确时间年限来描述数据的时效性，及数据记录时刻与使用时刻的延迟或时间差。

对于机器学习，数据时效性可以是适合机器学习任务的年限范围。与用于训练和测试模型的数据相比，由于生产数据的变化而产生的问题通常被称为数据漂移问题，可以通过确保数据的时效性来解决。

数据集的时效性可以根据数据集覆盖的总时间段（例如，从 2010 年到 2021 年收集的图像或句子）、数据项的最后日期和当前日期之间的时间段（如 8 个月）以及更新周期（如每 6 个月）来描述。时效性应被视为基于这三个特征的综合指标。

1.3.3.5 数据均衡性

对于数据集，均衡性是指数据集所有方面的样本分布。例如，如果数据集表示 X 个数据项类别，则每个类别的样本数应该均匀分布，以便使该数据集保持

汽标委智能网联汽车分委会

平衡。对于图像数据集，这些方面可以包括对业务逻辑有意义的标签、图形分辨率、图形亮度、标记边界框的宽高比、标记边界框的大小以及可能影响人工智能模型性能的任何其他方面。

1.3.3.6 数据多样性

数据集的多样性是指样本之间在目标数据方面的差异。在机器学习模型的数据集中，样本之间的差异性至关重要。如果数据集中的所有或大多数数据记录高度相似，那么从该数据集训练出的模型可能会出现过拟合现象，导致其泛化能力较弱。数据集的多样性反映了数据来源、类型和特征的广泛程度，即数据集中包含不同值域、标签、聚类和分布的程度。生成式机器学习中的数据增强方法可以提升数据多样性，但如果原始数据集本身缺乏多样性，这些方法可能无法奏效。多样性与数据的代表性和平衡性密切相关，是衡量数据质量的重要特征，可用于评估数据集的保真度。

多样性的测量可以在特定目标数据的上下文中完成，这由机器学习任务要求确定。

1.3.4 合成数据质量要求及评估

在数据稀缺或敏感场景下，合成数据的生成和使用变得日益重要。合成数据应模拟真实数据的特性，同时满足隐私保护要求。评估合成数据质量，确保其有效支持模型训练和验证。

合成数据是指通过计算机程序生成的数据，这些程序可以是生成式深度学习算法（例如包括但不限于生成对抗网络 GANs、变分自编码器 VAEs、扩散方法 Diffusion 等）或生成 3 维仿真场景、3D 模拟的 CGI 或游戏引擎（如 Unreal、Unity、Blender 等）。合成数据可以是 RGB 图像、分割图、深度图像、LiDAR 或红外图像、音频信号、自然语言文本信息等。合成数据的主要用途是在机器学习、自然语言识别、计算机视觉、计算机听觉、多模态等领域中，作为训练数据补充真实数据，以提高模型的性能和准确性，或模拟真实数据，从而提升人机交互的自然度和用户体验。

汽标委智能网联汽车分委会

合成数据有以下优点：

(1) 减少数据收集和标注的成本

合成数据可以自动化生成，无需人工干预，从而大幅降低数据收集和标注的成本。

例如，为自动驾驶汽车应用训练的模型需要从连接到汽车或无人机的传感器收集大量数据。这个数据收集过程很慢，可能需要数月甚至数年的时间。一旦收集了原始数据，就必须由人工手动标注，这也是昂贵且耗时的。此外，不能保证返回的标记数据作为训练数据是有益的，因为它可能不包含告知模型当前知识差距的示例。

标记这些数据通常涉及人工在传感器数据之上手绘标签。这是非常昂贵的，因为高薪的人工智能团队经常花费大量时间来确保标签是正确的，并将错误发回给标签商。合成数据的一个主要优势是可以根据需要生成任意数量的完美标记数据。所需要的只是一种生成高质量合成数据的方法。

比如，合成数据既可以是自动生成的仿真场景，包括道路、车辆、行人、建筑物等，又可结合真实世界的地图数据和 3D 模型，以及通过人工智能技术生成的动态元素，来创建逼真的驾驶环境，快速生成大量不同的场景数据，用于测试自动驾驶系统在各种条件下的性能。

(2) 提高数据多样性

合成数据可以模拟各种极端情况和罕见事件，使得数据集更加丰富和多样化，有助于提高模型的泛化能力。

例如：一个骑自行车的被遮挡的孩子从校车后面出现，在郊区风格的环境中骑自行车过马路的合成图像。自动驾驶汽车应用通常处理相对“不常见”（相对于正常驾驶条件）的事件，例如夜间行人或骑自行车的人在路中间骑行。模型通常需要数十万甚至数百万个示例来学习场景。一个主要问题是，收集的真实世界数据在质量、多样性（例如，天气条件、位置）和数量方面可能不能满足真实业

汽标委智能网联汽车分委会

务想要的。另一个问题是，对于自动驾驶汽车和机器人来说，与具有固定数据集和固定基准的传统机器学习任务不同，并不总是知道需要什么数据。

(3) 解决数据隐私问题

合成数据可以在不暴露真实数据的情况下进行数据分析、模型开发和算法测试，有助于保护个人隐私。

合成数据对于无法轻松获取真实数据的域中的应用程序非常有用。这包括某些类型的车祸数据和大多数具有隐私限制的健康数据（例如，电子健康记录）。需要强调的是，收集真实数据不仅需要时间和精力，而且实际上可能很危险。自动驾驶汽车等机器人应用的核心问题之一是它们是机器学习的物理应用。您不能在现实世界中部署不安全的模型，并且由于缺乏相关数据而崩溃。使用合成数据扩充数据集可以帮助模型避免这些问题。

(4) 克服数据稀缺性

对于某些领域，获取高质量的真实数据可能非常困难，甚至有些数据是人类无法完全解释和标记的，合成数据可以作为一种有效的替代品。以下是合成数据是唯一选择的一些用例：

- 从单张图像中准确估计深度和光流；
- 利用人眼不可见的雷达数据的自动驾驶应用；
- 生成可用于测试人脸识别系统的深度伪造。

合成数据面临的最大挑战是数据质量问题。合成数据的质量和真实性是一个关键问题，需要确保生成的数据能够准确地反映真实数据的特征和分布。因此，为了确保合成数据能用于车用人工智能任务，除了章节 1.4.4 涉及的十大维度的数据质量要求及评估以外，还需要从真实性、目标对齐性等维度来要求和评估合成数据的数据质量，同时这些指标在一定程度上也反映了生成模型的性能。

合成数据的真实性评估是确保生成数据能够真实反映真实世界数据特性的关键步骤。合成数据与真实数据在变量分布、相关性、模式、结构上等方面应该具有相似性。合成数据真实性是指合成数据与真实数据的相似程度，它反映了合

汽标委智能网联汽车分委会

成数据在多大程度上能够代表真实数据。同时，不同的数据类型，其合成数据的评估指标也不一样。比如图像合成数据不仅通过结构相似度（SSIM）、FID 衡量生成模型产生的图像分布与真实数据图像分布之间的差异，还常常通过 CLIP score 来衡量合成图像与其对应文本描述的对齐度。

1.3.5 数据应用及性能要求标准可行性分析

在车用人工智能领域，数据应用及性能要求标准的制定和实施，对于推动行业的健康发展、提升车辆智能化水平具有重要意义。以下是对数据应用及性能要求标准可行性的分析：

(1) 技术可行性：

当前，人工智能、大数据、云计算等技术在汽车上的使用及发展，为数据应用及性能要求标准的实施提供了坚实的技术支撑。智能网联汽车能够实时收集、处理和分析大量数据，满足性能要求标准中对于数据处理能力、实时性、准确性等方面的需求。

同时，随着技术的不断进步，数据安全和隐私保护技术也得到了显著提升，能够确保数据的安全性和合规性。

(2) 经济可行性：

虽然数据应用及性能要求标准的制定和实施需要投入一定的成本，包括技术研发、设备购置、人员培训等费用，但从长远来看，这些投入将带来显著的经济效益。

通过提升车辆智能化水平，可以降低事故率、提高行车效率、降低能耗等，为企业带来更大的利润空间。同时，也有助于提升品牌形象和市场竞争力。

(3) 市场可行性：

随着消费者对智能网联汽车的需求不断增加，市场对于数据应用及性能要求标准的需求也日益迫切。制定和实施数据应用及性能要求标准，可以满足市场需求，提升消费者的购车体验和满意度。

汽标委智能网联汽车分委会

同时，也有助于推动智能网联汽车行业的规范化发展，提升整个行业的竞争力和影响力。

(4) 法规和政策支持：

国家对于智能网联汽车的发展给予了高度重视，《汽车数据安全管理若干规定（试行）》明确了汽车数据处理的基本原则和安全要求，规范了数据收集、存储、传输等环节，为智能网联汽车的数据应用提供了合规框架，相关系列相关法规和政策来支持行业的发展。这些法规和政策为数据应用及性能要求标准的制定和实施提供了有力的保障和支持。

综上，数据应用及性能要求标准的制定和实施在技术、经济、市场和法规政策等方面都具有可行性。

1.4 总结与展望

本章节对数据应用及性能要求标准的可行性进行了深入分析，深入分析了车用人工智能数据获取及相关质量要求，并得出了积极的结论。随着技术的不断进步和市场的不断发展，数据应用及性能要求标准的制定和实施将成为智能网联汽车行业发展的必然趋势。同时，我们也需要注意在实施过程中充分考虑各种因素，确保标准的科学性和合理性。当前亟需通过数据应用及性能要求标准的制定和实施，构建车用人工智能数据生命周期，建立相关质量管理及质量评估体系，最终提升车辆智能化水平，为消费者带来更加安全、高效、舒适的出行体验。

汽标委智能网联汽车分委会

第二章 车用 AI 基础共性

2.1 车用 AI 风险等级分类标准化研究

2.1.1 车用 AI 风险等级分类标准化目的

随着人工智能(AI)技术的迅速发展，其在各个领域的应用越来越广泛，从自动驾驶汽车到智能家居设备，再到医疗诊断和金融服务等。然而，这种技术的快速发展也带来了许多问题，不仅威胁到了个人和社会的安全，也对现有的法律法规提出了新的挑战。因此，AI 治理成为了全球关注的焦点，以下为主要几个方面的背景：

(1) AI 技术的复杂性

AI 技术本身十分复杂，包括算法和模型、对大数据的需求、计算资源和硬件的要求、模型解释和透明度、跨领域的知识融合、持续学习和适应性，以及伦理规范等方面。

(2) 中国在全球 AI 治理体系中的的重要性

中国是全球 AI 技术发展的重要参与者之一，拥有庞大的数据资源和丰富的应用场景。中国在 AI 治理方面的态度和行动，将对全球 AI 治理产生深远影响。中国政府一直强调 AI 治理的重要性，并积极参与全球 AI 治理的讨论和实践。

(3) AI 治理的现实困境

AI 治理面临的现实困境主要包括技术快速发展的不确定性、隐私和安全性的担忧、以及法规和政策的滞后性。随着 AI 技术的飞速发展，其在各个领域的应用越来越广泛，但同时也带来了诸多挑战。例如，AI 算法可能因为偏见而导致不公平的结果，AI 系统的安全性也可能被恶意利用，这些都是 AI 治理需要解决的问题。

为了有效地管理 AI 技术，我们需要一套全面的治理框架。这包括但不限于以下几个方面：

汽标委智能网联汽车分委会

(4) 法律法规

制定和实施针对 AI 技术的法律法规，以保护个人隐私、防止数据滥用，并确保算法的公平性和透明度。

(5) 技术标准

制定和推广 AI 技术的技术标准，以促进兼容性和互操作性，减少技术壁垒。

(6) 伦理准则

制定和推广 AI 技术的伦理准则，以引导开发者和使用者负责任地使用 AI 技术，尊重人权和社会正义。

(7) 教育和培训

开展 AI 技术和伦理的教育和培训，提高公众对 AI 技术的了解和认识，增强他们的安全意识和自我保护能力。

(8) 隐私保护

随着 AI 技术的发展，个人数据的收集和分析变得越来越容易，这就需要有相应的法规来保护用户的隐私权。

(9) 数据安全

AI 技术的发展离不开大量的数据支持，而这些数据往往涉及国家安全、商业秘密和个人隐私等敏感信息，因此需要有相应的数据安全措施来防止数据泄露和滥用。

(10) 算法公平性

AI 技术的发展也带来了算法歧视的问题，例如，某些 AI 算法可能因为训练数据的偏见而产生歧视行为，因此需要有相应的法规来确保算法的公平性和公正性。

(11) 国际合作

由于 AI 技术的发展是全球性的，因此需要有相应的国际合作机制来共同应对 AI 带来的挑战。

汽标委智能网联汽车分委会

AI 治理的目标主要包括确保 AI 技术的安全性、公平性和透明性。这需要从多个方面进行努力，包括但不限于制定相关的法律法规、建立 AI 伦理准则、加强 AI 技术的监管等。

随着 AI 技术的不断发展，AI 治理也将面临更多的挑战和机遇。未来，我们需要更加深入地研究和探讨 AI 治理的问题，以期找到更加有效和可行的解决方案。同时，我们也需要加强国际合作，共同应对 AI 治理带来的挑战。

2.1.2 欧盟人工智能法案介绍

《人工智能法案》[9]包含了有关面部识别、生物识别监控和其他 AI 应用使用的新规则，并提出了严格监管 ChatGPT 等 AI 工具的新思路。法案严格禁止“对人类安全造成不可接受风险的人工智能系统”，包括有目的地操纵技术、利用人性弱点或根据行为、社会地位和个人特征等进行评价的系统等。

2023 年 12 月 8 日，欧盟成员国及欧洲议会议员就《人工智能法案》达成初步协议，其将成为全球首部人工智能领域的全面监管法规。

2.1.2.1 管辖范围和域外效力

《法案》主要监管人工智能系统的提供者和部署者，两个定义的范围都非常广泛，只要是提供人工智能服务的主体，无论是开发、发行，还是仅仅作为经销商或中间授权方，都属于监管的对象。对于人工智能系统的产业参与者，只要其针对欧盟市场提供产品或开展服务，或其产品、服务有可能在欧盟境内被销售、使用，甚至位于欧盟的用户可能因使用其产品受到不利影响，都需要遵守《法案》的相关规定。提供者和部署者应遵守以下义务：

- 在其组织内建立并实施质量管理体系；
- 起草并更新技术文件；
- 日志记录义务使用户能够监控高风险人工智能系统的运行；
- 进行合规性评估并在特定情形下重新评估系统（如存在重大修改）；
- 在欧盟数据库中注册人工智能系统；

汽标委智能网联汽车分委会

- 获取 CE 标志并签署符合性声明；
- 上市后监测；
- 与市场监督机构合作。

用户是指在其授权下使用人工智能系统的任何自然人或法人、公共当局、机构或其他团体，但人工智能系统用于个人非专业活动的情况除外。用户应遵守以下义务：

- 按照使用说明操作人工智能系统；
- 在使用人工智能系统时确保人为监督；
- 监控运营中可能存在的风险；
- 发生严重事件或故障时通知提供者或部署者；
- 遵守现有法律（例如 GDPR）。
- 基于风险识别的监管

《法案》对不同人工智能的风险程度进行分级并采取不同的监管措施，将人工智能的风险等级区分为不可接受的风险、高风险、有限风险、低或无风险，其中前三类将受到《法案》的监管。

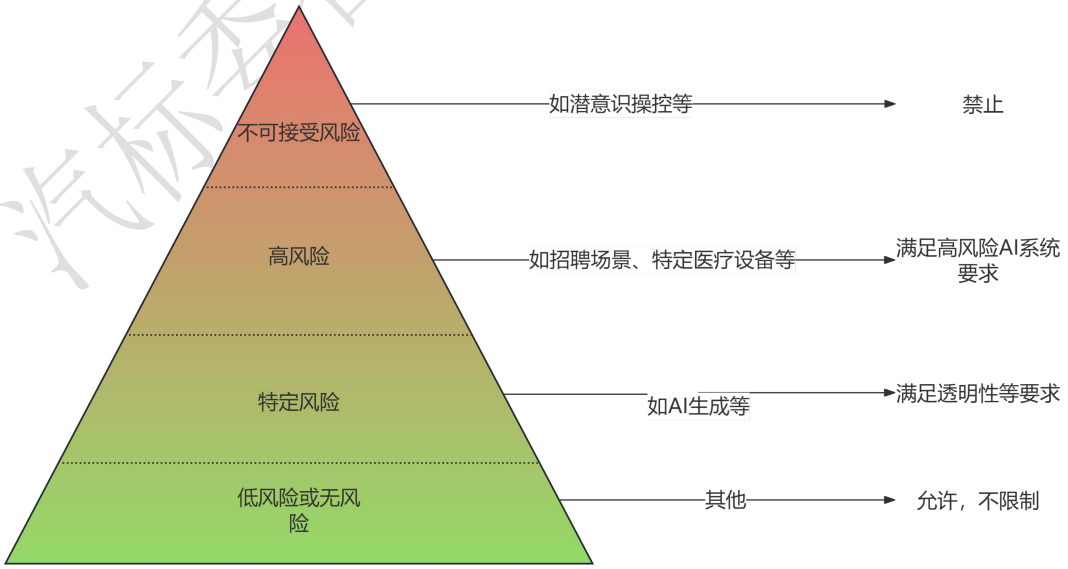


图 2-1 EU AI ACT 法案人工智能风险分级

汽标委智能网联汽车分委会

不可接受的风险的人工智能系统是禁止的。对人的安全具有不可接受的风险的人工智能系统将被禁止，例如：

- 人工智能系统使用潜意识技术，或操纵或欺骗技术来扭曲行为；
- 人工智能系统利用个人或特定群体的弱点；
- 基于敏感属性或特征的生物识别分类系统；
- 用于社会评分或评估可信度的人工智能系统（根据人们的社会行为、社会经济地位、个人特征对他们进行分类）；
- 用于预测刑事或行政违法行为的风险评估的人工智能系统；
- 人工智能系统通过无针对性的抓取创建或扩展面部识别数据库；
- 人工智能系统在执法、边境管理、工作场所和教育中推断情绪等。

高风险人工智能系统要重点监管。高风险人工智能系统主要包括：

- 生物识别和以生物识别为基础的分类
- 可能威胁人的生命和健康的关键基础设施的管理和运营
- 可能决定人的受教育机会和职业培训就业
- 就业、员工管理和获得自营职业
- 获得和享用基本的私人服务和公共服务及福利
- 可能影响人的基本权利的执法活动
- 移民、庇护和边境管理
- 司法和民主程序

对于高风险人工智能的要求需要建立和实施风险管理流程，识别已知的和可预见的投入使用后可能存在的风险，并确保人工可对人工智能系统进行监督，能通过对系统输出的存在“自动化偏见”的结果进行人为干预，以调整或终止其结果输出。此外，对于利用数据训练模式的人工智能系统，提供者还应确保其数据训练、验证和测试均符合欧盟数据立法的相关要求。

汽标委智能网联汽车分委会

同时，《法案》要求提供者将产品投入市场前，均应在欧盟委员会的数据库进行注册，并进行一般性事前评估：在使用“CE”标识（即安全合格标志）并投放市场前，均应遵守现有产品安全法规的相关要求。而对用于生物识别的高风险人工智能系统，则需由欧盟认证机构进行评估。《法案》还对人工智能产业链的不同参与者也提出了细化要求：提供者是人工智能系统的最终负责人，部署者要履行风险防范义务，其他参与者要履行以审查义务为核心的一系列的合规义务。

有限风险人工智能系统需遵循透明度要求。该类风险等级的人工智能系统无须取得特殊牌照、认证或履行报告、记录、留存等义务，但应遵循透明度原则，允许适当的可追溯性和可解释性，应让用户知悉其与人工智能系统的交流或互动。对于有限风险的人工智能系统，法案设定了透明度义务：

- 通知用户他们正在与人工智能系统互动（除非显而易见）；
- 通知用户情绪或生物识别分类系统已针对用户开启应用；
- 对深度造假行为进行标记（为了行使基本权利或自由或出于公共利益的原因除外）。

对于低风险或无风险的人工智能系统，欧盟提倡制定旨在促进对低风险人工智能系统自愿应用要求的行为准则。

2.1.2.2 通用型人工智能监管

《法案》审议过程中，以 ChatGPT、Midjourney 等为代表的通用型人工智能的风靡引起欧盟立法者的关注，故在其中对通用型人工智能系统和基础模型有特别监管措施。《法案》对通用型人工智能系统和基础模型提出了特别要求，即应当适用高风险人工智能系统的监管规则，其中通用型人工智能系统的提供者应履行相关的内部控制和合格性评估程序，产品投入市场后 10 年内向监管部门提供技术文件等。基础模型的提供者应在欧盟数据库中进行注册，制定详细的技术文件和易懂的使用说明，建立质量管理体系等。

汽标委智能网联汽车分委会

2.1.2.3 负责任的研究和创新

具体路径包括建立人工智能科学家和工程师的道德责任、研究人工智能的伦理设计、推动人工智能跨学科、多主体之间的对话、加强社会公众参与。《法案》还提倡监管沙盒，即由公共当局建立现实生活中的受控环境，并在有限的时间内促进创新人工智能系统的开发、测试和验证，然后再有计划地投放市场或投入使用。《法案》鼓励成员国研究和开发对社会与环境有益的人工智能方案，如为残疾人提供方便、解决社会经济不平等问题，这些体现了人本主义，即实现可持续发展、科技向善、多方共治和社会包容等人文目标。

2.1.2.4 严格执行和惩罚措施

《法案》规定，欧盟委员会负责促进《法案》的实施并协调各成员国监督机构之间的合作，各成员国监督机构具体负责对高风险人工智能系统合规情况的监督，有权要求系统提供者对不合规的人工智能系统进行改进，甚至禁止、限制、撤销或召回不符合《法案》要求的人工智能系统。

为制裁不遵守法案的行为，《法案》制定了不同的行政罚款：对不遵守禁止性人工智能系统相关规定的主体，将被处以最高 4000 万欧元或全球年总营业额的 7%的罚款；对不遵守高风险人工智能系统项下有关数据及数据治理、明确说明义务的主体，将被处以最高可达 2000 万欧元或全球年总营业额的 4%的罚款；对未遵守《法案》其他规定（包括对于基础模型的规定）的主体，将被处以最高可达 1000 万欧元或全球年总营业额的 2%的罚款。

2.1.2.5 面临挑战

(1) 人工智能迭代迅速，监管机制前瞻较难

欧洲立法者希望为人工智能领域的规范发展建立一套完整、系统的监管框架，但人工智能近年来的快速发展和更新迭代，为《法案》预期目标的实现带来了不小挑战，《法案》在审议过程中的争论以及《法案》推进到后期时，对通用型人工智能和基础模型不得不紧急“打补丁”，都反映出法律的监管滞后于人工智能行业的发展。

汽标委智能网联汽车分委会

(2) 超前制定监管措施，引发产业发展担忧

欧盟数字领域立法一直走在前列，但总体而言，属于偏规制型的立法，对产业促进方面的规定较少。《法案》超前和过度的监管让人工智能实务界充满担忧：如可能使人工智能企业在欧洲承担过高的成本，可能迫使创业者和创新者转移至监管环境宽松的地方等。OpenAI 公司首席执行官萨姆·阿尔特曼表示，如果《人工智能法案》对 AI 进行过度监管，他将带领团队撤出欧洲市场。

(3) 企业自治要求较高，规制有效性仍存疑

《法案》对人工智能系统提供者附加的合规评估义务仅涉及内部程序，外部监督不足，对批准入市后的人工智能系统监管过于宽松，可能降低规制的有效性和可执行性。人工智能技术发展仍处在探索期。欧盟此时制定《人工智能法案》，释放出欧盟正在强化其在数字领域内建规立制的强烈信号。如何更好地平衡与衔接技术创新和制度创新，还需在实践中不断修正。

2.1.3 车用 AI 风险等级分类的必要性

车辆的安全问题事关消费者人身财产安全及社会公共安全，一直是新技术应用需首要解决的问题，特别是对 AI 技术来说，随着车用 AI 技术的快速发展，其在车领域的应用也越来越广泛，智能驾驶、智能座舱、智能底盘等不同域控，在不同功能层面均有不同AI技术的使用，车用AI技术的安全治理也越发受到关注。

2022 年，德国 DIN/DKE 率先联合发布 AI 标准化路线图，其中汽车是仅次于工业制造的第二大关注领域，并制定了车用 AI 标准化七点要点，分别为：

- 设计的可信度及可测试性状态的记录；
- 技术、法律和组织基础的开发和实施；
- 通用化的需求目录；
- 在运行阶段持续的迭代与验证；
- 分析、模拟、测试方法及设施；
- 场景、数据集、接口、数据交换、数据质量及数字孪生；

汽标委智能网联汽车分委会

- 使用合成数据进行测试；

联合国 WP.29 智能网联汽车法规工作组自 2020 年开启 AI 相关主题讨论，国际汽车制造商协会 OICA 也在会议中提交了相关 AI 的倡议草案，目前建议车辆制造商不得直接修改/更新影响型式批准特性的软件，建议将 AI 纳入软件后，应冻结该软件版本。

ISO/PAS 8800《道路车辆 安全与人工智能》[11]从功能安全及预期功能安全角度，阐述了将安全需求分接到 AI 组件后的 AI 开发-设计-验证-运维的全生命周期安全活动及管理，对企业 AI 开发实践上有较好的指导作用。

我国是汽车大国，车用 AI 的安全治理更应该受到关注。参考欧盟人工智能法案，在车用 AI 领域，同样需要根据应用场景特点及潜在危害风险，划分风险等级，提出与风险等级相称的不同技术要求，在确保车用 AI 安全和伦理的同时，鼓励车用 AI 技术发展和创新，实现行业良性发展。

AI 技术的风险与其应用的具体场景密切相关。不同的应用场景下，AI 系统的功能、数据处理方式、交互对象和潜在影响等方面都存在显著差异。因此，需要针对 AI 应用场景开展风险评估，并基于不同的 AI 风险等级开展与风险等级相称的治理措施。

风险等级指示了 AI 在特定应用场景下的潜在风险，这使得 AI 产品的所有相关方能够就这一风险达成共识。对于产品的提供方（即 AI Act 中的 provider 和 deployer），他们需要根据产品的风险等级采取相应的治理措施，以确保产品的安全性和合规性。同时，监管和认证机构也应对 AI 产品和服务实施与该风险等级相应的监督和认证，以保障市场的安全和稳定。而对于使用者（车主/驾驶员）来说，风险等级的划分也具有重要价值，他们可以充分认识到 AI 技术的风险性，并在适当的场景下以正确的方式使用 AI 产品，以避免潜在的风险和问题。

汽标委智能网联汽车分委会

2.1.4 车用 AI 系统和应用场景举例

在风险等级分类与治理的进程中，服务提供方可以选择将整车视为一个完整的 AI 系统来执行相关操作；同时，汽车制造商亦可将整车细分为若干独立的子系统，并自主决定或由外部供应商受托，针对每个子系统分别实施风险等级的划分与治理工作。随后，服务提供方将基于各子系统的风险评估与治理成果，汇总并出具综合的风险等级评价及最终治理结果。

AI 应用场景的定义可以概括为一个特定情境下，某项 AI 技术、产品或服务被实际应用的场景。在智能网联汽车领域，AI 的应用场景极为广泛且深入，涵盖了从车辆设计、制造到实际行驶及服务的全过程。

在智能网联汽车领域，AI 应用的阶段包含：

- 服务阶段
- 行驶阶段
- 制造阶段
- 设计阶段（AI 作为设计工具，而不是 AI 作为被集成部件）

车用 AI 应用场景主要关注智能网联汽车行使阶段，主要的应用场景举例如下：

- 自动驾驶
- 辅助驾驶
- 智能座舱
- 驾驶员监控
- 自动泊车
- 自适应巡航等。

在界定服务或技术提供方所面向的应用场景时，该方可根据其特定需求与目标，选择采用较为宽泛（粗粒度）的划分方式，如聚焦于 L3 级别的自动驾驶整

汽标委智能网联汽车分委会

体能力；或采取更为精细（细粒度）的划分策略，具体至 L3 级自动驾驶下的特定功能，如自适应巡航控制（ACC）或自动泊车等独立子系统。针对每种粒度下定义的应用场景，均需执行详尽的风险评估与治理流程。

在风险定级阶段，不论所选粒度之粗细，均要求全面梳理并列出所有可能涉及的典型风险场景。若采取粗粒度划分，则倾向于对整个系统实施统一的风险管理与治理策略；反之，在细粒度划分下，各子系统将独立进行风险评估与治理，依据其特有属性与风险特征分别赋予不同的风险等级。

这两种策略各有其优势与局限性。粗粒度划分简化了管理复杂度，便于从宏观层面快速响应与决策，但可能忽略了子系统间的差异性风险；细粒度划分则提高了风险管理的精准度与针对性，有利于深入各子系统细节进行精准治理，但相应地增加了管理的复杂性与资源需求。因此，选择何种粒度进行应用场景定义及风险管理，需综合考虑项目需求、资源条件及风险承受能力等多方面因素。

为了方便各相关方减少风险等级评估工作量，可以建立智能网联汽车的通用 AI 应用场景目录，并认定每个 AI 应用场景的风险等级。

2.1.5 车用 AI 应用场景风险等级定义

AI 致力于模仿、延伸和拓展人的智能，考虑人类智能能力宽泛、水平不一而同，考虑 AI 技术水平的现状，考虑道路交通中异构智能网联汽车并存的现实，使用者可能对 AI 产生不恰当的期待和依赖。

考虑到智能网联汽车实现了更多的移动性和便利性，智能网联汽车的哪些使用和设计方式可能会导致个人发展自由、身心健康等人类核心价值模糊化，尚未形成清晰的、广泛共识。

智能网联汽车初步获得市场接纳，但人们对 AI 技术引发的风险事件认识和积累还不足，对风险事件发生的概率缺乏积累。

考虑到 AI 的以上特点，对车用 AI 应用场景的风险等级进行划分：

- L1：没有或仅有极小的潜在危害
- L2：有一定的潜在危害

汽标委智能网联汽车分委会

- L3：有相当大的潜在危害
- L4：有重大潜在危害
- L5：有不可接受的潜在危害

2.2 车用 AI 应用场景风险等级分类方法

传统风险分析方法，一般会从多个不同的维度来综合判断被分析对象的风险水平。例如在车辆功能安全 ASIL 等级划分中，会考虑从严重度（Severity）、运行场景的暴露概率（Exposure）及可控性（Controllability）三个层面对某一应用场景下的系统失效所造成的危害事件的安全等级进行划分。

考虑到上一节提到的 AI 的特点，从以下维度对 AI 应用场景风险等级分类：某应用场景下 AI 系统失效导致危害的严重度(Severity)、AI 系统失效影响的人群数(Number of people affected)、此场景下 AI 系统的可控性(Controllability)以及 AI 系统失效是否会带来歧视或隐私风险(Discrimination or privacy risks)。

在 AI 系统的风险评估中，必须重视人与 AI 的协同作用，这主要归因于 AI 的固有局限、人与 AI 之间的互补优势、对决策可控性和透明度的需求，以及应对复杂多变风险环境的挑战。AI 技术虽强大，但在直觉和道德判断方面仍显不足；而人类与 AI 各有专长，两者的结合能够显著提升评估的成效。此外，人的参与能够确保决策过程的可控性和透明度。在复杂多变的风险环境中，人与 AI 的合作更能增强评估的适应性和灵活性。

在某一应用场景下，不同类型的利益相关人将受到不同的影响。例如，企业决策者可能更看重评估结果的精确度和可控性，而公众则可能更加关注风险评估的透明度和道德考量。因此，深入剖析各利益相关人的影响变得尤为重要。

通过评估该应用场景下每一类利益相关人(People)最大可能受到伤害的严重度、影响人群数、可控性及歧视或隐私风险，可综合得出该类利益相关人在该应用场景的风险等级。

$$L_P = f(S, N, C, D)$$

汽标委智能网联专业分委会

上式中，S、N、C、D 分别表示严重度、影响人群数、可控性、歧视或隐私风险；Lp 表示在该应用场景中，某类利益相关人在某类危害中的风险等级，它是 S、N、C、D 的函数。

取该 AI 应用场景下所有利益相关人风险等级最高者作为该 AI 应用场景的风险等级。

$$\mathrm{L} = \max_{P}(L_P)$$

其中，P 表示利益相关人的类别。

具体实施时，为了评估某类利益相关人在该应用场景下的风险级别，同时考虑汽车领域的特殊性，一个应用场景可能会对应到多种危害类型，不同危害类型的影响不同，导致难以估计某个应用场景下利益相关人可能面临的风险。参考 ISO26262，可以列举出所有典型风险场景，并分析每个典型风险场景中利益相关人的风险等级，取风险等级最高者作为该应用场景的风险等级。通过如下二维表格来进行风险级别的衡量，其中括号为示例。

表 2-1 风险级别衡量表

	风险场景 1（高速向左变道）	风险场景 2（跟车距离不足）	风险场景 3（不遵守红绿灯）	……
利益相关人 1（司机）	1_0	1_3	1_6	……
利益相关人 2（乘客）	1_1	1_4	1_7	……
利益相关人 3（行人）	1_2	1_5	1_8	……

2.2.1 车用 AI 应用场景风险评估维度

某类利益相关人在某个风险场景下的风险等级，可以从以下 4 个维度综合评估得出：可控性、严重度、歧视或隐私风险、影响人群数。

汽标委智能网联汽车分委会

2.2.1.2 可控性

对于每个风险场景，应基于一个确定的理由预估用户或其他利益相关人员对该 AI 系统的可控制程度。对于智能网联汽车，可能的对 AI 系统进行控制或者交互的人员是：驾驶员、乘客、远程操作员、第三方监管系统和人员等。应取所有相关人员的最大控制等级作为可控性等级。

由于 AI 系统的复杂性、自主性以及在某些情境下可能产生的不可预测或不可解释行为，其被认为具有不可控制性，使得 AI 系统应用于部分场景时被认为具有潜在的安全风险。

AI 系统的可控制性旨在表明在进行决策时使用方与 AI 系统的协同程度和委派程度。将可控性纳入风险等级评估的框架之中，是确保评估全面性与精准性的关键步骤。可控性，作为衡量风险应对能力的重要指标，直接关联到风险事件发生时，个体（如驾驶员）或系统能够迅速、有效地采取预防措施或避免危险发生的能力。

如果某个风险场景的可控性较高，即驾驶员或系统能够较为容易地采取主动措施来避免或减少损害的发生，那么该事件的风险等级可能会相应降低。这有助于减少对系统安全性的过度要求，降低设计和实施的复杂性。对于开发者，通过评估可控性，可以识别出系统在设计、实施和运行中可能存在的安全薄弱环节（可能存在的 AI 系统失控），并采取相应的措施来加强这些环节。例如，在自动驾驶系统中，通过更容易的人类接管方式、及时的远程驾驶员接管、实时监管等手段来增强系统的可控性，在风险发生时更好地应对和避免损害的发生

当 AI 系统具有较高的智能水平时，使用方会委派该 AI 系统承担更多的决策责任，该 AI 系统的不可控制性也会更高。因此，对 AI 系统在某个风险场景下的可控性评估，也是评估该应用场景下，AI 系统对使用方的自主性限制。根据 AI 系统与使用方的协同关系，将 AI 系统在应用场景下的可控性分为 4 个等级。

汽标委智能网联专业分委会

表 2-2 AI 系统在应用场景下的可控性

可控性等级	C1	C2	C3	C4
描述	监督辅助	协同辅助	监管自主	完全自主
定义	AI 系统主要作为辅助工具，提供数据分析、预测和建议，但所有关键决策和控制指令最终由人类作出。人类能够显式地控制 AI 的行为，并通过设定规则或阈值来隐式地指导 AI 的操作	AI 系统与人类用户共同参与决策过程，AI 在日常操作中能够提出决策建议或直接执行任务，但在复杂或高风险情境下，人类用户拥有最终的决策权和控制权。在关键时刻，人类的控制仍然是决策过程的核心	AI 系统拥有高度的自主性，能够在特定的预定义范围内独立运行，但其行为受到明确规则或外部监管机制的约束。这意味着 AI 系统可以在限定的环境或任务中自主决策和行动，而无需人类的即时干预。尽管它们可以自主完成任务，但其运行范围和行为规范是由人为设定的规则所限定的	AI 系统拥有完全的自主性，能够在任何环境中独立做出决策并执行任务，无需人类的监督或介入。这种级别的 AI 能够理解和适应不断变化的环境，处理未曾预料的情况，并从中学习和改进。目前，完全自主的 AI 系统在实现之前，还需要克服技术、法律和伦理上的多重障碍
示例	驾驶过程中使用碰撞预警功能	开启 L2 辅助驾驶系统在路上行驶	在特定区域使用 L4 级无人驾驶出租车	无任何监控的 L5 无人驾驶车辆在路上行驶

2.2.1.3 严重度

对于每一个风险场景，应基于确定的理由来预估潜在伤害的严重度。应按照下表为严重度指定一个 S0、S1、S2 或 S3 的严重度等级。

表 2-3 严重度等级定义

等级	S0	S1	S2	S3
描述	无伤害	轻度和中度伤害	严重的和危及生命的伤害（有存活的可能）	危及生命的伤害（存活不确定），致命的伤害

汽标委智能网联汽车分委会

具体严重度等级之间的划分原则，可参考标准中使用的 AIS 分级来描述严重度。AIS 代表受伤的严重程度分级,它由汽车事故医学高级协会发布。具体分为七级:

- AIS0:无人员伤亡;
- AIS1:轻伤,例如皮肤表面伤口、肌肉疼痛、挥鞭样损伤等;
- AIS2:中度伤害,例如深度皮肉伤、脑震荡长达 15min 无意识、单纯性长骨骨折、单纯性肋骨骨折等;
- AIS3:严重,但未危及生命的伤害,例如无脑损伤的颅骨骨折、没有脊髓损伤的第四颈椎以下 脊柱错位、没有呼吸异常的超过一根的肋骨骨折等;
- AIS4:严重受伤(危及生命、有生存的可能),例如伴随或不伴随颅骨骨折的脑震荡引起的长 达 12h 的昏迷、呼吸异常;
- AIS5:危险伤害(危及生命,生存不确定),例如伴随脊髓损伤的第四颈椎以下脊柱骨折、肠道 撕裂、心脏撕裂、伴随颅内出血的超过 12h 的昏迷等;
- AIS6:极度危险或致命伤害,例如伴随脊髓损伤的第三颈椎以上脊柱骨折、极度危险的体腔 (胸腔和腹腔)开放性伤口等。

结合 AIS 等级确定严重度等级的示例如下表所示：

表 2-4 严重度等级划分

严重程度等级	S0	S1	S2	S3
描述	无伤害	轻度和中度伤害	严重的和危及生命的伤害（有存活的可能）	危及生命的伤害（存活不确定），致命的伤害
对单一伤害的参考(根据AIS分级)	AIS 0及AIS 1-6可能性小于10%；或不能被归为安全相关的损害	AIS 1-6可能性大于10%（不属于S2和S3）	AIS 3-6可能性大于10%（不属于S3）	AIS 5-6可能性大于10%

汽标委智能网联汽车分委会

表 2- 4 严重度等级划分（续）

严重程度等级	S0	S1	S2	S3
示例	冲撞路边设施；撞倒路边邮筒、围栏等；轻微刮痕损害；在进入或退出停车位置时损害；没有碰撞或者侧翻的情景下离开道路	侧面碰撞一个狭窄的静止物体，例如乘用车以非常低的速度撞上一棵树（影响到乘员舱）；以非常低的速度和其他乘用车后碰/正碰；没有乘员舱变形的正面碰撞（例如追尾其他车辆、半挂车）	侧面碰撞一个狭窄的静止物体，例如乘用车以低速撞上一棵树（影响到乘员舱）；以低速和其他乘用车后碰/正碰；以低速造成的行人或自行车事故	侧面碰撞一个狭窄的静止物体，例如乘用车以中速撞上一棵树（影响到乘员舱）；以中速和其他车辆后碰/正碰；有乘员舱变形的正面碰撞（例如追尾其他车辆、半挂车）

2.2.1.4 歧视或隐私风险

AI 系统的歧视和隐私风险是人工智能技术发展过程中不可忽视的重要问题。

AI 系统的歧视风险主要源于算法的不公平性和数据集的偏见。

算法不公平性：AI 系统通过算法进行决策，但这些算法可能受到设计者的偏见或历史数据中的偏见影响，导致决策结果不公平。例如，在招聘、贷款、司法等领域，AI 系统可能基于性别、种族、年龄等敏感特征做出歧视性决策，违背了平等和公正的原则。

数据集偏见：AI 系统的训练数据往往来源于现实世界，而现实世界中的数据往往存在偏见。如果训练数据中存在性别、种族、地域等偏见，那么 AI 系统在学习这些数据时也会继承这些偏见，并在决策过程中表现出来。这种偏见可能导致 AI 系统对某些群体产生不公平的对待。

AI 系统的隐私风险主要源于数据的广泛收集和不当使用。

汽标委智能网联汽车分委会

数据泄露风险：AI 系统需要处理大量的个人数据，包括姓名、地址、电话号码、电子邮件地址等敏感信息。如果这些数据没有得到妥善的保护，就容易被黑客或恶意软件攻击，导致个人隐私泄露。此外，AI 系统的算法和模型也可能存在漏洞或缺陷，导致个人隐私在处理过程中被意外泄露。

数据滥用风险：AI 系统可能在不经过用户同意的情况下收集和使用个人数据，用于商业广告、精准营销等目的。这种滥用行为不仅侵犯了用户的隐私权，还可能对用户造成骚扰和不便。

AI 系统的歧视和隐私风险是人工智能技术发展过程中必须面对的问题。为了确保 AI 技术的健康发展并为社会带来更大的价值，我们需要加强算法公平性设计、优化数据集选择和处理、建立透明的决策机制以及加强数据加密和隐私保护等措施。歧视或隐私风险可以分为两个等级。

表 2-5 歧视或隐私风险等级定义

等级	D0	D1
描述	不存在歧视或者隐私风险	存在歧视或者隐私风险

2.2.1.5 影响人群数

影响人群数(Number of people affected)是指被评估的风险场景，会影响的利益相关人数量。它表示利益相关人群受到伤害的可能性大小。它与两个因素有关：风险场景发生的概率和该风险场景下可能影响的利益相关人人数。在功能安全的 ASIL 安全等级中，风险场景发生的概率用暴露率表示。在这些风险场景中，会出现促成危害发生的相关环境因素。评估的结果会确定危害场景的暴露率级别，暴露率级别有 5 个，分别为 E0、E1、E2、E3、E4，具体如下表：

表 2-6 暴露率等级定义

等级	E0	E1	E2	E3	E4
描述	不可能出现	非常低的概率	低概率	中等概率	高概率

具体暴露概率等级的评估可基于场景持续时间及场景频率来判断，如下表所示：

汽标委智能网联车分委会

表 2-7 基于场景持续时间的暴露率等级划分

运行场景暴露概率等级	E1	E2	E3	E4
描述	极低概率	低概率	中等概率	高概率
持续时间（平均运行时间的百分比）	无定义	＜1%的平均运行时间	1%～10%的平均运行时间	＞10%的平均运行时间
道路类型示例	-	乡间道路交叉口；高速公路出口匝道。	单行道（城市道路）	高速公路；乡间道路；城市道路。
路面类型示例	-	冰雪路面；有很多光滑树叶的路面。	湿滑路面	-
车辆静止状态类型示例	车辆在跳线跨接启动期间；在维修厂。	连接挂车；装备车顶行李架；车辆在加油。	车辆在斜坡上（停在斜坡上）	-
驾驶操控类型示例	下坡时关闭发动机（山路）	倒车；超车；停车(有挂车连接)。	交通拥挤（频繁起停）	加速；减速；停在红绿灯前（城市道路）；变道 (高速公路)。

表 2-8 基于场景频率的暴露率等级划分

运行场景暴露概率分级	E1	E2	E3	E4
类型	极低概率	低概率	中等概率	高概率
场景发生的频	对大多数驾驶员而	对大多数驾驶员	对普通驾驶员而	平均几乎发生在

汽标委智能网联汽车分委会

运行场景暴露概率分级	E1	E2	E3	E4
率	言，一年发生的频率小于一次	而言，每年发生几次	言，基本上每个月发生一次或多次	每次驾驶中
道路类型示例	-	山路，带有不安全的陡峭斜坡	-	-
路面类型示例	-	冰雪路面	湿滑路面	-
车辆静止状态类型示例	停车,需要重新启动发动机(在铁路道口)； 车辆被拖的过程中。	装备车顶行李架	车辆在加油； 车辆在斜坡上（停在斜坡上）。	-
驾驶操控类型示例	-	避让动作，偏离预期路线	超车	换挡； 转弯（转向）； 使用指示器； 倒车行驶。

参考以上定义，为了便于计算影响人群数，我们对暴露率进行了量化，如下表所示：

表 2-9 风险等级发生概率等级划分及量化因子

概率分级	E1	E2	E3	E4
类型	非常低的概率	低概率	中等概率	高概率
场景发生的频率	对大多数驾驶员而言，一年发生的频率小于一次	对大多数驾驶员而言，每年发生几次	对普通驾驶员而言，基本上每个月发生一次或多次	平均几乎发生在每次驾驶中
量化因子	1	5	20	100

汽标委智能网联汽车分委会

对于每个风险场景的影响人数，可分为 4 个级别。

表 2-10 每次风险场景发生时影响人数

级别	n0	n1	n2	n3
影响人数	≤1	≤5	≤10	>10
示例	座舱语音控制，定时开启空调、音乐	有乘客时，车道保持撞上障碍物或单车事故	在城区辅助驾驶场景，影响行人通行	与大型营运车辆发生事故

不同类型的利益相关人，在同一场景受到伤害的可能性不同。例如在不同的风险场景，司机可能被影响的数量为 N0/N1；乘客可能被影响的数量为 N0/N1/N2/N3；行人可能被影响的数量为 N0/N1/N2/N3。

结合以上两个因素，影响人群数等于风险场景的暴露率乘以该风险场景的影响人数。

$$N = E * n$$

最终，影响人群数可以划分为以下等级：

表 2-11 影响人群数等级划分

级别	N0	N1	N2	N3
影响人群数	≤1	≤50	≤500	>500
含义/解释	极少 影响极少数人群	有限 影响有限数量人群	相当规模 影响相当规模人群	大规模 影响大规模人群

上表中的影响人群数的含义是：每年每车在该风险场景下对该利益相关人造成影响的人数，它等于每年该风险场景发生的次数（量化因子）乘以该风险场景每次会影响的利益相关人人数。具体分级可以从下表查出：

汽标委智能网联专分委会

表 2-12 影响人群数等级划分表

场景概率 影响人数	E1	E2	E3	E4
n0	N0	N1	N1	N2
n1	N1	N1	N2	N2
n2	N1	N1	N2	N3
n3	N2	N2	N3	N3

2.2.2 车用 AI 风险等级评估及示例

在某应用场景中，针对某类利益相关人，在某个风险场景中，车用 AI 的风险等级从无风险到不可接收风险，分为以下 5 类：

表 2-13 车用 AI 风险等级定义

等级	L1	L2	L3	L4	L5
描述	无风险	低风险	有限风险	高风险	不可接受风险

根据 AI 系统在风险场景的严重度(Severity)、风险场景的影响人群数(Number of people affected)和 AI 系统的可控性(Controllability)等级，同时也考虑歧视和隐私风险，按下述方法可以具体确定车用 AI 应用场景的风险等级。

表 2-14 车用 AI 风险等级判定表（可控性等级 1）

监督辅助C1						
严重度	致命伤害S3	严重伤害S2		轻中度伤害S1		无伤害S0
影响人群数		会有歧视和隐私风险	不会有歧视和隐私风险	会有歧视和隐私风险	不会有歧视和隐私风险	
N3	L3	L3	L2	L2	L2	L1
N2	L3	L2	L2	L2	L1	L1

汽标委智能网联汽车分委会

N1	L3	L2	L1	L1	L1	L1
N0	L3	L1	L1	L1	L1	L1

注：S3级对应的风险等级相同，且不区分是否导致特定人群伤害，是为了体现生命面前人人平等的理念

表 2-15 车用 AI 风险等级判定表（可控性等级 2）

协同辅助C2						
严重度	致命伤害S3	严重伤害S2		轻中度伤害S1		无伤害S0
影响人群数		会有歧视和隐私风险	不会有歧视和隐私风险	会有歧视和隐私风险	不会有歧视和隐私风险	
N3	L4	L4	L3	L3	L2	L1
N2	L4	L3	L3	L2	L2	L1
N1	L4	L2	L2	L2	L1	L1
N0	L4	L2	L1	L1	L1	L1

表 2-16 车用 AI 风险等级判定表（可控性等级 3）

监管自主C3						
严重度	致命伤害S3	严重伤害S2		轻中度伤害S1		无伤害S0
影响人群数		会有歧视和隐私风险	不会有歧视和隐私风险	会有歧视和隐私风险	不会有歧视和隐私风险	
N3	L4	L4	L4	L4	L3	L1
N2	L4	L4	L3	L3	L2	L1
N1	L4	L3	L2	L2	L2	L1
N0	L4	L2	L2	L2	L1	L1

汽标委智能网联汽车分委会

表 2-17 车用 AI 风险等级判定表（可控性等级 4）

完全自主C4						
严重度	致命伤害S3	严重伤害S2		轻中度伤害S1		无伤害S0
影响人群数		会有歧视和隐私风险	不会有歧视和隐私风险	会有歧视和隐私风险	不会有歧视和隐私风险	
N3	L5	L5	L5	L5	L4	L1
N2	L5	L5	L4	L4	L3	L1
N1	L5	L4	L3	L3	L3	L1
N0	L5	L3	L3	L3	L2	L1

2.3 车用 AI 治理等级的分类方法

2.3.1 车用 AI 治理要素

基于 AI 应用场景的风险等级，AI 系统的提供方需要开展与风险等级相称的治理措施。与风险等级为了 AI 使用方针对 AI 应用场景进行分级分类的方式不同，AI 治理等级是为 AI 的提供方针对 AI 系统（或者 AI 产品与服务）进行治理等级的分级分类。AI 系统的治理等级与 AI 应用场景风险等级不是一一对应的关系。一方面，不一一对应可以强化使用方对应用场景的认知，强化提供方对产品治理的责任，避免使用方可能会忽略做场景风险清晰界定或者忽视使用方责任；另一方面，这样给提供方和使用方留出足够的弹性和自主选择空间，使用者有权选择在某个风险等级的场景中使用合适的或者更高治理等级的产品，提供方也可以根据自身的治理水平，选择进入哪个风险等级的市场。

对 AI 的治理应始终坚持以“增强人类、社会和环境福祉”为核心治理目标。这意味着在推动 AI 技术发展的同时，要充分考虑其对人类生活、社会结构和自然环境的长远影响。通过制定合理的政策与法规，确保 AI 技术的应用能够提升人类的生活质量，促进社会公平与和谐，并减少对环境的不良影响。这样的治理

汽标委智能网联汽车分委会

目标有助于实现 AI 技术的可持续发展，使其更好地服务于人类社会，为全人类的福祉贡献力量。

车用 AI 治理要素包括透明性、可追溯性、多样性非歧视和公平性、鲁棒性、安全性、隐私保护、AI 与人协作服务于人等。

2.3.1.1 透明性

透明性有助于建立用户对车用 AI 系统的信任，确保系统的公正性和可靠性，并为可能发生的争议提供解决途径。在车用 AI 的治理中，是一个核心要素，它关系到 AI 系统的可解释性、用户信任、责任归属和监管有效性。关键方面包括决策过程透明、算法逻辑透明、数据使用透明、影响评估透明、更新和升级透明、错误和故障透明治理政策透明责任主体透明用户界面透明。

2.3.1.2 可追溯性

AI 系统的可追溯性指的是在整个 AI 系统或模型的生命周期中，能够清晰地记录和追踪其各个阶段的变化、影响以及决策过程的能力。这包括从数据收集、预处理、模型训练、验证、部署到后期维护等所有关键阶段的信息。具体来说，AI 系统的可追溯性涵盖以下几个方面。

- 数据可追溯性

能够清晰地记录和追踪数据的来源、变化、处理过程以及它们对模型结果的影响。这有助于确保数据的准确性和可靠性，以及在出现问题时能够迅速定位到数据层面的问题。

- 模型可追溯性

能够记录和追踪模型的结构、参数、训练过程、验证过程等信息。这有助于理解模型是如何工作的，以及在不同阶段模型性能的变化情况。同时，当模型出现问题时，可以通过追溯模型的历史记录来找出问题所在并进行修复。

- 决策过程可追溯性

汽标委智能网联汽车分委会

对于 AI 系统做出的每一个决策，都能够清晰地记录和追踪其决策过程、依据以及影响因素。这有助于确保决策的透明性和可解释性，提高用户对 AI 系统的信任度。

- 错误和异常可追溯性

在 AI 系统运行过程中，能够记录和追踪出现的错误和异常情况，包括错误的类型、发生的时间、影响的范围以及处理的结果等。这有助于快速定位问题并进行修复，同时也有助于分析问题的原因并防止类似问题的再次发生。

2.3.1.3 多样性、非歧视和公平性

指的是在 AI 系统的设计、开发、部署和监管过程中考虑到不同人群、文化、需求和应用场景的广泛性，是实现车用 AI 系统公平性、有效性和可接受性的关键要素，同时也是 AI 治理中不可或缺的一部分。关键方面包括不同应用场景的需求多样性、算法和模型的多样性、代表数据的多样性、考虑不同用户需求的多样性等。非歧视和公平性是指 AI 系统基于用户的种族、性别、年龄、残疾状态或其他个人特征等不应进行不公平的区分或歧视，应确保对所有用户都公正无偏。关键方面包括无歧视算法、伦理准则遵循、数据集多样性、用户参与和反馈机制等。

2.3.1.4 鲁棒性

指的是 AI 系统在面对各种预期之外的情况时，能够保持稳定性和安全性的能力，是车用 AI 系统安全性的重要保障，也是 AI 治理的关键要素之一。关键方面包括系统稳定性、安全冗余、攻击抵抗、错误容忍、长期可靠性、性能退化管理、异常检测、用户干预、灾难恢复等。

2.3.1.5 安全性

AI 系统的安全性在数据安全和信息安全两个层面至关重要。数据安全方面，AI 系统需确保数据采集、存储、处理和应用过程中数据的完整性、保密性和可用性，防止数据泄露、篡改或非法访问。信息安全层面，则涉及保护 AI 系统免受恶意软件、网络攻击、身份冒充等威胁，确保系统稳定运行且服务不中断。这

汽标委智能网联汽车分委会

要求 AI 系统具备强大的安全防御机制，包括加密技术、身份验证、访问控制等，以全面保障数据和信息安全。

2.3.1.6 隐私保护

AI 系统的隐私保护，指的是确保在 AI 技术的开发、部署和应用过程中，个人或组织的敏感数据不被非法获取、滥用、泄露或不当处理的一系列措施和技术手段。这包括在数据采集、存储、处理、传输和共享等各个环节中，采取加密、匿名化、访问控制、差分隐私等技术，以及建立严格的隐私政策和合规流程，以保护用户隐私权益不受侵害。隐私保护是 AI 系统发展中不可忽视的重要环节，对于维护用户信任、促进 AI 技术健康发展具有重要意义。

2.3.1.7 AI 与人协作、服务于人

AI 与人协作、服务于人至关重要。尽管 AI 在数据处理、模式识别等方面展现出卓越能力，但其决策过程常缺乏人类的直觉、道德判断与情感理解。人的参与能确保 AI 在复杂情境下做出符合伦理、可控的决策，避免潜在的风险与错误。此外，人的监督与指导有助于及时纠正 AI 的偏差，保持其行为的合理性和可预测性。因此，AI 与人协作不仅能提升任务执行效率与质量，更能确保 AI 技术安全、可控地服务于人类社会。

同时，AI 治理还需考虑伦理原则。确保 AI 系统的设计、开发和部署遵循道德规范和社会价值，增强人类、社会和环境福祉，确保 AI 应用的发展不会对社会造成不良影响，以促进公平、正义和人的福祉。关键方面包括隐私保护、公平公正、包容开放、敏捷治理、人类福祉等。

2.3.2 车用 AI 治理等级要求

按照划分的风险等级，AI 系统需要满足对应的 AI 治理要求，以确保车用 AI 系统在各个关键方面得到适当的治理和风险控制。

对 AI 系统的治理等级共分为 7 级，每一个等级比其较低等级采取了更多的治理措施。

AI 应用场景风险等级和 AI 系统的治理等级对应关系如下：

汽标委智能网联汽车分委会

表 2-18 车用 AI 风险等级和治理等级对应关系

<table>
<tr><td>风险等级</td><td colspan="2">L1</td><td colspan="2">L2</td><td colspan="2">L3</td><td colspan="2">L4</td><td>L5</td></tr>
<tr><td>治理等级</td><td colspan="2">L1</td><td>L2</td><td>L3</td><td>L4</td><td>L5</td><td>L6</td><td>L7</td><td>禁止</td></tr>
</table>

风险层级与治理层级的对应关系并非直接且唯一，这一现象可归因于多方面因素：首要原因在于治理维度的广泛性和多样性，导致治理层级的划分相较于风险层级更为细致且类别繁多；其次，风险层级通常与 AI 系统的提供方相关联，而治理层级则与开发者相对应，两者可能隶属于不同的组织或个体，这种非必然的一致性通过将风险层级与治理层级解耦，为提供方提供了依据不同治理需求选择开发者的灵活性，同时也便于具备特定治理能力的开发者能够服务于具有不同风险水平的 AI 系统；最后，这种设置方式旨在明确区分最低限度的治理要求与更为严格的治理标准，体现了治理框架内的层次性和差异性。

汽标委智能网联汽车分委会

AI 治理等级和详细的治理要求如下表所示。

表 2-19 车用 AI 的治理等级要求

	1 级	2 级	3 级	4 级	5 级	6 级	7 级	治理要求
透明	L1	L1	L1	L1	L1	L1	L1	L1：告知 AI 产品与服务的输出有概率性
		L2	L2	L2	L2	L2	L2	L2：告知影响 AI 产品与服务性能的因素及适用范围
					L3	L3	L3	L3：特别约定下可告知直至 AI 产品与服务的测试用例以及测试结果
							L4	L4：特别约定下可告知直至 AI 产品与服务使用的 AI 模型文件，模型的训练数据集、可能基于的预训练模型以及
AI 与人协作、服务于人	L1	L1	L1	L1	L1	L1	L1	L1：辅助和服务于人类，不伤害人类的自主性、学习和成长的乐趣
				L2	L2	L2	L2	L2：在人的授权下运行并能回落到人的控制
隐私保护	L1	L1	L1	L1	L1	L1	L1	L1：使用个人数据符合法律法规要求
			L2	L2	L2	L2	L2	L2：开发阶段不识别训练数据中的个人敏感信息
							L3	L3：使用阶段产品不识别个人敏感信息
多样性、无歧视和公平性			L1	L1	L1	L1	L1	L1：在目标应用场景中不因 AI 产品与服务造成歧视
鲁棒	L1	L1	L1	L1	L1	L1	L1	L1：能容忍既定法规/技术规则规定的扰动范围
					L2	L2	L2	L2：能容忍已知非对抗扰动引发的功能失效
						L3	L3	L3：能容忍已知对抗扰动引发的功能失效
安全	L1	L1	L1	L1	L1	L1	L1	L1：AI 产品与服务满足安全认证要求
				L2	L2	L2	L2	L2：AI 产品与服务遭到攻击后能履行最小系统功能
可追溯	L1	L1	L1	L1	L1	L1	L1	L1：可追溯到整车
		L2	L2	L2	L2	L2	L2	L2：可追溯到系统
				L3	L3	L3	L3	L3：可追溯到子系统
							L4	L4：可追溯到模块

2.3.3 车用 AI 安全等级分类标准化可行性

目前国内外没有已发布的关于车用 AI 的安全等级分类方面的标准，但欧盟《人工智能法案》的草案中提及了相关的要求，将 AI 系统的风险分为低风险、高风险和不可接受风险等几个等级，并提出了相应的监管措施。AI 系统的挑战，如算法的可解释性、数据隐私和网络安全等，需要额外的安全考虑。另外 AI 系统的决策可能依赖于大量训练数据，而这些数据的质量和代表性可能会影响系统的安全性，同时车用 AI 可能需要特定的安全措施，如实时性能监控、异常检测机制、以及对 AI 模型和算法的保护等，因此建议起草针对车用 AI 的相关标准，如《车用 AI 风险等级分类方法》、《车用 AI 风险治理的原则》等标准。

2.4 总结与展望

本章节对车用 AI 系统应用场景的风险等级进行分类，考虑严重度(Severity)、AI 系统失效影响的人群数(Number of people affected)、此场景下 AI 系统的可控性(Controllability)以及 AI 系统失效是否会带来歧视或隐私风险(Discrimination or privacy risks)四个因素综合定义和确定风险等级，从低到高分别为 5 个等级。同时要求建立风险管理体系，并给出了供参考的建立步骤。另外治理的技术要求，综合考虑了透明性、可追溯性、多样性非歧视和公平性、鲁棒性、安全性、隐私保护、AI 与人协作服务于人共 7 个要素，通过对 5 个不同等级的风险分析，提出了相应的技术要求。

车用 AI 的治理至关重要，可确保 AI 系统在汽车应用中的安全性、合规性、伦理性和透明度，直接关系到乘客和行人的安全，同时也涉及到数据保护、法律责任和公众信任等关键问题。未来，车用 AI 的治理需持续更新完善，以适应包括管理算法决策的公平性、防止歧视、确保系统的可解释性以及处理跨境数据传输等新兴的挑战，以适应包括 AI 大模型的应用、智能座舱的发展、智能服务的扩展以及智能驾驶的深化等技术创新的步伐，以匹配法规和标准的不断发展变化，最终形成统一的治理框架。

汽标委智能网联汽车分委会

第三章 车用 AI 基础系统平台标准化研究

3.1 车用 AI 基础系统平台研究概述

3.1.1 研究意义

汽车产业正处于向电动化和智能化的转型期，国内企业在这两方面的技术创新和产业化应用已处于世界领先地位。当前智能网联汽车的核心场景主要为自动驾驶和智能座舱。自动驾驶需处理多源异构传感器数据（包括激光雷达、摄像头及毫米波雷达），通过复杂算法实现环境感知与决策控制。随着端到端模型在高级别自动驾驶系统中的应用，自动驾驶对计算平台的并行处理能力与能效提出更高的要求。智能座舱在传统信息娱乐功能基础上，通过多模态大模型实现自然语言交互、用户行为预测及场景化服务。其数据处理规模已超越移动终端，需支撑大规模传感器融合（如视觉、语音及生物识别）与实时 AI 推理。

当前相关技术应用处于碎片化阶段，芯片企业提供异构硬件平台，车企主导上层应用开发，导致系统兼容性与算法复用性不足。建立标准化体系可降低开发难度，实现跨硬件平台迁移，同时强化功能安全及数据安全的合规。

3.1.2 研究内容

车用 AI 基础系统平台是支撑自动驾驶和智能座舱两大场景的核心，其架构包含计算资源、系统软件、开发框架三大核心模块，见图 3-1。计算资源负责异构计算单元的协同调度和管理；系统软件提供 AI 运行框架、端云协同框架，支持底层硬件屏蔽；开发框架则提供应用与车用 AI 基础系统平台的交互接口，定义系统和上层应用间数据交互和功能权限。

汽标委智能网联汽车分委会

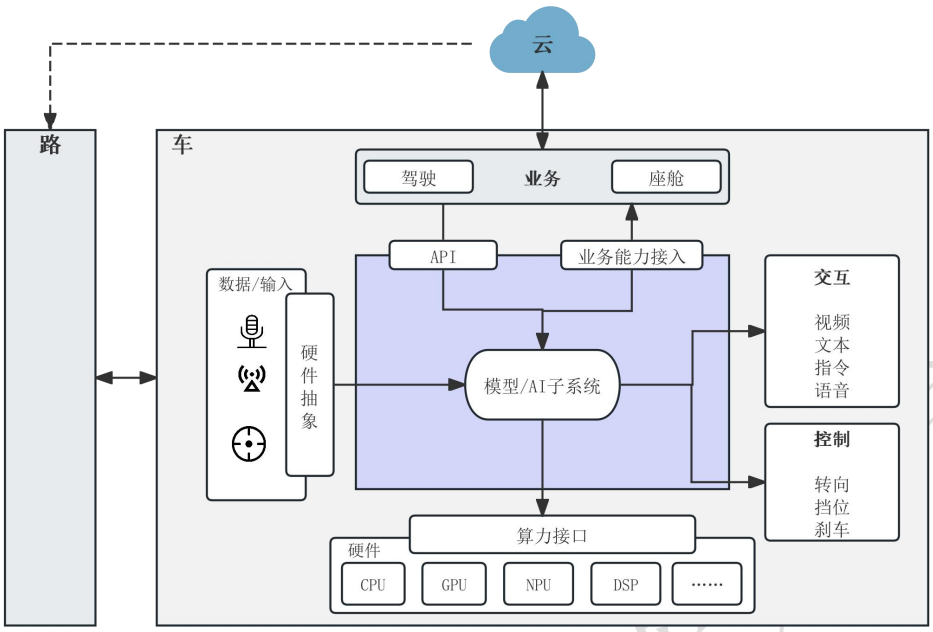


图 3-1 车用 AI 基础系统平台范围

3.2 车用 AI 基础系统平台标准化研究

3.2.1 概述

针对 AI 异构硬件平台和车云协同架构引发的开发复杂度问题，可通过算力资源标准化能力要求和统一抽象接口，将上层车端 AI 应用开发与底层计算单元解耦。同时，为了避免异构硬件平台多样性给算力平台能力评估带来难度，应有对算力平台计算、存储、资源协同等能力的评估标准，对利用算力资源提供针对性的指导。因此，需对算力抽象技术要求和能力评估两方面进行标准化需求识别。

3.2.2 车用 AI 基础系统平台产品现状

3.2.2.1 硬件平台

由于自动驾驶和智能座舱在功能需求和安全要求上的不同，车企往往采用不同的硬件平台。随技术发展，行业正向跨域融合发展，通过中央计算单元集成智驾域和座舱域，以实现硬件、软件和应用的全面打通，提升用户体验，缩短开发周期，降低整车成本。当前主流厂商在不同领域的布局情况如图 3-2 所示。

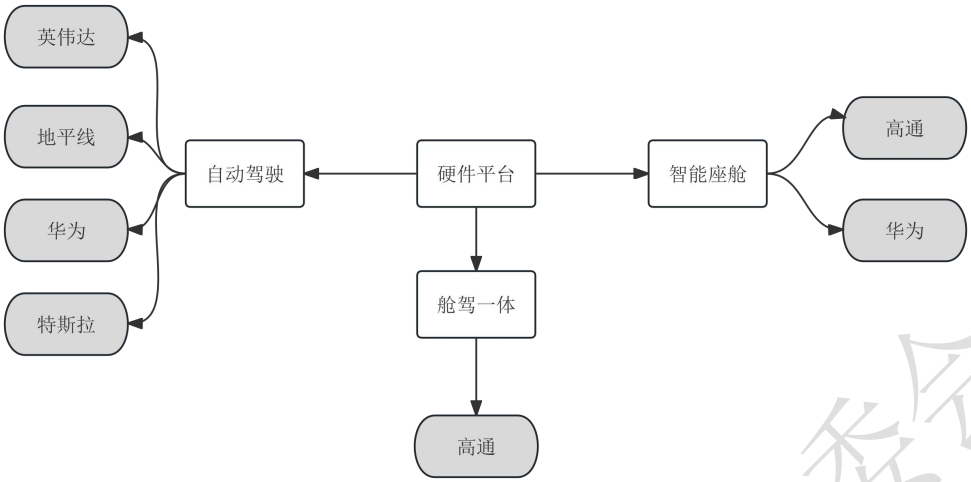


图 3-2 车用 AI 硬件平台产业现状

(1) 英伟达

英伟达（NVIDIA）在自动驾驶领域构建了强大的芯片解决方案。其 DRIVE Thor 系列 定位旗舰级自动驾驶芯片，基于 Blackwell 架构，面向 L4/L5 级高阶自动驾驶场景。顶配版采用双芯片配置，AI 算力达 2000 TFLOPS，集成 Arm Neoverse V3AE 服务器级 CPU，支持生成式 AI、大语言模型（LLM）和实时决策。该系列芯片于 2025 年 CES 期间发布。DRIVE Orin 系列 包括 Orin-X（高阶）、Orin-N（中阶）、Orin-Y（定制化版本），覆盖 L2-L4 级需求。Orin-X 算力达 254 TOPS，支持多传感器融合和冗余设计，符合 ASIL-D 安全标准。

(2) 高通

高通座舱芯片被业界广泛采用，在 2025 年国际消费电子展（CES）上，其推出了骁龙座舱至尊版与 SnapdragonRide 至尊版两大平台。骁龙座舱至尊版专注于数字化交互与多模态体验，其 CPU 基于自研 Oryon CPU 架构，性能较前代骁龙 8295 提升 3 倍，支持多应用无延迟并行运行；Hexagon NPU 算力提升 12 倍，可本地处理数十亿参数的大语言模型（如车辆维修助手），支持生成式 AI 驱动的实时交互；Adreno GPU 图形性能提升 3 倍，支持 16 块 4K 屏幕显示及实时光追 3D 渲染。SnapdragonRide 至尊版为面向 L2+至 L4 级智能驾驶的全栈计

汽标委智能网联汽车分委会

算平台，支持 40+传感器，专用 NPU 集成 Transformer 加速器，支持混合精度运算，时延降低 50%，适用于实时轨迹预测与决策，并符合 ASIL-D 功能安全标准

(3) 地平线

地平线征程系列芯片可覆盖L0到L4全场景整车智能芯片方案。其Matrix 自动驾驶计算平台（新一代）基于伯努利 2.0 架构，支持 12 路 4K 超高清视频输入，可同时处理 10 类车型分类、车灯检测和 6 类车道线检测，专为 L2+至 L4 级自动驾驶设计，已适配城市 NOA、高速领航等高阶辅助驾驶功能。 征程 6 系列芯片中旗舰型号单芯片算力达 560 TOPS（INT8），功耗仅 65W，能效比 8.6 TOPS/W，支持 12 路摄像头、5 颗雷达及 3 颗激光雷达的实时融合计算；中端型号算力 256 TOPS，专注高速领航与轻量化城区功能。

(4) 特斯拉

从 2017 年开始，特斯拉开始启动自研主控芯片，尤其是主控芯片中的神经网络算法和 AI 处理单元全部自己完成。2024 年 2 月，特斯拉推出了 AutoPilot HW4.0 平台，搭载了 Tesla FSD 自研版本的主控芯片。该芯片算力达 720TOPS，较 HW3.0 提升 5 倍，CPU 内核由 12 个增加到 20 个，以协助快速处理图像信息。

(5) 联发科

2025 年第二十一届上海国际汽车工业展览会期间，联发科发布新一代天玑汽车旗舰座舱平台 C-X1。该平台基于先进的 3nm 制程，采用 Arm v9.2-A 架构，主频最高3.62GHz，CPU算力达260K DMIPS，集成了 NVIDIA Blackwell GPU 与深度学习加速器，以双 AI 引擎构建弹性算力架构，提供满足未来智能座舱需求的强大 AI 算力。

3.2.2.2 软件和系统

车用 AI 软件平台采用垂直整合的分层技术架构，通过四层协同实现从芯片级资源管理到智能应用落地的全栈式支撑。硬件编程模型层方便用户管理和使用异构硬件资源； AI 加速库层通过深度车用场景优化，集成了覆盖目标检测、传

汽标委智能网联汽车分委会

感器融合等关键功能的 300+神经网络算子；训练推理框架提供深度学习器编译成硬件可执行模型以及运行任务调度负责模型和算子执行；AI 应用开发层提供智能应用开发所需数据和功能。车用 AI 软件平台在各层级的产业现状如图 3-3 所示。

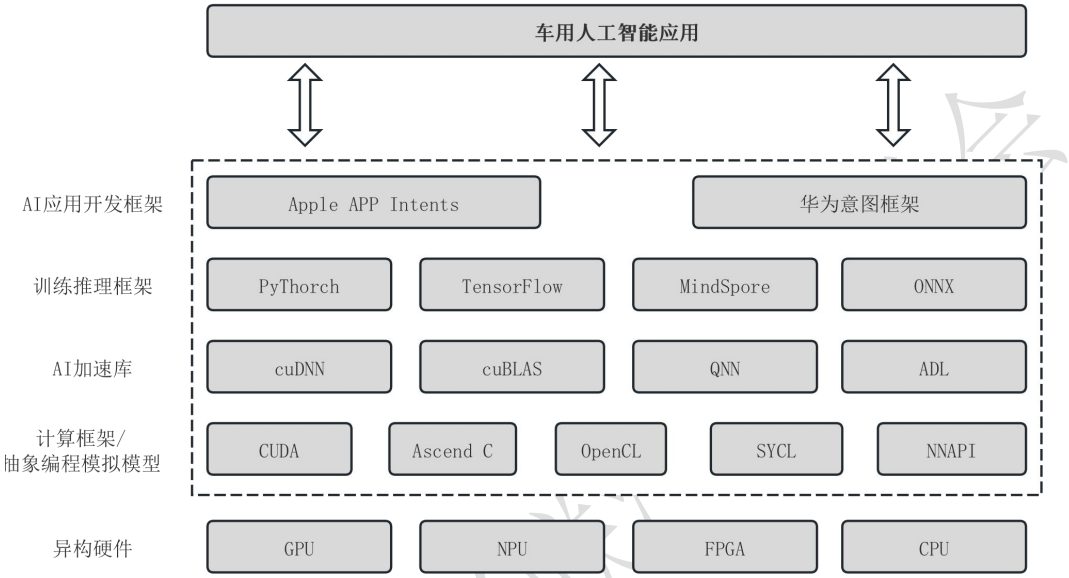


图 3-3 车用 AI 软件平台产业现状

(1) 抽象编程模型

1) oneAPI

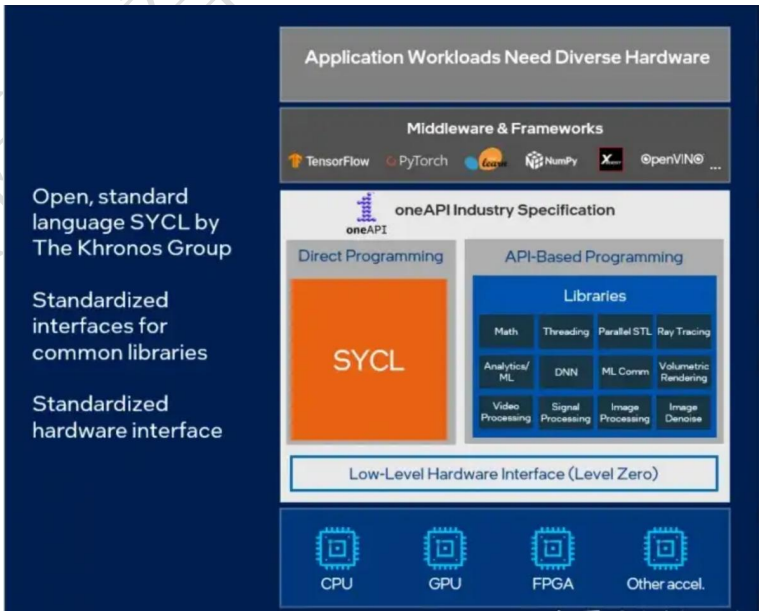


图 3-4 OneAPI 平台软件架构[12]

UXL 基金会（由高通、谷歌、英特尔等公司参与）主导开发的 oneAPI 是一套开放式跨平台工具集，旨在支持多种 AI 芯片的统一编程，使开发者能够通过同一套代码适配不同硬件（如 CPU、GPU、FPGA）。其核心架构基于英特尔推出的编程模型和工具集合，如图 3-4 所示。

A. 该开放规范编程语言与接口，跨架构的编程语言 Data Parallel C++（DPC++），用于 API 编程的函数库以及底层硬件接口（oneAPI Level Zero）。

B. 开发人员可使用相同的代码库优化不同的硬件，实现高性能及可移植性：

- oneDNN：用于深度学习框架的高性能基元实现。
- oneCCL：用于在多个设备上扩展深度学习框架的通信库。
- LevelZero：用于 oneAPI 语言和库的系统接口。
- oneDAL：用于加速数据科学的算法 。
- oneTBB：用于在多处理器上为复杂应用添加基于线程的并行性的库。
- oneMKL：用于科学、工程和金融应用的高性能数学例程。

2) Android NNAPI

汽标委智能网联汽车分委会

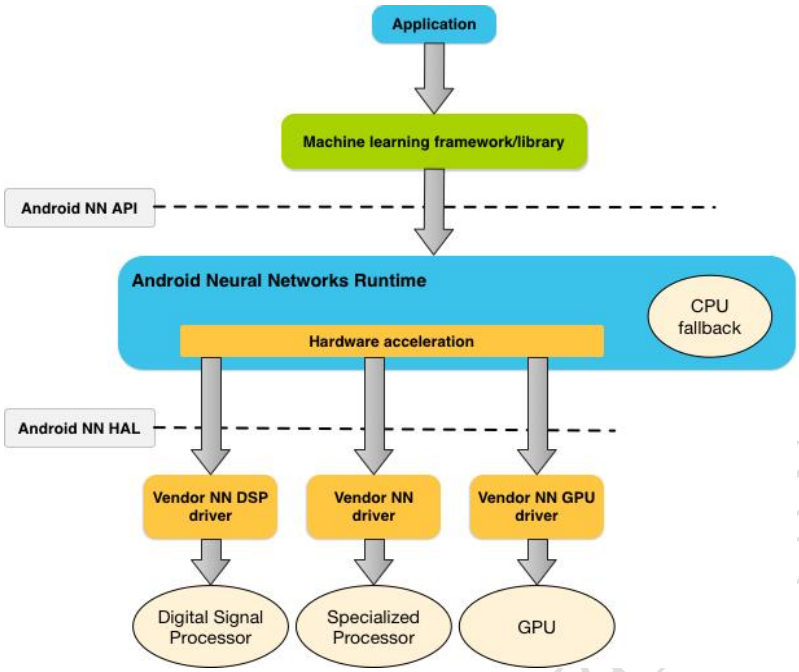


图 3-5 Android Neural Networks API 的系统架构[13]

Android Neural Networks API 的接口包括两部分：高层用户侧的 NDK 接口（图中 Android NN API）、和底层硬件设备侧的 HAL 接口（途中的 Android NN HAL），如图 3-5 所示。

高层 NDK 接口（Android NN API），提供 C 语言风格的 API，集成在 NDK 中，主要用于深度学习框架。支持网络定义、模型编译与执行等核心功能，接口正确性通过 CTS（兼容性测试套件）验证。内置 Runtime，负责对接后端硬件。

底层 HAL 接口（Android NN HAL），是一组用 Android Oreo 引入的 HIDL（硬件接口定义语言）描述的接口。硬件厂商只需实现该层接口（通过 VTS 验证），即可支持多种异构计算单元（如 NPU/DSP）。如图所示，允许多种硬件驱动并行存在，实现计算任务自动分配

(2) AI 加速库

硬件加速库在智能网联汽车中起到关键作用。它们通过充分利用硬件的计算能力，可显著提升数据处理和算法执行的效率。目前，主要的硬件加速库包括：

1）QNN（Qualcomm AI Engine Direct SDK）

Qualcomm AI Engine Direct SDK（又名 QNN）为 AI 开发提供统一 API 的底层框架。通过这个 SDK，可以在高通的 CPU/GPU 和 HTP 上对 AI 模型进行性能加速。

2）cuDNN（CUDA Deep Neural Network library）

NVIDI 专为深度学习设计的 GPU 加速库，提供了一系列针对深度神经网络的高效运算函数。cuDNN 通过对深度学习模型的计算进行优化，使得这些计算能够在 GPU 上更高效地执行。通过 cuDNN，开发者可以在不需要深入了解底层 GPU 编程的情况下，轻松利用 GPU 加速深度学习模型的训练与推理。

3）AOL（Ascend Operator Library）

AOL 是昇腾 AI 异构计算架构 CANN 的逻辑架构中的算子加速库，包括神经网络（Neural Network，NN）库、线性代数计算库（Basic Linear Algebra Subprograms，BLAS）等，为神经网络在昇腾硬件上加速计算奠定了基础。

(3) 训练/推理框架

推理框架是智能网联汽车系统中不可缺少的软件组成部分，负责将训练好的模型应用于实际数据进行推断。目前，主要的训练推理框架包括：

1）TensorFlow

TensorFlow 是由 Google 开发的开源深度学习框架，不仅支持训练，同时对推理进行了高度优化，适用于各种硬件平台，包括 GPU、TPU 和 CPU。

2）PyTorch

PyTorch 是由 Facebook AI 研究团队开发的深度学习框架，具有灵活性和便于调试的特性，支持张量计算和自动求导，在研究和生产应用中得到了广泛使用。

3）ONNX Runtime

ONNX Runtime 是 Open Neural Network Exchange（ONNX）生态系统中的一个高性能推理引擎，能够支持多种硬件平台，包括 CPU、GPU 和 FPGA。

汽标委智能网联汽车分委会

(4) AI 应用意图框架

Apple Intent 框架，使得应用开发者可以暴露应用内部的核心功能和信息，从而让 Siri 能够跨应用访问数据并执行操作。App Intents 框架提供的功能可将三方的 app 操作、内容与跨平台的系统体验（包括 Siri、Spotlight、小部件、控件等）深度集成。Siri 借助 Apple Intelligence 和 App Intents 增强功能，具备在 app 内和跨 app 执行操作的能力。

A. Intents 框架包含三个核心要素，定义了应用如何向系统暴露自己的功能或能力：

- Entities：应用内的一个对象，比如一个集合或一条路线，可以看作“名词”。
- Intents：负责执行某个动作，比如打开一个视图或开始一个徒步旅行，可以看作是“指令”或者“动词”。
- App Shortcuts：应用内的一个重要功能，可以看作是“事件”。

B. Transferable API 有三种数据对象，定义了应用如何向系统或其他应用暴露自己的数据：

- IntentFile：将 App Entities 抽象成文件、图片或富文本。
- FileEntity：App Entities 本身就是一个存储在磁盘的对象。
- Universal links：Entity、Enum、Intent 这些特殊内容都可以被抽象成一个 URL 供 Short、Siri 等调用。

C. Intent 域定义了基于哪些类型提供给开发者，这些 Intent 域已经过预定义，训练和测试，IOS18 中发布了 18 个域，100 多种操作。

D. 系统服务集成调用：Siri、控件和小组件、Apple Pencil、快捷指令、操作按钮。

汽标委智能网联汽车分委会

3.2.2.3 模型和业务

车用 AI 模型根据其功能可以分为图像分类、图像检测、点云目标检测、车道线检测等类别，如图 3-6。除企业自研模型外，一般是根据开源模型进行修改、优化，实现搭载应用。

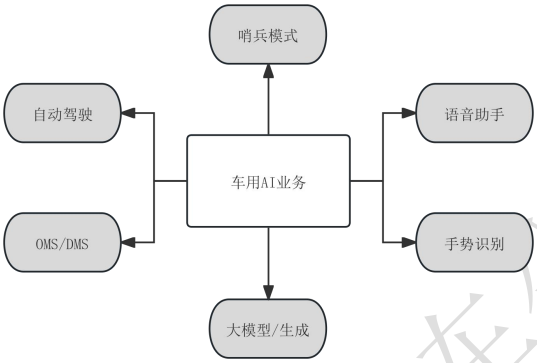


图 3-6 车用 AI 业务场景

自动驾驶、哨兵模式等通常基于模型感知，需要进行分类、检测、重建等任务。图像分类模型以改进版 ResNet 为代表。图像检测任务主要依靠 Yolo 系列模型完成，BEVFusion 支持雷达图像融合检测和点云检测采用 pointpillars/CenterPoint。车道线检测，U-Net、LaneNet 等模型适应不同场景需求。在自动驾驶中，分割模型可以精确地区分道路、车辆、行人、交通标志等不同的目标。常用的分割模型有用于图像分割的 SegFormer，用于点云分割的 RangeNet++等；三维重建在自动驾驶场景下主要用于环境感知和模拟仿真等任务，单目模型包括 SMOKE、MonoCon 等，双目模型包括改进 MVSNet 等，以及基于 BEV 视角的占用网络模型 OccNet，VoxFormer 等。

在智能汽车行业中，智能座舱系统相较于自动驾驶系统，具备更为多样化的功能集成能力，包括音频处理、图像处理、语音识别、手势识别以及自然语言交互等多种任务。在语音识别（ASR）方面，常用的模型包括 Whisper、DeepSpeech 和 MMS。在语音合成（TTS）领域，GPT-SoVits、VALL-E、VITS 和 WaveNet 是业内常用的网络架构。对于文本生成，尤其是在设备端侧，需要一些轻量化且高效的模型，例如 MiLM、MiniCPM 和 Gemini Nano。视觉语言理解同样是智能

汽标委智能网联汽车分委会

座舱的一项重要功能，涉及集成区分图像与文本的能力。模型如 MiniCPM-V、Phi-3-vision、Ferret-UI、ScreenAI 和 CogVLM 被用于处理和理解多模态信息。通过这些多样化的技术和模型集成，智能座舱系统能够实现更加复杂和人性化的交互功能，提供给驾乘人员更为便捷和愉悦的体验。近年来，吉利、广汽等车企陆续将 AI 大模型应用上车，如语言大模型、多模态大模型、数字孪生大模型等基础模型，相关单位已启动汽车行业大模型行业标准制定工作。

3.2.2.4 云平台

AI 车用云平台是指基于云计算技术，专为汽车行业打造的智能云服务系统，该平台整合了大数据、人工智能算法、物联网技术以及边缘计算能力，通过提供强大的计算能力和智能算法，支持车辆数据的收集、处理、分析和应用，从而实现更高级别的驾驶辅助和自动驾驶功能。

车用云平台作为智能网联汽车的核心基础设施。当前主要应用在包括智能汽车制造商、4S 店、车联网服务提供商等汽车销售及售后服务。制造商通过云平台提供的数据存储和分析功能，优化车辆设计、改进驾驶体验，例如通过高性能计算和大规模训练平台，加速自动驾驶算法的研发与迭代。相关服务商可以利用高算力的云平台实时为购车、维护保养或者客户体验提供服务，包括利用 3D 渲染技术、大语言模型等 AI 技术，动态展示相关车辆信息，为客户提供各类讲解和答疑，在客户允许的情况下，通过智能车载应用，提供远程控制、车辆诊断、智能导航等服务，丰富用户体验。部分保险公司也将目标投向 AI 车用云平台，与相关公司合作，基于车辆数据进行风险评估和保险定价，推出更智能化的车辆保险产品，提高服务质量。此外，城市交通管理机构也在利用 AI 车用云平台的数据分析功能，优化交通流量管理、改善交通拥堵情况，并提升城市交通运行效率。智能交通系统开发商利用 AI 车用云平台提供的数据接入和算法支持，开发智能交通系统，实现车辆之间的信息共享和协同。同时，利用 AI 相关技术，通过人脸识别、手机远程操控等为驾驶员及车内乘客提供舒适安全的驾驶出行服务。当前随着市场竞争加剧，各大云服务提供商、汽车厂商以及科技公司纷纷布局车用云平台市场，推动市场进入差异化竞争阶段。并且云平台通过融合城市交通业

汽标委智能网联汽车分委会

务数据、社会车辆数据、自动驾驶车辆数据、地理信息数据、个人数据和设施设备数据等多维度信息，利用大数据和人工智能技术，为城市的出行和服务提供依据。包括大语言模型实时为私家车辆提供云端服务，AI 技术引发的多模交互为驾驶员和车舱乘客提供各类服务。车用 AI 云平台还可以运用大模型收集与分析汽车与路侧的全量感知精细化数据，为车辆提供车联网信号灯消息、路侧交通信息消息等信息汇集到车机系统，实现向车辆推送 C-V2X 车路协同功能场景，从而优化出行体验。随着自动驾驶技术的快速发展，车用云平台在数据处理方面也发挥着重要作用。它们依托云端高算力和广泛的后台处理经验积累，实现更高效的智能汽车大数据处理等服务。这有助于提升自动驾驶的安全性、可靠性和效率，推动自动驾驶技术的商业化进程。一部分车企在车载信息系统方面，提供了基于 AR 技术的内容服务，比如导航增强、行车安全提示等，提升驾驶者的用户体验。在智能监控和关怀服务方面，云平台依然能发挥高效的作用，智能监控服务，借助 AI 视频分析技术，关注车内乘客的安全与健康状况，尤其是在长途旅行、老人儿童单独乘车等场景中提供安全保障。他们也会自己搭建相关云平台以增加数据安全、管控等风险。也有一部分车企利用华为云、百度云等公有云，有效降低成本的同时又能确保数据、信息等安全问题。

AI 车用云平台的发展仍面临一些挑战，如数据安全和隐私保护等问题。法规政策日趋严格，对数据安全、隐私保护、网络安全等方面提出了更高的要求，促使云平台在保障功能的同时必须达到相应的合规标准。随着车联网和自动驾驶技术的普及，如何确保数据的安全性和合规性成为了一个亟待解决的问题。因此，车用云平台需要在保障数据安全的前提下，不断提升服务质量和技术水平，以满足市场和用户的需求。包括先进的加密技术和访问控制策略，确保数据的机密性、完整性和可用性，以及安全审计、漏洞扫描等安全机制。AI 云平台技术要求相当高的算力服务、无线网络传输速率等相关技术。智能汽车每天会产生巨量的数据，云平台需要有足够的存储容量和高效的数据处理能力，包括数据清洗、分析和模型训练等。包括 CPU、GPU、内存及存储空间等，直接影响到了 AI 云平台的计算速度。宽带和无线网络的连接能确保传输速率及其稳定性，如果无法保证网络传输，那么 AI 相关运行只能在车端运行，否则无法满足相关要求。高性能

的 AI 计算需要强大的基础设施支撑，这涉及到数据中心建设、带宽租赁、硬件维护等方面的高昂费用，以及随之而来的节能减排和绿色可持续发展问题。当前行业内尚缺乏统一的标准规范，不同的车厂、供应商和云服务提供商之间的接口和数据格式可能存在差异，这给数据共享和云平台间的协作带来了困难。需要建立统一的标准，从而实现各方产品之间的可操作性。

2024 年是大模型应用的爆发之年，大模型以其强大的计算力、卓越的学习能力和泛化能力，正在重新定义汽车的核心竞争力。这一技术的发展推动了汽车智能化的进步，通过高级的感知、决策与执行能力，使得智能化汽车从实验室走向现实，真正开启了智能汽车时代的大门。大模型在智能座舱、人机交互、车载服务等领域发挥关键作用，大幅提升用户体验，促使汽车成为集出行、娱乐、生活于一体的智能移动空间。这种技术跃迁不仅改变了汽车产品的形态和功能，更催生出全新的技术标准和知识产权体系，引领全球汽车产业步入技术密集型的崭新阶段。大模型云平台为车用 AI 应用场景搭建提供模型推理、模型微调、智能体搭建提供平台性服务，发展十分迅速。目前国内主要的大模型云平台提供的核心能力包括：

(1) 模型推理能力

大模型云平台基于云平台的优势和优化服务调度策略，提供极致的大模型推理性能。同时大模型云平台提供开放模型集成能力，集成并提供不同参数量、不同来源的模型，提供模型推理服务。

(2) 模型训练工具

提供模型数据准备-模型微调-模型测评-模型部署全生命周期一站式工具集。在模型数据管理层面，提供数据管理、清晰工具，方便构建训练集及测试集。在模型微调方面，提供 SFT (Supervised Fine-Tuning)、Lora 等多种微调方式，并且展示训练过程的信息，方便及时监控训练过程。在模型测评方面，提供单/多模型对比效果评测的能力。在模型部署方便，提供一键部署工具，支持模型快速部署到算力资源。

汽标委智能网联汽车分委会

(3) 应用构建能力

大模型云平台提供丰富的应用构建能力，支持车用 AI 应用 agent 快速搭建。通过开放插件集成，给大模型增强“手、脚”能力，内置大量官方预置插件，同时支持自定义业务插件的开发。提供预置 prompt 模版及优化能力，提升模型应用效果。提供大模型应用编排流程画步，通过可拖拽的画布形式，便捷操作，方便基于大模型及插件进行应用的构建，同时支持智能体 Agent 构建。

(4) 内置丰富的开箱即用的大模型应用

内置文档问答-RAG、多模态生成、应用写作、NL2SQL 等开箱即用的应用能力。

未来随着智能汽车市场的快速发展和自动驾驶技术的成熟，AI 车用云平台将在未来继续发挥至关重要的作用，推动汽车产业朝着更加智能化、网联化、电动化和服务化的方向迈进。

3.2.2.5 主流厂商的方案

目前多屏交互、智能语音、车联网、OTA、VR/AR 已经成为智能座舱的标配。智能座舱对于主控芯片的算力要求越来越高，智能座舱 SoC 芯片正是其核心部分。近两年来车载座舱 SoC 的市场竞争越发激烈。在中高端领域，竞争入局的企业包括传统的车载 SoC 厂商如恩智浦、瑞萨、德州仪器，还有消费电子领域的芯片厂商如 AMD、高通、英伟达、华为等。在智能座舱 SoC 领域，高通骁龙的 8155 和 8295 几乎占领了大部分新势力厂商市场。

智能座舱 OS 主打座舱娱乐、座舱体验、座舱感知及人车交互等功能，因此对于车载应用生态的多样性、丰富性以及用户体验的需求更为注重。以 QNX 为代表的实时操作系统在应用生态上较为薄弱，因此当前座舱 OS 使用 QNX、VxWorks 及 WinCE 的 OEM 较少，主流则是 Android 以及基于 Linux 系统的定制型及 ROM 型系统。由于国内 Andriod 应用生态更好，国内传统车企与造车新势力等大多基于 Andriod 定制车载操作系统，比亚迪 DiLink、蔚来 NIO OS、小鹏 Xmart OS 等，部分 OEM 车载 OS 现状如表 3-1 所示。

汽标委智能网联汽车分委会

表 3-1 部分 OEM 车载 OS 现状[14]

分类	OEMs	系统名称	底层 OS
传统车企	奥迪	MMI	QNX+Android
	奔驰	MBUX	Linux
	宝马	iDrive	QNX+Android
	沃尔沃	VolvoCars	QNX+Android
	丰田	G-BOOK	Linux
	大众	VW	Linux
	福特	SYNC	QNX+Android
	比亚迪	DiLink	QNX+Android
	吉利	Flyme Auto	QNX+Android
	荣威	维纳斯	AliOS（基于 Linux）
新势力	特斯拉	Version	Linux+QNX
	蔚来	Banyan 榕	QNX+Android
	小鹏	天玑 AIOS	QNX+Android
	理想	Li	Linux+Android
	智己	IM	AliOS（基于 Linux）
	塞力斯	鸿蒙 OS	鸿蒙 OS（基于 Linux）
	小米	HyperOS	澎湃 OS（基于 Linux）

除了自研操作系统，目前各新势力厂商的智能驾驶操作系统内核主要有两条开发路径。一种选择是利用 Linux 庞大的开源生态系统，基于其功能强大的宏内核，特别加强安全性和实时性，从而开发出符合 ASIL-B/D 等级的 Safety Linux 操作系统。虽然 Linux 拥有丰富的生态系统，但要实现所需的功能安全等级认证却具有一定挑战。另一种选择是以功能安全为核心目标，专注于实现 ASIL-D 功能安全级别，通过 POSIX 标准构建微内核实时操作系统（RTOS），例如 QNX 操作系统。与 Linux 相比，微内核 RTOS 在开发上难度较大，因为它没有类似的开源生态系统支持。各新势力厂商采用的底层操作系统与芯片的组合如表 3-2 所示。

表 3-2 国内外主流的系统内核 OS：实时车控操作系统[15]

厂商	智驾底层 OS	芯片适配	应用情况
中兴	微内核 RTOS 和 Safety Linnx 的双内核智能驾驶操作系统	已与地平线、黑芝麻智能、芯驰科技等国内主流芯片企业建立合作	中兴通讯正与国内主要 OEM 主机厂联合推动双内核智驾系统量产落地
斑马	AliOS Drive，AliOS	已与 10 余家主流芯片	斑马智行 AliOS

汽标委智能网联汽车分委会

	RTOS 与 AliOS Safety Linux 双核	厂商建立合作	Drive+地平线 J5+禾多科技中间件和算法
理想（预研项目）	RT Linux（预研）、RTOS（Vxworks、QNX）	理想自研智驾 SOC（预研）	Li OS 基于 Linux 内核深度定制开发，将搭载理想纯电车型首发
蔚来	[整车全域操作系统天枢 SkyOS]，基于 Linux 内核	支持多种硬件平台，包括 NVIDIA、高通、英特尔等	搭载 NT3 平台车型
哪吒汽车	黑莓 QNX OS for Safety	地平线 J 系列	/

3.2.3 车用 AI 基础系统平台相关标准现状

当前，智能网联汽车正处于从示范验证进入到量产应用的关键时期。总体来看，围绕智能网联汽车的标准化工作正在稳步推进。虽然车载计算平台的标准化进程在智能网联汽车的标准体系中已有进展，但仍然缺乏系统和完整的标准指导车用场景 AI 基础平台。

中关村新兴科技服务业产业联盟发布《人工智能算力资源池技术规范》团体标准。该规范适用于人工智能算力资源池的设计、建设、运维和验证，对于利用算力资源池实现算力资源的虚拟化、整合、调度和高效利用具有重要的现实意义。该规范规定了人工智能算力资源池的总体架构、总体要求、功能要求、通信网络、资源池部署与集成、兼容性要求、可靠性要求和安全要求，并给出了验证方法和过程[16]。

车用智能算力平台与其他智能算力平台如云智能算力平台有一定的相似性，其特征主要表现在异构、高性能、兼容性和安全性等方面。

云智能算力平台和车用智能算力平台虽然在规模上和 AI 应用场景上有较大差异，但其中涉及的功能和性能要求，以及验证方法和流程对车用智能算力平台有一定的参考意义。

中国电子工业标准化技术协会发布《信息技术算力服务能力成熟度评估模型》团体标准。该标准面向算力服务提供方从人员、技术、资源和流程四个方面提出

汽标委智能网联汽车分委会

了算力服务能力在计算力、存力、运力、资源协同、数据可信处理及管理、算法和模型、云化及服务保障能力和绿色低碳等要求[17]。

该标准重点关注于服务能力的成熟度，因此会涉及人员组织保障、服务运营运维等流程的规定与要求，车用智能算力平台只关注平台能力本身即算力本身的成熟度，因此不涉及。但该标准中具备技术　技术（包括计量、计费、监控、资源池化、数据工具集和算力调度 6 个能力域），以及资源（包括计算、存储、网络和能源 4 个能力域）的能力域划分和相关要求，这种划分方式提供了一种有价值的参考框架。但车用智能算力平台有其特定的应用场景和需求，在计算精度和效率、功耗与安全可靠性等方面的具体要求都有明显的区别。

中国电子工业标准化技术协会发布的 T/CESA 1228—2022《人工智能计算设备调度与协同第 2 部分：分布式计算技术要求》　目的在于确立人工智能计算设备分布式计算的参考架构，规定软硬件系统的功能和性能技术要求[18]。其中有涉及到对云边端协同的技术要求，包括云边端协同管理框架、云边端模型部署、AI 任务协同、数据协同。云边端协同管理框架规定了应具备的功能。云边端模型部署规定了模型部署的流程以及功能要求；AI 任务协同主要关注数据加密等端边侧安全要求等。但车用人工智能平台涉及智能座舱、自动驾驶等专用场景，需要对算力、存储、网络、AI 数据、大小模型（包括智能体）等的车云协同进行标准化，对功能性能的要求差异较大。

中国通信标准化协会（CCSA）与应用委员会（TC1）立项的两项异构硬件相关标准，规定了智能算力硬件的统一编程模型及范式技术要求，其主要目的是针对各类异构硬件提供一套统一的编程模型及范式，屏蔽开发生态的复杂多样性，提升应用的开发效率，降低开发复杂度。《智能算力硬件统一编程模型及范式技术要求》中平台模型主要描述异构计算系统的结构，针对多种异构 API 进行编程，实现不同后端支持这些目标 API。内存模型包含内存类型、内存分配方式和内存一致性，描述主机和硬件设备的内存管理和映射方式。执行模型对程序如何配合平台模型在异构计算系统上执行的技术要求，包括任务划分和执行策略。编程范式描述开发程序编写规范，完成并行执行的任务间的同步操作等[19]。

车用智能算力硬件和其他智能算力硬件从集成的异构硬件类型，到主机和各异构硬件的协同关系基本一致，因此该标准对车用智能算力同样适用。但车用 SoC 芯片与其他 SoC 在芯片架构设计上存在多方面区别，如车用 SoC 更注重多核性能以应对复杂环境和多种任务处理需求，在内存分配和执行策略等方面会有相应的差别。编程模型和范式也需要更侧重于满足汽车特定应用的开发需求，例如针对实时图像处理、传感器数据融合等方面进行优化。可能会提供一些特定的库、函数或接口，方便开发者快速构建与汽车相关的智能应用程序，使开发者能更高效地利用硬件资源来实现汽车领域特有的功能和算法。

国家标准《人工智能算子接口第 2 部分：神经网络类》计划制订标准化的软件接口，有望将上述映射的复杂度降低，从而有力地促进人工智能软硬件生态的良性发展[20]。该标准规范了人工智能领域的神经网络类算子的基本功能和参数要求类，可以覆盖常用 AI 算法中的大部分神经网络类算子。本标准支持不同维度、不同密度、不同精度和不同类型的数据处理，适用于不同类型的基础数学操作，不限制软硬件的实现方式，给予接口实现者较大的优化空间，能满足云服务器、边缘设备、终端设备等不同神经网络处理器设备的计算和存储需求，并提供相关软硬件优化扩展以及用户自定义的扩展等功能。

3.3 车用 AI 平台架构

车用智能算力平台需具备高可靠性、实时计算能力、低功耗及安全性，以支撑自动驾驶核心功能与座舱个性化交互。其架构由两部分构成：异构分布式硬件提供算力基础；系统软件层分为车端（实时控制）与云端（数据处理与升级）两大模块。总体架构图如图 3-7 所示。

汽标委智能网联汽车分委会

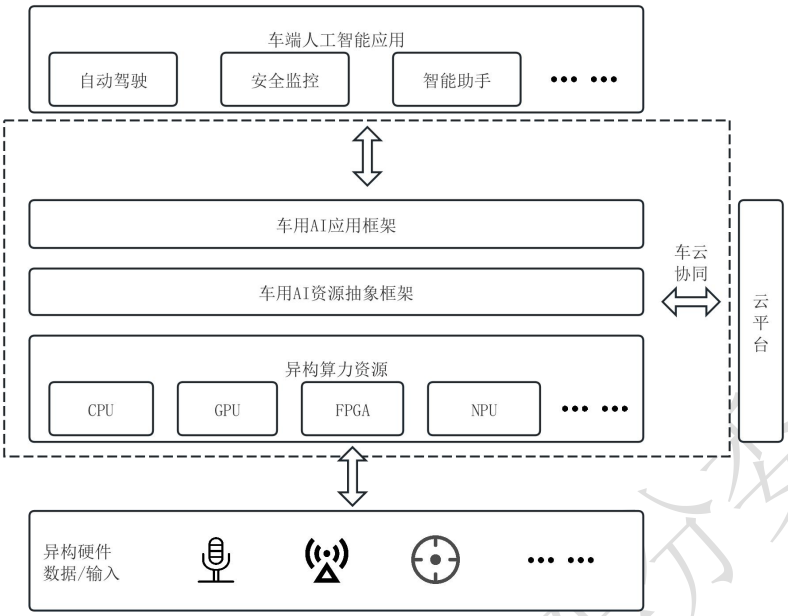


图 3-7 车用 AI 基础系统平台总体架构

3.3.1 硬件平台

3.3.1.1 AI 基础硬件平台处理单元的基本介绍

车用 AI 基础系统平台通过异构计算架构优化任务分配，根据处理单元特性匹配计算需求。硬件平台通常包括 AI 计算单元、通用计算单元、控制单元以及安全处理单元等。

(1) AI 计算单元

AI 计算单元基于并行计算架构，负责图像处理、深度学习推理等数据密集型计算。GPU、NPU、TPU、BPU 等专用加速器等正逐渐引入到 AI 计算单元中，并负责不同的计算任务。随着数据量的增加、复杂模型的推理和部署、实时性要求的提高、AI 应用领域的丰富以及硬件技术的进步，对 AI 计算单元的算力需求持续增长，需要通过芯片制程升级以及内存访问、数据传输、电源管理、时钟管理、电路设计的优化提升 AI 计算单元的能效。同时，运用更灵活的任务划分和卸载机制、动态任务调度和资源管理等技术，实现通用计算单元与 AI 计算单元的协同，进一步提高处理效率。

汽标委智能网联汽车分委会

(2) 通用计算单元

通用计算单元负责运行复杂的逻辑串行任务。通用计算单元由多个车规级多核 CPU 组成，各单核主频高、计算能力强。通用嵌入式 CPU 通常采用 ARM 架构，近年来业界也在尝试基于 RISC-V 架构进行设计。

(3) 控制单元

控制单元是智能网联汽车子系统的核心控制组件，采用高安全和强实时性 MCU（Microcontroller Unit），包含实时多核 CPU、嵌入式存储及必要 IO/通信接口。其通过优化指令与数据存取通路设计、提升核心工作频率、部署实时任务调度器等方式降低任务切换延迟，满足实时性需求；同时集成 Ethernet/CAN-FD 等高速接口，提供硬件级数据包转发与路由功能，减少 CPU 资源占用并提升数据交换吞吐量，从而保障系统高效低延迟运行。

(4) 安全处理单元

安全处理单元负责安全业务的处理。硬件层面依据功能安全等级设计，采用内建自测（BIST）电路监测电路工作状态。部分执行单元采用冗余电路设计以实现高功能安全等级要求。数据存储与数据通信采用奇偶校验编码保证端到端数据传输安全。

(5) 扩展单元

扩展单元通过扩展槽或接口提供网络通信、数据输入输出等功能。在自动驾驶或高精度导航等应用中，可能需要集成摄像头、雷达、GNSS 或惯导等传感器，以支持高速数据传输或特定输入需求。

3.3.1.2 架构差异

虽然 AI 硬件架构存在差异，，但需在不确定性和异常情况时保持性能稳定。因此，需综合考量性能、功耗、可靠性、兼容性、安全性等多方面，以确保满足多样化场景的应用需求。

汽标委智能网联汽车分委会

(1) 性能

指计算性能，包括速度、吞吐量、AI 性能（AI 性能：包含任务覆盖度、模型支持率、实现精度、算力、精度、帧率等关键指标测试）等指标。通过各种工作负载测试，评估在不同任务下的性能表现。

(2) 功耗

指硬件平台在不同负载下的功耗消耗情况，以评估其节能性能。这涉及静态功耗和动态功耗的测试，以及在不同工作负载下的功耗曲线。

(3) 可靠性

检验在连续工作状态下的稳定性和可靠性。通过模拟实际工作环境，检测在长时间运行中是否出现故障或性能下降。

芯片在长期使用中的可靠性和耐久性。通过加速老化实验或持续运行测试，检验在不同工作条件下的可靠性表现。

(4) 兼容性

指 AI 相关硬件对不同操作系统、软件框架和硬件设备的兼容性，确保可以与各种软件组件无缝集成。

(5) 安全性

车用 AI 硬件平台不同单元应具有不同的功能安全等级，或者支持不同功能安全等级应用的设计。且满足在完整性、机密性、可用性等方面的信息安全需求。

3.3.1.3 AI 计算设备

车用 AI 用例需求呈现多样化特征，且计算需求覆盖垂直领域持续扩展。需依托异构计算架构整合不同处理单元优势，以应对复杂场景。常用的 AI 计算资源种类如图 3-8 所示。

汽标委智能网联汽车分委会

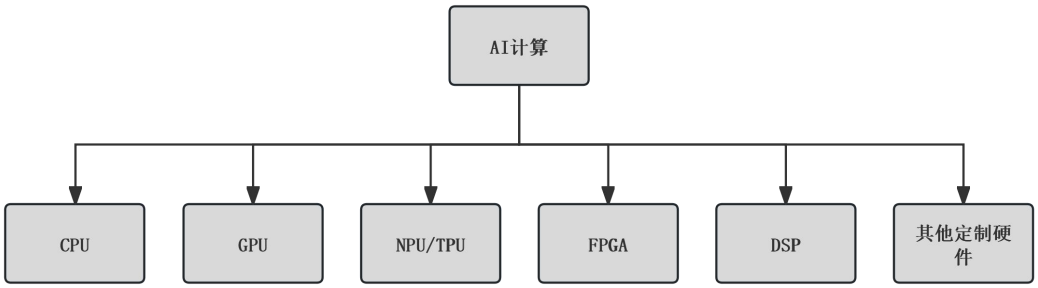


图 3-8 AI 计算设备不同类型

对于异构硬件架构，需考虑制定差异化测试指标并基于统一测试方法执行，实现测试结果的横向对比，支撑计算设备能力评估。包含的指标有 CPU 算力（Dhrystone（DMIPS）、CoreMark、SPEC 或 Stream 指之一）、GPU 算力、AI 算力（帧率、时延，推荐 INT8 精度）；最大分辨率、吞吐量和时延；编解码格式、最大分辨率、最大帧率、时延；DDR（Double Data Rate）。

3.3.1.4 输入设备

比如摄像头、麦克、雷达等。

3.3.1.5 输出设备

涉及车外信息的处理需考虑屏幕、方向盘、制动系统等输出的信号反馈，车内座舱的输出结果可通过语音系统、屏幕等反馈。

3.3.2 系统软件

3.3.2.1 AI 基础软件

大模型是人工智能迈向通用智能的里程碑技术，标志着在互联网时代与云计算时代之后，正式进入以模型为核心的时代。大规模数据集和模型训练算力等领域取得的突破性进展，离不开 AI 基础软件的底层支撑，但行业应用仍面临技术和成本的双重门槛。AI 基础软件通过三个层面加速大模型落地应用。开发层面提供大模型能力萃取；在训练层面提供高效硬件层支持的分布式训练框架；运维层面作为企业基础设施强化模型管理与部署能力。

汽标委智能网联汽车分委会

作为大模型生态系统的中坚力量，AI 基础系统将会成为大模型应用落地的最主要的效率支撑。因此，为简化开发过程和降低门槛，降低各层级软件之间的耦合度，依次约定各层级软件的具体分工，各软件层级主要分工如图 3-9 所示：

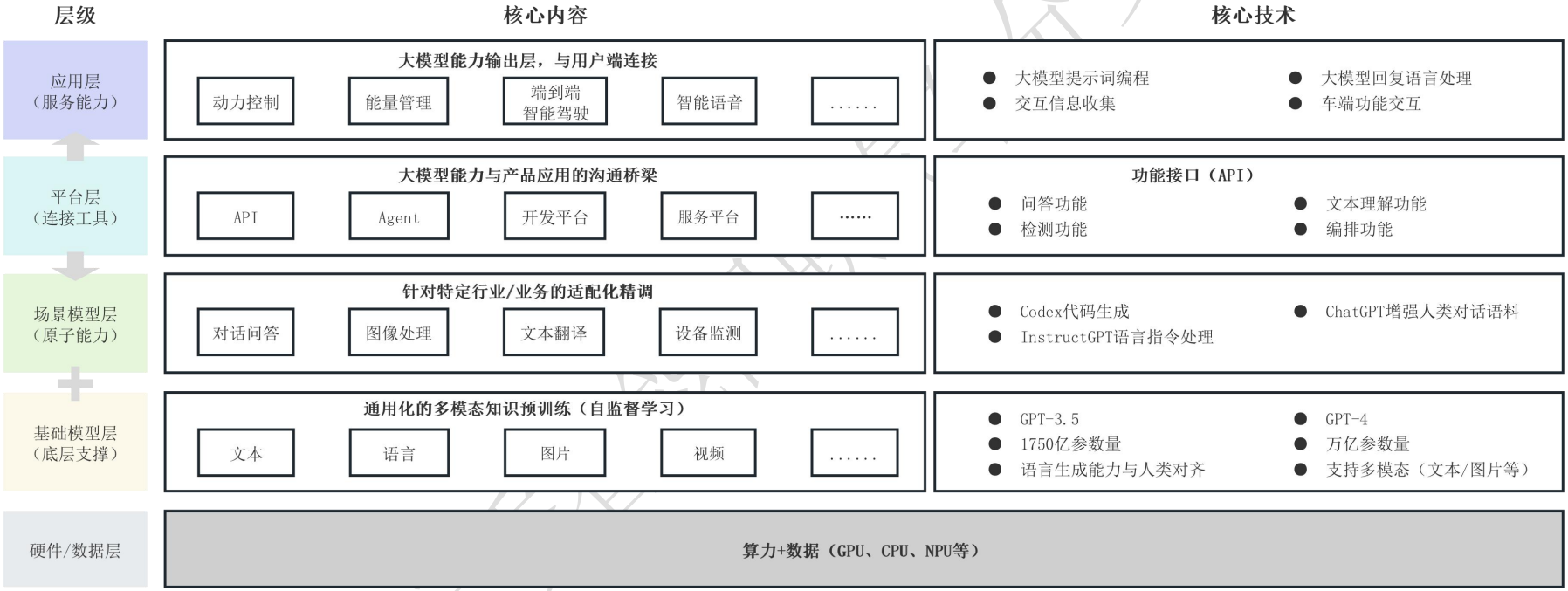


图 3-9 大模型产业应用范式[21]

(1) 硬件/数据层：为大模型的开发和训练提供硬件和数据基础。

(2) 基础模型层：通过大规模的数据集进行训练，能够处理多种不同类型任务、在多个领域和应用中表现出良好的效能的人工智能模型，通常包含千亿甚至万亿参数，能够执行自然语言处理、图像识别、语音识别等多方面的任务。

(3) 场景模型层：是在通用大模型的基础上，结合特定行业的数据和知识进行进一步训练和优化，针对特定行业的需求，能够更好地理解和处理行业特有的复杂问题和任务。

(4) 平台层：将场景模型层的能力进行封装，使得应用层软件开发者能够直接利用场景模型软件的功能来构建新的应用程序。

(5) 应用层：利用通用基础大模型的智慧能力，结合车企自身产品特色，打造智能化、差异化的产品。

3.3.2.2 车云协同

车端的计算资源受限导致车云协同成为车用 AI 基础系统的关键之一。传统的车用人工智能交互系统聚焦车窗控制、加油站搜索等基础场景。近年来，大模型技术推动车用交互系统升级，支持触控交互、语音交互、面部识别算法交互、手势识别交互、生物（指纹、虹膜、静脉、心率等）识别交互等。鉴于驾驶场景对驾驶者注意力要求高，众多车企率先从语音助手模块突破，结合大模型技术提升准确性和功能丰富性。典型应用场景包括多样化意图识别和理解、个性化服务生成、多意图路径规划及车用全场景智能体等。

(1) 车云协同架构

针对车端计算资源受限、车辆是否可以预装、是否部署大模型等情况，可分为三种架构：类型 I 模型纯云端协同架构（见图 3-11）、类型 II 车云模型协同架构（见图 3-12）和类型 III 大模型协同架构（见图 3-13），提供车云协同架构、功能和接口。对车用 AI 基础系统平台的车云协同架构达成共识，将有助于相关

汽标委智能网联汽车分委会

方遵循统一的框架进行多场景研发和建设，避免重复开发，降低开发成本和资源利用率。

类型 I 模型纯云端协同架构：自然语言理解和生成模型均部署在云端，适用于车端资源较少的情况，例如现有车型 OTA 升级或低配新车前装等情况。

类型 II 车云模型协同架构：自然语言理解和生成模型分别部署在车端和云端，适用于车端资源较多的新车前装等情况。

类型 III 大模型协同架构：包括纯云端大模型协同架构和车云大模型协同架构。传统的车云协同 AI 系统广泛应用于打开车窗等简单场景。结合大模型技术可实现更准确、更丰富的汽车智能化语音助手能力，支持多样化意图识别和理解、个性化音乐律动、多意图路径规划、车用全场景智能体等复杂应用场景。

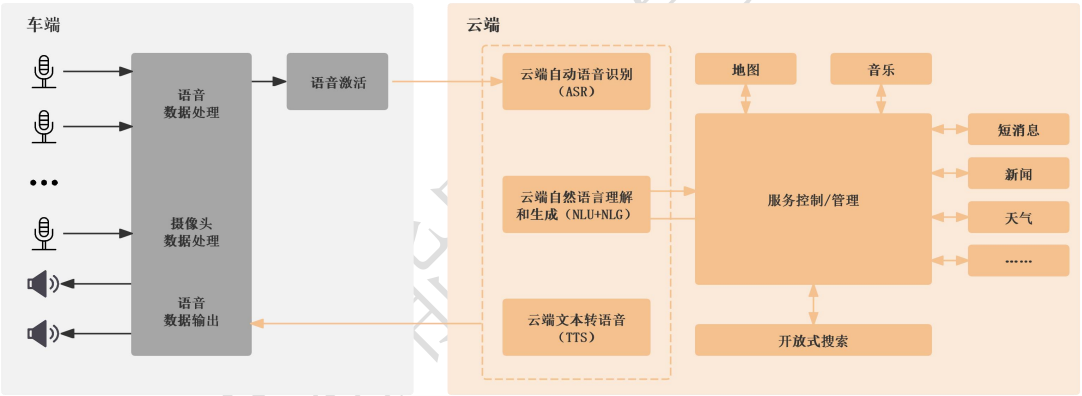


图 3-10 车云协同架构类型 I（模型纯云端协同）

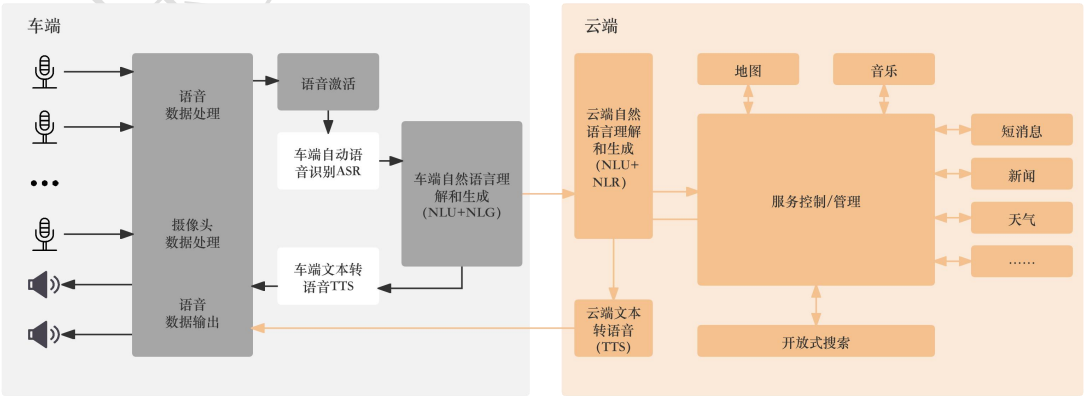
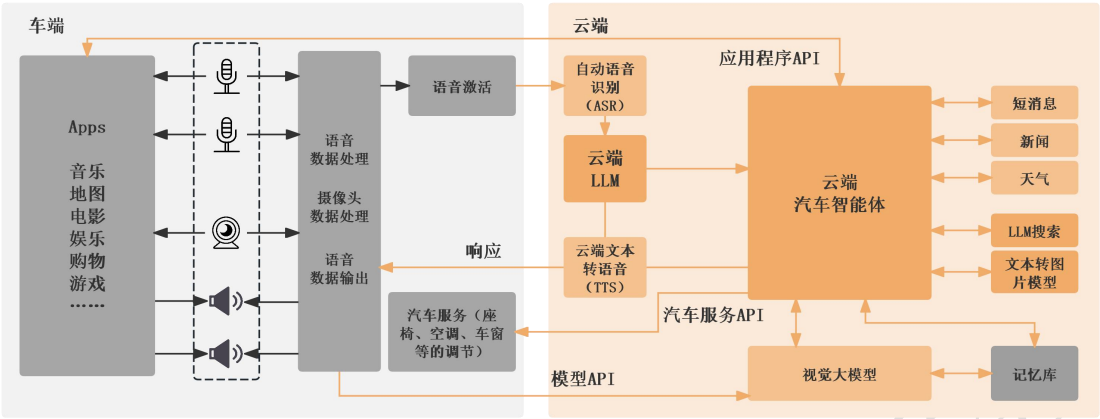
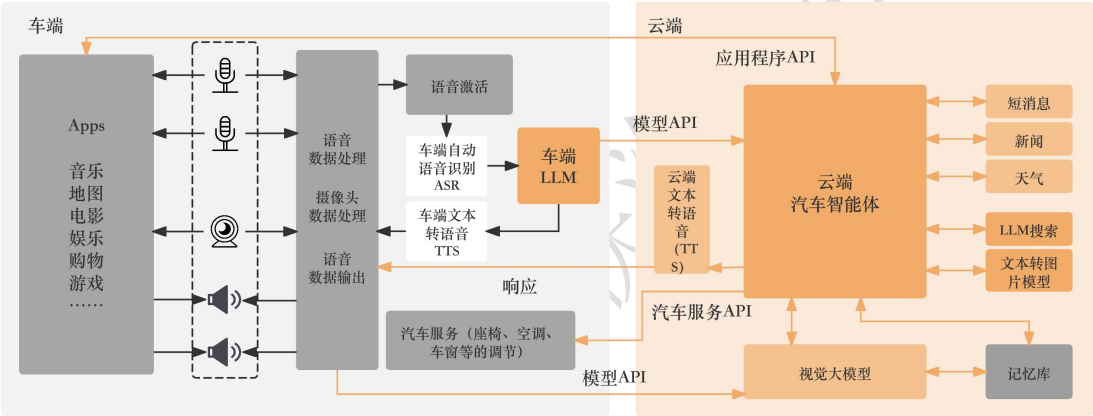


图 3-11 车云协同架构类型 II（车云模型协同）



（a）纯云端大模型协同



（b）车云大模型协同

图 3-12 车云协同架构类型 III（大模型协同）

(2) 模型纯云端协同

模型纯云端协同的云端支持对用户的意图理解、思考与规划，实现多模型协同，并将处理结果返回车端；车端可支持车内设备与应用的协同控制；云端提供大规模存储服务，支持各业务系统的汽车数据流转至存储系统，支持高性能存储、对象存储、块存储、并行文件存储、全闪高速存储和混闪容量存储，支持对接多类计算引擎和多种存储介质选择，支持检索能力，具备存储容量扩展到 PB 级能力；云端支持 RDMA 高性能网络、虚拟专用云（VPC）网络和超大规模网络，国产化场景应支持 RoCE 网络，支持负载均衡和流量控制机制，确保数据传输速

度满足实时交互需求；云端具备高可用性和容错能力，确保计算、存储、网络服务的持续稳定运行。

(3) 车云模型协同

为了降低车用 AI 场景模型推理成本，同时保障在弱网、断网场景下仍具备 AI 能力，充分利用车机算力资源，车用 AI 场景构建需充分考虑端云多模型协同。云端大模型负责对用户复杂任务的理解、思考与规划，实现多模型协同，在车端通过一个相对尺寸较小的模型，实现用户的意图理解和车内设备与应用协同控制。云端提供大规模存储服务，支持各业务系统的汽车数据流转至存储系统，支持高性能存储、对象存储、块存储、并行文件存储、全闪高速存储和混闪容量存储，支持对接多类计算引擎和多种存储介质选择，支持检索能力，具备存储容量扩展到 PB 级能力；云端支持 RDMA 高性能网络、虚拟专用云（VPC）网络和超大规模网络，国产化场景应支持 RoCE 网络，支持负载均衡和流量控制机制，确保数据传输速度满足车云实时交互需求；云端具备高可用性和容错能力，确保计算、存储、网络服务的持续稳定运行。

(4) 计算资源协同

在车云协同中，云端服务器承担着大规模数据存储和复杂计算任务，而终端设备则负责实时感知、数据采集以及低延迟的响应。车用 AI 场景涉及大量端侧服务需求，需要强大的算力支持进行实时推理和模型更新，为了保障用户在车用场景的使用体验，对于模型反馈响速度应提出了非常高的要求，应重点关注大模型推理的 QPS (Queries Per Second)、首包延时、每秒 token 生成量等关键指标。计算资源协同的云端支持对任务进行计算资源分配，由云端全部承担，或将一部分计算任务分配给车端完成；云端支持复杂非实时计算任务的处理，并将结果传输回车端；车端支持云端分配的简单实时计算任务的处理，并将结果传输回云端；云端支持计算资源的管理，包括跨集群的资源开通、释放、升配、降配等；支持弹性计算和多任务计算资源调度，采用负载均衡等技术实现跨集群调用资源，包

汽车智能网联汽车分会

括批处理、交互式应用、实时流处理等；支持流量管控，按照任务的不同权重或阈值使用计算资源，实现高效的车云协同。

(5) 个性化辅助驾驶

在汽车智能化领域，车辆需要实时感知周围环境（交通流量、天气、空气质量），并做出决策。车云协同可以使车辆实现实时数据处理和决策，同时将关键数据上传到云端进行更深入的分析和学习，同时可基于历史信息，为用户提供符合心意的智能决策。个性化辅助驾驶的车端支持驾驶数据的实时处理，并将关键数据上传到云端；云端支持对驾驶数据的深度分析和学习，基于用户和车辆的相关历史信息，为用户提供个性化的辅助驾驶决策。

(6) 多模态数据处理

车用 AI 场景交互涉及语音、文本、图像等多种数据类型。云平台支持车用图像、语音等多模态数据的统一接入；支持多模态数据的预处理，包括多模态数据的混合标注、图像变换及数据增强、多模态数据集管理和协同等；支持多模态数据的融合分析；提供针对不同模态数据的专门处理框架，便于大模型高效地理解和响应用户多维度的交互请求。

(7) 模型更新与版本管理

由于 AI 技术发展迅速，大模型需要定期更新以保持其先进性和准确性。云平台应提供便捷的模型版本管理功能，支持无缝升级、灰度发布和回滚操作，确保新模型上线不影响用户体验，且能快速响应潜在问题。

(8) 车用大模型协同

车用大模型协同支持车云意图识别与响应，使用车云意图识别来理解用户意图，并基于车云协作生成响应；支持车云多意图规划生成，基于多意图识别生成规划；对于模型纯云端协同架构，车端支持车用图像、语音等多模态数据统一接入云端，云端提供车用 AI 慢应用相关的多模态交互信息生成；对于车云模型协同架构，车端提供车用 AI 快应用相关的多模态交互信息生成，云端提供车用 AI

汽标委智能网联汽车分委会

慢应用相关的多模态交互信息生成，例如，基于感知的语音和图片信息，生成相关的交互信息；支持个性化驾驶辅助信息生成，结合用户与大型模型之间的历史对话，分析用户偏好、实时主动提醒和车辆控制设置；支持车用全场景智能体，提供不同应用程序之间的系统级交互，智能体能够基于语音交互自主执行下游任务，支持远程故障提醒、车辆信息问答、多意图理解与控制、推荐与提醒、抖音聊天等车用全场景。

(9) 接口

车云协同接口主要包括三类：车用人工智能汽车服务 API 在云端智能体和汽车服务之间提供接口，如车窗、座椅、空调等的车内设置；车用人工智能模型 API 为车端算法模型与云端小模型、智能体、多模态模型等提供调用功能；车用人工智能应用程序 API 在结合用户意图识别和理解的基础上，通过云端智能体为车用生态应用服务提供 APP 接口，支持车端和云端的各类 APP 的调用，如导航、支付、出行、送餐等 APP。

(10) 低延迟通信

为了保证用户与车用 AI 之间的交互流畅无阻，云平台需提供低延迟、高带宽的网络连接。特别是在需要实时反馈的场景中，确保数据传输速度满足实时交互需求，减少延迟对用户体验的影响。

(11) 数据安全与隐私保护

智能座舱涉及大量用户个人数据、车辆状态数据以及地理位置信息，云平台必须符合严格的数据安全标准，并采用加密传输、访问控制、数据脱敏等手段保障数据的安全存储与传输。此外，应具备完善的身份认证机制和权限管理功能，确保只有授权用户和设备才能访问相关服务。

汽标委智能网联汽车分委会

3.3.3 开发框架

3.3.3.1 应用

随着大模型的发展，AI 在理解和生成能力上有了大幅提高，国内头部车厂在积极把这种能力引入智能座舱领域。比如理想汽车在 2023 年 12 月 10 号发布的 OTA 5.0 版本，把自研的 MindGPT 引入智能座舱，2024 年以来，华为、蔚来等也完成了大模型上车。大模型的引入使得原有的语音助手在感知、理解、规划、执行等各个环节都有了大幅增强，从而逐渐出现了一种新的业务形态：即用户通过语音表达意图，大模型识别用户意图并调用来自系统和不同应用的功能完成用户意图，具体流程如图 3-13 所示。与传统的业务场景不同，用户不再直接与应用进行交互，而是由大模型来代替用户完成。这样节省了用户操作，提高了交互效率，使得语音助手成为了用户的个人管家或者秘书，提高了座舱的智能化水平。

为了实现上述场景，系统和应用需要把自己的“功能”贡献给大模型，让大模型理解各个功能的输入、输出、具体的功能等。由于应用可能来自于第三方，所以这里有必要对这一机制进行标准化，使得不同的应用或者同一个应用面对不同的座舱系统，都可以统一的方式进行能力的贡献，而不必在不同系统直接进行适配，促进应用在不同系统间的复用。

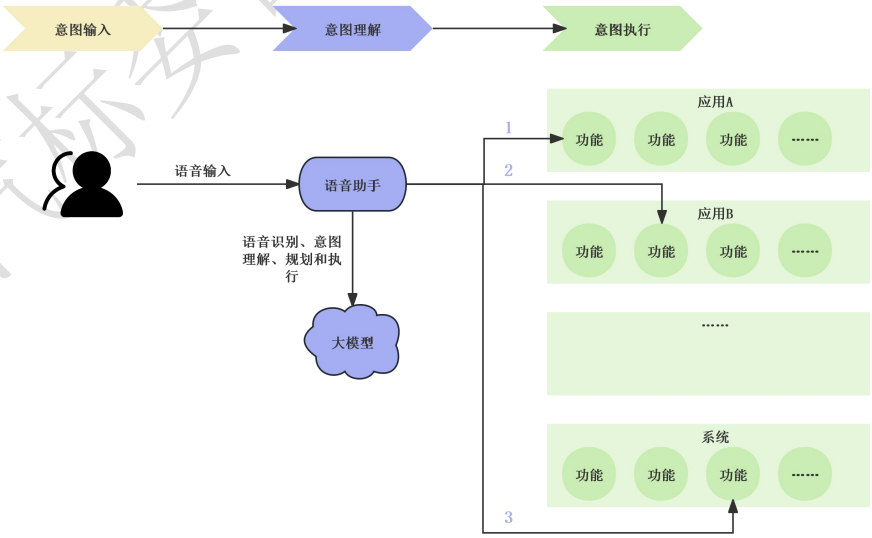


图 3-13 车用 AI 应用新形态

3.3.3.2 数据

这里的数据指的是车里用来推理的数据，包括车辆状态和各种传感器的输入（比如摄像头、各种雷达、传感器、车辆信号）、存储于车辆的历史信息和用户信息等。

不同业务类型需要不同的数据作为输入，但是随着智能化的发展，尤其在智能座舱中，车辆也在逐渐引入第三方的开发者。出于安全和隐私保护的考虑，需要对这些业务对车辆和用户数据的使用进行管控，包括访问授权、访问记录、用户告知等。

3.4 总结与展望

随着智能网联汽车电动化和智能化转型加速，车用 AI 平台作为支撑自动驾驶与智能座舱的核心平台，存在异构硬件兼容难、开发碎片化和安全合规要求高等挑战，搭建车用 AI 技术标准化体系是降低研发门槛、实现跨平台迁移、保障功能安全与数据安全的有效路径。本章节主要围绕车用 AI 基础系统平台标准化研究，系统梳理了车用 AI 架构（计算资源、系统软件、开发框架）、核心模块（硬件平台、软件系统、模型业务、云平台）的技术现状与产业实践，对已有标准情况进行梳理和分析。

通过对车用 AI 架构（计算资源、系统软件、开发框架）、核心模块（硬件平台、软件系统、模型业务、云平台）的技术现状、产业实践和已有标准梳理和分析，车用 AI 基础系统平台标准化主要方向有：

- 平台算力资源标准化，针对异构硬件（如 GPU/NPU/CPU）制定统一算力评估指标（如 AI 算力、功耗、时延）与抽象接口标准（如 oneAPI、Android NNAPI 适配优化），支持跨硬件平台的算法复用；
- 车云协同标准化，围绕车端、云端、车云协同和大模型协同的模型部署、计算资源分配、低延迟通信及数据安全等制定相关标准，平衡及时性与成本；

- 大模型车端部署标准化，针对多模态交互、个性化服务和智能体构建等场景，规范大模型的车端部署要求（轻量化、能效）、数据调用权限及更新管理机制，推动大模型与车用场景深度融合。

随着车用 AI 基础系统平台标准体系完善与生态协同深化，车用 AI 基础系统平台可有效助力智能网联汽车向高阶自动驾驶与全域智能化发展。

参考文献

[1] 艾克热木江·赛买提, 数据安全管理,智能网联汽车不可逾越的考验

[2] 任利敬. 面向自动驾驶的大规模图像安全传输与可信人工智能技术研究[D]. 石家庄铁道大学, 2022.

[3] 李升波, 刘畅, 殷玉明, 等. 汽车端到端自动驾驶系统的关键技术与发展趋势[J]. 人工智能, 2023, (05): 1-16.

[4] 赵明姝. 车载导航到自动驾驶电子地图发展综述[J]. 测绘通报, 2023, (06): 6-10+92.

[5] 祖国强, 何俏君, 张志德, 等. 无人驾驶环境感知中多源数据融合应用综述[J]. 汽车文摘, 2022, (08): 8-13.

[6] Han, Yushan, et al. Collaborative perception in autonomous driving: Methods, datasets, and challenges. IEEE Intelligent Transportation Systems Magazine (2023).

[7] 《中国公路学报》编辑部. 中国汽车工程学术研究综述·2023[J]. 中国公路学报, 2023, 36(11): 1-192.

[8] 何文思.汽车智能座舱人机交互系统性设计与研究[D]. 河南工业大学, 2022.

[9] Artificial Intelligent Act (Regulation (EU) 2024/1689), Official Journal version of 13 June 2024.

[10] ISO/IEC 42001:2023 Information technology - Artificial intelligence - Management system[S/OL].[2025-04-28].https://www.iso.org/standard/81230.html

汽标委智能网联汽车分委会

[11] ISO/PAS 8800:2024 Road vehicles - Safety and artificial intelligence[S/OL].[2025-04-25].https://www.iso.org/standard/83303.html

[12] oneAPI Specification[DB/OL].[2024-12-18].https://oneapi-spec.uxlfoundation.org/specifications/oneapi/v1.3-rev-1/

[13] Neural Networks API[DB/OL].[2025-06-11]https://developer.android.com/ndk/guides/neuralnetworks?hl=zh-cn

[14] Wu, P., Chen, L., Li, H., Jia, X., Yan, J., & Qiao, Y. "Policy pre-training for autonomous driving via self-supervised geometric modeling." In ICLR, 2023.

[15] Tian, X., Gu, J., Li, B., Liu, Y., Wang, Y., Zhao, Z., et al. "DriveVLM: The Convergence of Autonomous Driving and Large Vision-Language Models." 2024.

[16] T/STSI 43-2023 人工智能算力资源池技术规范[S/OL].[2023-11-06].https://qxb-img-osscache.qixin.com/standards/7ecdfcfb6348bffc582f66f4586f2431.pdf

[17] T/CESA 1268-2023 信息技术算力服务能力成熟度评估模型[S/OL].[2023-07-27].https://www.ndls.org.cn/standard/detail/350c124579a6fba36459bb04d6916f6e

[18] 人工智能 计算设备调度与协同第 2 部分： 分布式计算框架[S/OL].[2025-03-28].https://std.samr.gov.cn/gb/search/gbDetailed?id=31DA5F377BDE8F08E06397BE0A0A4CFB

[19] 智能算力硬件统一编程模型及范式技术要求[R/OL].[2024-04-12].https://www.c114.com.cn/news/118/a1259689.html

[20] 人工智能 算子接口 第 2 部分：神经网络类[S/OL].[2023-08-06].https://std.samr.gov.cn/gb/search/gbDetailed?id=02DDA540F04FAACCE06397BE0A0A0A50

[21] 盖斯特研报：大模型的本质及其对汽车行业的影响[R/OL].[2024-07-22].http://www.gast-group.com/insightDe_599.html

汽标委智能网联汽车分委会

二、车用 AI 技术应用及标准化

第四章 车用 AI 智能网联汽车感知标准化研究

4.1 车用 AI 感知标准化研究概述

感知（perception）功能是自动驾驶车辆实现环境理解的核心基础，也是实现自主决策的前提条件。感知系统是模块化的自动驾驶系统中的核心子系统（端到端系统的感知功能不在本章进行讨论），需要执行障碍物检测、车道线检测等多项任务，其通过从车辆行驶环境中获取和理解信息，进而为后续的决策和控制提供可靠的输入。深度学习等 AI 算法是目前实现自动驾驶感知的主流技术手段，基于已知数据样本训练的 AI 感知模型能够从复杂的传感器数据中提取关键特征，并推理得到结构化的环境信息，从而实现对环境的理解。本章的内容将突出“车用 AI”的主题，弱化对传感器及非 AI 感知方式（如通过 V2X 等方式获得环境信息），着重探讨智能网联汽车的关键感知任务、AI 感知模型、测试场景及其标准化方向。

4.1.1 车用 AI 感知标准化研究目的与意义

当前，车用 AI 感知在环境理解的精确性、稳定性和适应性方面仍存在显著差异，多源异构传感器的数据融合与协同机制尚不完善，导致感知系统的整体效能受限；感知数据的采集规范、同步接口、标注准则缺乏统一框架，不同厂商的感知数据兼容性与可对比性难以保障；在复杂动态交通场景下，感知算法在目标检测、语义分割、轨迹预测等关键任务的性能评估体系尚未形成行业共识，测试验证环节的场景覆盖度与边界条件设定亦缺乏系统性指导。本标准化研究旨在通过对于车用 AI 感知技术现状的分析与归纳，探索未来的标准化方向与可能性。其意义在于打破技术壁垒，推动行业在数据互通、算法优化和场景适应性上的协

汽标委智能网联汽车分委会

同发展，为提升自动驾驶系统的可靠性与安全性提供参考，为技术迭代、产品认证及规模化应用提供统一基准，最终促进智能网联汽车的高质量发展。

4.1.2 车用 AI 感知标准化现状

4.1.2.1 国内标准化现状

当前我国人工智能标准化建设已取得阶段性成果，GB/T 42888-2023、GB/T 42755-2023 等通用型国家标准为 AI 技术的基础框架提供了规范支撑。然而，在汽车智能化加速落地的产业背景下，车用 AI 感知技术仍面临标准体系专项化不足的挑战。现有通用标准虽能覆盖数据治理（如 GB/T 42755-2023 对数据标注流程的标准化要求）、算法安全（如 GB/T 42888-2023 提出的抗攻击设计规范）等基础维度，但针对自动驾驶场景中复杂环境感知的特殊需求——包括极端工况下的算法鲁棒性、多传感器融合的实时性要求、持续学习机制的可靠性等——尚未形成针对性技术规范。传统汽车电子标准体系中的感知系统要求（如计算机视觉规范）更难以适配深度学习驱动的现代 AI 感知技术特性，尤其在神经网络可解释性（GB/T 43782-2024 虽有涉及但未细化）、动态场景泛化能力（GB/T 45225-2025 的评估体系需进一步场景化）等核心维度存在明显断层。这种针对车用 AI 感知标准化研究的滞后性可能导致风险控制与监管不全面与技术应用碎片化，因此亟待行业主管部门联合产学研机构，在现有标准基础上构建兼顾技术创新与安全可控的车用 AI 感知标准生态。

4.1.2.2 国际标准化现状

当前国际车用 AI 感知标准化研究呈现多维度协同推进态势，核心聚焦感知系统的安全性、可靠性与合规性。ISO/IEC 42001 构建的全生命周期治理框架强调模型可解释性与结果可追溯性，要求感知算法逻辑透明且输出符合伦理规范，这与欧盟《人工智能法案》形成联动。ISO 26262 从功能安全角度提出宏观要求，感知系统需通过危害分析（HARA）和 ASIL 等级验证，确保在故障时维持安全输出。ISO 21448 针对预期功能安全（SOTIF）提出更细化的感知系统要求：需覆盖边缘场景识别、动态环境适应、冗余容错等能力，并通过系统化验证流程保

汽标委智能网联汽车分委会

障复杂环境下的鲁棒性。ISO/PAS 8800 整合前两项标准方法论，扩展出覆盖数据质量、模型开发到监控更新的全流程安全框架，其中数据可靠性（如高覆盖度独立数据集）被列为感知系统的基础要求。ISO/IEC 24029 则专门规范神经网络鲁棒性，要求感知模型通过对抗攻击测试和形式化验证确保稳定性。德国向车用 AI 技术标准化与合规性评估的指导性文件 AIMobilityAuditPrep 进一步提炼五大安全维度：性能保障（如恶劣环境下的识别精度）、鲁棒性验证（覆盖极端场景测试）、数据可靠性（高覆盖度与独立性数据集）、可解释性决策（模型逻辑透明化）及实时监控（输入/输出合理性校验），形成从算法设计到运行维护的安全框架，最终为失效可容错、决策可追溯、性能可验证的技术目标提供标准支撑。

4.2 车用 AI 感知系统架构、关键任务与主要实现方法

车用 AI 感知系统需要理解复杂的车辆行驶环境，并为后续的智能决策提供多维度、多粒度的感知信息，因此其中涉及多个不同的关键感知任务。本节将对车用 AI 感知系统基本架构进行介绍，进而对关键感知任务在感知系统中发挥的功能进行梳理。同一感知任务通常具有多种实现方式，本节也将针对各项关键感知任务分别介绍几种典型 AI 模型，从而为车用 AI 感知技术的标准化提供参考。

4.2.1 基本架构

车用 AI 感知系统的基本架构如图 4-1 所示。数据源为车用 AI 感知系统提供输入数据，主要包含视觉传感器采集的图像数据、激光/毫米波雷达采集的点云数据，以及 GPS、车速表、加速度计等传感器提供的位置、速度、加速度、方向等自车实时状态数据等。这些数据通常会经过时空同步、数据压缩、数据清洗等处理，以保证有效可靠的数据被输入给后续的 AI 感知模型。

AI 感知模型是实现感知任务的关键，实现不同感知任务（如目标检测、语义分割等）的模型通常具有不同的结构，它们接收输入数据，依据其具体任务进行特征提取与推理，得到相应的推理输出。这些推理输出将被进一步组织成能够结构化表示的感知结果数据，服务于后续的预测、决策与控制环节。

汽标委智能网联汽车分委会

感知融合通过前、中、后三种方式协同提升系统的感知能力：前融合（数据级）整合多模态数据源信息，实现高精度时空同步与多维属性综合表示；中融合（特征级）通过 AI 模型提取不同数据源输入的中间特征并进行融合，兼顾计算开销与多模态互补；后融合（目标级）对不同 AI 模型感知结果进行整合，增强感知置信度与容错性。

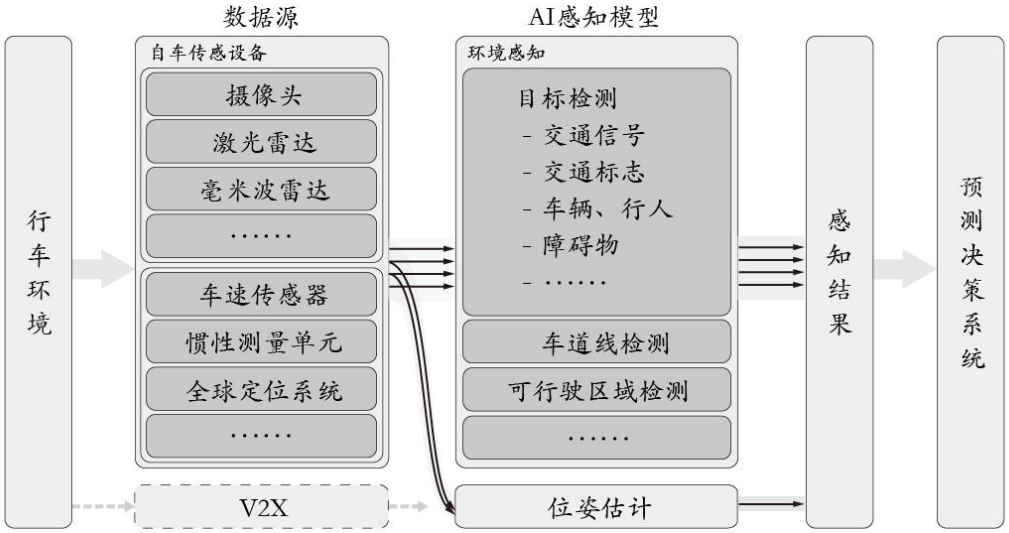


图 4-1 车用 AI 感知系统架构示意图

4.2.2 车用目标检测

车用目标检测任务旨在行车过程中基于传感器采集的实时数据，确定车辆行驶环境中物体的位置、形状、类别以及运动信息（如图 4-2）。车用目标检测 AI 模型通常需要支持对交通信号灯、交通标志、车辆、非机动车与行人等目标物的识别，模型输出包含目标的位置（中心坐标）、大小（边界框尺寸）和类型的信息；对于车辆、非机动车与行人以及其他动态障碍物，还应输出其速度、加速度、朝向角等运动信息。以下列举几种实现车用目标检测的典型 AI 模型。

图 4-2 车用目标检测结果可视化

Yolo（You Only Look Once）[1]是一系列基于深度学习的实时目标检测模型，以整幅图像为输入并直接在输出层回归预测目标位置和类别，将检测任务转化为单阶段的回归问题以提高效率（如图 4-3）。其核心原理是将输入图像划分为网格，每个网格负责预测多个边界框的坐标、置信度及类别概率。当目标中心位于某网格内时，该网格承担检测职责，通过综合边界框定位损失、置信度误差和分类损失进行训练，这种模型结构赋予 Yolo 极快的处理速度，平衡了识别效率与准确性的需求。经过持续迭代优化，Yolo 系列通过对骨干网络、引入锚框机制、多尺度预测和注意力模块、训练策略等的改进，在维持实时性的同时显著提升检测精度，使其在自动驾驶的实时目标检测场景中保持领先地位。

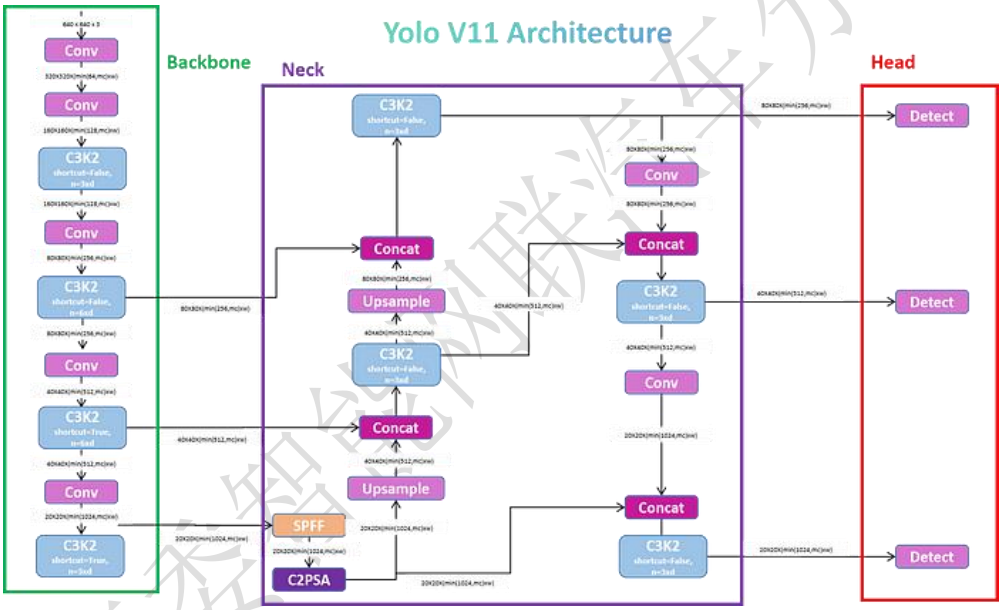


图 4-3 Yolo V11 架构图[1]

CenterNet[2]是一种基于关键点检测的单阶段目标检测模型，由主干网络、中心点预测网络和尺寸预测网络三部分构成。主干网络首先提取输入图像的深层特征，为后续任务提供基础信息。中心点预测网络通过生成高斯热图定位目标中心，真实热图中每个目标中心对应高斯分布的峰值，通过最大化预测与真实热图的相似性实现精准定位。尺寸预测网络则基于中心点特征回归目标的宽高值。该模型摒弃传统锚框设计，每个目标仅对应一个中心点作为正样本，因此无需非极大值抑制（NMS）后处理，显著简化检测流程。其结构简洁高效，在保持较高检测

汽标委智能网联汽车分委会

精度的同时实现实时处理，在目标检测领域展现出了强大的应用潜力，适用于自动驾驶等时效性敏感场景。

FCOS3D[3]是基于全卷积单阶段检测框架（Fully Convolutional One-Stage Object Detection）的 3D 目标检测模，通过融合图像与 3D 点云数据实现高效检测。其核心思路是将 3D 检测转化为类似 2D 检测任务，利用成熟的 2D 技术简化 3D 检测流程。算法采用 2D 引导的多层级策略，先将 3D 目标投影至图像获取中心点，结合三维信息预测目标的位置、尺寸及类别。特征提取阶段使用 ResNet 等骨干网络提取二维特征，同步融合点云数据的三维特征，支撑后续回归与分类任务。创新性引入基于 3D 中心的 2D 高斯分布重构中心度指标，更精准定位目标三维中心，有效提升检测精度。FCOS3D 优势体现在多模态数据融合能力，通过二维图像的空间信息与点云深度信息互补，在保持实时性的同时，提升复杂场景下的检测鲁棒性。其框架灵活，可适配不同网络结构与数据输入形式，为车用感知任务提供了轻量高效的解决方案。

4.2.3 车道线（路沿）检测

车道线（路沿）检测旨在根据传感器采集的车辆周围的环境数据，实时识别道路标线的位置与类别，从而为自动驾驶系统提供车道保持、横向控制及避障决策的核心依据，以保证车辆在行驶的过程中不偏离或者超出车道范围（如图 4-4）。车道线检测 AI 模型输出包括车道线点集坐标（像素坐标/实际位置坐标）、置信度、车道线状态类别标签。以下列举几种实现车道线检测的典型 AI 模型。

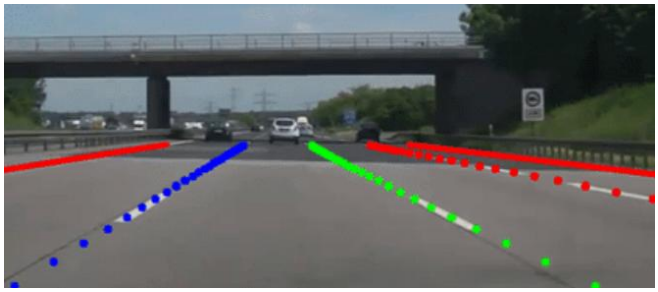


图 4-4 车道线检测结果可视化

CondLaneNet[4]是一种基于条件卷积和逐行公式的车道实例检测框架，它摒

汽标委智能网联标准分委会

弃了传统方法对语义分割和后处理的依赖，实现单阶段的车道线识别。其骨干网采用标准 ResNet 和 FPN 进行多尺度特征提取，增加了变压器编码器模块以更有效地提取上下文特征。提案头负责检测位于线路起点位置（提案点），同时对包含动态卷积核的参数映射进行预测。条件形状头预测行方向位置、垂直范围和偏移映射，以描述每一行的形状。此外，CondLaneNet 设计了循环实例模块（RIM）来区分重叠线，因此支持分叉线、密集线等复杂拓扑结构车道线的感知。

GANet[5]是一种基于关键点全局关联的车道检测模型，其核心思想是将图像中检测到的车道线关键点直接回归到对应车道线的起点，替代传统逐点关联或锚点方法，以提升检测效率并增强对复杂拓扑结构的处理能力。模型利用主干网络预测关键点位置及置信度，并计算其到所属车道线起点的偏移量，结合车道特征聚合器（LFA）建立的局部相关性，将离散关键点依据偏移量全局关联至对应起点，从而构建完整车道线信息。该方法通过并行处理关键点与全局偏移预测，减少传统后处理的计算冗余，在交叉、密集等复杂场景中实现高效推理。

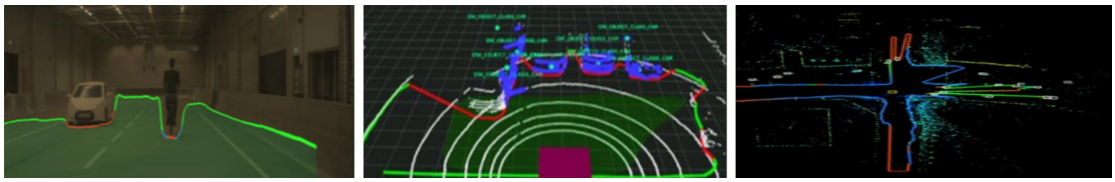
CurveFormer[6]利用 Transformer 架构中的稀疏查询表示和交叉注意机制来有效地回归 3D 车道线的多项式系数。LATR 基于 CurveFormer 的查询锚建模，构建了车道线感知查询生成器和动态 3D 地面位置嵌入。CurveFormer++提出了一种单阶段 Transformer 检测方法，该方法无需图像特征视图变换，能够直接从透视图像特征中推断出 3D 车道线检测结果。准确估计车道线的 3D 位置需要较强的深度感知，且需要大量计算资源，因此在复杂的实际场景中进行应用时，需要对其泛化性与实时性进行考量。

4.2.4 可行驶区域检测

可行驶区域（Drivable Space 或 Free Space）检测本质上可视为一种语义分割任务，目标是感知车辆周围被占据的空间，从而得到允许行驶的空间（如图 4-5）。可行驶区域检测为感知语义类别不清晰的障碍物而存在，对自动驾驶的行车安全起到“兜底”作用。可行驶区域检测 AI 模型输出栅格（以笛卡尔坐标或极坐标划分的栅格）或点集形式描述的被占用位置，以及置信度、占用类别等其他辅助

汽标委智能网联汽车分委会

信息。以下列举几种实现可行驶区域检测的典型 AI 模型。



(a) 视觉　　(b)激光雷达　　(c) 毫米波雷达

图 4-5 可行驶区域检测结果可视化

基于视觉的可行驶区域检测可以视作对图像作语义分割。图 4-6 给出了两种不同的可行驶区域检测模型[7, 8]。前者将原始的透视图像作为输入，通过卷积神经网络得到分割后的二值图。经过逆透视映射（IPM），该二值图可以转换成对应的鸟瞰图（BEV）。后者则直接使用原始透视图像经 IPM 变换后的 BEV 视图，基于 Encoder-Decoder 模型进行检测，Encoder 利用多层卷积提取高层次语义信息得到特征图（Feature Map），Decoder 通过残差连接与 Encoder 各个分辨率的特征图进行拼接并逐步上采样得到与输入图像相同分辨率的语义分割图。

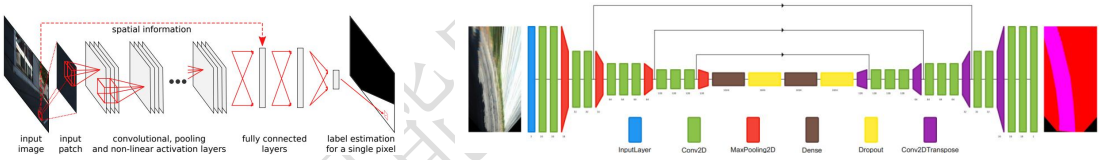


图 4-6 两种不同图像输入的可行驶区域检测模型[7, 8]

基于 3D 点云的语义分割也可实现自动驾驶中可行驶区域的检测，将被语义标签修饰的点云向 BEV 视图中映射即可得到栅格占用情况。图 4-7 给出了一种 BEV 视图下基于点云的可行驶区域检测模型[9]。此外，如 Point Transformer [10] 等许多基于 Transformer 架构的点云语义分割网络也可以用于该任务。点云具有准确的几何信息，且不受光照等影响，可一定程度弥补视觉检测模型失效的情况。然而，LiDAR 采集的点云近密远疏的特性导致远处检测精度受限，而缺乏语义信息也使其难以分辨非机动车道等区域。

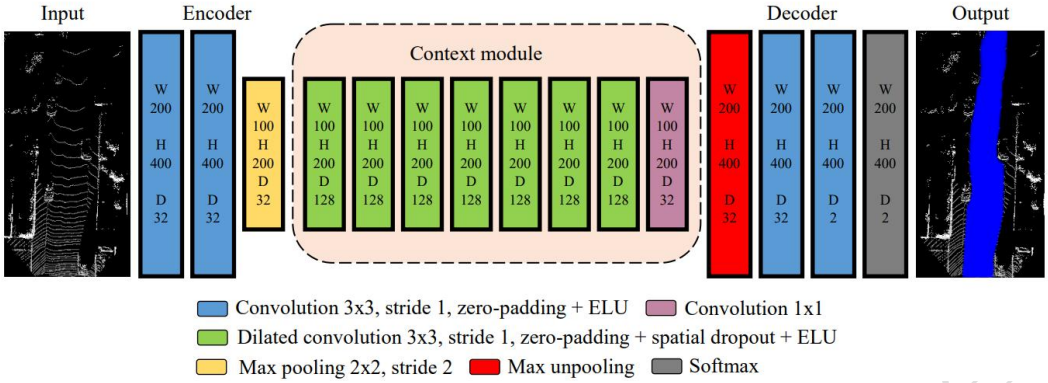


图 4-7 基于点云的可行驶区域检测网络[9]

基于多模态的可行驶区域检测旨在融合图像和点云信息与其各自优势从而提升检测的可靠性。SNE-RoadSeg[11]是一种结合点云和图像信息的可行驶区域检测的模型（如图 4-8）表面法线估计模块使用点云生成法向量图，进而与视觉图像一同经过 Swin Transformer 生成相应的特征图，经特征融合后通过 UNet 网络实现像素空间下的可行区域检测，最后经投影得到 BEV 视图下的占据栅格图。

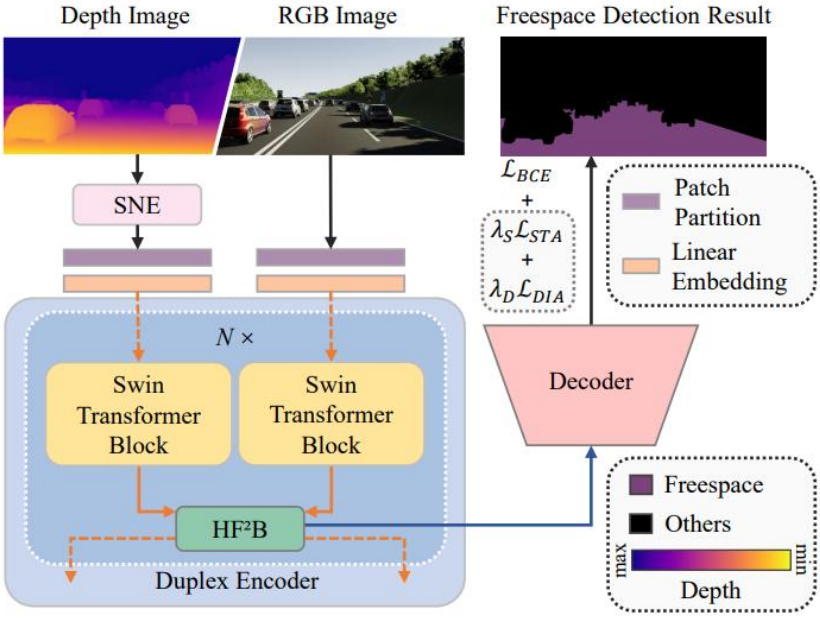


图 4-8 一种融合点云和图像信息的可行驶区域检测网络[11]

BEVFormer[12]是一种支持可行使区域检测与其他感知任务共享一个模型进行训练与推理的整体架构，利用空间注意力机制与时间注意力机制实现多视角图像的空间特征和时序特征的融合，其任务头包括目标检测与地图分割，后者包含

了可行驶区域检测。此外，利用 BEV 空间在多模态信息融合上的特性优势，BEV Fusion[13]等模型基于图像和 3D 点云融合实现了对可行使区域的检测（如图 4-9）。

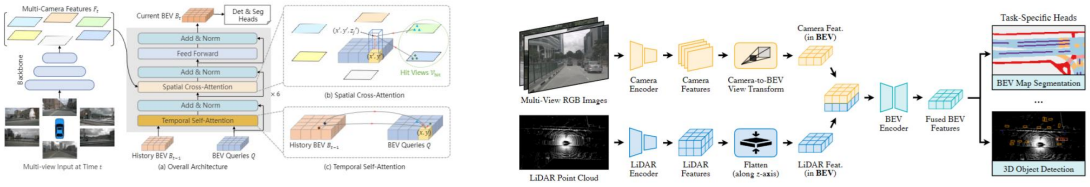


图 4-9 BEVFormer 与 BEV Fusion 架构图[12,13]

SurroundOcc[14]是一种实现 3D 占用语义预测（Occupancy Semantic Prediction）的模型，其能够对三维空间中高度信息的感知，可以视为 BEV 感知的升维，在语义类别中增加可行使区域，就可实现可行驶区域检测（如图 4-10）。模型通过注意力机制融合多个视角下的特征图，并升维至 3D 特征图，进而利用 3D 卷积神经网络实现预测。SurroundOcc 感知性能得益于稠密的 3D 空间内的多尺度监督训练，因此需要高精度标注数据的支撑。

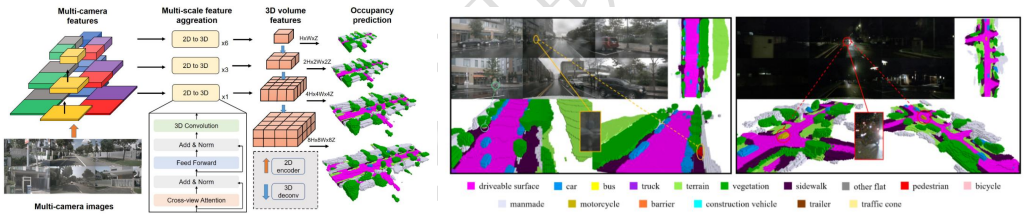


图 4-10 SurroundOcc 模型架构及感知结果[14]

4.2.5 位姿估计

位姿估计旨在基于传感器采集到的数据实时推理自车在环境中的精确位置与姿态，从而为车道级导航等规划决策任务体用有效信息。基于深度学习的 AI 位姿估计方法近年来迅速发展，在惯性导航失效或无法保持精度的特殊场景，AI 位姿估计模型能够通过对视觉图像或 3D 点云等其他感知数据的推理提供高精度的位姿信息。位姿估计 AI 模型的输出主要包含自车在三维空间中实时的相对运动参数（三轴平移量、三轴旋转量）以及对于位姿的不确定性估计（置信度或协方差矩阵）。以下列举几种实现位姿估计的典型 AI 模型。

视觉里程计（Visual Odometry，VO）在自动驾驶定位中通过视觉传感数据实现高精度、低漂移的实时运动估计。DeepVO[15]是一种基于深度学习的单目 VO 模型（如图 4-11），其利用卷积神经网络（CNN）自动提取图像几何特征，并结合长短时记忆网络（LSTM）对时序运动信息进行推理。相较于传统方法，DeepVO 避免了传统 VO 方法中手工特征提取、特征匹配等复杂流程，并有效解决单目视觉的尺度不确定性问题。

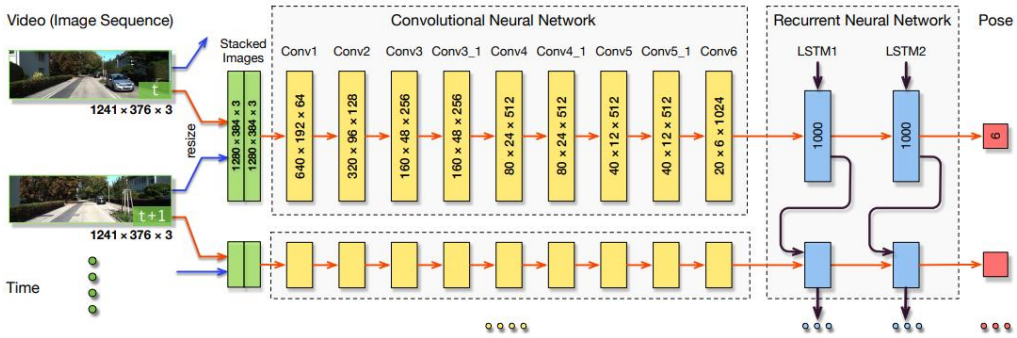


图 4-11 DeepVO 架构图[15]

激光雷达里程计（LiDAR Odometry，LO）则是基于 LiDAR 采集的 3D 点云数据实现位姿估计的技术。DeepLO[16]是一种基于无监督学习的 LO 模型（如图 4-12），其中的特征提取网络从相邻两帧点云数据中提取几何特征，位姿网络（PoseNet）则利用特征差异预测六自由度位姿变换。特别地，DeepLO 通过引入一种类似迭代最近点算法的无监督损失函数，避免了模型训练对真实位姿标签的依赖。

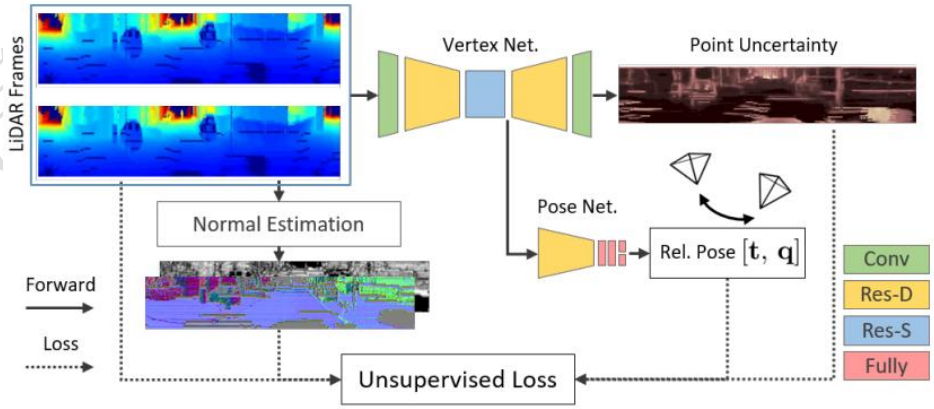


图 4-12 DeepLO 架构图[16]

4.2.6 感知融合

感知融合是一种通用的感知任务，该任务旨在综合利用不同传感器数据或不同 AI 感知模型进行联合感知推理，从而提升准确性、鲁棒性、实时性等感知性能。根据在 AI 感知系统中实现信息融合的不同阶段，感知融合可以细分为前融合、中融合、后融合及不对称融合。以下列举几种实现感知融合的典型 AI 模型。

前融合是一种通过在原始数据级别进行空间对齐和投影直接融合多模态感知数据的方法，目前主要用于对视觉图像数据与 3D 点云数据的融合，旨在同时利用激光雷达的几何精度与图像的语义细节以提升感知性能（如图 4-13）。MVX-Net[17]的核心是通过 PointFusion 和 VoxelFusion 两种方法实现多模态融合：PointFusion 在点云数据中直接嵌入图像语义特征，VoxelFusion 则在体素化处理后的特征空间进行跨模态融合。Pointpainting[18]的核心是将语义信息和点云数据融合，将图像和点云基于传感器参数确定位置关系，进而基于图像生成的语义信息为点云数据赋予语义类别。

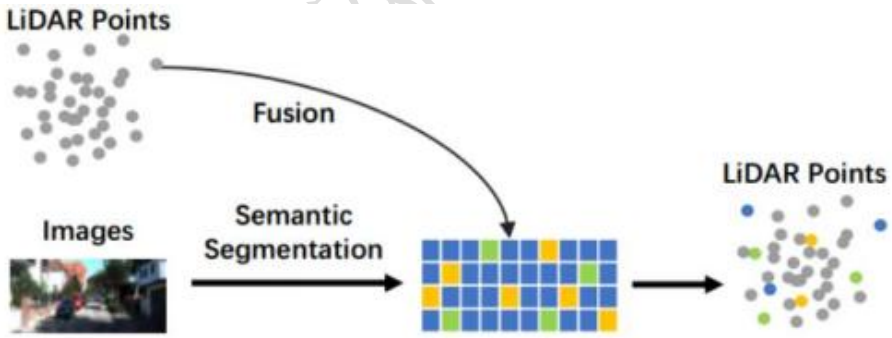


图 4-13 多模态数据的前融合示意图

中融合又称为深度融合，旨在特征层次对感知信息进行融合。基于中融合的 AI 感知模型使用特征提取器分别获取视觉图像与 3D 点云的嵌入表示，进而通过一系列下游模块将两种特征进行融合（如图 4-14）。3D-CVF[19]将图像特征投影成与 BEV 域点云特征具有最大对应性的平滑空间特征图，进而基于空间注意力图利用自适应门控融合网络实现特征融合。ContFuse[20]使用 ResNet18 对图像与点云进行特征提取，然后对图像特征进行多尺度融合并将其投影到 BEV 视图的

点云特征上，由此融合图像语义信息以及点云空间位置信息。

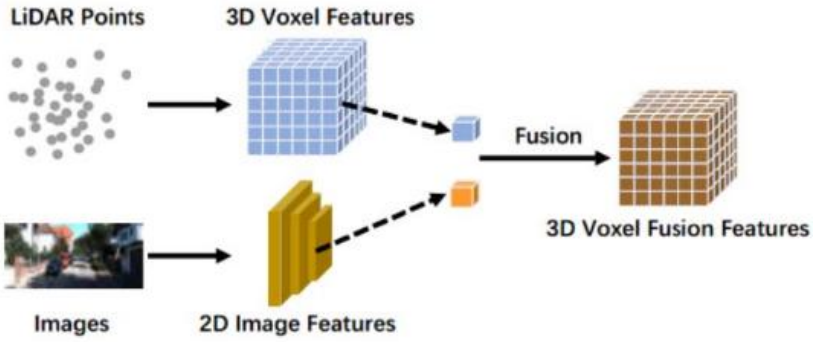


图 4-14 多传感器深度融合示例

后融合又称为目标级融合，旨在对各传感器的独立感知输出进行综合处理以提升感知性能（如图 4-15）。其核心在于通过时空对齐与目标关联算法匹配同一物体在不同感知输出中的表征，最终通过状态估计方法进行融合以输出稳定、全面的环境感知结果。该技术优势在于模块化解耦性强，但需解决多源数据延迟补偿和冗余信息整合等挑战。

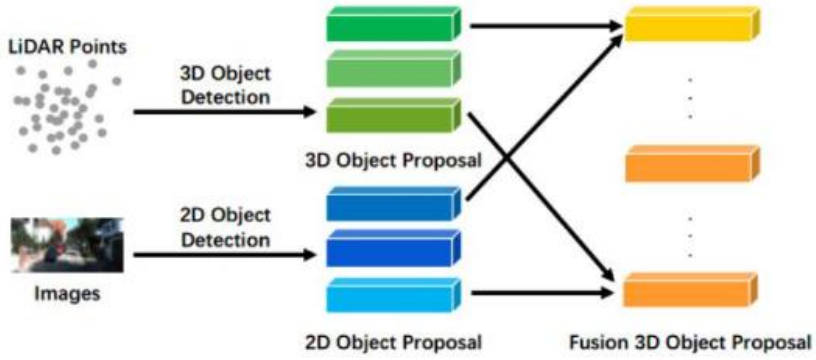


图 4-15 多传感器目标级融合示例

非对称融合指的是融合一个分支的目标级信息和其他分支的数据级或特征级信息。前、中、后三种融合方法将多模态的各个分支平等对待，不对称融合则强调至少有一个分支占据主导地位，其他分支则提供辅助信息预测最终结果（如图 4-16）。

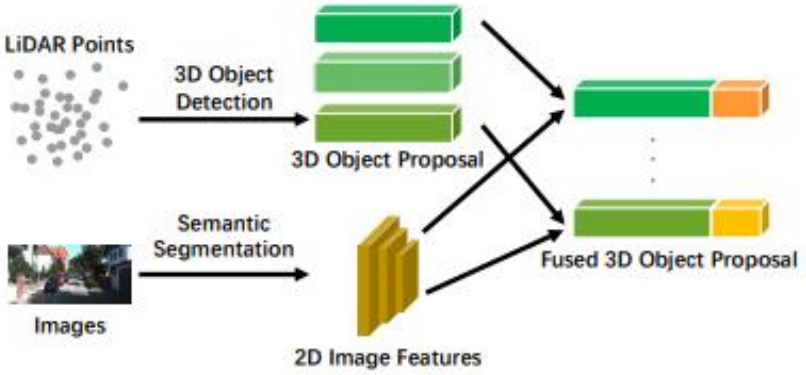


图 4-16 多传感器非对称融合示例

SIFRNet[21]是一种基于非对称融合的感知方法，其以基于 3D 点云的感知网络为主体，在云数据中提取的视椎体中融入图像特征，以增强视椎体所包含的信息量，从而提升其感知性能。

4.3 车用 AI 感知测试

对于软硬件系统的测试旨在分析系统的实现是否能够按照其设计要求正确运行，是系统研发的重要环节。AI 感知模型作为自动驾驶核心组件，对其进行测试是保障自动驾驶系统安全可靠的重要途径。因此，对于 AI 感知模型的测试也是车用 AI 标准化的主要研究对象。本节将重点对给出不同感知模型的关键性能指标，归纳不同方法所涉及的测试用例来源，并梳理与分析测试中所需覆盖的场景，为自动驾驶 AI 感知模型的测试标准化提供参考。具体测试方法及相关细节将会在“第八章：车用 AI 测试与评估标准化研究”章节进行讨论。

4.3.1 关键性能指标

关键性能指标（Key Performance Indicators）能够以定量的形式客观反映车用 AI 感知模型的性能。通过测试得到模型的关键性能指标，并对其进行分析对比是评价模型优劣、发现模型潜在缺陷的重要手段。针对车用 AI 感知模型的关键性能指标主要包含准确性与实时性两个维度：准确性维度的关键指标直接反映了模型对环境感知与理解的可靠程度，显示模型误检、漏检的风险；实时性维度的关键指标则直接反映了模型的感知效率，显示延迟导致系统失效的风险。二者

协同衡量了车用 AI 感知模型在复杂环境中的稳定感知能力，更为模型与系统迭代提供了优化方向。

4.3.1.1 准确性维度

针对感知模型的准确性，表 4-1 中给出了评价目标检测任务 AI 模型的关键性能指标并根据实际测试中的重要程度进行了分类。表 4-2 中给出了评价其他感知任务 AI 模型的关键性能指标。

表 4-1 目标检测模型在准确性维度的关键性能指标

优先级	指标	基本定义	指标含义
高	混淆矩阵	行代表预测值，列代表真实值，矩阵元素为测试样本数。	基于混淆矩阵可以计算其他评估指标，如准确率、召回率、F1 分数等。
	准确率 Accuracy / Acc	定义为正确预测数除以总测试样本数。	通常 Acc 越高越好，但样本不平衡时存在局限。mAcc 是对各分类类别准确率的平均，用于评估多分类器或检测器的性能。
	精确率 Precision / P	定义为预测结果正样本中真实正样本的比例。	用于衡量模型预测结果的可信度，其数值精确率越高，预测中的误报率越低。
	召回率 Recall / R	定义为预测结果中正确识别的实际正样本占全部正样本的比例。	值越高漏检越少。精确率 P 和召回率 R 常需平衡，提高一方可能降低另一方，因需在减少误报和漏报间权衡。
	P-R 曲线	曲线横轴为召回率，纵轴为精确率，通过改变置信度阈值绘制出两者的权衡关系。	P-R 曲线展示了精确率随召回率的变化趋势，帮助评估模型在不同召回率下的精确率表现。
	漏检率	定义为未检测到的真实目标占实际正样本的比例。	漏检可能导致严重事故，需与召回率联合优化。
	误检率	定义为错误预测为目标的样本占实际负样本的比例。	直接关联系统安全性，需在复杂场景中严格控制。
	F1 分数 F1-score	定义为精确率和召回率的调和平均数。	F1 分数综合评估模型性能，高分表示精确率和召回率良好。
	平均精确率 AP	定义为 P-R 曲线下的面积，衡量准确率与召回率的权衡，反映目标检测结	数值越高表示模型在不同召回率下的平均精确率越好。AP 综合了精确率和召

汽标委智能网联汽车分委会

		果与真实值的平均匹配程度。	回率。mAP 是对多类别模型 AP 的平均，用于评估多分类器或检测器的整体性能。
	交并比 IoU	定义为预测框与真实框的交集面积与并集面积的比率，用于衡量检测结果的重叠程度。	IoU 衡量检测结果与真实标签的重叠程度，值越高模型性能越好。mIoU 是对多类别 IoU 的平均，用于评估多目标检测模型的整体性能。
中	平均召回率 AR	定义为 IoU 在[0.5,1.0]范围内所有召回率的平均，通过面积积分计算平均召回率，评估检测框的有效性。	AR 衡量目标检测器对某一类别检测框的准确性，值越高表示检测越准确。mAR 是对多个类别 AR 的平均，用于评估多分类器或检测器的整体性能。
	平均属性精度 AAE	AEE=1-Acc，AAE 定义为所有属性 AEE 的平均值。	该指标用于评估图像分类模型对图像属性的预测准确性，适用于属性为离散值的情况。
	平均平移误差 ATE	定义为检测框与真实框中心点的欧式距离偏移，用于评估目标定位的准确性。	用于评估检测模型对目标位置的预测准确性。
	平均缩放误差 ASE	ASE=1-IoU，IoU 是 3D 目标检测中检测框和真值框在方向与位置对齐后的三维交并比。	对齐后的检测框与真值框的非重叠部分比例，用以评估模型预测与真实结果的重叠误差。
低	平均速度误差 AVE	定义为预测物体速度与真实物体速度之间的二维 L2 范数差。	/
	平均角度误差 AOE	定义为预测航向角与真实航向角之间最小偏差，计算各类别的平均值。	/
	检测得分 NDS	将多种错误类型（如 ATE、ASE、AOE、AVE、AAE）整合为单一标量分数。	旨在简化评估流程，平衡检测性能与质量，使评估体系更全面、准确。
	平均朝向精度 AP with Heading	对 AP 的扩展，考虑目标检测框朝向。	衡量模型目标预测的平均精度和运动航向精度。

汽标委智能网联汽车分委会

表 4-2 其他 AI 感知模型在准确性维度的关键性能指标

任务	指标名称	基本定义	指标含义
语义分割	交并比 IoU	参见目标检测评价指标中的 IoU 部分。	用于衡量预测的语义分割区域与真实区域的重叠程度，评估模型的准确性。
	像素准确率 PA	定义为预测正确的像素点数量除以总像素点数量占总像素的比例。	数值越接近 1 表示分割性能越好。适用于多种场景，但不适用于类别不平衡或需评估区域重叠程度的场景。
	骰子系数 Dice	定义为两倍的交集除以像素和。	衡量预测结果与真实标签的重叠程度，越接近 1 越好。适用于不平衡数据和小目标分割。
实例分割	Jaccard 系数（又名 IoU）	参见目标检测指标中的 IoU 部分。	用于衡量预测实例分割区域与真实实例分割区域的重叠程度。
	平均精确率均值 mAP	参见目标检测指标中的 AP 部分。	衡量目标定位和分类的准确性，而不直接评估分割 mask 的质量。
	对象级精度	定义为正确分割的物体实例数与总物体实例数的比率。	用于评估每个物体实例的分割质量，越接近 1 表示算法性能越好。
	边界漂移误差（BDE）	定义为边界像素偏离真实边界的平均距离。	用于评估物体边界的分割效果，数值越小表示算法性能越好，边界的预测越准确。
车道线检测	点级别准确率/召回率	准确率定义为模型正确预测为车道线的点占所有预测为车道线点数的比例；召回率定义为正确预测为车道线的点占所有真实存在的车道线点数的比例。	/
	交并比 IoU	参见目标检测指标中的 IoU 部分。	用于评估图像中预测车道线与真值车道线之间的匹配程度。
	Dice 系数	$\mathrm{D}=\frac{2\sum_{i=1}^{N}p_i g_i}{\sum_{i=1}^{N}p_i^{\ 2}+\sum_{i=1}^{N}g_i^{\ 2}}$ 其中 p_i 为第 i 个像素的预测结果（1 为目标，0 为背景），g_i 为当前像素的真值（1 为目标，0 为背景），N 为像素个数。	用于衡量图像中车道线像素点的预测正确性，值越高模型性能越好。Dice 系数和线级别准确率、召回率可用于评估基于分割的 2D 车道线性能。若车道线非单像素级别，Dice 系数

汽标委智能网联汽车分委会

			更适用；否则，线级别指标更优。
位姿估计	平均位置误差 APE	$\mathrm{ATE} = \frac{1}{N}\sum_{i=1}^{N}\lVert \vec{x}_i^{\mathrm{est}} - \vec{x}_i^{\mathrm{true}} \rVert_2$	其中，N 是时间序列长度；$\vec{x}_i^{\mathrm{est}}$ 和 $\vec{x}_i^{\mathrm{true}}$ 分别是定位算法在第 i 个时间点输出的位置坐标及其对应的真值坐标。
	平均角度误差 AAE	$\mathrm{AAE} = \frac{1}{N}\sum_{i=1}^{N}\lVert \theta_i^{\mathrm{est}} - \theta_i^{\mathrm{true}} \rVert_2$	其中，N 是时间序列长度；θ_i^{est} 和 θ_i^{true} 分别是定位算法在第 i 个时间点输出的航向角及其对应的真值航向角。
	相对位置偏移	$\mathrm{RPO} = \frac{1}{N}\sum_{i=1}^{N}\lVert \Delta P_i^{\mathrm{est}} - \Delta P_i^{\mathrm{true}} \rVert_2$	其中，N 是时间序列长度；$\Delta P_i^{\mathrm{est}}$ 和 $\Delta P_i^{\mathrm{true}}$ 分别是定位算法在第 i 个时间点输出的相对坐标系下的位移量与及其对应的真实位移量。
	相对角度偏移	$\mathrm{RAO} = \frac{1}{N}\sum_{i=1}^{N}\lVert \Delta Q_i^{\mathrm{est}} - \Delta Q_i^{\mathrm{true}} \rVert_2$	其中，N 是时间序列长度；$\Delta Q_i^{\mathrm{est}}$ 和 $\Delta Q_i^{\mathrm{true}}$ 分别是定位算法在第 i 个时间点输出的相对坐标系下的角度变化量与及其对应的真实角度变化量。
	定位精度（置信度）	1σ、2σ和 3σ表示定位误差在 68.27%、95.45%和 99.73%的概率范围内，用于描述数据分布的离散程度。	定位精度指标用于衡量定位技术的准确性和可靠性，涉及 GPS、组合定位和激光定位等系统。

此外，AI 感知模型的鲁棒性可视为其在扰动条件下的准确性体现，因此可以在给定类型/强度的扰动场景中，对上述准确性维度指标进行测评，从而对模型的鲁棒性进行分析。

4.3.1.2 实时性维度及其他

车用 AI 感知模型的实时性主要通过以下两个关键性能指标进行测评：

推理延迟（Inference Latency）：AI 感知模型执行单次推理（处理一帧数据）所需的时间开销，该指标直接决定感知模型的响应速度。

帧率（FPS）：AI 感知模型每秒钟所处理的数据帧数，该指标直接影响后续

模块接收感知结果的更新频率。

此外，由于车端硬件算力限制，车用 AI 感知模型规模及其对计算资源的占用也会对部署后模型的实际性能造成巨大影响，因此需要在进行模型测试评估时得到充分分析与考量，其中涉及的关键性能指标如表 4-3 所示。

表 4-3 AI 感知模型的其他关键性能指标

指标	基本定义	指标含义
硬件算力需求	部署模型所需的最低硬件算力，Required TOPS = FPS ×单帧浮点运算次数。	示例：若模型单帧需 50 GFLOPs，目标 FPS=30 → 需 1.5 TOPS 算力。
浮点运算次数 FLOPs	处理单帧输入所需的浮点运算总次数。	衡量模型计算复杂度，值越低说明计算效率越高，适用于横向对比不同模型的计算需求。
参数量	模型中可训练参数的总数量。	参数越多，模型存储和内存占用越高。
吞吐量	指单位时间内处理的输入数据总量。	强调系统并行处理能力，在批量处理场景下，吞吐量比单帧延迟更重要。
乘加操作数 MACs	单次推理中所有乘法-累加操作的总次数。	反映硬件实际执行的计算量。
内存占用	模型运行时占用的显存（GPU）或内存（CPU）大小。	嵌入式设备（如 Jetson AGX）的显存有限，高内存占用可能导致系统卡顿或崩溃。
峰值内存带宽	模型运行期间对内存带宽的最大需求。	内存带宽可能成为计算瓶颈。
模型压缩率	压缩后模型大小与原模型大小的比值（例如剪枝、量化后的压缩率）。	量化（如 FP32→INT8）可减少模型大小和计算量。压缩率越高，部署效率越高，但可能损失精度。
内存占用	模型运行时占用的内存或显存（GPU）大小。	嵌入式设备（如 Jetson AGX）的显存有限，高内存占用可能导致系统卡顿或崩溃。
峰值内存带宽	模型运行期间对内存带宽的最大需求。	内存带宽可能成为计算瓶颈。
模型压缩率	压缩后模型大小与原模型大小的比值（例如剪枝、量化后的压缩率）	量化（如 FP32→INT8）可减少模型大小和计算量。压缩率越高，部署效率越高，但可能损失精度。

汽标委智能网联车分委会

4.3.2 车用 AI 感知测试场景

4.3.2.1 测试场景的重要性

与预测、决策模块的测试中主要关场景中交通参与者行为逻辑与交互关系不同，车用 AI 感知模型的测试旨在验证其对场景复杂语义信息的精准理解能力。由于车辆行驶在开放的真实环境中，车用 AI 感知模型需要应对各种复杂场景及其中多维度的干扰因素。因此，尽管车用 AI 感知模型的测试方法与数据来源是多样的，但都需要尽可能广泛、全面地覆盖模块设计支持的不同的场景类型，并在扰动维度上覆盖关键影响因素的变化，这对系统性验证车用 AI 感知模型的关键性能、泛化性、鲁棒性具有重要意义。

4.3.2.2 基础场景

自动驾驶车辆的行驶环境有时存在巨大差异，而车用 AI 感知模型的泛化性要求其能够应对满足其设计运行范围的各种不同场景，因此场景选择与覆盖对车用 AI 感知测试十分关键。从交通参与要素和车辆行驶速度两个维度考虑，我们可以将其划分为多个类别的基础场景，具体包括高速场景、城市道路场景、泊车场景、特殊封闭场景（如图 4-17）。

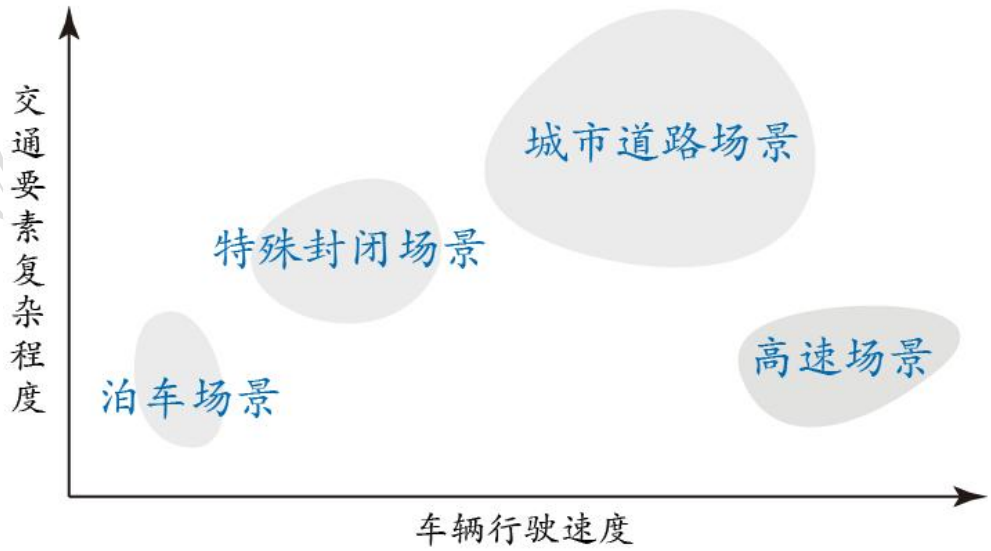


图 4-17 车用 AI 测试基础场景分类

(1) 高速场景

高速场景是指高速公路、城市快速路等互通枢纽上执行自动驾驶的场景，此类场景具有自车与其他交通车辆车速较快，但交通要素相对简单的特性。由于车速较快，此类场景对自动驾驶系统中 AI 感知模型的实时性及感知范围（距离）提出了更高的要求，同时对于感知融合的时空对齐精度也产生了更大的挑战。此外，尽管高速场景中的交通流复杂性相对较低，但其中存在一些该场景特有的交通元素如特殊交通标志（如出口预告、最低限速、禁止超车与解除、里程定位、服务区等）、特殊标线（如匝道隔离线、车距确认线等）等，这些元素也需要在 AI 感知模型的测试中得到覆盖。

(2) 城市道路场景

城市道路场景是在城市中除了高速公路和快速路外的道路中执行自动驾驶的场景，此类场景虽然车速低于高速场景，但其复杂程度与差异性远高于高速场景，这对自动驾驶中 AI 感知模型的泛化性与鲁棒性提出了极高的要求。一方面，城市道路场景中存在多样的交通元素，除了行人、各类机动车与非机动车等交通参与者，还存在各种复杂的交通信号与标志及道路标线，对 AI 感知模型的测试需要全面地覆盖上述内容。另一方面，城市道路光源复杂（特别是在夜间）、场景元素空间关系复杂，这导致光照、遮挡等因素对 AI 感知的干扰在城市道路场景尤为突出，因此需要在测试中重点关注。

(3) 泊车场景

泊车场景属于城市场景中一类相对独立的场景，其特点在于需要车用 AI 感知模型能够对泊车场环境进行识别与理解，特别是对泊车位与障碍物的三维空间位置进行精准感知。除了车位线，车用 AI 感知模型所涉及的识别目标还包括细杆、低矮圆柱等泊车场景特有的障碍物。此外，泊车场景中的空间相对狭小，因此需要车用 AI 感知具有对近距离障碍物及空间的感知能力。同时，对于地下停车场，车用 AI 感知需具备在昏暗、车灯光源等特殊光照条件的感知能力。上述因素均应该被纳入车用 AI 感知模型对于泊车场景测试的考虑。

汽标委智能网联汽车分委会

(4) 特殊封闭场景

特殊封闭场景是区别于高速和城市道路等半封闭及开放道路的低速自动驾驶场景，一般包括矿区自动驾驶、港口自动驾驶等特殊细分领域。在矿区场景中，AI 感知模型需克服扬尘干扰、地面颠簸等挑战，实现不规则、小尺寸障碍物的稳定感知，因此该场景的车用 AI 感知模型测试中需要充分考虑环境能见度、图像抖动模糊、识别目标尺度等复杂因素对感知的影响；港口场景则存在多变的天气条件与逆光、夜间等光照条件，并需要精准识别密集分布的锥桶、隔离墩等障碍物，及对挂车角度、桥吊位置等场景特有目标的空间信息进行准确感知，因此该场景的车用 AI 感知模型测试中需要充分覆盖港口的特殊识别对象，并考虑密集车挡、光照等因素对感知的影响。

4.3.3.2 鲁棒性与干扰因素

鲁棒性是对 AI 模型能够抵御外界扰动、稳定输出正确结果的重要能力要求。比起基于图像处理和规则判定的传统感知方法，基于深度学习等 AI 模型的感知方法具有更高的能力上限（如一个参数足够的深度神经网络理论上能够逼近任意的非线性函数），其计算图像、点云高维数据获得感知结果的能力大大超越了传统方法，而 AI 感知模型在鲁棒性方面也表现出更加优异的性能。尽管如此，仍然存在诸多能够对 AI 感知性能产生严重影响的因素，大量研究表明光照[21-25]、天气[26-31]、遮挡[32-37]、噪声[38-41]、模糊[42-48]、几何[49,50]、色彩[51,52]等因素会对基于视觉成像（即图像）的感知或基于距离成像（即点云）的 AI 感知结果造成较大的影响，而许多针对这些影响因素的数据增强、模型构建与训练技术也是 AI 感知的重要研究主题。

此外，与传统基于规则的感知算法不同，自动驾驶系统中的感知任务通常基于深度神经网络等数据驱动的机器学习 AI 模型实现。由于模型自身结构的数学特性、训练数据的多样性与平衡性、训练的充分性等多方面因素，AI 感知模型往往存在鲁棒性的潜在缺陷[53,54]。尽管 AI 感知算法在上述影响因素下的总体性能大大优于传统方法，但由于 AI 感知算法模型的黑盒性，这些因素对模块的影响存在更多的不确定性。换言之，在什么具体条件下错误的发生更加难以预料，

汽标委智能网联汽车分委会

而这些风险更加难以被发现。故当车辆运行在复杂开放的现实环境中时，AI 感知模型易受上述因素影响发生错误，如识别模块的错检、漏检、误检，从而导致后续预测、决策模块接收错误的感知结果，最终影响整个高级驾驶辅助系统或自动驾驶系统的安全性。

因此，在对车用 AI 感知模型进行构建与测试的时候，我们需要充分考虑（或部分考虑所涉及的）易对车用 AI 感知模型的运行结果造成影响的重要因素。特别地，为了保证测试的充分与全面，我们应基于不同的基础场景，通过现实采集、数据处理、仿真模拟、对抗生成等技术手段产生包含下列因素的差异化场景，从而有效发现被测车用 AI 感知模型的鲁棒性缺陷。

(1) 光照：良好光源、弱光源、强光源、复杂光源

光照是影响车用 AI 感知模型性能的重要因素，大致可分为以下四类。良好光源指光源充足、稳定、简单的情况（如天气良好的白昼、照明合适的夜晚或室内等），是车用 AI 感知模型最为适宜的运行光照条件；弱光源（如夜晚、黎明、黄昏等）会导致视觉传感器所采集的图像明暗细节信息的丢失，同时也可能产生由视觉传感器感光度升高引起的的影像噪声；此外，强光源（如太阳光或远光灯直射产生的逆光环境）以及复杂光源（如道路灯光、车辆灯光、频闪、反光）可能导致眩光、亮斑、鬼影等情况，或因感光范围受限、曝光不准确导致极端明暗差异下的整体或局部图像信息丢失。

图 4-18 不同类型光照：良好光源、弱光源、强光源、复杂光源

(2) 天气：晴、阴、雨、雪、雾、霾

不同天气（特别是极端天气条件）除了会引起环境光照强度的变化外，还会对车用 AI 感知模型输入数据产生其他严重的影响。一方面，由雨、雪、雾、霾、扬尘等引起的光线折射、散射或遮挡会导致的能见度下降，从而使得视觉传感器无法清晰捕捉环境细节；另一方面，雨、雪与视觉传感器镜头的接触及低温下的

再结晶会形成透明至半透明的阻塞物，从而导致所采集图像中物体或其边缘失真与不完整。此外，雨雪、沙尘导致对信号的阻挡、散射、衰减等效应也会一定程度上导致毫米波雷达与激光雷达探测范围减小以及探测精度的下降，进而影响感知模型的性能。

图 4-19 不同类型天气：大雨、大雪、雾霾、雨雪引起镜头阻塞

(3) 遮挡：类内遮挡、类间遮挡

当物体被部分遮挡时，AI 感知模型无法通过传感器获得物体的全部特征。然而，车用 AI 感知模型需要面对包含大量复杂物体的实际场景，部分遮挡的情形无法回避，故应该具备在一定程度遮挡条件下准确感知环境的能力。根据遮挡物的类别，遮挡大致可分为类内遮挡与类间遮挡。类内遮挡，产生于场景中的同类物体，也称为密集遮挡（如行人之间的遮挡、车辆之间的遮挡）；类间遮挡，不同类别物体之间的遮挡现象（如树木对交通指示牌的遮挡、地面障碍物对车道线的遮挡）。此外，对于上述两类遮挡，遮挡位置与遮挡物形态也是影响感知性能的重要因素，应给与充分的考虑。

图 4-20 不同类型遮挡示意图：类内遮挡（车辆）、类间遮挡（绿化-车辆）、类间遮挡（树木-信号灯）、类间遮挡（涂鸦-路锥）

(4) 噪声：高斯、泊松、椒盐、伽马、瑞利

噪声的存在是降低图像质量进而影响车用 AI 感知模型性能的重要干扰因素。噪声可能产生于信号采集、信号压缩、信号传输等各个感知过程中，而不同的干扰来源也会导致不同的噪声类型。高斯噪声，通常由传感器采集与传输信号过程中的随机误差引起；泊松噪声（散粒噪声），由图像传感器的不均匀性或光照不

均造成的颗粒状噪声；椒盐噪声，通常由图像压缩引起，表现为图像中随机分布的黑点和白点；伽马噪声，通常由信道特性随机变化引起；瑞利噪声，通常出现在如激光雷达等距离成像采集数据中。

图 4-21 不同类型噪声：高斯噪声、泊松噪声、椒盐噪声、伽马噪声

(5) 模糊：低通滤波、动态模糊、散焦模糊

模糊的本质是图像信息丢失，具体表现为图像纹理细节或物体结构的不清晰，极大影响了车用 AI 感知模型对于图像内容的辨识。这种信息丢失可能是由于多种原因造成的，故在实际应用中须考虑以下几种重要的模糊类型。低通滤波，通常同于图像去噪，通过过滤图像高频部分使其变得平滑，均值模糊、中数模糊、高斯模糊等均属于此类；动态模糊，也称运动模糊，是视觉传感器在成像过程中由于车辆运动、场景变化或焦距改变所造成的不同空间位置信息的混叠，通常涉及线性、旋转、缩放三类；散焦模糊，也称光学模糊，通常是由视觉传感器对焦不准确及透镜焦距限制产生的图像不清晰。

图 4-22 不同类型模糊：清晰图像、低通滤波（高斯模糊）、动态模糊（缩放）、散焦模糊

(6) 几何：仿射、视角、扭曲

车辆与感知目标物位姿的不同、传感器部署位置的不同、传感器的物理特性的不同等诸多因素都可能导致车用 AI 感知模型的输入发生不同类型的几何变换。因此，对于输入一定范围内的几何变换，如仿射（平移、缩放、旋转、错切、镜像）、视角、扭曲（球面、切变等）等，车用 AI 感知模型需要具备保持计算输出正确感知结果的能力。

汽标委智能网联标准分委会

图 4-23 不同类型几何变换：原始图像、仿射变换、视角变换、扭曲变换

(7) 色彩：亮度、曝光度、对比度、饱和度、色温、色调、暗角

图像的色彩变化通常是由外部环境光照条件变化及视觉传感器参数变化等因素所引起的，这些色彩变化往往会对 AI 感知模型的性能产生影响。而车用 AI 感知模型运行在复杂开放环境中，将面临多样化的外部环境光照条件及视觉传感器工作状态，故需要具备应对亮度、曝光度、对比度、饱和度、色温、色调、暗角等一定范围内的色彩变换。

图 4-24 不同类型色彩变化：原始图像、亮度、曝光度、对比度（第一行）
饱和度、色温、色调、暗角（第二行）

(8) 客体：语义多样性

除了上述影响因素外，感知客体（即环境中感知模型需要识别与理解的目标物）本身的复杂程度同样是影响 AI 感知模型性能的重要因素，因此在构建场景时需要其语义的多样性给予充分的考虑。例如，应考虑感知客体在场景中的位置、姿态、密度等特性。对于不存在特定规格的感知客体，还应该注意个体差异化，充分考虑其造型、尺寸、色彩等特性。

4.4 车用 AI 感知标准化需求与挑战

4.4.1 标准化需求分析

数据作为车用 AI 感知模型的底层支撑，其标准化直接影响算法的训练效率和泛化能力。通过统一数据格式（如标注规则、传感器参数）、存储结构（如多模态数据同步机制）和共享协议，可减少跨平台数据融合的复杂度，提升数据集的复用率，避免因数据歧义导致的模型偏差，加速多场景泛化能力建设。

接口是访问车用 AI 模型的数据通道，明确模型输入、输出的接口规范（如数据类型、格式）可增强不同厂商模块间的兼容性，减少因接口变化而导致的系统风险，简化后续决策与控制系统对数据的解析流程，降低数据适配与集成开发成本，并为模型的标准化测试提供有力基础。

测试是验证车用 AI 模型性能的重要手段，关键评估指标（如精度、效率）的标准化能保障分析结果有效性，推动横向对比，促进行业良性竞争。同时，构建统一的测试场景/用例库及验证流程可有效提升测试效率与覆盖率，发现模型的潜在缺陷，缩短模型优化与产品落地周期。

4.4.2 标准化的挑战

数据标准化的核心挑战在于如何实现多源异构数据的统一整合。由于不同数据来源的传感器参数（如激光雷达点云密度、摄像头分辨率）、数据分布和质量存在显著差异，若不能有效解决数据集融合问题，这些数据差异将直接影响模型训练的性能表现。

接口标准化的主要难点在于平衡动态适配与兼容性需求。一方面，不同系统架构导致模块间传递的信息内容存在差异，使得数据接口格式难以完全统一；另一方面，感知模型需要适配多种算力平台，这导致其输入输出接口在带宽、时序等参数上存在固有差异。

测试标准化的最大挑战来自场景覆盖的完备性和指标体系的科学性。首先，开放的系统运行环境包含大量复杂场景和干扰因素，使得全面覆盖极端测试场景

智能网联汽车分委会

变得困难；其次，性能指标不仅与具体任务和场景紧密相关，各指标间（如精度与实时性）还存在相互制约的关系，导致合理阈值的确定颇具难度。

4.5 总结与展望

以深度学习为核心算法的车用 AI 感知技术已在智能驾驶系统构建中得到广泛应用，并在目标检测等关键感知任务中展现出显著性能优势。然而，该技术发展仍面临诸多挑战，包括性能评估体系缺乏统一标准、数据质量规范不明确等问题。目前，尽管我国在人工智能标准化建设方面已取得阶段性成果，但针对车用 AI 感知技术的专项标准体系仍有待完善。

未来车用 AI 感知技术的标准化研究应致力于构建涵盖数据采集、算法开发、测试验证的全流程标准体系，其中将优先聚焦于模型与系统评估的标准研究，着力突破复杂环境下的性能评估技术瓶颈，建立全面覆盖应用场景的测试框架及数据集，并制定适应不同干扰条件的鲁棒性测试方法与指标，从而为智能驾驶技术的安全商用和规模化部署提供标准化支撑。

汽标委智能网联汽车分委会

第五章 车用 AI 自动驾驶预测与决策标准化研究

5.1 车用 AI 预测与决策标准化研究概述

5.1.1 自动驾驶预测与决策功能简介

车用 AI 预测与决策系统是自动驾驶算法架构中的核心功能模块之一。该系统通过融合感知系统获取的实时环境数据和地图系统提供的高精度道路先验信息，运用先进的预测算法对周围交通参与者的未来运动轨迹进行多维度预测，包括行人横穿概率、相邻车辆变道意图、交叉路口冲突点等。决策系统基于预测结果并结合车辆运动学约束、交通规则、安全约束和行驶舒适度等因素进行全局参考路径生成和局部避障轨迹生成，最终输出符合车辆动力学特性的动作决策以及最优轨迹路径规划，为底层控制系统提供精确的行驶指令。

5.1.2 预测与决策标准化研究目的与意义

目前车用 Al 预测与决策测试结果的全面性、真实性和可重复性并不明确，数据需求前后关联也并能不清晰区分也并不明确。车用 Al 预测与决策的标准化研究将建立针对相关功能的数据收集和处理统一标准，以确保不同系统之间的数据可比性和一致性。在预测算法的准确性、鲁棒性和实时性等方面，制定统一的评估标准和测试方法。在决策算法的安全性、合理性和可靠性等方面，制定统一的决策规则和约束条件。同时构建不同类型的交通场景、设置不同的交通条件和障碍物等，以测试自动驾驶汽车在不同场景下的表现，通过场景测试来验证预测与决策算法的有效性。车用 Al 预测与决策的标准化可以使行业在预测与决策相关术语与定义上达成相关共识，明确车用 Al 预测与决策的评价指标能为行业提出应遵循的基本原则。

汽标委智能网联汽车分委会

5.1.3 预测与决策标准化研究现状

目前车用 AI 决策与预测相关的标准法规较少，相关度较高的主要分布在与自动驾驶场景设计高度相关的标准。这些与场景相关的法规规定了自动驾驶系统测试场景分类的相关标准和评价体系，明确了自动驾驶系统在不同道路、环境、交通状况下的测试需求。

(1) ISO 34501：2022-道路车辆-自动驾驶系统测试场景-词汇

定义了 ADS 自动驾驶系统（automated driving systems）测试场景中的术语，适用于 ISO/SAE PAS 22736 中定义的 L3 级及以上 ADS。

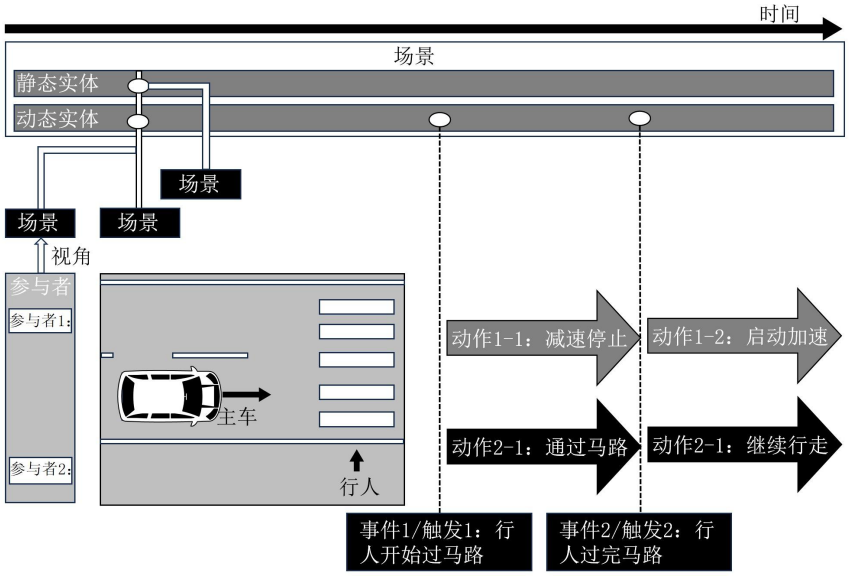


图 5-1 场景相关术语关系示例[55]

图 5-1 说明了场景中几个相关术语之间的关系，图中两个参与者，参与者 1 是受试车辆，参与者 2 是即将过街的行人。该场景的目的是指定受试车辆（参与者 1）响应行人给定动作的行为（参与者 2）。此场景中有两个触发器，触发 1 表示场景的开始，对应于行人开始横穿街道的时刻。同时，受试车辆以预设速度行驶，以进行测试。在这种情况下，从目标车辆的角度来看，行人出现在街道上的场景变成了场景。ADS 控制下目标车辆的预期行为是减速停车，直到行人安全穿过道路。行人到达对面路边的事件可能会触发场景中的下一系列行动。

(2) ISO 34502：2022-道路车辆-自动驾驶系统测试场景-基于场景的安全评估框架

该框架阐述了在产品开发过程中应用的基于场景的安全评估过程，为自动驾驶系统（ADS）的基于情景的安全评估框架提供指导。交通相关的关键场景分为一般车辆场景（包括四轮车和摩托车）、摩托车特定场景和其他交通参与者或障碍物特定场景。其中，针对一般车辆交通相关的关键场景，通过系统地分析道路结构、本车车辆行为、周围车辆位置和行动的不同组合，生成了一个总共包含24 种可能的组合矩阵如图 5-2 所示。道路类别被分为三类：主路、汇入和汇出。本车车辆行为分为车道保持和车道变换。周围车辆的位置被定义为目标车辆周围八个方向的相邻位置，可能的动作分为四组：切入、切出、加速和减速。从安全评估的角度来看，可以通过关注目标参与者可能阻碍目标车辆轨迹的动作来最大限度地减少评估测试的次数。

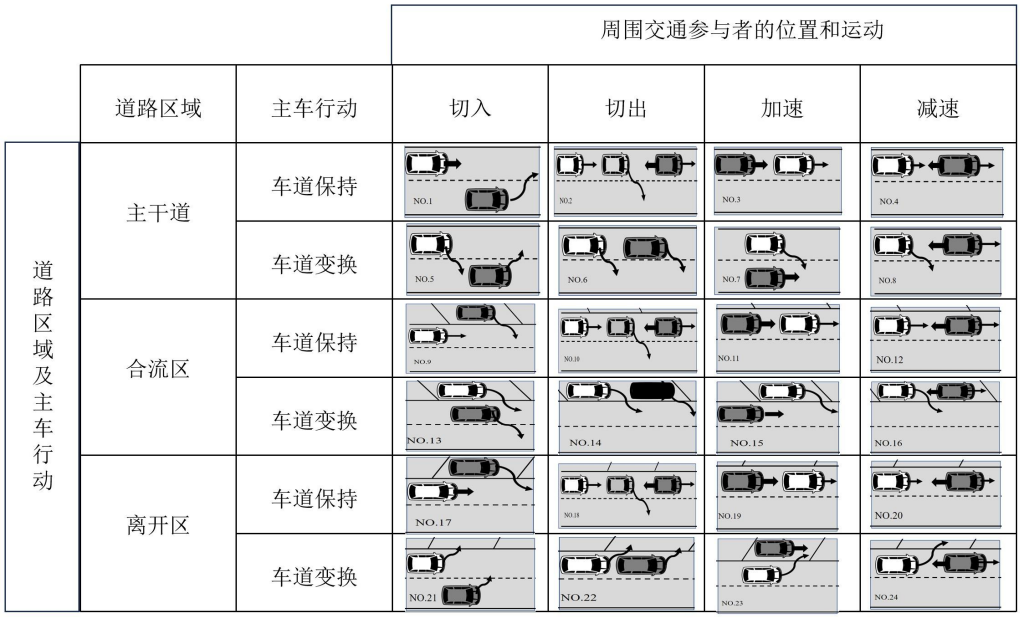


图 5-2 一般车辆的具体交通关键情景[56]

(3) ISO 34503:2023-道路车辆-自动驾驶系统测试场景-操作设计领域规范

规定了分级分类法的要求，用于指定能够定义自动驾驶系统（ADS）的运行设计域（ODD）的运行条件。

ODD 的实例化以及期望的行为以及根据事件和触发器的交通参与者行为的描述将为 ADS 提供场景定义然后，可以将这样的定性场景进一步细化为功能场景、抽象场景,逻辑场景和具体场景，以创建场景库，具体关系如图 5-3 所示。

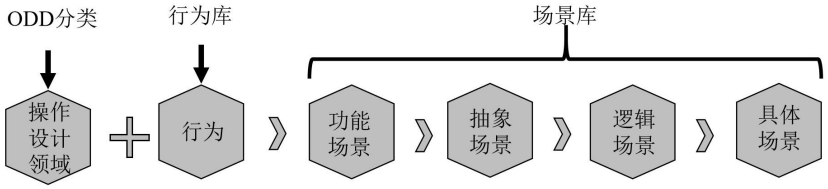


图 5-3 ODD、行为和场景之间的事例关系[57]

(4) ISO 34504:2024-道路车辆-自动驾驶系统测试场景-场景分类

提出一种对场景进行分类的方法,提供了一种协调场景分类方式的方法。

为了实现场景分类，“标签”被定义为元属性，为每个场景提供额外的信息源。场景类别使用标签定义，不同标签组如图 5-4 所示，这样所有共享相同标签的场景都被视为属于该场景类别。

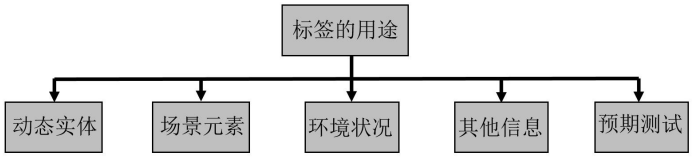


图 5-4 不同标签组[58]

对于上述场景描述的相关标准，对其主要内容和与车用 AI 预测与决策的标准化联系关系总结如下：

表 5-1 自动驾驶场景相关的标准内容及关联总结

标准号	标准名称	标准与车用 AI 标准化研究的联系
ISO 34501:2022	道路车辆-自动驾驶系统测试场景-词汇	定义了 ADS 自动驾驶系统（automated driving systems）测试场景中的术语，同样适用于车用 AI 的预测决策场景中。
ISO 34501:2022	道路车辆-自动驾驶系统测试场景-基于场景的安全评估框架	规定了自动驾驶系统测试场景分类的相关标准和评价体系，明确了自动驾驶系统在不同道路、环境、交通状况下的测试需求。但是测试主要面向整车测试，车用 AI 预测与决策的模块测试的输入输出主要考虑是车辆的运动轨迹和输出动作。此标准的测试场景可以为车用 AI 预测与决策的测试场景搭建提供参考。
ISO 34503:2023	道路车辆-自动驾驶系统测试场景-操作设计领域规范	此标准规定了分级分类法的要求，用于指定能够定义自动驾驶系统（ADS）的运行设计域（ODD）的运行条件，适用于 L3 级和 L4 级自动驾驶系统。但不涉及 AI 相关部分，也不单独涉及决策与预测模块，其描述的相关操作规范内容可作为车用 AI 预测与决策测试做一定输入和支撑。
ISO 34504:2024	道路车辆-自动驾驶系统测试场景-场景分类	提出一种对场景进行分类的方法,提供了一种协调场景分类方式的方法。主要面向整车的测试，场景分类的思路可以为车用 AI 预测与决策分级仿真测试提供参考。

5.2 车用 AI 预测与决策系统简介

5.2.1 车用 AI 预测与决策系统基础架构

5.2.1.1 AI 预测框架

在车用 AI 预测与决策系统中，预测框架如图 5-5 所示，预测算法聚焦于解决目标未来运动趋势判断的问题，由于目标与目标及目标与道路环境之间存在深度的耦合和博弈关系，因此可采用基于深度学习的方法在高维空间中对交互关系进行提取处理并进行预测。

汽标委智能网联汽车分委会

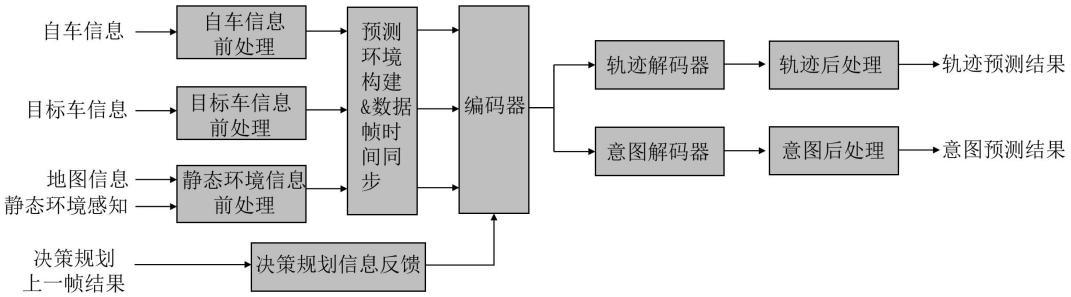


图 5-5 AI 预测框架

AI 预测从应用框架上来讲主要分为四部分：数据预处理、预测环境构建、预测模型、数据后处理。下面分别对各部分功能进行简要说明：

数据预处理环节从预测上游模块获取所需数据后，针对自车时序和属性信息、目标时序和属性信息、道路环境信息等进行解析和筛选。通过后继搜索等方法构建局部拓扑地图和局部时空环境。同时，部分方法也可以利用前序决策规划结果使预测考虑的交互信息更加丰富。

预测环境构建环节将不同来源的上游数据进行时间同步，保证用于模型输入的信息实时性。同时，将不同坐标系下的位置和点数据转化到以自车位置为中心或特定位置为中心的坐标系下，构建特征序列和环境矢量，生成模型训练及推理所需特征。

AI 预测模型应用最为广泛的结构为编码器-解码器（如 RNN、transformer 等）架构，模型根据环境构建信息提取不同对象之间的交互信息，将特征编码映射到高维空间后再利用不同任务的解码器进行解码，最终输出目标未来轨迹或意图。其中，模型输出一般为多模态形式，用于覆盖同一环境状态下预测的不同横纵向意图及其概率。

数据后处理从模型原始输出中提取轨迹和意图信息，结合实际应用中的车道约束、目标筛选、轨迹平滑等条件对信息进行进一步优化处理，最终输出为决策规划可用的预测结果。

汽标委智能网联汽车分委会

5.2.1.2 AI 决策框架

在车用 AI 预测与决策系统中，决策框架如图 5-6 所示，决策算法的输出直接影响自动驾驶系统的驾驶表现，因此需要对驾驶环境具有充分的理解和认识。与预测类似，基于深度学习的决策算法一般也采用编码器-解码器架构，在高维特征空间中处理车辆和环境的交互行为。

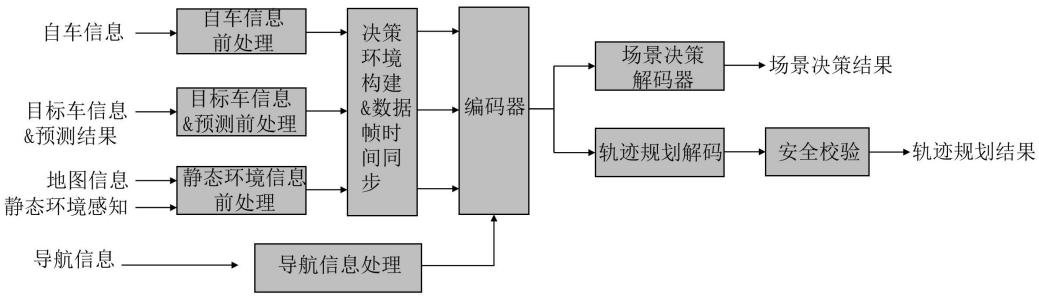


图 5-6 AI 决策框架

AI 决策框架一般可分为四部分：数据预处理、决策环境构建、决策模型、数据后处理。下面分别对各部分功能进行简要说明：

(1) 数据预处理环节从上游感知和预测模块中获取自车历史数据、目标历史及未来预测数据、道路环境、目的地导航等信息，从原始信息中对数据进行过滤和筛选，并初步构建所需的状态序列及道路特征。

(2) 决策环境构建将不同来源的数据进行时间同步，保证用于模型输入的信息实时性。同时，将不同坐标系下的位置和点数据转化到以自车位置为中心的坐标系下，构建特征序列和环境矢量，生成决策模型训练及推理所需特征。

(3) AI 决策模型架构一般采用基于 RNN 或 transformer 的编码器、解码器架构，模型以导航路径结合道路信息生成参考路径，同时通过编码器在高维空间中取自车与目标、环境的交互关系，然后根据不同的决策任务采用不同的解码器输出场景决策、轨迹规划等结果。

(4) 数据后处理从模型原始输出中提取向量特征，结合实车应用的必要规则

如安全校验、车辆动力学约束等进行进一步优化，最终输出可用于车辆控制的决策规划结果。

5.2.1.3 预测与决策联合框架

预测和规划联合的深度学习方法（Integrated Prediction and Planning, IPP）是指在自动驾驶系统中，将车辆轨迹的预测（Prediction）和规划（Planning）结合在一个模型中进行处理的方法。这种方法的核心思想是考虑到自动驾驶车辆（ego-vehicle）的行为会对未来周围交通参与者（surrounding vehicles）产生影响，同时交通参与者的行为也会对自动驾驶车辆产生影响，因此在一个联合的框架中处理预测和规划任务。

相比于将预测和规划作为单独模块处理的方法，AI 预测与决策联合方法允许预测和规划阶段共享信息和上下文，使得规划者可以考虑其他交通参与者的预测行为，并影响周围交通参与者的行为。通过联合考虑预测和规划，系统能够更好地处理不确定性和复杂场景，识别潜在的冲突并提前规避，从而提高整体的安全性和在各种交通情况下的鲁棒性。此外，基于 AI 预测与决策方法可以减少独立模块间接口不匹配而导致的信息损失，提高计算效率。

目前， AI 预测与决策联合方法有多种网络结构设计，但其总体架构可以分为整体式架构和分块式架构两种，如图 5-7 所示。

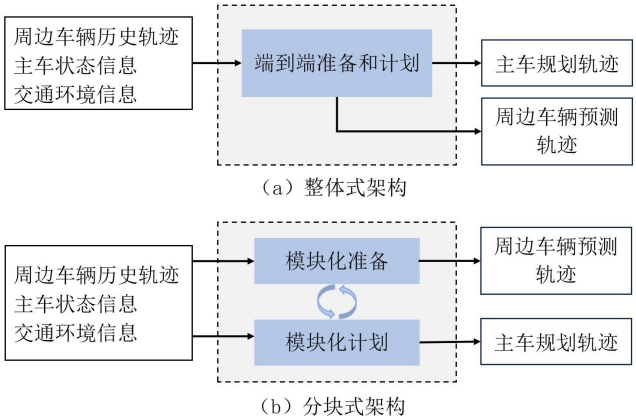


图 5-7 AI 预测与决策联合框架

整体式预测与规划框架将预测和规划集成在一个端到端（E2E）的系统中，

通常由单个神经网络实现，因此两个任务之间不需要额外的工程化的接口。在这种框架中，预测可以作为一项重要的学习任务，在模型应用时候独立输出给自动驾驶系统作为行车预警信息或者其他系统的输入，也可以作为规划任务的一个辅助学习目标与规划任务联合训练，预测的目标作为正则化器，可以提高学习过程的样本效率并增强泛化能力。整体式预测与规划网络通常有一个共享的特征提取主干（backbone）来编码用于预测和规划的深度特征。相比于纯端到端规划，整体式框架增加了模型的解释性。

分块式预测与规划框架使用模块化的预测和规划子系统，每个子系统处理特定的任务，其中预测模块和规划模块之间通过手动设计的接口进行交互。预测模块和规划模块通常采用顺序连接的方式，先对周边交通参与者进行预测，然后将预测结果传递给规划模块，或者在迭代过程中交替进行预测和规划。相比于整体式框架，由于手动集成的系统结合了领域知识，分块式框架通常具有更高的安全性和可解释性，可以根据特定需求灵活设计预测和规划模块以及它们之间的交互。

总体而言，整体式预测与规划框架倾向于简化模型设计，而分布式框架则允许更多的定制化和领域知识的应用。

5.2.2 车用 AI 预测与决策技术发展现状

5.2.2.1 传统预测与决策方法

(1) 预测

传统的车辆预测方法主要有三大类：基于规则、基于运动模型和基于统计模型的方法。这些方法各有优缺点，但在实际应用中普遍面临一些挑战。

首先是基于规则的方法，它就像一套固定的交通规则手册，根据预先设定的规则和经验来预测其他车辆或行人的行为。这种方法在简单的场景下还行得通，但在复杂的、动态的环境中就显得力不从心。比如，文献[59]提到，这种方法需要结合多种传感器信息（比如摄像头、雷达等）来做决策，甚至还得引入额外的数据（比如视频图像分割或目标检测的结果），这不仅增加了系统的复杂性和成

汽标委智能网联标准分委会

本，还无法应对现实中那些不确定、难以解释的边界问题。说白了，这种方法太死板，缺乏灵活性。

其次是基于运动模型的方法，它就像用物理学公式来预测车辆的运动轨迹。文献[60]提到，这种方法用智能驾驶员模型来模拟车辆的反应，但问题在于，这个模型的输出是固定的，而现实中的车辆行为往往带有随机性。比如，驾驶员可能会突然变道或急刹车，这些都是模型无法准确捕捉的。所以，这种方法在模拟现实中的不确定性时表现不佳。

第三类是基于统计模型的方法，它通过分析历史数据来预测车辆的行为。比如，文献[61]提到，可以用马尔可夫决策过程来模拟两辆车的可能行为，并通过模拟找到最可能的交通场景。文献[62]则提到，可以用高斯过程来建模车辆的轨迹，并考虑车辆之间的相互影响。但这种方法也有问题，它需要大量的数据来训练模型，而且在面对新事件或动态环境时，效果也不太理想。简单来说，这种方法依赖历史数据，但现实中的交通场景往往变化莫测，历史数据未必能完全覆盖。

最后，还有一些方法尝试考虑不确定性，比如状态估计模型[63]和状态可达集模型[64]。状态估计模型就像通过观测车辆的状态来推测未来的轨迹，并用贝叶斯滤波器来量化不确定性。状态可达集模型则是通过计算车辆理论上能达到的极限状态，来定义控制约束。这些方法虽然考虑了不确定性，但计算复杂度较高，实际应用中也有一定局限性。为了解决车辆姿态不确定性的问题，研究者们提出了几种不同的方法，核心思路都是通过更好的建模和规划来减少不确定性，提高预测和控制的精度。

Vinay[65]提出了一种基于 EKF（扩展卡尔曼滤波）的方法，就像一个智能的“预测器”，通过随机动力学模型来预测车辆未来的状态。为了减少不确定性，他在采样规划时，把更多的样本放在置信度高的区域，就像在确定的地方多放一些“探测点”，这样一来，不仅降低了不确定性，还减少了地图中的样本密度，从而加快了路径规划的速度。简单来说，这种方法通过“重点区域重点采样”来提高效率。

汽标委智能网联汽车分委会

Farhad[66]则设计了一个基于可观测性滤波器的位姿估计器，就像一个“定位专家”，确保车辆的位置和方向能够被全局观测到，并且能够量化不确定性。也就是说，这种方法让车辆的位姿估计更加准确，同时还能明确地知道估计的误差有多大。

Vadim[67]提出了基于模型预测控制（MPC）的方法，这种方法就像一个“双重规划器”。它分为两层：上层负责预测车辆在给定控制动作下的未来状态，同时把未来的测量值当作随机变量来处理，而不是假设一个最可能的值；下层则根据预测结果计算最优的控制策略。这种方法的好处是能够更灵活地应对不确定性，避免因为假设错误而导致规划失败。

Matthias[68][69]则基于可达集的思想，就像画一个“可能性的范围”，把所有可能的车辆状态都囊括进来。这种方法考虑了传感器噪声、干扰等不确定输入，以及初始状态的不确定性，通过可达性分析来构建一个包含所有可能行为的集合。不过，随着预测时间变长，这个集合会变得非常大，占用的空间也会急剧增加。为了解决这个问题，文献[70]提出用概率分布来随机描述这个集合，就像用一个“概率地图”来代替一个固定的范围，从而减少计算负担。

(2) 决策

传统的决策方法主要分为基于规则的方法和基于优化的方法。基于规则的方法就像一套“交通规则手册”，通过历史经验和专家知识制定一系列规则，在实际场景中根据规则匹配来做出决策。这种方法的核心思路是通过预先定义的规则库，快速解决交通场景中的冲突问题。在匝道合流场景中，研究者发现，基于规则的方法可以有效提高通行效率。文献 [72]提出了三种基于 CAV（网联自动驾驶车辆）技术的合流规则，通过分布式控制策略缓解交通拥堵和速度波动。文献[73]则考虑多车道场景，允许车辆在进入冲突区域时进行变道，并提出了一种基于规则的协同变道控制模型（CLCC），确保变道安全。文献[74]以四辆车为一个协调单元，根据车辆可能的运动状态组合制定并线规则，实现整体车流的协同。

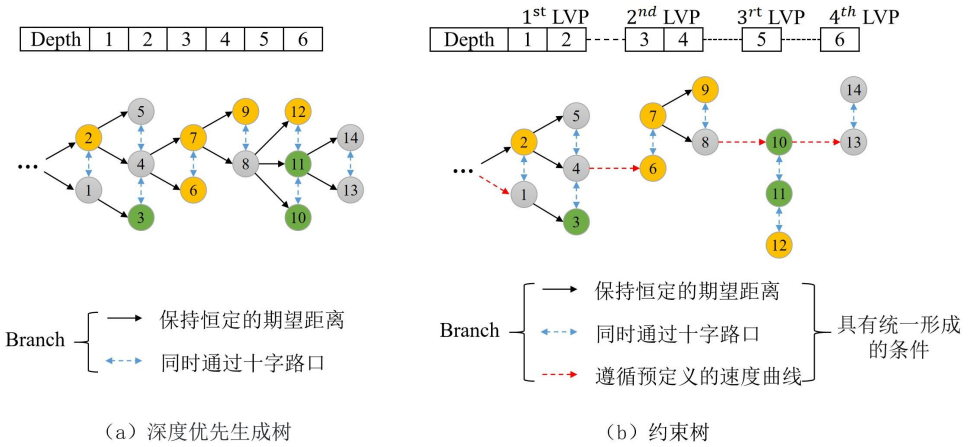
这些方法的优点是计算效率高，能够快速生成接近最优的通行决策。但缺点

汽标委智能网联汽车分委会

是在复杂、动态的场景中，规则的灵活性和适应性有限，可能无法完全满足需求。为了弥补基于规则方法的不足，研究者引入了博弈理论，让车辆之间通过“谈判”来实现协同。博弈理论的核心思路是让每辆车通过观察其他车辆的状态，预测其行为，并规划自己的动作，从而在多车交互中找到最优解。文献[75]在无信号交叉口中用博弈论建模车辆交互行为，让每辆车通过预测和规划实现高效通行。文献[76]提出了一种集中式博弈方法，根据交通密度、公平性等规则设计调度决策序列。文献[77]则提出了一种分层控制策略，通过分布式架构降低计算负担，同时保证最优性，使用帕累托最优集[24]来寻求最优的通行序列。

然而，博弈理论也有局限性。当车辆数量增多时，求解过程会变得复杂，甚至可能不存在最优解。特别是在高动态的连续交通流中，博弈模型的效果可能会大打折扣。为了应对连续交通流的问题，研究者引入了图论，将车辆的冲突和跟随关系用图的方式表示，并通过构建虚拟队列来协调交通流。这种方法的核心思路是将每辆车视为一个节点，通过优化算法寻找最优的无冲突通行序列。文献[80]将每辆车视为一个节点，将不同车道的车辆投影到虚拟车道中，构建无冲突的几何拓扑，形成一个“虚拟车队”进行控制。文献[81]则提出了一种基于虚拟队列的协同方法，通过局部虚拟引导式约束队列解决车流间距过大的问题。这些方法类似于动态规划，将每辆车看作一个节点，寻找最优的无冲突通行序列，并对车辆进行分布式协同控制。针对连续交通流下的问题，文献[85]不局限于接近路口的范围，而将范围扩大至能够容纳车流，并提出构建一个虚拟队列计算得到最优无冲突几何拓扑的方法如图 5-8 左边构建拓扑所示，文献[86]在其基础上进行改进，构建领导约束队列，这两种方法都类似于动态规划，将每一辆车都看作是一个节点，寻优一条最优的无冲突通行序列，并按照决定的通行序列对控制区域内的车辆进行分布式协同控制。

北京交通大学智能网联汽车分委会



（a）深度优先生成树　　（b）约束树

图 5-8 两种几何拓扑决策寻优模式对比

优化类决策算法的核心目标是通过混合整数优化、模型预测控制等优化方法寻找车辆到达路口的最优时间，从而避免冲突并提高通行效率。这种方法的核心思路是通过数学建模和优化计算，为每辆车制定最合理的通行策略，从而实现交通流的协同控制。混合整数优化是一种将连续变量和离散变量结合起来的优化方法，常用于解决复杂的交通控制问题。文献[87]将无冲突通行模式的求解问题转化为混合整数线性规划问题，其中每辆车的目标动作（如直行、左转、右转）作为离散变量进行优化。这种方法通过数学建模，将交通冲突问题转化为可计算的优化问题。模型预测控制（MPC）是一种基于预测和优化的控制方法，常用于动态交通场景。它的核心思路是通过预测未来一段时间内的交通状态，优化当前的控制决策。文献[88]提出了一种双层模型预测控制器，用于十字路口场景下的自动驾驶车辆协同控制。上层（协调层）负责分配车辆在路口的占用时间段，下层（车辆层）计算每辆车的具体控制命令。这种方法在保证一致性和稳定性的同时，提高了通行效率。文献[75]在给定通行顺序后，针对每辆车构造有限时域优化问题，并添加避撞约束，实现分布式协同控制。

5.2.2.2 基于 AI 的预测与决策方法

(1) 预测

传统驾驶预测方法在处理静态、确定性问题上表现较好，但在面对动态、不确定的实际驾驶环境时，表现出明显的不足。相比之下，AI 驾驶预测算法通过

深度学习从大量数据中学习复杂驾驶模式，具有较高的准确性和鲁棒性，能够处理高维和多样化的数据，利用并行计算满足实时性要求，并能融合来自不同传感器的非结构化数据，从而提高预测的准确性和可靠性，能更好地应对复杂和多变的驾驶环境。当前主要的 AI 预测算法模型如下：

VectorNet[35]是一种用于学习高维数据中局部结构的深度学习模型，它的基本原理是基于图神经网络（GNN），通过在时间和空间上的交互来捕捉数据的动态和局部结构，通过矢量化的表示来编码高清地图（HD Maps）和代理（如车辆、行人）的动态。其网络结构分为两个层次。第一层是局部图网络，处理单个多边形内部的向量节点，聚合局部信息。第二层是全局图网络，在更高层面上模拟所有多边形和轨迹之间的高阶交互。VectorNet 的核心优势在于它能够直接处理矢量数据，避免了传统基于栅格图像的表示方法中的信息损失和计算密集型问题。此外，VectorNet 引入了一个自监督的图完成目标，通过随机遮蔽输入节点特征并让模型重建这些特征，从而鼓励图网络更好地捕捉代理动态和场景上下文之间的交互。

QCNet[90]是一种用于量子计算的神经网络架构，它是一个量子机器学习模型，旨在通过量子超越性能实现更高的准确性和效率。其主要用于解决现有轨迹预测方法中的一些关键挑战，比如多代理（如其他车辆、行人等）的未来运动时的效率和准确性问题。QCNet 引入了一种查询中心的场景编码方法，与传统的以代理为中心的编码方案不同，它允许模型通过学习与全局时空坐标系无关的表示来重用过去的计算。通过在局部时空参考帧中处理所有场景元素，并学习独立于全局坐标的表示。QCNet 能够在观察窗口滑动时重用之前计算的编码，从而减少在线预测期间的冗余计算。

TNT[91]是一种用于自动驾驶行为预测的框架，它通过三个阶段来预测移动代理（如车辆和行人）的未来轨迹。TNT 首先预测代理可能的目标状态，这些目标状态是代理在未来某个时间步长内可能到达的位置。这一步骤通过编码代理与环境和其他代理的交互来实现。一旦确定了潜在的目标状态，TNT 的第二阶段会为每个目标生成条件化的轨迹状态序列。第三阶段对每个预测的轨迹进行评分，

然后选择最终的轨迹集合。TNT 的三个阶段都是端到端训练，模型从输入到最终的轨迹预测是连续优化的。

HiVT[92]是一种用于多代理运动预测的深度学习模型，采用分层的方法来处理多代理运动预测问题，通过局部上下文提取和全局交互建模来有效地处理场景中的大量代理。在第一阶段，HiVT 将场景划分为多个以代理为中心的局部区域，并在这些局部区域内提取与中心代理相关的上下文特征。第二阶段通过全局消息传递在局部区域之间聚合信息，以捕获场景中的长距离依赖性和全局动态。此外，HiVT 提出了一种翻译不变的场景表示方法，使用相对位置来描述所有矢量化实体，从而使得模型对全局坐标系的平移不变，并引入旋转不变的空间学习模块，HiVT 能够学习对场景旋转不变的局部和全局表示，能够快速进行多代理运动预测，在 Argoverse 运动预测基准测试中取得了最先进的结果。

LaneGCN[93]是一种用于自动驾驶车辆运动预测的图卷积网络模型。它的核心原理是利用图结构来表示和处理道路网络，特别是车道之间的关系。LaneGCN 通过从矢量化的高精地图数据构建车道图来避免栅格化过程中的信息损失。在这个图中，每个车道被表示为节点，车道之间的拓扑关系（如前后连接、左右邻居）通过边来表示。LaneGCN 引入了多尺度和扩张卷积技术，这允许模型在不同的尺度上捕捉信息，并能够跨越多个车道节点来获取更广泛的上下文信息。LaneGCN 通过图卷积层来提取车道节点的特征，这些特征编码了车道的几何形状、方向以及与其他车道的关系。除了车道图的特征，LaneGCN 还考虑了交通参与者（如车辆）与地图之间的交互。LaneGCN 使用一个融合网络来模拟参与者自身之间以及与地图的交互，包括参与者到车道、车道到车道、车道到参与者和参与者到参与者的四种交互类型。利用融合后的特征，LaneGCN 的预测头部负责生成多模态的未来轨迹预测，并为每种可能的轨迹分配置信度分数。

(2) 决策

相较于传统决策算法，AI 决策模块在自动驾驶系统中通过深度学习、强化学习和生成对抗网络等技术，实现了高效、精确的驾驶决策，主要包括行为决策

模块、运动规划模块及控制模块。

行为决策模块利用强化学习和深度学习等算法，通过策略优化和经验学习，从驾驶数据中学习高层次行为决策和复杂驾驶行为模式，能够处理复杂的驾驶环境和动态交通状况，可根据实时数据和环境变化进行自适应调整，提高驾驶安全性和效率。文献[96]提出了一种基于先验知识和数据驱动的高速公路自动变道策略。通过先验知识优化变道特征，利用 AttentionBiLSTM 模型将变道决策转化为时间序列预测，显著提高了变道决策的准确性和抗干扰能。文献[97]旨在解决现有基于规则决策方法的适应性和泛化性不足的问题，提出了一种基于 DQN 生成对抗模仿学习(DGAIL)的智能车辆行为决策方法。

运动规划模块通过 AI 技术（如生成对抗网络 GAN 和贝叶斯优化）生成详细的动态路径规划，能够处理复杂的交通环境和障碍物。此外，高级导航系统利用 AI 大模型预测未来交通流态，为自动驾驶汽车提供实时路线更新，从而降低延误并提升整体行驶效率[98]。文献[99]提出了一种基于深度强化学习的智能车辆轨迹规划方法，通过深度确定性策略梯度(DDPG)与车辆动力学模型结合，优化智能驾驶策略，并使用模型转移方法将实际场景转换为等效的虚拟抽象场景。

控制模块将高层次行为决策和低层次运动规划结合起来，通过深度强化学习等优化控制策略，提高驾驶的平稳性和舒适性。文献[100]提出在自动驾驶汽车的控制层使用基于深度学习(DeepLearning,DL)的算法。更具体地说，实现深度强化学习(DRL)算法，如深度 Q-Network(DQN)和深度确定性策略梯度(DDPG)，以便比较它们之间的结果。这项工作的目的是获得一个经过训练的模型，应用 DRL 算法，能够向车辆发送控制命令，使其按照确定的路线正确有效地导航。文献[101]在 Pomerleau 的基础上提出了一种道路跟随系统，利用强化学习来设计控制器。它的优点是能够从以前的经验中学习，在新的环境中驾驶，并通过在线学习不断学习和提高其道路跟随能力。文献[102]将监督学习和强化学习相结合，开发了一种高速公路环境下的车道选择策略。

当前主要的 AI 决策算法模型如下：

BEVGPT[105]是一种集成了驾驶场景预测、决策和运动规划的生成式预训练大模型。如图 5-9，该模型以鸟瞰（BEV）图像作为唯一输入源，并根据周围的交通场景做出驾驶决策。 为了确保驾驶轨迹的可行性和平滑性，作者开发了一种基于优化的运动规划方法。 在 Lyft Level 5 数据集上实例化 BEVGPT，并用 Woven Planet L5Kit 进行真实驾驶模拟。 所提出框架的有效性和鲁棒性得到了验证，在 100% 决策指标和 66% 运动规划指标上优于以前的方法。

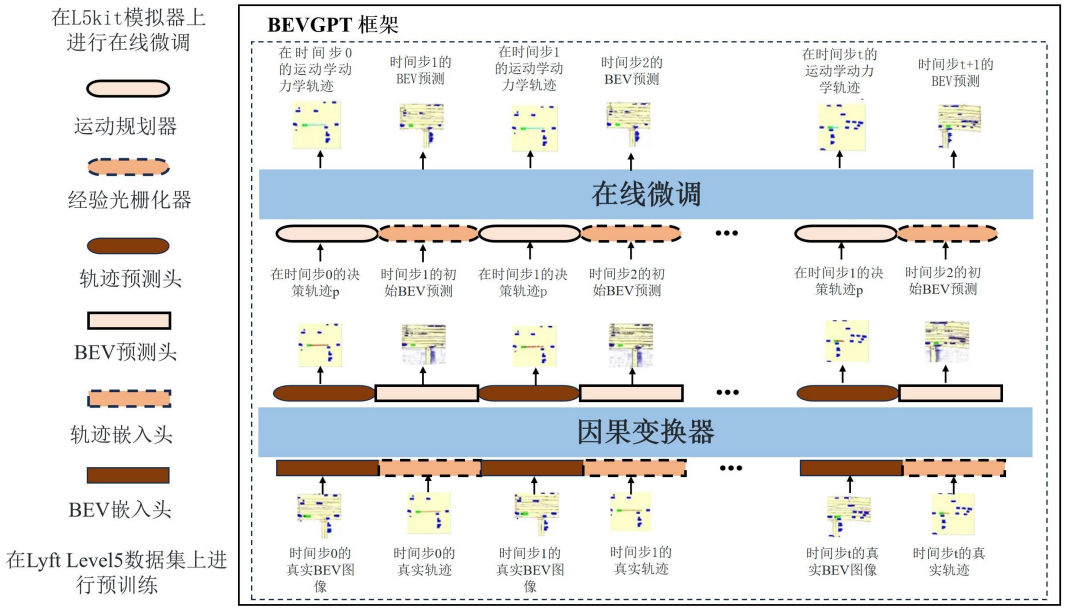


图 5-9 BEVGPT 整体框架[105]

DriveGPT 模型的输入是感知融合后的文本序列，输出是自动驾驶场景文本序列。如图 5-10，DriveGPT4[106]是一个可解释的端到端自动驾驶系统，利用大型语言模型。DriveGPT4 将前视 RGB 摄像头捕获的视频序列作为输入，以及车辆的历史控制信号，进而预测下一步的控制信号。与此同时，人类用户可以与 DriveGPT4 进行对话，该系统可以提供自然语言响应，例如描述车辆的行为并解释其行为背后的原因。为了训练 DriveGPT4 像人一样交流，其遵循 LLaVA，并使用 ChatGPT 基于 BDD-X 数据集创建视觉指令调优数据集。

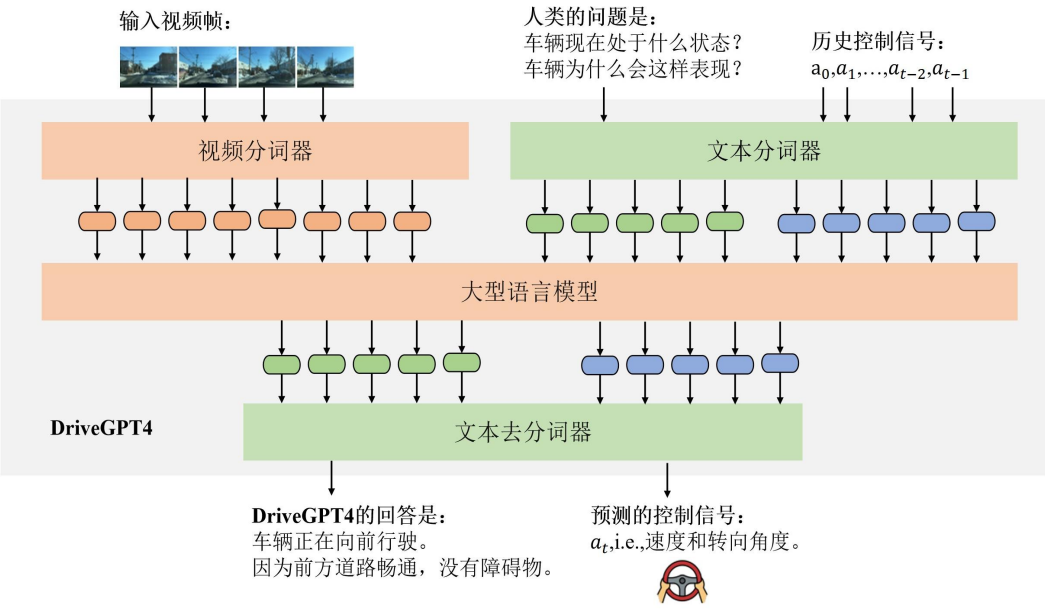


图 5-10 DriveGPT4 整体框架[105]

5.2.3 车用 AI 预测与决策技术应用现状

表 5-2 决策与预测方法使用类别调查结果表（2024 年 7 月）

	决策算法使用类型	预测算法使用类型
高校 1	机器学习类(混合 A*，树搜索)	机器学习类(混合 A*，树搜索)
高校 2	神经网络类(CNN、RNN、Transformer)	神经网络类(CNN、RNN、Transformer)
高校 3	神经网络类(CNN、RNN、Transformer)	神经网络类(CNN、RNN、Transformer)
企业 1	规则和上述学习类方法结合类	规则和上述学习类方法结合类
企业 2	决策与预测结合端到端类	决策与预测结合端到端类
企业 3	决策与预测结合端到端类	决策与预测结合端到端类
企业 4	规则和上述学习类方法结合类	规则和上述学习类方法结合类

基于上述对现有车用 AI 预测与决策技术框架和方法类别的分析，为了解目前技术实际研究和应用的最新情况，本报告对现有国内高校和企业做了预测与决策技术算法类别的问卷调查，调查的结果如表 5-2 所示。从下表中可以得出以下结论：（1）目前国内高校普遍采用机器学习或神经网络的方法来展开对预测与决策方法的研究，基于学习的方法在学术研究领域是热点。（2）业界普遍采用了学习类方法作为预测与决策应用落地的技术方案，其中采用大模型端到端的学习类方法同样成为目前车企主流的应用路线。总的来说，AI 技术的强泛化性和多场景适用性能使得自动驾驶预测与决策功能更加贴近人类驾驶，同时从调查的结果上来看，目前基于 AI 的自动驾驶预测和决策无论是在学术研究领域和企业端的实际应用，也逐渐成为主流的技术研究方案。

5.3 车用 AI 预测与决策测试及其评价方法

5.3.1 测试数据格式与内容

车用 AI 预测与决策系统是一种高度复杂的技术，它涉及到对车辆行驶环境的实时感知、数据处理、预测分析以及做出快速而准确的决策。该系统的数据格式和内容需要能够支持机器学习模型的运行，确保系统的高效性和准确性。车用 AI 预测与决策系统数据包括感知数据、环境数据、预测数据、决策数据等。

5.3.1.1 感知数据类型简介

(1) 传感器数据

类型：雷达、激光雷达（LiDAR）、摄像头、GPS、IMU（惯性测量单元）等。

格式：通常是二进制格式，如`.bin`或`.log`，也可以是视频流`.mp4`或图像序列`.jpg`。数据内容：雷达和 LiDAR 数据可以提供周围物体的距离、速度、方位角。摄像头数据可以视觉图像，用于物体检测和识别。GPS 数据可以提供车辆的精确地理位置。IMU 数据提供车辆的加速度和旋转速率。

(2) 车辆状态数据

类型：车辆的当前速度、加速度、转向角度、发动机状态等。

格式：通常是结构化数据，如`.csv`或`.json`。

内容：速度和加速度用来描述车辆的动态状态。转向角度可以描述车辆的行驶方向。发动机状态参数包含转速、油量、冷却液温度等。

5.3.1.2 环境数据类型简介

(1) 地图数据

类型：高精度地图，包括道路、交通标志、信号灯等。

格式：`.osm`（OpenStreetMap）或自定义格式。

数据内容：道路网络数据内容主要包括道路的几何形状、连接关系。交通标志数据主要包括限速、禁止通行等信息。信号灯状态数据包括当前信号灯的颜色和变化规律等

(2) 天气数据

类型：当前和预测的天气状况。

格式：`.json`或`.xml`。

内容：天气数据主要包括温度、湿度、降水概率，以及能见度和风速情况。能见度是影响驾驶视线的重要条件。风速是可能影响车辆稳定性的因素。

5.3.1.3 预测数据类型简介

(1) 交通流量预测数据

类型：基于历史数据和实时数据的交通流量预测。

格式：时间序列数据，如`.timeseries`或`.csv`。

数据内容：交通流量预测数据主要包括预测的交通拥堵点和车流量变化。

(2) 行为预测

类型：对周围车辆和行人行为的预测。

格式：通常是模型输出，如`.prediction`或`.json`。

数据内容：行为预测数据内容包括周围车辆的预计行驶轨迹、行人的预计移动路径等。

5.3.1.4 决策数据类型简介

(1) 路径规划

类型：基于当前位置、目的地和交通状况的最优路径。

格式：可能是图形数据结构，如`.graph`或`.xml`。

内容：在路径规划中，路径节点数据包括交叉口、转弯点等。路径权重数据包括时间、距离或拥堵情况。

(2) 驾驶策略

类型：根据预测结果和路径规划生成的驾驶策略。

格式：决策树或规则集，如`.rules`或`.json`。

数据内容：驾驶策略内容包括行驶速度建议、变道和超车时机、应对突发情况的策略等。

5.3.2 公开测试数据集简介

车用 AI 研究工作主要集中在提高车辆的感知、预测和决策能力领域。公开数据集包含了从真实世界中收集的传感器数据、车辆状态信息以及环境数据，在测试和评估 AI 系统的性能方面为研究人员提供了标准化的测试基准和比较平台。

5.3.2.1 数据集的重要性

基准测试：公开数据集提供了标准化的测试基准，允许研究人员和开发者比较不同算法的性能。

算法开发：公开数据集为开发新的预测和决策算法提供了丰富的数据资源。

研究共享：公开数据集促进了研究成果的共享和交流，加速了技术进步。

5.3.2.2 相关数据集介绍

常用的轨迹预测开源数据集有 nuScenes、Argoverse Motion Forecasting Dataset、NGSIM、CQSkyEyeX 等数据集，相关介绍如表 5-3 所示。

表 5-3 轨迹预测公开数据集简介

数据集名称	数据规模	传感器数据	应用场景	优势
nuScenes	1000 个场景，20 秒/场景	摄像头、激光雷达、雷达、GPS/IMU	多模态感知、轨迹预测、目标检测	数据丰富、标注详细，支持多传感器融合
Argoverse Motion Forecasting	300,000+个 5 秒轨迹片段	无传感器数据，仅轨迹数据	车辆轨迹预测，复杂交互场景	数据量大、场景多样，适合交互行为研究
NGSIM	多个场景，15-45 分钟/场景	无传感器数据，仅轨迹数据	交通流分析、车辆轨迹预测、交通仿真	数据真实、场景典型，适合交通流动力学研究
CQSkyEyeX	多个城市道路场景	无传感器数据，仅轨迹数据	多目标轨迹预测，复杂场景分析	数据多样、场景复杂，适合多目标交互研究

5.3.3 典型测试场景分类

5.3.3.1 城区道路单车智驾测试场景

(1) 交通信号灯控制场景

验证自动驾驶系统在面对实际道路条件和复杂的交通信号变化时的表现。测试场景包括：

汽标委智能网联汽车分委会

1) 绿灯起步/红灯停车测试

在交通灯交叉口，自动驾驶车辆需精准感知交通信号灯的颜色变化。当红灯转为绿灯时，车辆应迅速而安全地启动并融入车流；而当绿灯转为红灯时，车辆则需提前减速并在停止线前平稳停车，确保不违反交通规则且不影响其他道路使用者。

2) 无保护交叉口穿行测试

面对无交通信号灯或信号灯失效的城市交叉口，自动驾驶系统需准确评估周围交通状况，并自主决策是否安全通过或安全停车等待。这一测试旨在验证系统在没有明确信号指示情况下的预测与决策能力。

3) 左转等待绿箭头信号

在设有专用左转信号灯的复杂交叉口，自动驾驶车辆需准确识别并响应绿箭头信号。当绿箭头亮起时，车辆应迅速而安全地完成左转动作，同时确保不影响直行和右转车辆及行人的通行安全。

(2) 行人和非机动车辆场景

模拟自行车和行人在不同情况下突然出现的场景，测试自动驾驶系统的识别能力、反应时间和策略。测试场景包括：

4) 城市交叉口行人穿越

在城市的繁忙交叉口，自动驾驶系统需实时监测行人信号灯及行人的实际行为。当绿灯亮起时，系统应确保车辆给予行人足够的通行空间；同时，对于红灯时违规穿越的行人，系统需迅速识别并作出安全避让，保障行人与车辆的安全。

5) 非机动车并行与切换车道

在城市街道的非机动车道上，自动驾驶车辆需时刻警惕非机动车的动态。当非机动车试图切入机动车道时，系统需快速响应，调整车速和行驶路径，确保与非机动车保持安全距离，避免碰撞事故的发生。

汽标委智能网联汽车分委会

(3) 复杂交叉路口场景

全面评估系统在复杂交通环境中的性能，发现潜在问题并改进系统，测试场景包括：

6) 多车道交叉口穿越测试

在拥有多个车道和复杂交通流的大型交叉口，自动驾驶系统需具备良好的预测和决策能力。车辆需精准识别并遵循各车道的通行规则，在穿越交叉口时保持稳定的行驶状态，同时灵活应对其他车辆的变道和加减速行为。

7) 转弯车道选择测试

在城市中心的高流量交叉路口，自动驾驶车辆需准确判断并选择合适的转弯车道。系统需提前识别并规划行驶路径，确保在到达转弯点前顺利完成车道变换，并在转弯过程中保持对周围环境的持续监测和响应。

(4) 车流密集场景

确保系统在高压力和高复杂性的交通环境中能够可靠地运行，测试场景包括：

1) 高峰时段城市高速公路测试

在交通高峰时段的城市高速公路上，自动驾驶车辆需应对高密度的车流和频繁的车道变换。系统需保持高度的稳定性和响应能力，确保车辆在多车道高速公路上稳定行驶，同时灵活调整车速和车距，避免与前后车辆发生碰撞。

2) 城市交通堵塞模拟

在极端交通堵塞的城市区域，自动驾驶系统需具备在缓慢移动和频繁停车-前进的交通流中保持安全行驶的能力。车辆需精确控制与前车和侧车的距离，确保在拥堵路段中安全通行，同时遵守交通规则并减少对其他车辆的影响。

3) 紧急情况下的车流穿插

在紧急车辆（如救护车）需要快速通过的车流密集环境中，自动驾驶系统需迅速识别并响应紧急信号。车辆需根据紧急车辆的行驶路径和速度调整自身的行

汽标委智能网联汽车分委会

驶轨迹或停车让路，确保紧急车辆能够顺畅通行，同时避免增加交通混乱和安全隐患。

4) 停车场和商业区驾驶

在大型购物中心的停车场和商业区街道，自动驾驶系统需精确判断车位大小、周围障碍物的位置及动态变化，同时避免与行人、其他车辆发生碰撞，该场景主要测试自动驾驶系统在密集、混合使用的环境中的性能，验证车辆在处理多重动态障碍时的预测和决策能力。

5.3.3.2 高速道路单车智驾测试场景

(1) 高速跟车测试

在高速公路上的长直路段，自动驾驶车辆需要在高速公路上与前车保持安全跟车距离，并根据前车速度进行适时的加速和减速。

(2) 高速变道测试

面对高速公路上多车道、高车流量的复杂环境，自动驾驶车辆需准确判断变道时机，平稳变换车道以应对车流变化或超车需求。

(3) 高速上匝道测试

在高速公路上的入口匝道，自动驾驶车辆需要在入口匝道上加速，并安全地与高速公路上的车流合并。

(4) 高速下匝道测试

在高速公路上的出口匝道，自动驾驶车辆需要在出口匝道上适时减速，并安全地退出高速公路。

(5) 高速公路紧急停车测试

在高速公路上的应急车道，自动驾驶车辆需要在高速公路上识别并响应紧急情况，安全地停靠在应急车道上。

(6) 高速公路交通堵塞绕行测试

汽标委智能网联汽车分委会

在高速公路上的交通堵塞区域，自动驾驶车辆需要识别交通拥堵并选择最合适的绕行路线以避开拥堵区域。

(7) 高速公路隧道行驶测试

在高速公路上的隧道区域，自动驾驶车辆需要在隧道内保持稳定行驶，并根据环境变化自动调整灯光和行驶策略。

5.3.3.3 城区道路车路云协同场景

车路协同系统通过路侧系统和联网车辆的互动，实现高精度地理信息、交通规则、环境状态等数据的实时交换。路侧系统不仅提供定位辅助和交通信息，还能进行区域协同控制，提升交通系统的整体服务水平。联网车辆通过实时获取周边交通状态，完成合理的路线和轨迹规划，以提升车辆的行驶安全性和通行效率。

城区道路 V2X 车路云协同的测试场景主要关注于路线与车速引导、变道与超车协同控制、信号交叉口协同控制、无信号交叉口协同控制、车路协同行车安全预警等。

(1) 路线与车速引导：

在信息共享下车辆可以进行路线与轨迹规划以提升通行效率，重点测试车辆的信息整合能力与轨迹优化效果。测试内容包括：测试车辆如何结合车路云协同系统共享的路网中交通状态信息来完成最优的路线规划方案；测试车辆如何通过车路云协同系统获取路面障碍信息来合理规划轨迹以避免在接近障碍时紧急刹车，实现更加平顺的驾驶体验。

(2) 变道与超车控制：

信息共享和协同控制系统可以支持车辆优化变道与超车的行为和轨迹，重点测试自动驾驶车辆通行安全、效率、舒适性等指标。测试内容包括：测试自动驾驶车辆如何结合实时获取的周围车辆行驶状态进行分布式控制，智能地确定变道与超车的时机，优化驾驶行为和行驶轨迹；测试多个自动驾驶车辆进行协同控制时，如何根据多车的综合优化目标，为每个车辆规划出合理的通行轨迹。

汽标委智能网联汽车分委会

(3) 信号交叉口协同控制：

车路云系统中，通过信号交叉口与联网车辆的信息共享，可以同时优化信号配时方案和车辆轨迹。测试内容包括：测试信号交叉口控制器如何根据实时获取的来自不同进口道的车辆到达信息和实时交通状态，对交通口进口道进行需求估计，从而完成合理规划交叉口的信号配时方案；测试自动驾驶车辆如何利用共享的信号交叉口控制方案来合理规划通行轨迹，减少停车次数，并提升通行效率。

(4) 无信号交叉口协同控制：

车路云系统中，允许无信号交叉口实现分布式的多车通行协同控制，重点测试自动驾驶车辆在无信号交叉口的协同控制效果。测试内容包括：测试根据获取得到的周围车辆交通状态预测每个车辆到达停车线或者冲突点的时间，并根据博弈论、混合整数优化等路权分配方法确定车辆的通行先后顺序；测试车辆在通行优先权确定之后的轨迹优化策略，以避免车辆间冲突，提升通行效率。

5.3.3.4 高速道路车路云协同场景

高速道路的车路云协同测试场景主要关注于高速通信组网和匝道合流等。

(1) 高速通信组网：

高速场景下，车路云系统在通信组网方面面临链路稳定性、信号质量、网络覆盖、实时性、安全性等挑战，需验证车辆的通信性能、系统响应与兼容性，确保自动驾驶系统的高效与安全。测试内容包括：测试通信链路的稳定性和切换效率；测量数据传输延迟和系统响应时间；评估在不同信道条件下的信号强度和传输质量；测试网络覆盖的广度和盲区的识别；模拟攻击场景，测试系统的抗干扰和数据保护能力；验证不同设备和系统在各种场景下的兼容性和协同工作能力。

(2) 匝道合流：

涉及到来自不同方向的车辆汇聚到同一车道，重点测试车辆的通行安全和效率。测试内容包括：评估自动驾驶系统在匝道合流区域生成最优轨迹的能力，需要考虑最小化合流时间和车辆加速度，同时保证车辆动力学和安全约束；评估如

汽车委智能网联汽车分委会

何通过协同博弈理论和成本函数优化车辆的合流次序，以实现高效和节能的通行轨迹；评估匝道合流后自动驾驶车辆对于整体车流稳定性的影响。

5.3.4 仿真与测试平台

5.3.4.1 AI 软件仿真介绍

自动驾驶系统中的 AI 软件仿真平台是一个集成了虚拟环境仿真和自动驾驶系统算法的综合性平台。其中虚拟环境仿真平台提供一系列具备一定精度的仿真模型，包括控制算法、交通模型、环境模型、车辆动力学模型、车辆感知系统模型，这些仿真模型能够根据测试场景的不同以及测试车辆的不同进行实时切换，同时虚拟环境仿真平台还具备应良好稳定的接口，实时存储功能，较高的运算性能，可调整的仿真步长等能力，以满足长时间不间断自动化测试的要求。AI 预测及决策模块专注于模拟驾驶决策、规划和控制过程。它集成了先进的 AI 算法和机器学习模型，能够模拟自动驾驶车辆在各种复杂交通环境和场景下的行为。这些算法能够接收来自虚拟环境仿真平台的数据，包括传感器信息、交通状况等，并基于这些数据进行实时决策和路径规划。

5.3.4.2 AI 软件仿真平台要求

环境仿真平台所包含的传感器模型种类丰富，涵盖摄像头、超声波雷达、毫米波雷达以及激光雷达等。摄像头模型能确保图像采集的准确性和低畸变。超声波雷达模型则模拟真实探测功能，其探测角度、宽度范围和最大量程均达到高精度要求。毫米波雷达模型精准捕捉前方障碍物的位置和速度信息，满足严格的精度标准。激光雷达模型则通过激光测距技术，确保飞行时间和三角测距的精确性。车辆动力学模型模拟汽车在不同运动状态下的动力学特性，包括车辆通过不平路面的平顺性、横向和纵向加速度以及俯仰、侧倾等动态行为。在人-车-路闭环仿真系统中，驾驶人模型能够部分或全部替代真实驾驶员在仿真环境中的操作行为，其控制响应与真实驾驶员高度一致。驾驶人模型主要分为三类：补偿跟随模型，基于实际与预期轨迹的偏差进行方向盘决策；预瞄跟随模型，模拟真实驾驶员的驾驶行为；智能控制模型，则基于智能控制原理构建。

汽标委智能网联汽车分委会

环境仿真平台的基本要求包括：强大的场景搭建功能，提供静态环境、动态环境和天气模型等，并能根据真实路网或高精地图生成大规模虚拟道路环境；交通流生成功能，可根据实际路测数据或参数化模型生成微观和宏观交通流；完善的环境传感器模型，支持提供原始数据和真值；开放的输入输出接口，兼容不同语言的自动驾驶算法；高精度的车辆动力学模型，支持闭环测试；分布式案例存储和运行平台，通过硬件扩展显著提升自动驾驶测试里程；数据收集功能，支持仿真案例回放；开放的软件接口，便于接入第三方软件进行联合仿真；以及直观的动画显示模块，实时展示车辆在虚拟环境中的行驶状态。

5.3.4.3 AI 预测与决策主流软件仿真平台

目前国内外仿真软件众多，PanoSim、VTD、Prescan、Carmaker、SCANeR、CARLA、51Sim-One 和腾讯 TAD Sim 是目前主流的自动驾驶仿真平台，它们通过集成高精度车辆动力学模型、环境建模、传感器仿真和交通流模拟等功能，为自动驾驶系统的开发与测试提供了全面的虚拟环境支持。这些平台不仅支持从离线到实时的全流程仿真测试，还能通过丰富的场景库和 OpenDRIVE、OpenCRG、OpenSCENARIO 等标准地图格式，精确还原复杂交通场景。此外，它们兼容 Matlab/Simulink、ROS 等第三方工具，支持感知、决策、控制等模块的验证与优化，为智能驾驶算法的研发提供了高效、安全的测试环境。

在车用 AI 领域，仿真平台如 TraNS、Veins、iTETRIS、NCTUns 和 VSimRTI 通过深度耦合交通仿真器（如 SUMO）和网络仿真器（如 NS3、OMNeT++），实现了车辆行为与通信网络的联合仿真。这些平台支持多种无线通信协议（如 IEEE 802.11p、WiMAX），能够模拟复杂的车联网应用场景，包括 V2X 通信、数据交换、紧急警报和防撞系统等。它们通过实时信息交互和动态车辆轨迹生成，为车用 AI 技术的研究与验证提供了全面的技术支持。

5.4 车用 AI 预测与决策评价指标

车用 AI 预测与决策算法的性能评价一方面通过实车体验来进行，另一方面可结合仿真、数据挖掘等方法通过量化指标进行评价。表格 5-4 总结了车用 AI

汽标委智能网联汽车分委会

决策、车用 AI 预测的评价指标，同时注明了相关指标在仿真或实车场景中是否可测。

表 5-4 车用 AI 决策评价指标总结

评价指标\测试方法		仿真	实车	公式/定义
预测	准确率	√	√	$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$ #
	精确率	√	√	$\text{Precision} = \frac{TP}{TP+FP}$ #
	召回率	√	√	$\text{Recall}=\frac{TP}{TP+FN}$ #
	F1 值	√	√	$\text{F1_Score}=2\times\frac{Precision\times Recall}{Precision+Recall}$ #
	宏精确率	√	√	$\text{Macro Precision} = \frac{1}{n}\sum_{i=1}^{n}\text{Precision}_i$ #
	宏召回率	√	√	$\text{Macro Recall} = \frac{1}{n}\sum_{i=1}^{n}\text{Recall}_i$ #
	宏 F1 值	√	√	$\text{Macro F1} = \frac{1}{n}\sum_{i=1}^{n}\text{F1}_i$ #
	平均位移误差（ADE）	√	√	$ADE = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\sqrt{(\hat{x}_{ij}-x_{ij})^2+}$
	最终位移误差（FDE）	√	√	$FDE = \frac{1}{m}\sum_{i=1}^{m}\sqrt{(\hat{x}_{n,i}-x_{n,i})^2+(\hat{y}_{n,i}-}$
	分时 ADE	√	√	针对预测的时长，将 ADE 指标细化为第 t 秒的 ADE、FDE
	分时 FDE	√	√	针对预测的时长，将 FDE 指标细化为第 t 秒的 FDE
	横向 ADE	√	√	$ADE_Lat_Frenet = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\hat{d}_{ij}$
	横向 FDE	√	√	$FDE_{Lat_{Frenet}} = \frac{1}{m}\sum_{i=1}^{m}\hat{d}_i - d_i$ #
	纵向 ADE	√	√	$ADE_Lat_Frenet = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\hat{s}_{ij}$

汽标委智能网联汽车分委会

	纵向 FDE	√	√	$FDE_{Lon_{Frenet}} = \frac{1}{m}\sum_{i=1}^{m} \hat{s}_i - s_i \#$
	minADE	√	√	ADE 代表平均位移误差，minADE 则是在多个预测轨迹中，选择与实际轨迹平均距离最小的那个轨迹所对应的 ADE
	minFDE	√	√	FDE 代表最终位移误差，minFDE 是在多个预测轨迹中，选择与实际轨迹最终位置距离最小的那个轨迹所对应的 FDE
	丢失率	√	√	$MR = \frac{N_{miss}}{N} \times 100\% \#$
	可行驶区域合规性	√	√	$DAC = \frac{k-p}{k} \#$
	重叠率	√	√	$Overlap = \frac{N_{overlap}}{N} \times 100\% \#$
	均方根误差	√	√	$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(\hat{x}_{ij} - x_{ij}\right)^2 +}$
决策	MPCI		√	$MPCI = \frac{TotalAutonomousMiles}{TotalCriticalInterventions} \#$
	跟车纵向安全距离		√	$d_{min} = max\left(0, v_r\rho + \frac{1}{2}a_{max,accel}\rho^2 + \frac{(}{}\right.$
	横向安全距离		√	$d_{min} = \mu + max\left(0, \frac{v_1 + v_{1,\rho}}{2}\rho + \frac{v_{1,}^2}{2a_{min,}^{lat}}\right.$
	平均通行时间	√	√	$T_{pass} = \frac{1}{N}\sum_{i=1}^{N} t_i \ \#$
	换道准确率	√	√	$\text{acc_change} = \frac{N_{change}}{N_{total}} \ \#$
	换道回退率	√	√	$\text{change_back_rate} = \frac{N_{change_back}}{N_{total}} \ \#$
	匝道通过率	√	√	$\text{ramp_pass_rate} = \frac{N_{ramp_pass}}{N_{total}} \ \#$
	路口通过成功率	√	√	$\text{intersection_pass_rate} = \frac{N_{intersection_pa}}{N_{total}}$
	绕障成功率	√	√	$\text{obstacle_pass_rate} = \frac{N_{obstacle_pass}}{N_{total}} \ \#$

汽标委智能网联汽车分委会

	交互一致率	√		$\text{agreement_pass_rate} = \frac{N_{agreement_pass}}{N_{total}}$ #
	误制动平均发生里程	√		$\text{MPFB} = \frac{M}{N_{False_brake}}$ #
	急加速平均发生里程	√		$\text{MPRA} = \frac{M}{N_{Rapid_acceleration}}$ #
	抖动平均发生里程	√		$\text{MPJ} = \frac{M}{N_{Jerk}}$ #
	急转弯平均发生里程	√		$\text{MPHT} = \frac{M}{N_{Hard_turn}}$ #
	行驶压迫感平均发生里程	√		$\text{MPDPE} = \frac{M}{N_{Driving_pressure_experience}}$ #
	极限换道平均发生里程	√		$\text{MPELC} = \frac{M}{N_{Extreme_lane_change}}$ #

5.4.1 预测评价主要指标概述

车用 AI 预测评价指标可分为意图预测和轨迹预测两大类。意图预测主要表达当前或未来很短时间内的离散的行为意图，不同的智能体的意图通常需要单独定义。轨迹预测主要表达智能体一段时间以内的具体运动趋势。

5.4.1.1 驾驶意图预测评价指标

驾驶意图评价指标基本沿用机器学中分类问题中的二分类或多分类指标。常见的驾驶意图预测如下图所示：

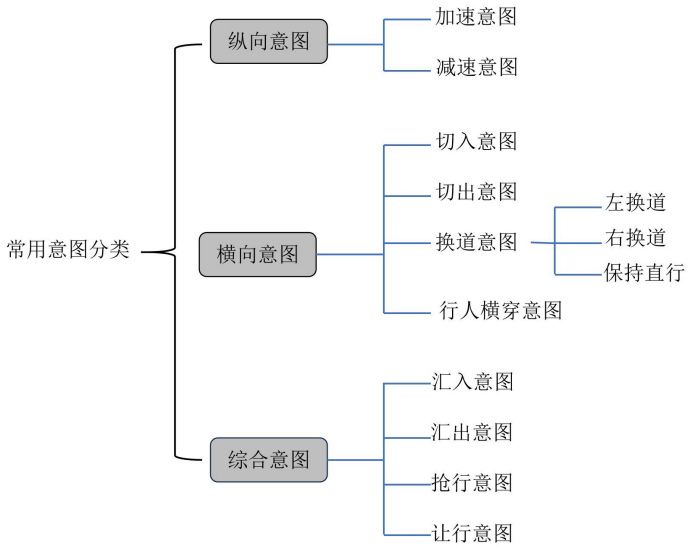


图 5-14 常见驾驶意图分类

以上意图中，每一类意图均可采用二分类评价指标（例如将切入与不切入进行二分类），也可以根据具体的预测任务，将有关联的意图进行组合，再进行二分类评价（例如将左换道/右换道组合成换道意图，再与直行不换道进行二分类）。评价指标采用通用的二分类评价。

首先，基于该行为发生与否（发生为正样本，不发生为负样本）对比真实意图与预测意图，计算以下 4 项基础指标：

(1) 真阳性 TP(True Positive)：真实意图为发生，且预测意图为发生的数量（对正样本预测正确）

(2) 假阴性 FN(False Negative)：真实意图为发生，但预测意图为不发生的数量（对正样本预测错误）

(3) 假阳性 FP(False Positive)：真实意图为不发生，但预测意图为发生的数量（对负样本预测错误）

(4) 真阴性 TN(True Negative)：真实意图为不发生，且预测意图为不发生的数量（对负样本预测正确）

在基础指标的基础上，分别结算准确率、精确率、召回率、F1 分数，作为意图预测的评测指标，计算方法如下：

(1) 准确率（Accuracy）：模型分类正确的样本数占总样本数的比例。

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \#(5-1)$$

意义：准确率越高，表示模型在整体上分类得越准确。但是，当数据集中各类别样本分布不平衡时，准确率可能会产生误导，因为它可能会受到多数类样本的影响。

(2) 精确率（Precision）：模型预测为正样本的实例中，真正为正样本的比例。

$$\text{Precision} = \frac{TP}{TP + FP} \#(5-2)$$

意义：精确率越高，说明模型在预测为正样本的实例中，真正为正样本的比

汽标委智能网联汽车分委会

例越高，即模型预测为正样本时越可靠。

(3) 召回率（Recall）：衡量真正为正样本的实例中，被模型预测为正样本的比例。

$$\text{Recall}=\frac{TP}{TP+FN}\ \#(5-3)$$

意义：召回率越高，说明模型能够找出更多的正样本，即模型预测为正样本的实例中，真正为正样本的比例越高。

(4) F1 分数（F1-Score）：精确率和召回率的调和平均数，用于衡量模型在保持精确率和召回率之间的平衡时的性能。

$$\text{F1_Score}=2\times\frac{Precision\times Recall}{Precision+Recall}\ \#(5-4)$$

意义：F1 值越高，说明模型在精确率和召回率之间取得了较好的平衡。它提供了一个单一指标来评估模型的整体性能，特别是在需要同时考虑精确率和召回率的场景中。

驾驶意图预测的指标还可以扩展到多分类，例如换道意图可以定义为左换道、保持直行、右换道路。在这种情况下，可以计算混淆矩阵来表示意图预测的性能。

表 5-5 驾驶意图混淆矩阵表

混淆矩阵		真实值				
		意图 1	意图 2	意图 3	…	意图 N
预测值	意图 1	TP_{11}	TP_{12}	TP_{13}	…	TP_{1N}
	意图 2	TP_{21}	TP_{22}	TP_{23}	…	TP_{2N}
	意图 3	TP_{31}	TP_{32}	TP_{33}	…	TP_{3N}
	…	…	…	…	…	…
	意图 N	TP_{N1}	TP_{N2}	TP_{N3}	…	TP_{NN}

汽标委智能网联汽车分委会

混淆矩阵是一个 n 行 n 列的表格，其中 n 是类别的数量。其中：

(1) 行：代表真实类别，即数据集中样本的实际标签。

(2) 列：代表预测类别，即模型对样本的预测结果。

矩阵的对角线上的元素（TPii）表示模型正确预测的样本数，非对角线元素则表示模型错误预测的样本数。

混淆矩阵的四个主要元素是：

(1) 真正例（True Positives, TP）：模型正确地将某个样本预测为其实际类别的数量。例如，模型预测一个实际上是左换道的样本为左换道。

(2) 假正例（False Positives, FP）：模型错误地将一个样本预测为某个类别，但实际上该样本属于其他类别。例如，模型预测一个实际上是保持直行的样本为左换道。

(3) 真反例（True Negatives, TN）：模型正确地将一个不属于某个类别的样本预测为不属于该类别。在多分类问题中，TN 的计算方式可能不太直观，因为它涉及到不属于当前类别的所有其他类别的样本。但在二分类问题中，TN 是指模型正确地将负样本预测为负样本的数量。

(4) 假反例（False Negatives, FN）：模型错误地将一个属于某个类别的样本预测为不属于该类别。例如，模型预测一个实际上是左换道的样本为保持直行或右换道。

在基础指标的基础上，分别结算准确率、精确率、召回率、F1 分数，宏平均，作为多分类意图预测的评测指标，计算方法如下：

(1) 准确率(Accuracy)：计算所有类别的正确分类样本数占总样本数的比例。

$$\text{Accuracy} = \frac{(TP_{11} + TP_{22} + \cdots + TP_{nn})}{Num_{\text{总样本}}} \#(5-5)$$

(2) 精确率（Precision）：计算模型正确预测为第 i 类的样本数占所有预测为第 i 类的样本数的比例。

汽标委智能网联汽车分委会

$$\text{Precision}_i = \frac{TP_{ii}}{TP_{i1} + TP_{i2} + \cdots + TP_{iN}} \#(5-6)$$

(3) 召回率（Recall）：计算模型正确预测为第 i 类的样本数占所有实际为第 i 类的样本数的比例。

$$Recall_i = \frac{TP_{ii}}{(TP_{1i} + TP_{2i} + \cdots + TP_{Ni})} \#(5-7)$$

(4) F1 分数（F1-Score）：综合考虑了精确率和召回率，用于平衡二者之间的关系，特别适用于不平衡数据集。

$$F1_i = \frac{2 \times Precision_i \times Recall_i}{Precision_i + Recall_i} \#(5-8)$$

(5) 宏平均（Macro-average）

宏平均是对每个类别的指标进行平均，不考虑类别的样本数量。

宏精确率（Macro Precision）：

$$\text{Macro Precision} = \frac{1}{n}\sum_{i=1}^{n} \text{Precision}_i \ \#(5-9)$$

宏召回率（Macro Recall）：

$$\text{Macro Recall} = \frac{1}{n}\sum_{i=1}^{n} \text{Recall}_i \#(5-10)$$

宏 F1 值（Macro F1 Score）：

$$\text{Macro F1} = \frac{1}{n}\sum_{i=1}^{n} \text{F1}_i \#(5-11)$$

其中 n 是类别的数量。宏平均从整体角度更加全面地了解模型在各个类别上的性能表现，有助于评估模型的泛化能力。

5.4.1.2 车辆轨迹预测评价指标

车辆轨迹预测是指预测交通参与者在未来时刻内的物理坐标点序列，车辆包含自行车、摩托车、轿车、客车、货车等机动车。车辆轨迹预测的评价方法将实际轨迹与预测轨迹进行对比，并计算以下指标：

汽标委智能网联汽车分委会

(1) 平均位移误差（Average Displacement Error，ADE）

ADE 用于衡量预测轨迹与实际轨迹之间的平均距离，其定义为预测轨迹与实际轨迹点序列在每一时间步长内的 L2 欧氏距离的平均值：

$$ADE = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\sqrt{\left(\hat{x}_{ij} - x_{ij}\right)^2 + \left(\hat{y}_{ij} - y_{ij}\right)^2} \#(5-12)$$

式中，i为第i个预测样本；j为第j帧预测轨迹；m为总样本目标；n为预测轨迹总帧数；$\hat{x}$、$\hat{y}$为目标真实横、纵坐标；x、y为目标预测横、纵坐标。

ADE 越小，说明预测轨迹与真实轨迹的平均偏离程度越小，预测精度越高。

平均位移误差的一个变种是均方根误差(Root Mean Square Error，RMSE)。RMSE 用于衡量预测轨迹与实际轨迹之间差异，其定义为预测轨迹与实际轨迹点序列之间差值的平方和的平均值的平方根：

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(\hat{x}_{ij} - x_{ij}\right)^2 + \left(\hat{y}_{ij} - y_{ij}\right)^2} \#(5-13)$$

RMSE 越小，说明预测轨迹与实际轨迹的差异越小，预测精度越高。

(2) 最终位移误差（Final Displacement Error，FDE）

FDE 用于衡量预测轨迹的最终位置与真实轨迹的最终位置之间的距离，其定义为预测轨迹与实际轨迹序列在最后一个时间步的 L2 欧氏距离：

$$FDE = \frac{1}{m}\sum_{i=1}^{m}\sqrt{\left(\hat{x}_{n,i} - x_{n,i}\right)^2 + \left(\hat{y}_{n,i} - y_{n,i}\right)^2} \#(5-14)$$

式中，x_n、y_n分别为目标预测时长最后一帧的横、纵坐标，m 定义同上。

FDE 越小，说明预测的最终位置越接近真实的最终位置，预测的准确性越高。

(3) 分时 ADE、FDE

可针对预测的时长，将 ADE、FDE 指标细化为第 t 秒的 ADE、FDE，例如通常情况下，长时域轨迹预测的时长为 8s，可分别计算 1s、2s、3s、5s 等的误差。其中前几秒的指标结果对车用 AI 预测的性能表现尤为重要。

汽标委智能网联汽车分委会

(4) 横纵向 ADE、FDE

可针对车辆的横向、纵向行为，将 ADE、FDE 指标细化为：

横向平均、最终位移误差为

$$ADE_Lat = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\hat{x}_{ij} - x_{ij} \#(5-15)$$
$$FDE_{Lat} = \frac{1}{m}\sum_{i=1}^{m}\hat{x}_{n,i} - x_{n,i} \#(5-16)$$

式中，x、m、n、x_n定义同公式（5-12）。

纵向平均、最终位移误差为

$$ADE_Lon = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\hat{y}_{ij} - y_{ij} \#(5-17)$$
$$FDE_{Lon} = \frac{1}{m}\sum_{i=1}^{m}\hat{y}_{n,i} - y_{n,i} \#(5-18)$$

式中，y、m、n、y_n定义同公式（5-12）。

在包含道路中心线的结构化道路场景中，使用 Frenet 坐标系可以获得更直观，对下游规划任务更有指导意义的评价指标。Frenet 坐标系以道路中心线为参考，纵坐标s代表沿道路中心线的距离，横坐标d表示垂直于道路中心线的位移。用横坐标d、纵坐标s分别代替横坐标x、纵坐标y，则

横向平均、最终位移误差为

$$ADE_Lat_Frenet = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\hat{d}_{ij} - d_{ij} \#(5-19)$$
$$FDE_{Lat_{Frenet}} = \frac{1}{m}\sum_{i=1}^{m}\hat{d}_{i} - d_{i} \#(5-20)$$

式中，d为目标预测 Frenet 横坐标，d_n为目标预测时长最后一帧的横坐标，m，n定义同公式（5-12）。

纵向平均、最终位移误差为

$$ADE_Lat_Frenet = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\hat{s}_{ij} - s_{ij} \#(5-21)$$
$$FDE_{Lon_{Frenet}} = \frac{1}{m}\sum_{i=1}^{m}\hat{s}_{i} - s_{i} \#(5-22)$$

汽标委智能网联汽车分委会

式中，s为目标预测 Frenet 纵坐标，s_n为目标预测时长最后一帧的纵坐标，m，n定义同公式（5-12）。

相对于笛卡尔坐标系，Frenet 坐标系中的横纵向 ADE、FDE 指标更能表征预测轨迹与实际轨迹在驾驶意图上的差异，如两者是否具有一致的加速、制动、换道意图等，因此，也就更能评价预测模型对交通参与者行为的理解和预测能力。

(5) 针对多模态轨迹的 minADE、minFDE

针对多模态轨迹预测，评价时可在 K 条轨迹中找与真实轨迹偏差最小的轨迹进行位移误差的评测，对应的指标为ADE_1/FDE_1（K=1）、ADE_6/FDE_6（K=6）等，因此可以衍生出以下指标：

最小平均位移误差(minADE)：ADE 代表平均位移误差，minADE 则是在多个预测轨迹中，选择与实际轨迹平均距离最小的那个轨迹所对应的 ADE。反映了预测算法在最佳情况下的平均预测精度。

最小最终位移误差(minFDE)：FDE 代表最终位移误差，minFDE 是在多个预测轨迹中，选择与实际轨迹最终位置距离最小的那个轨迹所对应的 FDE。反映了预测算法在预测轨迹终点位置方面的最佳准确性。

(6) 未命中率/丢失率（Miss Rate，MR）

MR 用于衡量预测轨迹终点落在真值终点附近的比例，一般通过计算预测和实际终点距离偏差后与所设阈值比较进行判断。

当预测输出的 K 条轨迹在给定时间步 T 与目标真实轨迹偏差在横、纵向超出一定阈值时，可认为预测结果是一次丢失，计算预测丢失的比例即为丢失率：

$$MR = \frac{N_{miss}}{N} \times 100\% \#(5-23)$$

式中，N_{miss}为预测丢失的样本数量，N是全量预测样本数量。

MR 越大，说明预测算法预测的轨迹终点与真实的轨迹终点偏离越小，预测的准确性越高。

汽标委智能网联汽车分委会

(7) 可行驶区域合规性（Drivable Area Compliance, DAC)

DAC 是指轨迹落在可行驶区域的比例，即如果预测模型生成了 k 条预测轨迹，其中有 p 条在某一时刻超出了可行驶区域，则该预测模型的 DAC 为

$$DAC = \frac{k-p}{k} \#(5-24)$$

DAC 最大值为 1，指标越高，预测模型性能越好。与 MR 指标相比，MR 评价的是预测轨迹与实际轨迹终点的偏离程度，而 DAC 是在不考虑真值的情况下，评估预测的轨迹是否合理，惩罚预测模型出现少数最好轨迹，其余预测轨迹质量很差的情况。

(8) 重叠率

当预测输出的概率最高的目标轨迹在预测时长范围内与其他目标的对应时刻预测轨迹距离小于一定阈值，则可认为预测结果是一次重叠，计算预测重叠的比例即为重叠率：

$$Overlap = \frac{N_{overlap}}{N} \times 100\% \#(5-25)$$

式中，$N_{overlap}$为预测重叠的样本数量，N是全量预测样本数量。

5.4.1.3 行人轨迹预测评价指标

行人轨迹预测与车辆轨迹预测的部分指标基本一致，由于行人运动不受车辆运动学的约束，故通常不采用横向和纵向分离的评价指标。

5.4.2 决策评价主要指标简介

车用 AI 决策可分为行车安全、通行效率、舒适性三方面进行评价，每个方面均包含通用的评价指标及针对典型场景的指标。行车安全主要从影响自动驾驶系统在行驶过程中的安全性维度进行评价，通行效率主要评价自动驾驶系统能否高效执行驾驶任务及通过必要场景，舒适性包含引起用户体验下降但未出现安全事故的性能评价。指标计算的数据一方面可通过实车测试进行收集，另一方面也可通过仿真设计场景的方法进行专项测试。

汽标委智能网联汽车分委会

5.4.2.1 面向行车安全的评价指标概述

(1) MPCI （Miles Per Critical Intervention）

1) 指标含义：

两个安全接管之间自动驾驶 AI 系统能安全行驶的平均里程数。

2) 指标表达式：

$$MPCI = \frac{TotalAutonomousMiles}{TotalCriticalInterventions} \#(5-26)$$

解释：

Total Autonomous Miles 表示整个测试过程行驶的英里数。

Total Critical Interventions 表示整个测试过程中发生的危险接管次数。

3) 指标相关性：

该指标数值越大，则系统在 AI 决策规划方面的安全性方面大概率越高。

(2) 跟车纵向安全距离

1) 指标含义：

为避免碰撞，自车需与前车保持的理论安全距离下界。

即使出现前车突然急刹停情况，本车在及时反应后，一定可以在纵向上避免本次碰撞。

2) 指标表达式：

$$d_{min} = max\left(0, v_r\rho + \frac{1}{2}a_{max,accel}\rho^2 + \frac{\left(v_r + \rho a_{max,accel}\right)^2}{2a_{min,brake}} - \frac{v_f^2}{2a_{max,brake}}\right) \#(5-27)$$

解释：v_f 前车速度。v_r 后车速度。ρ反应时间。$\alpha_{min,brake}$ 最小刹车加速度。$\alpha_{max,brake}$ 最大刹车加速度。$\alpha_{max,accel}$ 最大加速度。

说明：

上述公式中的参数应该根据法规制定合理取值范围。

汽标委智能网联汽车分委会

人类驾驶汽车和自动驾驶汽车的参数可以不同。比如自动驾驶汽车的反应时间一般会比人类短，而且自动驾驶汽车可以比人类驾车的刹车更有效。因此，自动驾驶汽车的$\alpha_{max,brake}$ 可以设置得更大些，且不同路况下可以设置不同的参数（湿滑路面、冰、雪等）。

3) 指标相关性：

自动驾驶决策规划系统应该能长期保持其跟车距离在该指标指示的 dmin 范围之外，如果 AI 系统有较高频率或长时间不满足该 dmin 距离，那么决策规划系统的安全度可能较低。

(3) 横向安全距离

1) 指标含义：

为避免横向碰撞，自车需与并行车辆保持的横向距离。

即使双方同时并线情况下，本车即使反应后，一定可以在横向上避免本次碰撞。

2) 指标表达式：

$$d_{min} = \mu + max\left(0, \frac{v_1+v_{1,\rho}}{2}\rho + \frac{v_{1,\rho}^2}{2a_{min,brake}^{lat}} - \left(\frac{v_2+v_{2,\rho}}{2}\rho - \frac{v_{2,\rho}^2}{2a_{min,brake}^{lat}}\right)\right) \#(5-28)$$

解释：μ 波幅系数。v_1 本车横向速度。$v_{1,\rho}$ 时间 ρ 后的本车横向速度。v_2 对方车横向速度。$v_{2,\rho}$ 时间 ρ 后的对方车横向速度。ρ 反应时间。$a_{min,brake}^{lat}$ 最小横向刹车加速度。

说明：上述公式中的参数应该根据法规与车况制定合理取值范围。

3) 指标相关性：

自动驾驶决策规划系统应该能长期保持其跟车距离在该指标指示的 dmin 范围之外，如果 AI 系统有较高频率或长时间不满足该 dmin 距离，那么决策规划系统的安全度可能较低。

汽标委智能网联汽车分委会

(4) 碰撞事件 TTC

1) 指标含义：

该指标表示车辆在当前速度下达到前车/前方障碍物位置的时间。

2) 指标表达式：

$$TTC = \frac{d}{v} \#(5-29)$$

解释：d 当前自动驾驶车辆与前方车辆/障碍物之间的距离。v 当前车辆相对于前方车辆/障碍物的速度。

(5) 碰撞率 CR

1) 指标含义：

碰撞率表示为每百万英里行驶里程内发生的碰撞次数。

2) 指标表达式：

$$CR = \frac{Number\ of\ Crashes}{Total\ Miles\ Driven} \times 10^6 \#(5-30)$$

解释：$Number\ of\ Crashes$ 在整个测试过程中自动驾驶车辆发生的碰撞总次数。$Total\ Miles\ Driven$ 在整个测试过程中自动驾驶车辆发生的碰撞总次数。

另还有车道偏离事件、紧急制动频率、事故严重度、法规遵从性等偏向于整体性测试的安全指标，将在后续第八章中详细介绍

5.4.2.2 面向通行效率的评价指标概述

(1) 平均通行时间

1) 指标含义：

指自动驾驶系统多次通过相同或相近的路况条件下的平均每次完成通行的时间。可通过实车选择具有代表性的路况或搭建特定仿真场景后测试的方法计算指标。

2) 指标表达式：

汽标委智能网联汽车分委会

$$T_{pass} = \frac{1}{N}\sum_{i=1}^{N} t_i \ \#(5-31)$$

其中 N 为测试总次数；t_i为每次测试通行时间。

3) 指标相关性：

在相同或相近路况下，平均通行时间越短，代表自动驾驶系统通行效率越高。

(2) 换道准确率

1) 指标含义：

指以专业驾驶员换道行为作为参照评价自动驾驶系统执行换道的准确性，一般可通过收集人为驾驶数据后离线进行评价，也可在线开启自动驾驶系统影子模式与驾驶员换道行为对比。

2) 指标表达式：

$$\text{acc_change} = \frac{N_{change}}{N_{total}} \ \#(5-32)$$

其中：N_{change} – 自动驾驶系统与驾驶员换道行为一致的次数；N_{total} - 驾驶员执行换道总次数。

3) 指标相关性：

换道准确率越高，代表自动驾驶系统的换道决策拟人化程度越高。

(3) 换道回退率

1) 指标含义：

指自动驾驶系统执行换道行为过程中回退到换道前车道的比例，一般可通过实车测试或仿真测试后统计行为次数进行计算。

2) 指标表达式：

$$\text{change_back_rate} = \frac{N_{change_back}}{N_{total}} \ \#(5-33)$$

其中N_{total}-自动驾驶系统执行换道的总次数

汽标委智能网联汽车分委会

3) 指标相关性：

换道回退率越低，代表自动驾驶系统执行的有效换道效果越好。

(1) 匝道通过率

1) 指标含义：

指自动驾驶系统在上、下匝道路况下的通行能力，一般可通过实车测试或仿真测试后统计通行次数进行计算。

2) 指标表达式：

$$\text{ramp_pass_rate} = \frac{N_{ramp_pass}}{N_{total}} \quad \#(5-34)$$

其中N_{total}-自动驾驶系统测试过程中处理的上、下匝道总次数

3) 指标相关性：

匝道通过率越高，代表自动驾驶系统在上、下匝道路况下的通行能力越强。

(2) 路口通过成功率

1) 指标含义：

指自动驾驶系统在路口路况下的通行能力，一般可通过实车测试或仿真测试后统计通行次数进行计算。该指标根据评测需求可进一步细分为路口直行通过成功率、路口左转通过成功率、路口右转通过成功率等。

2) 指标表达式：

$$\text{intersection_pass_rate} = \frac{N_{intersection_pass}}{N_{total}} \quad \#(5-35)$$

其中N_{total}-自动驾驶系统测试过程中处理的路口通行总次数。

3) 指标相关性：

路口通过成功率越高，代表自动驾驶系统在路口路况下的通行能力越强。

(3) 绕障成功率

1) 指标含义：

汽标委智能网联汽车分委会

指自动驾驶系统在针对障碍物进行绕障的通行能力，一般可通过实车测试或仿真测试后统计通行次数进行计算。

2) 指标表达式：

$$\text{obstacle_pass_rate} = \frac{N_{obstacle_pass}}{N_{total}} \#(5-36)$$

其中N_{total}-自动驾驶系统测试过程中处理的绕障总次数

3) 指标相关性：

绕障成功率越高，代表自动驾驶系统在绕障场景下的通行能力越强。

(4) 交互一致率

1) 指标含义：

指自动驾驶系统在处理与周围交通参与者的交互博弈如先行、让行等行为上双方达成一致通行意见的比例。一般可通过实车测试统计次数进行计算。

2) 指标表达式：

$$\text{agreement_pass_rate} = \frac{N_{agreement_pass}}{N_{total}} \#(5-37)$$

其中N_{total}-自动驾驶系统测试过程中处理的交互总次数。

3) 指标相关性：

交互一致率越高，代表自动驾驶系统处理与交通参与者的交互博弈行为的能力越强。

5.4.2.3 面向驾乘舒适性的评价指标概述

(1) 误制动平均发生里程(Miles Per False Brake)

1) 指标含义：

指自动驾驶系统在行驶过程中两次发生误制动间隔的平均里程数。一般通过实车体验结合场景进行计算。

2) 指标表达式：

清标委智能网联汽车分委会

$$\mathrm{MPFB} = \frac{M}{N_{False_brake}} \#(5-38)$$

其中M-自动驾驶系统行驶总里程；N_{False_brake}-发生误制动的次数。

3) 指标相关性：

误制动平均发生里程越高，代表自动驾驶系统发生误制动的频率越低，决策规划舒适性越好。

(2) 急加速平均发生里程(Miles Per Rapid Acceleration)

1) 指标含义：

指自动驾驶系统在行驶过程中两次发生异常急加速间隔的平均里程数。一般通过实车体验结合场景进行计算。

2) 指标表达式：

$$\mathrm{MPRA}=\frac{M}{N_{Rapid_acceleration}} \#(5-39)$$

其中M-自动驾驶系统行驶总里程；$N_{Rapid_acceleration}$-发生异常急加速的次数。

3) 指标相关性：

异常急加速平均发生里程越高，代表自动驾驶系统发生异常急加速的频率越低，决策规划舒适性越好。

(3) 抖动平均发生里程(Miles Per Jerk)

1) 指标含义：

指自动驾驶系统在行驶过程中两次发生异常横、纵向抖动间隔的平均里程数。一般通过实车体验结合场景进行计算。

2) 指标表达式：

$$\mathrm{MPJ} = \frac{M}{N_{Jerk}} \#(5-40)$$

其中M-自动驾驶系统行驶总里程；N_{Jerk}-发生异常抖动的次数。

3) 指标相关性：

汽标委智能网联汽车分委会

抖动平均发生里程越高，代表自动驾驶系统发生异常抖动的频率越低，决策规划舒适性越好。

(4) 急转弯平均发生里程(Miles Per Hard Turn)

1) 指标含义：

指自动驾驶系统在行驶过程中两次发生异常急转弯的平均里程数。一般通过实车体验结合场景进行计算。

2) 指标表达式：

$$\mathrm{MPHT} = \frac{M}{N_{Hard_turn}} \quad \#(5-41)$$

其中M-自动驾驶系统行驶总里程；N_{Hard_turn}-发生异常急转弯的次数。

3) 指标相关性：

急转弯平均发生里程越高，代表自动驾驶系统发生异常急转弯的频率越低，决策规划舒适性越好。

(5) 行驶压迫感平均发生里程(Miles Per Driving Pressure Experience)

1) 指标含义：

指自动驾驶系统在跟车或与邻车道车辆并行时用户感受到压迫、不安全但未实际发生碰撞的平均里程数。一般通过实车体验结合场景进行计算。

2) 指标表达式：

$$\mathrm{MPDPE} = \frac{M}{N_{Driving_pressure_experience}} \quad \#(5-42)$$

其中M-自动驾驶系统行驶总里程；$N_{Driving_pressure_experience}$-发生行驶压迫感的次数。

3) 指标相关性：

行驶压迫感平均发生里程越高，代表自动驾驶系统使用户感受到压迫的频率越低，决策规划舒适性越好。

汽标委智能网联汽车分委会

(6) 极限换道平均发生里程（Miles Per Extreme Lane Change）

1) 指标含义：

指自动驾驶系统在换道时道路、车流等换道条件过于极限从而使用户感受到不安全，但未实际发生碰撞的平均里程数。一般通过实车体验结合场景进行计算。

2) 指标表达式：

$$\mathrm{MPELC} = \frac{M}{N_{Extreme_lane_change}} \quad \#(5-43)$$

其中M-自动驾驶系统行驶总里程；$N_{Extreme_lane_change}$-发生极限换道的次数。

3) 指标相关性：

极限换道平均发生里程越高，代表自动驾驶系统在换道时使用户感受不安全的频率越低，决策规划舒适性越好。

5.5 总结与展望

5.5.1 车用 AI 预测与决策的标准化分析

统一数据格式：车用 AI 预测与决策系统需要处理多源异构数据，需要统一感知系统、地图系统与预测决策模块之间的数据交互格式，包括动态障碍物信息、道路先验信息以及预测结果的表达方式，以确保不同模块之间的高效协同。

(1) 定义统一的动态障碍物数据格式，包括对象类型、位置坐标、运动状态、尺寸信息以及置信度等，确保感知系统与预测决策模块之间的数据交互无碍。

(2) 制定高精度地图数据的统一表达规范，包括车道线类型、交通标志、道路拓扑结构以及特殊区域，为预测与决策提供准确的先验知识。

(3) 确保所有数据具有统一的时间戳和坐标系，以解决多传感器数据的时间对齐和空间对齐问题，避免因数据不同步导致的预测误差。

统一测试流程：为评估车用 AI 预测与决策系统的性能，需要建立统一的测试场景库和评估指标体系，涵盖典型交通场景和极端场景，以量化评估预测与决

汽标委智能网联汽车分委会

策算法的准确性、鲁棒性和实时性。

(1) 构建覆盖典型交通场景和极端场景的标准化测试场景库，包括城市道路、高速公路、交叉路口、环岛等常见场景，以及恶劣天气（雨雪雾）、传感器失效、突发障碍物等极端场景，确保评估的全面性和代表性。

(2) 定义量化评估指标，实现驾驶意图预测、车辆轨迹预测、行人轨迹预测评价标准化，面向行车安全、通行效率、驾乘舒适性评价标准化。

(3) 制定统一的测试流程和方法，包括仿真测试、封闭场地测试和开放道路测试，明确测试数据的采集方式、测试环境的配置要求以及测试结果的记录与分析方法，确保评估结果的可重复性和可比性。

5.5.2 车用 AI 预测与决策的标准化总结与展望

车用 AI 预测与决策系统的标准化工作涉及技术、工程实际应用等多个层面，除了上文所述数据格式和评价标准的统一需求，还包括接口标准化、数据标注标准化和测试标准化等方面的挑战。

首先，车用 AI 系统通常由多个硬件模块（如传感器、计算平台）和软件模块（如感知算法、决策算法）组成。不同厂商提供的硬件和软件可能采用不同的接口标准，导致系统集成困难。想要在保证硬件与软件接口的兼容性的前提下，设计完全统一的接口存在难度。同时车用 AI 预测与决策系统需要实现车内各组件之间以及与云端、其他车辆、路端基础设施的高效通信，还要考虑不同组件可能采用不同的通信协议。统一通信协议，确保数据传输的高效性和可靠性，是车用 AI 预测与决策系统接口标准化中的一大挑战。

此外，车用 AI 预测与决策系统的训练和测试依赖于大量标注数据，需要处理来自不同传感器和地图系统的多源数据，这些数据在格式、精度、更新频率等方面存在显著差异，不同数据集可能采用不同的标注规范，标注的质量和精度也不统一。数据标注的标准化需要建立一套标注质量控制机制，确保标注数据的准确性和完整性，制定一套统一的标注规范，确保不同数据集之间的一致性。

汽标委智能网联汽车分委会

第六章 端到端系统标准化研究

6.1 研究背景与范围

传统的自动驾驶架构更多是从机器人架构衍生出来的，因此，在机器人领域的感知-定位-规划三大模块以及基于此种模块化架构进行的功能衍生成为了自动驾驶架构发展的基调。如下图 6-1 展示了传统自动驾驶框架所涵盖的典型模块划分。

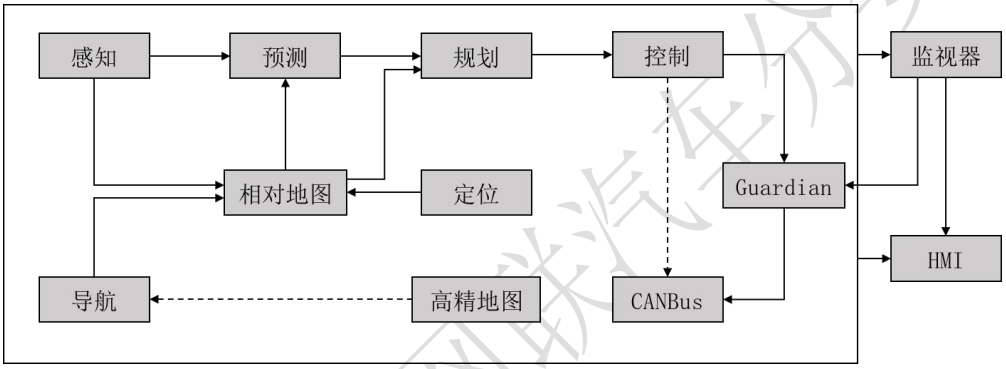


图 6-1 传统自动驾驶架构的典型代表——Apollo 3.0 软件架构[107]

典型的自动驾驶系统通常包含以下几个核心模块：感知模块（Perception）、定位模块（Localization）、预测模块（Prediction）、决策与规划模块（Decision and Planning）、控制模块（Control）。感知端输入摄像头、激光雷达等传感器信息，能够感知到车辆周围的障碍物、车辆、车道线、红绿灯等元素；然后感知信息会进一步汇总传递给规划模型，规划模型规划出车辆最佳行驶路线；再交给控制模块，实现车辆最终的控制。传统方案在模块上的划分使得每个模块的独立开发更加容易，在进行问题追溯时也更加简单快速。然而，这种方案的代价是限制了模块及整个系统所能达到的灵活性和泛化性的上限。

自动驾驶需要能够处理各种复杂和不可预测场景。传统的模块化方法侧重于将驾驶任务划分为多个标准化模块，使用基于规则的方法将这些不同的模块有机结合。而这些现行的传统方法需要花费大量精力来设计并组合所有系统组件于系统中，其中每个算法、功能，如定位、感知、预测以及规划，都是单独开发并集

成上车的。例如目前市面上大部分智驾系统中的感知模块使用的是神经网络，处于下游的规划模型需要依赖工程师编写大量代码去制定行驶规则。

这种传统自动驾驶系统中，各模块方法可解释，但应用中依然存在一些问题。首先，基于人为制定规则的方式无法穷尽所有的道路状况，特别是极端状况（Corner case）下，会提升决策规划控制的难度。其次，分模块方案的信息流转的层级较多、传递的链路较长，所以信息失真情况会更严重，各模块之间传递的信息会出现有损压缩，传递时会累计误差，还会面临并行模块信息相互干扰的问题。越来越多的技术迭代方向表明其趋势是利用大规模数据驱动，并将基于学习的不同指令结果作为一种可行的替代方案[108]。如下图 6-2 所示为传统自动驾驶与端到端自动驾驶方法系统框架对比。

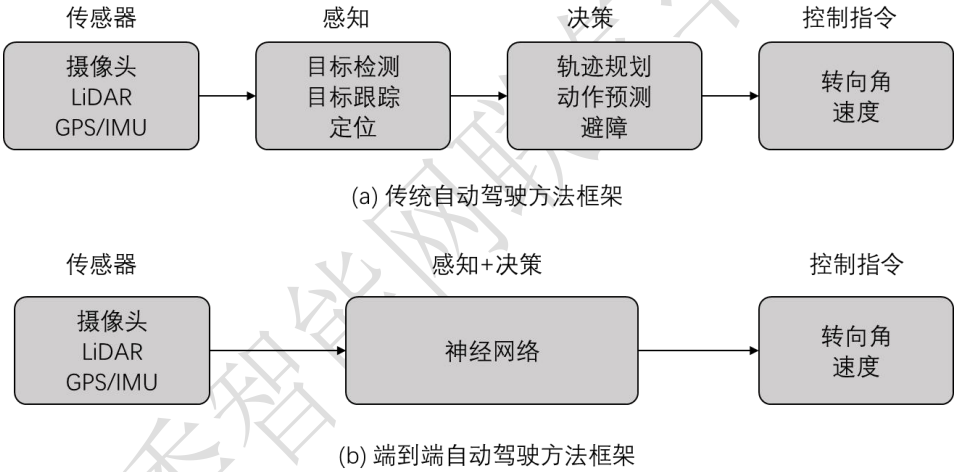


图 6-2 端到端方法和传统方法的架构对比

6.1.1 端到端系统介绍

与传统自动驾驶算法结构相比，端到端定义尚未统一，其主要特点在于深度学习的全面使用与数据驱动。深度学习是一种机器学习方法，其目标是通过构建多层神经网络来模拟人脑的工作原理，让机器能够从大量的数据中自动学习和提取特征，从而实现智能化的数据处理和决策。神经网络是深度学习的核心组成部分，深度学习采用多层神经网络模型，其中包含输入层、多个隐藏层和输出层。

“端到端”概念根据其应用和实现方式，主要分为以下四类：

感知“端到端”：自动驾驶系统架构在此类别中被分为感知模块与预测决策规划模块两大部分。感知模块借助多传感器融合的 BEV（Bird's Eye View，鸟瞰视角）技术，实现了模块内部的“端到端”处理。通过引入 Transformer[109]架构以及跨传感器的交叉注意力机制，相比传统感知方案，检测结果的精度和稳定性得到了显著提升。然而，规划决策模块在这一阶段仍主要依赖规则驱动。

决策规划模型化：这一类别的自动驾驶架构依然保持感知和预测决策规划两大模块的划分。感知模块沿用上一代技术方案，而预测决策规划模块则发生了重大变化：预测、决策、规划这三个功能被集成到同一个神经网络中。尽管感知和预测规划决策都采用了深度学习技术，但两者之间的接口仍基于人类定义的理解，如障碍物位置和道路边界等信息。此外，在这一阶段，各个模块仍然进行独立训练。

模块化端到端：从整体结构上看，这一类与上一类相似，但在网络结构和训练方案上存在显著差异。感知模块不再输出基于人类理解的结果，而是提供特征向量。相应地，预测决策规划模块的综合模型基于这些特征向量输出运动规划结果。在训练方面，这类模型必须支持跨模块的梯度传导，两个模块无法独立训练，需同时进行。

One Model 单一模型端到端：这是“端到端”技术的高级形式，不再明确区分感知、决策规划等功能模块。从原始信号输入到最终的规划轨迹或控制指令输出，整个过程由同一个深度学习模型完成。根据实现方案的不同，这一阶段的 One Model 可以是基于强化学习或模仿学习的端到端模型，也可以通过世界模型这类生成式模型衍生而来。

我们将端到端自动驾驶系统定义为：“端到端自动驾驶系统是一种智能计算系统，它以原始传感器数据（如相机图像、激光雷达数据等）为输入，完整地利用感知到的信息，通过一个统一的神经网络模型，输出规划轨迹或控制指令，实现自动驾驶的全局优化。”端到端自动驾驶系统减少了多模块间的冗余信息传递

汽标委智能网联汽车分委会

和误差叠加，在动态应对复杂场景时反应更迅速，并在开发流程上实现不同部门与功能的高度整合。

端到端系统在保持整体性与高效性的同时，也呈现出两面性：一方面，通过将感知、决策与规划控制深度耦合，减少延迟和误差累积，达到比传统模块化方法更高的性能上限；它依托多样化数据的持续学习，能够迅速适配复杂场景并带来更贴近人类驾驶风格的用户体验，研发流程也因此更为精简。另一方面，该模式在可解释性上往往面临挑战，网络内部的判定过程难以追踪，当系统发生决策失误时难以迅速定位根源，且对高质量多样数据的依赖度极高，数据采集和大规模训练的成本也可能成为瓶颈。要在实用层面充分释放端到端自动驾驶的潜能，如何在性能、可解释性和数据需求之间取得平衡，仍是产业与研究领域亟待突破的关键命题。

6.1.2 端到端自动驾驶技术发展和现状

6.1.2.1 技术发展

受复杂场景和长尾问题等因素的影响，近年来采用端到端算法框架方法的方案数量正快速增长，这些方法利用原始传感器输入来生成车辆运动计划，而不是专注于检测和运动预测等单个任务。其技术发展的历程中每一部分都有显著的范式转变或者性能提升。

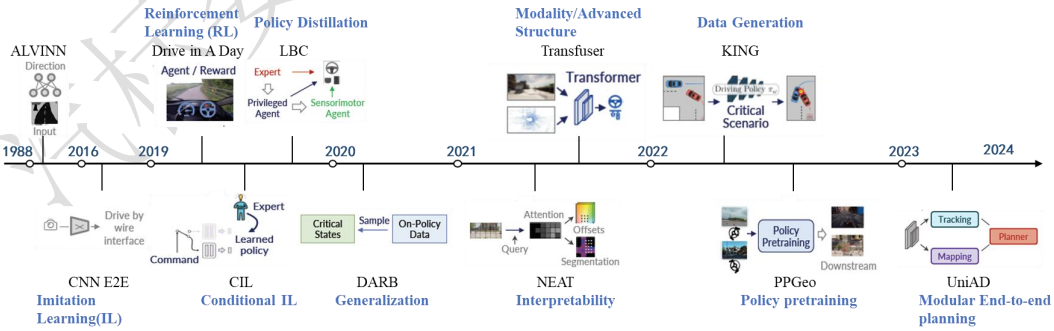


图 6-3 端到端自动驾驶方法的发展历程[110]

端到端自动驾驶的历史最早可以追溯到 1988 年的 ALVINN[111]，当时的输入是来自相机和激光雷达的两个外部输入传感器，以及一个简单的神经网络生成的

方向盘转向作为输出。从模型和训练方法的角度，ALVINN 标志着端到端自动驾驶的起点。随着时间的推移，模仿学习成为端到端自动驾驶的关键技术之一。IL 通过模仿专家的行为来进行训练，其中包括行为克隆和逆向最优控制等方法。随后，强化学习在自动驾驶中应用起来，尤其是在虚拟环境中，为模型提供了通过试错学习的机会。为了提高学习效率，策略蒸馏技术被提出，通过训练一个具有访问环境状态的特权代理，然后将策略蒸馏到最终的学生网络。多任务学习（MTL）联合执行多个相关任务，MTL 利用共享表示来提高模型的泛化能力、鲁棒性和一致性。未来包括零样本和少样本学习、模块化端到端规划、大型基础模型的应用，以及车辆到一切（V2X）系统的集成也将快速发展。

从信号输入源及模态来看，随着技术的发展，对比单一模态输入，多种传感器如雷达（LiDAR）、立体相机、IMU 和 GPS 等被集成到自动驾驶系统中，提供了更丰富的环境信息。随后，更高级的模态，如 3D 对象检测、语义分割和运动预测的加入提高了决策质量。后来出现的统一空间建模发展了将不同模态融合在统一空间（如 BEV）的技术，以更好地理解和预测环境。未来，自动驾驶系统将开始探索结合使用自然语言指令、图像场景理解等作为输入的端端系统。

ALVINN 属于模仿学习的范式，尽管并没有直接应用在商业产品中，但它是一个具有里程碑意义的自动驾驶项目，它在早期阶段就探索了神经网络在端到端自动驾驶中的应用。随后的 CNN E2E 设计了一个用于模拟和道路测试的端到端卷积神经网络原型系统，在 GPU 计算的新时代重新确立了端端的技术方向。

2019 年的 Learning to Drive in A Day[112]展示了深度强化学习在自动驾驶中的首次应用。采用一个单目图像作为输入，在少数训练集中就能够学习到车道跟踪的策略。方法提供了一个通用且易于获得的奖励或者说是回报（Reward），即：车辆在没有安全驾驶员控制的情况下行驶的距离。方法使用的是一种连续的、无模型的深度强化学习算法，所有的探索和优化过程都在车端进行，是自动驾驶摆脱以定义好的逻辑规则、高精地图和直接监督学习作为依赖的一个新框架。

汽标委智能网联汽车分委会

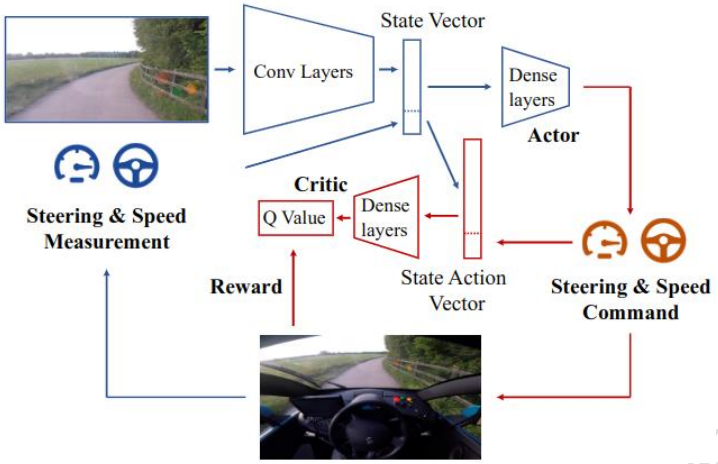


图 6-4　Learning drive in a day 架构[112]

2021 年的 TransFuser[113]以及他的后续 NEAT[114]实现了跨模态端，并通过网络结构改进，即 transformer 的引入和其他网络层的结构调整优化从而取得了更好的成绩。首先它们证明了基于现有传感器融合方法的模仿学习策略在存在高密度动态代理和复杂场景的情况下表现不佳，例如处理从多个方向迎面而来的交通不受控制的交叉路口这种工况，必须引入全局上下文推理才行。而 NEAT 提出了神经注意场 Neural Attention fields，用一个连续函数把 BEV 场景坐标的位置映射到航路点和语义，用中间注意力映射（Intermediate attention maps）将高维 2D 图像特征迭代压缩。这套方法允许模型选择性地关注相关输入区域，同时忽略与驾驶任务无关的信息，这样有效地将图像与 BEV 表征相关联。

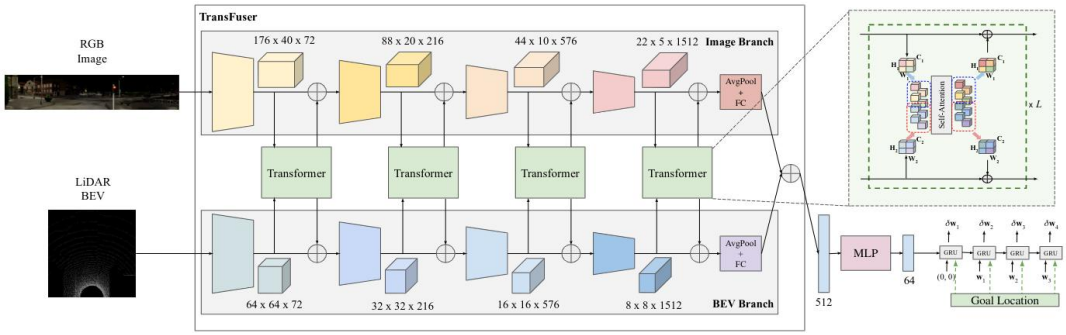


图 6-5 TransFuser 架构[113]

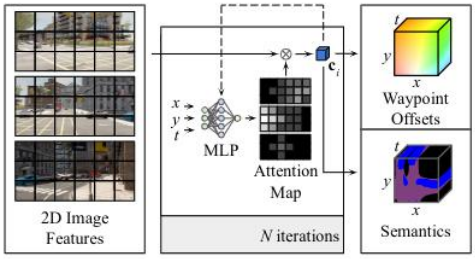


图 6-6 NEAT 使用 MLP 基于 BEV 查询位置将高纬输入压缩为低纬表示[114]

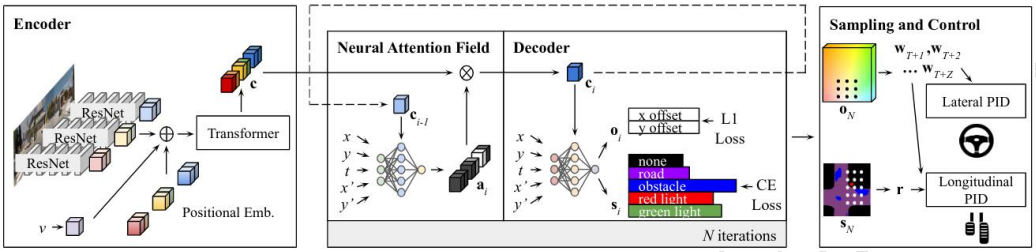


图 6-7 NEAT 模型架构[114]

2022 年的 PPGeo[115]是由 OpenDriveLab 团队提出的基于自监督几何建模的端到端自动驾驶策略预训练框架。PPGeo 利用大量无标注的网络驾驶视频，以自监督的方式对驾驶策略进行预训练。比如，静态障碍物和背景建筑物与驾驶决策无关；视觉输入中的交通信号仅占图片的很小一部分，但对于控制输出而言却是确定性的；端到端驾驶策略模型必须对不同的光照和天气条件具有鲁棒性。PPGeo 分为两个阶段，在第一阶段，根据时序前后帧图像训练位姿和深度预测网络；在第二阶段，利用第一阶段训练的深度网络和位姿网络监督单图像输入的视觉编码器，完成视觉编码器的预训练。于是，将使用 PPGeo 预训练后的视觉编码器进行微调，就可以适配于下游不同的端到端自动驾驶任务。

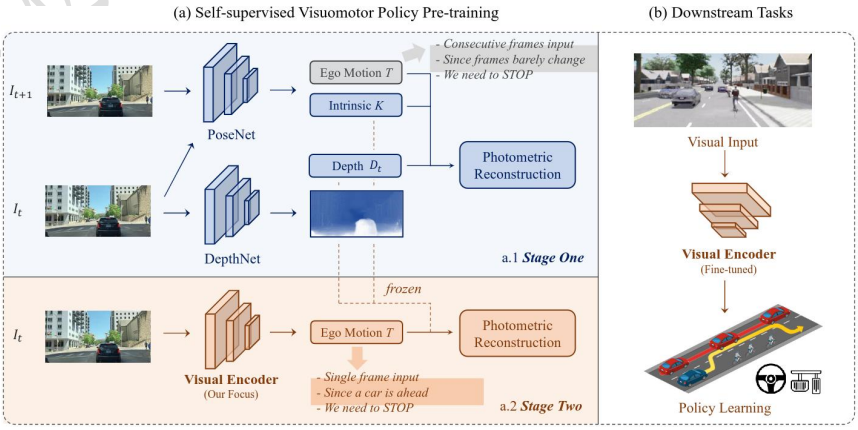


图 6-8 PPGeo 架构[115]

2023 年以上海人工智能实验室为学术界代表，UniAD[116]是一个模块化端端综合框架，它将全栈驾驶任务整合到一个基于 Transformer 的端到端网络中。除了感知、预测和规划之外，还包括六个子任务，包括目标检测、目标跟踪、场景建图、轨迹预测、栅格预测和路径规划。UniAD 是由四个基于 Transformer 解码器的感知预测模块以及一个规划模块组成。多组查询向量 (Query) 用于连接整个架构以及学习多个智能体与周围环境的交互。协调具有多个任务的综合系统是一项非常复杂的任务，并且需要大量的计算资源，尤其是在具有时间历史训练的情况下。因此，对于 UniAD 而言如何设计和优化该系统以实现轻量级部署是值得进一步探索的问题。此外，是否将更多任务（如深度估计、行为预测）纳入系统中以及如何将它们嵌入到系统中，也是值得进一步研究的方向。

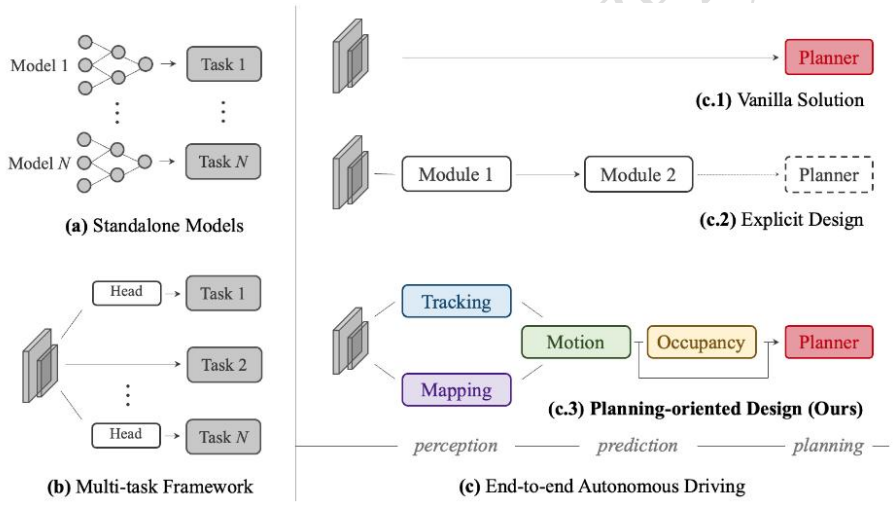


图 6-9 UniAD 总结的模型架构区别[116]

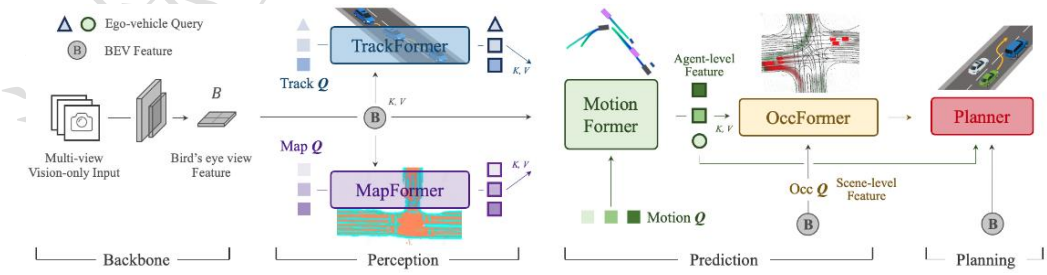


图 6-10 UniAD 的模型 Pipline 示意图[116]

2024 年清华大学的 DriveVLM-Dual[117]，是一种利用视觉语言模型来增强场景理解和规划能力的自动驾驶系统，并集成了用于场景描述、场景分析和分层规

划的独特组合。DriveVLM 接受图像序列作为输入，并通过基于推理的思维链 (CoT) 机制输出分层规划预测。它可以选择性地结合传统的 3D 感知和轨迹规划模块，以实现空间推理能力和实时轨迹规划。双系统分快慢系统，模型本身分为三大模块，场景描述、场景分析和分层规划。场景描述模块从语言上描绘了驾驶环境并识别场景中的关键物体；场景分析模块深入探讨关键物体的特性及其对 ego 车辆的影响；分层规划模块逐步制定计划，从元动作和决策描述轨迹点。而快慢系统则结合了传统系统和模型系统的优势。

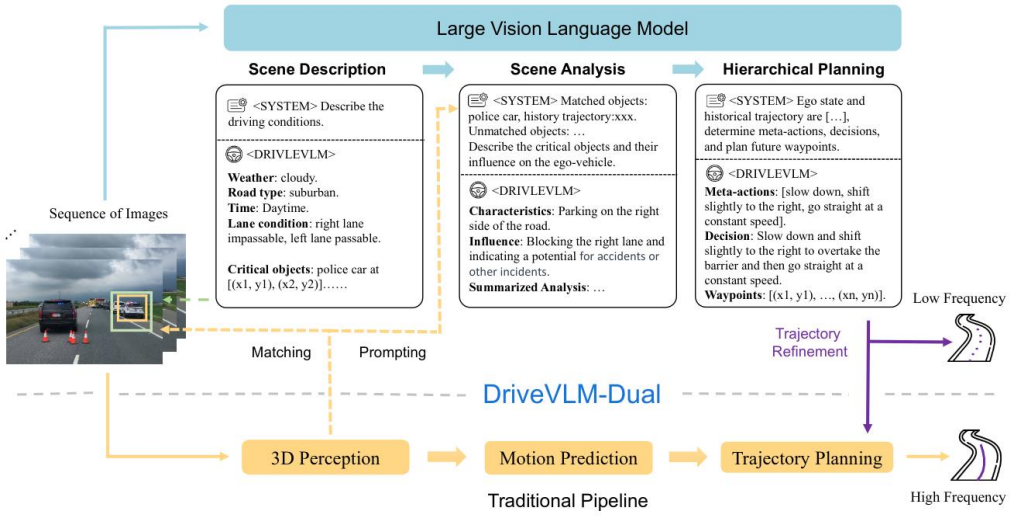


图 6-11 DriveVLM 双系统运行流程[117]

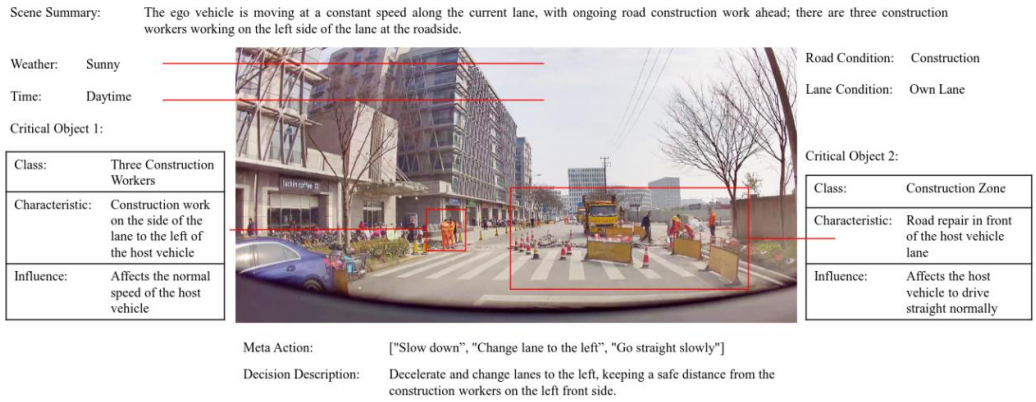


图 6-12 来自 SUP-AD 的标注示例[117]

2024 年开源的 OpenEMMA[118]是一个基于多模态大型语言模型的端到端自动驾驶框架。它结合车辆历史数据和前置摄像头图像，运用思维链（Chain-of-Thought，CoT）推理预测未来速度与曲率，并融入轨迹规划。借助微

调的外部视觉专家模型，OpenEMMA 实现了 3D 路面物体的精确检测。该框架相比零样本学习（Zero-shot learning）基线有显著改进，在多种挑战性驾驶场景中展现出有效性、泛化能力和鲁棒性。作为使用现成预训练模型开发端到端自动驾驶框架的初步尝试，OpenEMMA 在推理时仅采用基本 CoT 推理。尽管这是基础方法，但将更高级的推理技术（如 CoTSC 和 ToT）集成到框架中，仍可大幅提升框架实用性，有望为自动驾驶提供更有效的方法。

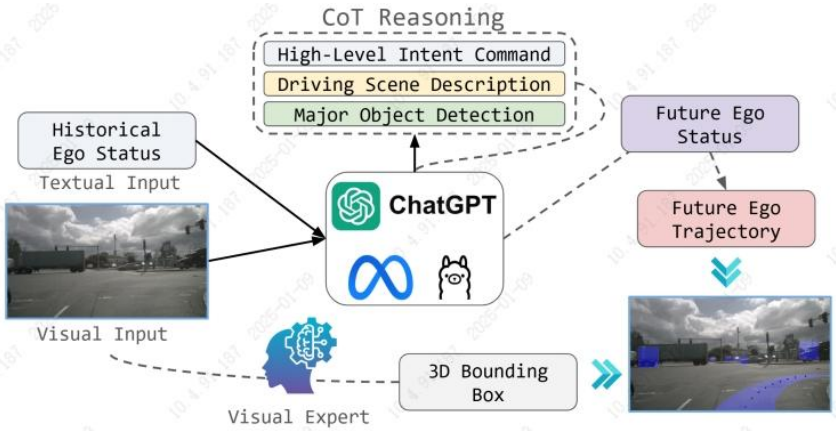


图 6-13 OpenEMMA 框架示意图[118]

2024 年，中科院和哈工大联合提出了 SparseAD[119]稀疏架构。该架构通过稀疏查询全面表征跨时空的驾驶场景，完全避免了使用稠密的 BEV 表示。SparseAD 集成了检测、跟踪、在线建图等感知任务，并引入了一个更为合理的运动规划框架。在 nuScenes 数据集上，该架构在各项任务上均超越了 UniAD，达到了最新的 SOTA（State-of-the-Art，最先进技术水平）。

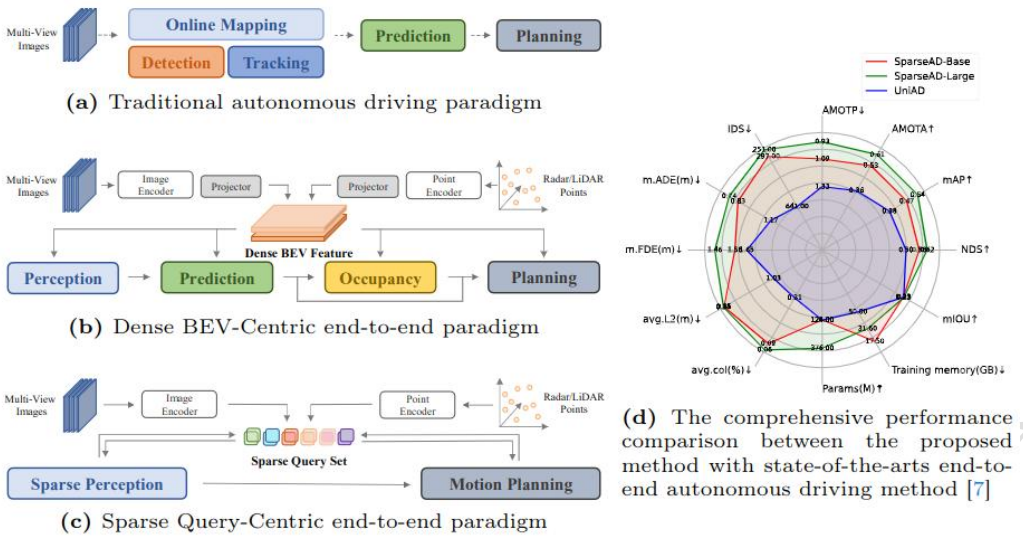


图 6-14 主流自动驾驶流程与 SparseAD 的比较[119]

SparseAD 的端到端架构简化为稀疏感知和运动规划器两个部分。在稀疏感知中，感知任务（如检测、跟踪和在线地图）通过通用时间解码器进行统一处理，使用多传感器特征和历史记忆作为 Token，分别通过对象查询和地图查询来表示驾驶场景中的障碍物和道路元素。运动规划则利用稀疏感知查询作为环境表示，对自车及周围车辆进行多模态运动预测，生成各种初始解。随后，系统结合多维度的驾驶约束，生成最终的运动规划。此外，该架构通过端到端多任务存储库统一存储驾驶场景的时间信息。

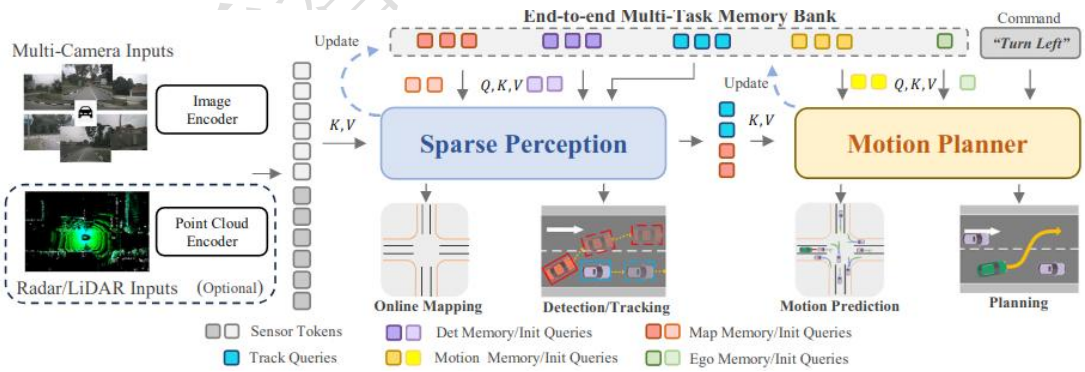


图 6-15 SparseAD 流程示意图[119]

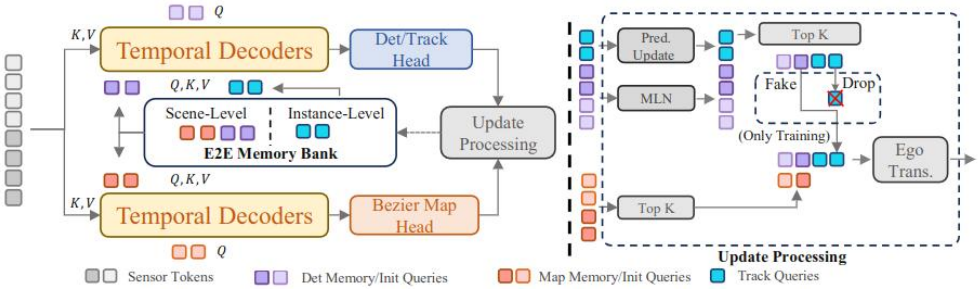


图 6-16 Sparse Perception[119]

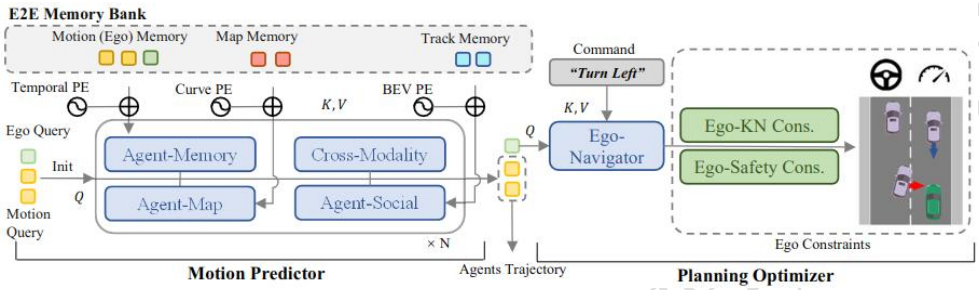


图 6-17 Motion Planner 和 Optimizer[119]

总而言之，端到端自动驾驶技术的发展轨迹展示了自动驾驶系统从基于规则的决策向更智能化、自适应算法架构逐步演进的趋势。通过多传感器融合以及深度学习与强化学习的结合，显著提升了系统的精度和鲁棒性，为应对复杂场景和自主决策奠定了坚实基础。

6.1.2.2 技术现状

2024 年初，特斯拉发布了 FSD V12.3 版本。端到端自动驾驶方案受到国内主机厂和自动驾驶企业的广泛关注。相比于传统的多模块方案，端到端方案将感知、预测和规划功能整合为单一模型，简化了系统结构，可以有效解决传统模块化方案中的长尾场景问题，显著提升模型的训练效率和性能上限。

表 6-1 传统多模块方案与端到端方案的对比

方案类型	模块化自动驾驶	端到端自动驾驶
驱动类型	规则驱动（Rule-based)编写代码制定规则	数据驱动（Data-driven）用海量数据去训练系统
任务方式	多任务学习	单任务学习
模型	分为多个模块，每个模块采用独立模型	采用端到端一体化大模型
泛化性	较差	良好

表 6-2 部分主机厂对端到端方案落地量产的规划

OEM	上车日期	实施现状	方案特点
特斯拉	2023 年 9 月，测试版 V12（FSD Beta V12）	FSD V12（Supervised）是全新的“端到端自动驾驶”，模型架构发生了重大变化。	完全采用神经网络进行车辆控制，该神经网络由数百万个视频片段训练而成，取代了超过 30 万行的 C++ 代码
蔚来	2024 上半年	2024 下半年开发研发，预计 2024 年量产	基于世界模型生成数据进行训练和仿真
小鹏	2024 年	2024 年 1 月公布上车计划	构建 600FLOPS 的 GPU 集群用于训练
理想	2024 年 10 月	OTA 推送端到端+VLM 方案	双流架构快慢智驾系统
小米	2024 年	2023 年底公布端到端感知决策大模型、2024 年 3 月小米 SU7 搭载	实时生成道路拓扑实时识别静态智能体
吉利汽车	2024 年	与鉴智机器人合作，预计 2024 年量产	采用动态场景图预测智能体碰撞可能
极越	2024 年	迭代 VTA 大模型，开发训练 BEV 端到端感知模型	实现道路元素完全覆盖实时生成道路拓扑

FSD 在美国和加拿大都已经开通，Tesla 的 FSD Beta v12 端到端实现输入图像，输出转向、刹车、加速等车辆控制信号的能力。工业界之所以能够在 Tesla 首先看到端端技术落地，是由于其内部强大的科学研究资源及足够的数据量等综合因素所取得的优势，而国内的自动驾驶总体还是跟随者的形式，业界过去几年都会关注 Tesla AI day 以获取启发。FSD 是一个纯视觉方案，简单来说，将“感知”和“决策”两个模块融合到一个模型中，通过纯视觉感知，利用影子模式实时收集数据并训练模型，来实现仅依靠车载摄像头感知道路和交通情况，并做出相应决策。

2024 年 5 月 15 日，百度 Apollo 发布了支持 L4 级自动驾驶的大模型 Apollo ADFM（Autonomous Driving Foundation Model）。百度的端到端架构属于两段式端到端，感知和决策规划有两段网络，后续通过隐形传递和联合训练实现端到端无人驾驶。在感知端，通过摄像头得到视觉图像数据，并通过激光雷达和 4D 毫米波雷达得到主动光传感器数据。之后利用感知网络对视觉和主动光特征进行识别，融合 BEV 特征判断道路特征、障碍物等。再通过显式三维向量空间（使用

三维坐标系来精确地表示物体的位置和方向）和隐式 BEV 特征（在 BEV 视角下，不直接给出三维空间中的所有信息，而是通过一些间接的方式，如特征编码或者深度学习模型来隐式地表示和推断这些信息），以联合训练的方式实现端到端无人驾驶。

华为乾崑 ADS 3.0 端到端是基于少量人工规则+端到端模型，并逐渐减少人工规则的数量。乾崑 ADS3.0 主要包括 GOD 和 PDP。华为的 ADS 是从感知 BEV 网络开始的，可识别白名单障碍物；2023 年 4 月发布最新的 ADS3.0，引入 GOD（通用障碍物检测）和 PCR（道路拓扑推理）网络，现在开始进入场景理解阶段，构成了 GOD 大网。PDP（预测决策规控）网络：乾崑 ADS3.0 的全新架构采用 PDP 网络，实现了预决策和规划一张网。

2024 年 5 月 21 日，小鹏汽车发布量产上车的端到端大模型。此次小鹏量产上车的端到端大模型包括：感知神经网络 XNet+规控大模型 XPlanner+大语言模型 XBrain。从结构来看，小鹏的端到端架构仍分为感知、规控两大功能模块，其中最大的变化应该在于将过去基本完全由规则组成的规控模块，切换为神经网络为主。

2024 年 7 月 5 日，理想公开了其端到端自动驾驶技术架构，该架构主要由端到端模型、VLM 视觉语言模型、世界模型三部分共同构成，从技术架构上看，比分段式端到端更近一步。

当前国内自动驾驶公司在量产车中，模块化架构依然占据主导地位。这反映出端到端架构在实际应用中仍面临一些关键挑战尚未解决，例如对大规模高质量数据的依赖以及对复杂计算资源的高要求。例如，地平线的 VAD v2、理想的 DriveVLM 和毫末的 DriveGPT 等项目在端到端自动驾驶领域进行了深入探索。这些项目均强调，数据资产的转化（如将原始数据转化为可用于模型训练的结构化数据）和强大的数据处理能力是实现端到端架构的核心需求。

在研发过程中，很多端到端自动驾驶方法依赖仿真环境（如 Carla）进行验证，以模拟真实世界的驾驶场景。这种方式能够在一定程度上弥补真实数据的不

汽标委智能网联汽车分委会

足，但也暴露出端到端架构在实际部署中的局限性，例如难以完全适应复杂多变的现实环境。因此，对于云端服务提供商而言，构建强大的数据智能工程能力尤为重要。这包括高效的数据闭环系统（数据采集、标注到模型优化的循环）、完善的测试评估机制以及先进的仿真训练平台。国内外相关企业，如地平线、理想和毫末等，都在积极建设算力中心，以满足数据智能的需求。这些努力不仅有助于解决端到端架构的现有挑战，也为自动驾驶技术的未来发展奠定了基础。

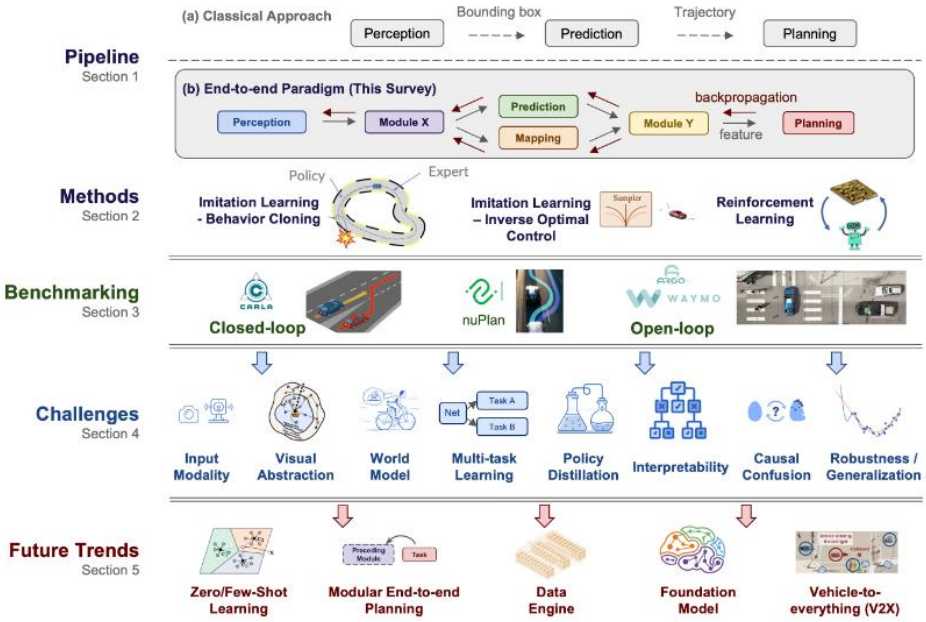


图 6-18 端到端自动驾驶技术现状示意图[120]

6.1.3 问题与挑战

6.1.3.1 应用与技术层面

在端到端自动驾驶系统中，面临着一系列复杂且相互关联的挑战。这些挑战不仅独立存在，而且它们之间相互影响，共同塑造了自动驾驶技术的发展轨迹。

首先，数据的输入处理构成了系统的基础。有效的输入模态融合是获取全面环境信息的关键，这直接影响到视觉抽象的质量。视觉抽象通过将原始数据转换为紧凑的特征表示，为系统的决策和控制提供支持。

接着，在数据处理与先验知识方面，确保数据的多样性和有效引入先验知识对于模型的训练效果至关重要。数据多样性提高了模型对不同场景的适应能力，而先验知识的引入能够加速模型的训练过程并提升性能。

在模型的学习和优化方面，如何处理因果混淆是一个关键问题。因果混淆可能导致系统对因果关系的误解，从而影响决策的准确性。与此同时，构建准确的世界模型可以帮助系统更好地预测未来状态，从而改善因果关系的处理，并优化决策策略。

在系统性能和评估方面，稳健性和评测手段密切相关。有效的评测方法能够帮助识别和改进系统在各种场景下的表现，从而提高系统的稳健性和稳定。此外，端到端系统在解决边缘情况时的路径清晰性也是一个关键问题，这关系到系统在面对不常见或极端情况时的应对能力。最后，域自适应性是确保系统在真实世界中有效运行的关键。有效的域自适应策略需要结合数据多样性和评测手段，以处理不同环境条件下的挑战。这包括从模拟到真实的转换、不同地理位置的适应、不同天气条件下的传感器输入变化、昼夜变化对传感器输入的影响，以及不同传感器特性的适应。

综合来看，解决这些挑战需要一个多维度的策略，包括技术创新、跨学科合作、以及对现有问题深入理解的系统性解决方案。通过这些努力，可以实现一个既安全又高效的端到端自动驾驶系统，它能够在真实世界的复杂多变环境中提供无缝的驾驶体验，并为未来的智能交通系统的发展做出重要贡献。

6.1.3.2 安全性层面

端到端控制技术涉及到对车辆行为的全面控制和优化，整个生命周期都存在被恶意篡改、窃取或泄露的可能。一旦系统出现故障或者遭受到恶意攻击，可能会影响模型的准确性和安全性，导致严重的安全问题，如交通事故、车辆损坏等，因此需要加强系统的安全性设计和防护机制，确保系统的稳定性和可靠性。

汽标委智能网联汽车分委会

6.1.3.3 隐私保护层面

端到端控制技术涉及到对车辆行驶过程的全面监控和控制，可能会涉及到车辆驾驶员和乘客的隐私信息，如个人身份信息、行驶轨迹、驾驶习惯等，因此需要采取有效的隐私保护措施，确保用户的隐私权不受侵犯。

6.2 端到端系统技术概述

6.2.1 系统概述

根据 2020 年汽标委发布的《自动驾驶系统设计运行条件白皮书》，定义自动驾驶系统的第一步是明确其设计运行条件。这是指自动驾驶系统可以启动并安全执行动态驾驶任务的条件范围。通过设定设计运行条件，可以明确自动驾驶系统的功能和局限性，并使用户或系统知晓相应信息。这表明，自动驾驶系统的运行必须严格在其设计运行条件内，否则将增加风险。

端到端自动驾驶系统的目标是在这一框架下，通过整合传感器输入、规划与控制来优化系统效率。尽管现阶段的端到端模型在传感器输入和轨迹输出与传统基于规则的智能驾驶方案没有显著差异，但它们在内部的决策与规划过程中采用了更为复杂的深度学习和强化学习算法。因此，传统的系统设计方法仍适用于端到端系统的某些方面，但在调试和测试中需要特别考虑端到端系统的独特性和复杂性。

下图展示了一种传统智能驾驶系统的架构示例，其中包括应用层、软件平台层、硬件平台层和车辆平台层。这种分层设计适用于模块化的传统系统，但对于端到端系统而言，其架构将更加简化和一体化。

汽标委智能网联汽车分委会

图 6-19 智能驾驶系统设计框架

（1） 应用层

- 感知：负责处理来自车辆传感器的数据，以识别和理解周围环境，如行人、车辆、道路标志等。

- 定位：提供车辆的精确位置信息，通常通过 GPS、IMU 和地图匹配技术实现。

- 融合：负责融合不同类型的传感器数据，形成统一的环境模型，提高感知的准确性和可靠性。

- 规划：负责生成车辆的行驶路线和动作计划，以使车辆在保证安全的前提下尽可能快地到达目的地。

- 人机交互 (HMI)：提供驾驶员和车辆之间的交互界面，包括显示、控制和反馈。

- 控制：执行规划模块生成的路径和动作，通过控制加速、制动和转向等来实现车辆的实际运动。

- 安全监控：监控系统的运行状态，检测和处理可能的故障和异常情况。

- 接管方案：在系统发生故障时，提供安全可靠的应急处理方案。

（2） 软件平台层

- Autosar 自适应平台：支持灵活和动态的软件架构，适用于高级自动驾

汽标委智能网联汽车分委会

驶功能。

- Autosar 标准化平台：提供标准化的软件基础架构，适用于传统汽车电子控制单元（ECU）的功能。

（3） 硬件平台层

- 性能 SoC：高性能系统级芯片，处理复杂计算任务，如感知、融合和规划等。

- ADCU (Autonomous Driving Control Unit)：自动驾驶控制单元，集成多种传感器和处理单元，负责执行自动驾驶任务。

- 安全 MCU：安全微控制器，确保系统的安全性和可靠性，处理安全相关的任务。

（4） 车辆平台层

- 传感器 / 车辆：包括各种传感器（如摄像头、雷达、激光雷达）和车辆本身，用于感知环境和执行控制命令。

端到端自动驾驶系统的主要特征是将传统自动驾驶系统中的感知、决策、规划和控制等多个独立模块整合为一个整体。这种整合旨在减少模块间的误差累积和信息传递延迟，从而提高系统的整体性能和效率。与传统的基于规则的系统相比，端到端系统通过深度学习模型直接从传感器输入生成控制输出，减少了人为设计的中间环节。这使得系统具备了更强的自适应能力和复杂场景处理能力。

6.2.2 端到端系统通用模型要求

6.2.2.1 端到端模型与大模型概念的区别

在自动驾驶领域，“端到端模型”与“大模型”这两个概念经常被混淆，但实际上它们有着不同的技术内涵与应用目标。

端到端模型强调的是系统结构的完整性和全局优化能力。其主要特征是通过一个统一的神经网络架构，从感知到控制的整个流程都在同一模型中完成，从而

汽标委智能网联汽车分委会

实现梯度的端到端传播和全局优化。这种架构的目的是简化信息交互流程、减少误差累积，并提高车辆自主导航和决策的能力。大模型则侧重于模型参数规模及其涌现能力，通常具有大量的参数，且能处理多种复杂任务，如自然语言处理、图像识别等。大模型的广泛应用使其对数据多样性要求更高，但对特定领域的可解释性和可靠性要求可能相对较低。

在自动驾驶领域，所谓的“大模型”与自然语言处理或通用 AI 中的大模型存在显著差异。自动驾驶中的“大模型”通常指的是相对于原有感知系统的“小模型”而言的更大规模模型，而不是具有数十亿参数的超大模型。这主要是因为自动驾驶系统需要考虑车载算力的限制以及实时性要求。

当前车载的算力资源通常只有几百 TOPS（每秒数万亿次运算），难以支撑传统大模型如 GPT-3 级别的运算需求。因此，自动驾驶领域的大模型规模通常在 1 亿参数以下，这与其他领域的大模型存在明显区别。随着 Nvidia 等厂商不断推动更高算力芯片的发展，未来车端模型的大型化将是一个重要趋势，但短期内，这种“大模型”仍将主要服务于特定任务，而不是像通用 AI 那样覆盖广泛的应用场景。

未来，随着端侧算力的持续提升，端到端模型和大模型的交叉将越来越多，特别是在更复杂的决策和规划任务中。类似于 Transformer 架构的模型可能会逐渐进入自动驾驶领域，以进一步提升系统的智能化水平。然而，这种技术趋势也带来了新挑战，如何在有限的算力条件下提升模型的推理能力，并确保系统的可靠性和安全性，将是自动驾驶产业需要重点关注的方向。

总体而言，端到端模型和大模型在自动驾驶领域各自扮演着不同的角色，随着技术的发展，两者的结合有望推动自动驾驶技术的进一步突破。

6.2.2.2 端到端模型的通用要求

端到端自动驾驶模型的设计旨在通过统一感知与决策控制过程，解决传统模块化自动驾驶方案中常见的级联误差，并有效提升视觉信息处理的能力。这种模型以认知决策为核心，直接优化整体性能和用户体验。

汽车智能网联汽车分委会

端到端模型的通用要求包括：

完整性和一体化：端到端模型应确保从传感器输入到控制信号输出的全流程在一个连续的学习框架内完成。通过减少信息转换中的分段处理，端到端设计能够有效降低误差累积，确保决策流程的整体性。这个一体化设计对实现系统的全局优化至关重要。

性能和效率：端到端模型需要高效处理来自多种传感器的大量数据，并在有限的计算资源下，实时生成准确的驾驶决策。性能优化的关键在于提升模型的计算效率并减少延迟，从而确保自动驾驶系统在各种复杂环境中的安全性与平稳性。这不仅涉及到硬件层面的支持，还要求软件层面的高度优化。

可解释性与安全性：尽管纯端到端模型天生缺乏透明度和可解释性，安全性仍然是自动驾驶技术的核心要求。因此，端到端模型设计应包含增强系统可解释性的机制，例如通过设计中间可视化或分解性方法提高模型的透明度。利用强化学习或模仿学习方法时，应特别关注算法的可解释性，确保系统在处理复杂场景时的安全性和可信度。

适应性和灵活性：端到端模型必须具备应对各种驾驶场景的适应性，包括不同的交通环境、天气条件和道路类型。这意味着模型不仅需要从训练数据中学习，还应具备自适应能力，能够在遇到新场景时调整决策。这种适应性是解决自动驾驶领域长尾问题（例如极端或罕见的驾驶情境）所不可或缺的能力。

测试和验证方法：端到端自动驾驶模型的测试与验证需要综合多种场景和条件，确保模型的性能在现实环境中同样有效。测试方法应结合仿真测试与实际路测，全面评估模型在复杂和动态环境中的表现。只有通过系统化的测试与验证，才能确保模型在应用中的安全性和可靠性。

6.2.3 端到端自动驾驶模型的典型设计方案

2023 年以来，模块化端到端和 OneModel 端到端自动驾驶技术都取得了重大进步。UniAD[116]提出了全新的自动驾驶技术方案，成为模块化端到端技术的基

汽标委智能网联汽车分委会

准范式；Wavye 提出的生成式世界模型 GAIA-1[121]，以及视觉-语言-动作模型 LINGO-2[122]可能使 OneModel 端到端技术实现的重要基础；FSDv12 作为一款实车应用的产品让端到端技术更广为人知。本节将呈现这几个重要进展的技术方案。

6.2.3.1 一体化端到端模型设计

一体化端到端模型设计通过一个统一的网络模型连接，从输入到输出实现一体化处理，最大化模型的连贯性。模型直接从数据中学习，实现端到端的优化，这类方案通常依赖大规模数据集的支持，并且通过回归学习映射输入数据到规划/控制输出。这类设计倾向于减少中间模块的依赖，从而提升模型在多样化驾驶场景中的适应性和鲁棒性，在处理复杂场景时则需要更多的训练数据和更复杂的训练策略来确保其稳定性。

简单的一体化端到端方案在自动驾驶发展的早期阶段就已出现。2016 年 4 月，英伟达团队发表了题为《End to End Learning for Self-Driving Cars》[123]的论文，介绍了基于卷积神经网络（CNN）的端到端自动驾驶系统 DAVE-2。该系统使用 CNN 处理车辆前方的摄像头图像，直接输出车辆的转向角度。该系统的模型训练以驾驶数据为基础进行模仿学习。DAVE-2 在不同类型的道路（如城市道路、乡村公路等）都有着出色的表现，展现出车辆在复杂路况下的平稳驾驶能力和对不同交通状况的适应能力。

DAVE-2 系统展示了神经网络在自动驾驶领域的巨大潜力，打破了传统的模块化自动驾驶系统框架，是端到端自动驾驶领域的开创之作。这项研究证明了从数据中直接学习驾驶策略，可大幅简化系统设计，并有助于提高系统的适应性和鲁棒性。DAVE-2 为后续端到端自动驾驶研究奠定了基础，推动了该领域的快速发展。

汽标委智能网联汽车分委会

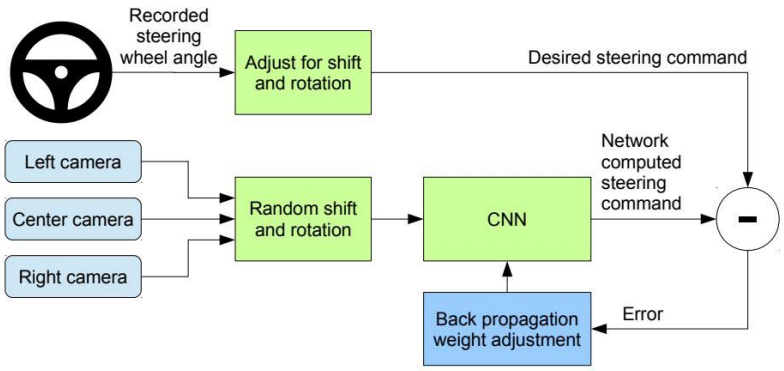


图 6-20 典型架构一：DAVE-2 的网络结构[123]

6.2.3.2 模块化端到端系统设计

模块化端到端系统设计将模型划分为多个功能模块，每个模块具备独立性和灵活性，强调局部优化与全局目标的平衡。各个模块通过特定的机制（如 Transformer 中的查询向量）实现信息流动，从而协同优化感知、预测和规划等多任务。模块间的信息交互通常依赖于注意力机制，通过对特征的权重分配，筛选有效信息并优化决策过程。

以 VAD 架构[124]为例，使用 BEV 编码器对多视角摄像头的视频流进行特征提取工作，然后建立矢量化的场景地图，并预测其他交通参与者的运动轨迹，基于这两个信息以及自身状态进行路径规划，实现端到端自动驾驶。

VAD 架构将整个自动驾驶划分为了场景建图、运动轨迹预测和驾驶规划三个部分，其中各个子模块都是基于 Transformer 结构的：BEV Encoder 用于编码输入的图像信息，将其转化为鸟瞰图视角下的特征图；Vectorized Map Transformer 提取场景中的道路静态元素信息；Vectorized Motion Transformer 提取场景中的动态目标信息，实现动态目标检测和轨迹预测；Planning Transformer 以动态和静态场景的特征以及自身状态为输入，完成自动驾驶的轨迹规划。

汽标委智能网联汽车分委会

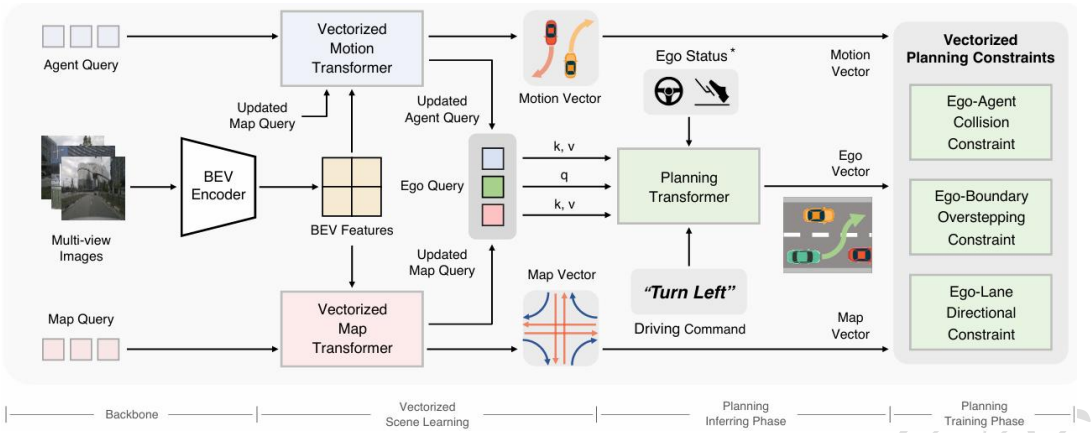


图 6-21 典型架构二：VAD 架构[124]

6.2.3.3 基于大语言模型的端到端系统设计

基于大语言模型（LLM）的端到端系统[125]能够处理多种输入类型（如视频、图像、文本），并通过大语言模型实现多模态信息的融合与推理。此类系统不仅能够对驾驶场景进行深度语义理解，还具备一定的语言交互能力，能够根据描述性输入进行分析和决策。大语言模型在各种信息处理任务中表现出了非凡的熟练程度。这些任务从提取数据和总结文献到生成内容、预测建模、决策和系统控制。此外，视觉语言大模型（VLM）和多模态大模型（MLLM）可以结合和集成许多具有语言理解能力的数据模式，从而推进了多模态信息的系统，如自动驾驶系统。事实上，通过将语言交流与多模态传感器输入相结合，例如全景图像和激光雷达或雷达数据，可以得到更准确的驾驶行动。通过预训练的大语言模型，系统可以将复杂的驾驶任务解耦为多个子任务，并通过不同层次的推理来应对复杂场景。

这种基于大语言模型的端到端设计将自动驾驶的推理能力提升到了新的高度，系统能够通过自然语言的方式理解并生成驾驶决策。这一设计的优势在于其强大的场景理解能力以及处理复杂任务的能力。然而，其面临的挑战在于如何保证实时性，如何处理对大规模数据的需求，以及确保多模态信息融合的稳定性。

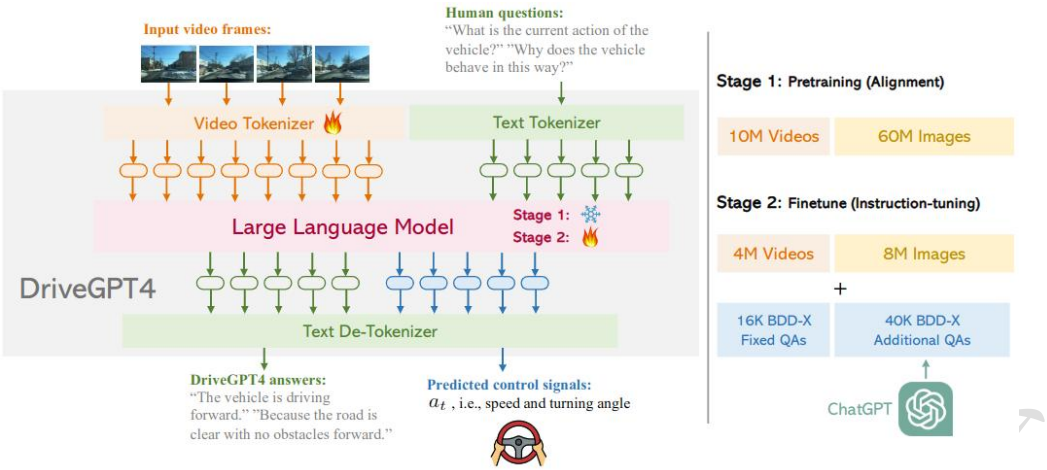


图 6-22 典型架构三：基于 LLM 的架构[125]

pipline 过程举例：

- 人：这是 8 帧图像. 总共 2.1s. 在这个视频中，你就坐在车上，现在这个车正在执行什么动作?
- 模型:车子减速并停车了。
- 人：为什么这辆车要采用这样的行为？
- 模型：因为前面的交通灯已经变红。
- 人：预测这个车辆的速度和转向角。
- 模型：速度: 9.86 9.1 8.18 7.24 6.18 5.21 4.22 3.11 2.09；转向角: 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0

prompt 可以分为描述、分析、控制决策三大类，具体示例见下表。

表 6-3 prompt 设计举例

prompt	示例
描述	车辆当前在做什么？ 车辆此时的表现如何？ 车辆正在进行的动作是什么？ 车辆目前正在采取什么行动？ 描述车辆的当前状态。
分析	为什么这辆车会有这种行为？ 这辆车的行为背后的原因是什么？ 你能解释一下这辆车发生故障的原因吗？ 什么因素导致了这辆车的行为方式？ 这辆车的行为背后的理由是什么？ 为什么车辆会以这种特殊的方式运行？ 是什么情况导致了这辆车的行为？ 此车辆行为的根本原因是什么？ 是什么驱使车辆做出这种行为？
控制决策	请帮我规划接下来的轨迹动作 这辆车的动作是什么？ 为什么这辆车在这条路上？ 预测自车的速度。 预测自车的转弯角度。 要超过前面的车辆，应该怎么操作？ 车辆当前速度：2.15m/s。此视频的长度为 2.10 秒。这辆车目前的动作是什么？ 这辆车是否转弯？ 预测下一帧中车辆的速度和转弯角度。

尽管 LLM 在自然语言处理（NLP）任务上有很强的推理性能，但无法直接处理图像任务，因为 LLM 仅限于输入离散文本。相比之下，视觉模型可以很好地处理图像任务，但与 LLM 相比，推理能力通常较弱。然而，视觉和大语言模型的互补为多模态大模型 MLLM 或者视觉语言大模型 VLM 的新领域铺平了道

汽标委智能网联汽车分委会

路。实际上，MLLM 可以将图像、视频和音频数据与 LLM 的高级推理能力相结合，因此可以较好地执行广泛的任务，诸如语音识别、图像处理和多模态识别。

在自动驾驶背景下端到端应用大模型的技术路线。具体来说，将自动驾驶中 LLM/VLM/MLLM 的使用分为 2 个主要类别：基于提示工程的方法和基于预训练模型的微调方法。

(1) 基于提示工程的方法：

提示工程使用查询（提示）来指导 LLM/VLM/MLLM 的输出。基于大模型的提示工程以实现自动驾驶任务供应的相关工作，其中路径规划和控制是决定车辆导航和真实的决策能力的关键组成部分。

LLM 对于问答，重点是设计和优化提示词来回答各种问题。 VLM 使用提示工程来提高视觉理解。这些方法侧重于增强车辆对环境中视觉线索的解释和响应能力。LLM-driver[126] 这一框架将数值向量模态整合到预训练的 LLM 中。该方法通过目标级的 2D 场景表示，将向量数据与预训练的 LLM 相融合，使模型能够理解驾驶情境并生成相应的行动指令，从而实现对驾驶场景的合理应对。

(2) 基于微调的方法

DriveCoT[127]中开发了一个使用 CARLA 模拟器的端到端驾驶数据集。它包括复杂的驾驶场景（例如：换道和高速驾驶），并纳入了 CoT，为驾驶决策提供推理过程。微调预训练 VLM 的方法，主要是为了提高特定视觉感知任务的准确性，例如检测和识别物体、道路标志和障碍物。微调还可以提高模型在视觉问答任务中的性能，提供更多上下文感知的视觉理解。微调 MLLM 使模型适应自动驾驶特定的任务，如轨迹规划、多模态感知和综合决策。这种方法确保 MLLM 能够有效地处理复杂多样的数据输入。

6.2.3.4 结合大语言模型的快慢系统架构

快慢系统架构结合了传统自动驾驶系统的实时性与大语言模型的高级推理能力，弥补了各自的短板。通过多层次的规划与执行，确保驾驶策略在各种场景

汽标委智能网联汽车分委会

下的适用性和安全性。快慢系统架构的核心优势在于其兼具复杂场景下的高级推理能力和实时性规划能力。通过将高延迟的推理过程与低延迟的传统规划器结合，系统能够在复杂场景中实现更强的应对能力。主要挑战在于如何有效协调快慢系统间的信息交互，以及在实时性和性能之间找到平衡。

以 DriveVLM[127]为例，它结合了自动驾驶技术和大型视觉语言模型，旨在提升自动驾驶车辆在复杂和长尾场景中的理解与规划能力。该系统通过引入“思维链”推理过程，将场景描述、场景分析和层级规划三个模块结合起来，类似于传统的感知、预测和规划流程。DriveVLM-Dual 则是其混合系统，结合了 DriveVLM 的推理能力与传统自动驾驶流程的高效性，弥补了 VLM 在空间推理和计算需求方面的局限性。

(1) DriveVLM 架构概述

DriveVLM 通过处理一系列图像，利用视觉编码器生成图像标记，再通过基于注意力的提取器将这些标记与 LLM 对齐，最终进行 CoT 推理，得出驾驶规划结果。CoT 推理包括三个模块：场景描述、场景分析和分层规划。

场景描述模块：通过环境描述和关键目标识别，系统能够理解驾驶环境（如天气、道路状况）及重要物体。这与传统感知模块不同，它专注于识别对当前驾驶最有影响的物体。

场景分析模块：进一步分析关键物体的特性及其对自车的潜在影响，包括物体的静态属性、运动状态和行为模式，最后生成场景级的综合描述？。

规划模块：基于场景的综合信息，DriveVLM 生成元动作、决策描述和轨迹航点。元动作为短期决策，如加速、转弯等；决策描述则进一步细化驾驶策略，涉及互动物体、行动和持续时间；轨迹航点则提供车辆未来行驶路径。

尽管 VLM 在复杂场景理解和长尾目标识别方面表现出色，但在物体空间位置和运动状态的精确理解上存在挑战，同时由于模型庞大，存在高延迟问题，影响实时性。为应对这些问题，DriveVLM-Dual 结合了 DriveVLM 的推理能力与传统自动驾驶系统的高效性。

汽标委智能网联汽车分委会

(2) DriveVLM-Dual 的改进

3D 感知融合：通过 3D 检测器识别关键物体的 3D 位置与动作，并将其作为语言提示，增强物体分析的准确性。

高频轨迹细化：结合 DriveVLM 生成的低频轨迹，利用传统规划器进行高频细化，确保规划轨迹的质量与实时性兼具。

通过这种快慢结合的方式，DriveVLM-Dual 能够在复杂场景下提供高级推理能力，同时保持实时响应的要求。

6.2.4 功能开发与数据处理

6.2.4.1 功能适配与性能验证新思路

在端到端自动驾驶系统中，功能实现的逻辑与传统模块化系统相比发生了显著变化。功能适配方面，激活和退出功能、变道时机、导航信息以及人机交互均需重新布局，不再主要依靠预定义规则与明确逻辑，而是通过整体模型的学习与优化来完成；故障诊断则更具复杂性，当车道线不清晰时，系统需依赖其他感知数据或已学得的特征，而非单纯的车道线输出，虽然模型对车道线模糊的容忍度更高，但对这种容忍度的量化评估与标定仍是难点。对于性能边界的确定，传统系统往往通过固定规则（如速度限制或弯道半径）来约束，端到端模式则依赖大量数据的训练来学习此类边界，如何保证这一过程的稳定性与可信度仍是研究热点。最后，安全冗余的设计与验证也需要转变思路，传统系统往往在不同模块各自实现冗余，而端到端系统则需从整体架构层面考虑规划、感知等多方面的冗余布局，以确保系统安全性与稳健性。

端到端系统在性能指标的定义和验证上与传统自动驾驶系统存在显著差异，其核心在于对感知、决策和控制等模块独立性的重新界定、对关注元素的拓展以及整体协同效率的提升。由于模块间关联更加紧密，需通过整体调试与优化来替代传统的单模块调试方式，同时网络元素和性能要求的演变也带来对上下游链路影响的更精细考量，尤其在大语言模型等新兴范式中可借鉴自然语言处理领域的

汽车委智能网联汽车分委会

评价指标。为了验证鲁棒性，系统必须在延迟环境下确保整体可靠性，并在级联网络结构中适配各种输入输出关系，以免任何局部故障放大到全局。最终，端到端系统对感知信息的无损传递同样提出了更高的要求，下游模块需对输入数据进行有效筛选与过滤，使决策与控制环节能够基于更准确的环境感知来作出安全可靠的行动。

端到端系统在迭代优化和调试方式上与传统系统也有显著差异：在问题调试方面，由于控制、规划、决策和感知等模块的依赖性增强，必须以整体协调为基础来保证各部分协同运作；同时，尽管传统系统依托基于规则的逻辑验证确保性能，端到端系统却因模型解释性较弱，在优化模型时需防范对其他模块产生负面影响，这也成为其迭代调试过程中的一项新挑战。

6.2.4.2 数据处理流程

端到端自动驾驶模型的性能高度依赖于大规模高质量数据。数据的采集、存储、标注和处理构成了数据处理框架的核心。一个高效的数据处理框架需要在多样性、自动化和闭环反馈等方面优化，确保模型能够适应复杂多变的驾驶场景。

(1) 数据采集

端到端模型需要覆盖多样化场景的数据采集，包括不同的天气、光照条件和道路环境。车辆上的多模态传感器（如摄像头、雷达和激光雷达）是数据采集的主要来源。这些数据主要包括：驾驶行为数据，比如记录方向盘转动、油门和刹车操作，帮助优化模型对驾驶行为的学习；车辆状态数据，如车速、加速度和转向角度，用于评估车辆的运动状态和控制效果；场景数据，如复杂驾驶条件（如雨雪天气、密集交通）中的数据尤其重要，有助于提升模型对极端工况的适应能力。

(2) 数据传输与存储

在数据传输和存储阶段，采集到的数据需要在本地存储设备和云端之间合理分配。一部分关键数据会通过车载通信模块实时上传到云端，用于即时分析和验证，而其他数据则保存在本地以供后续处理和离线分析。为了保证数据的高效利

汽标委智能网联汽车分委会

用，存储过程中通常采用分级管理策略，根据数据的重要性和访问频率进行优化。

(3) 数据标注

标注是数据处理的关键环节，也是数据流转的瓶颈之一。为了提高标注效率，许多公司采用自动化标注工具，利用深度学习算法对图像和视频中的车辆、行人、交通标志等进行识别和标注。然而，自动标注的准确性有限，尤其是在复杂场景中。因此，人工审核与修正依然不可或缺。近年来，智能标注工具的引入为解决这一问题提供了新思路。例如，通过生成模型可以合成稀缺场景数据，并自动完成标注，从而显著提高效率。

(4) 数据预处理

数据标注完成后，需要对数据进行进一步的预处理。数据清洗是预处理的重要步骤，可以去除无效或错误数据，确保输入模型的数据质量。与此同时，数据增强技术通过模拟不同的天气、光照和动态条件，生成更多样化的训练样本。这些样本有助于提高模型的泛化能力，尤其是在应对未见过的驾驶场景时。

在处理数据的过程中，还必须高度关注数据的合规性和安全性。这包括对敏感信息（如车牌和行人面部）的匿名化处理，以及在数据传输和存储时采用加密技术以防止数据泄露或被篡改。随着法律法规对数据保护的要求日益严格，数据合规性已成为端到端自动驾驶系统落地的基础。

数据处理框架的另一个关键特点是闭环机制。通过在实际道路测试中收集的数据，系统可以发现模型的不足，并将这些数据重新引入到训练流程中。以特斯拉的 Data Engine[128]为例，它构建了一个完整的数据流转闭环：车辆在道路上实时上传视频数据，完成采集、标注、训练和反馈优化的全过程。在一个季度内，特斯拉便处理了 1000 万个视频片段，用于不断提升模型对复杂场景的适应能力。这种高效的数据闭环机制不仅显著加快了模型迭代速度，也增强了系统在真实场景中的稳定性和可靠性。

端到端数据处理框架的一个显著优势在于其场景驱动的特点。相比传统的自动驾驶系统，它更关注特定场景的数据需求，能够针对不同驾驶场景进行精准的

数据采集和训练。同时，自动化和闭环优化也使得数据处理更加高效。通过减少人工干预和优化数据流转，端到端框架为自动驾驶技术的快速迭代和规模化落地提供了坚实支持。

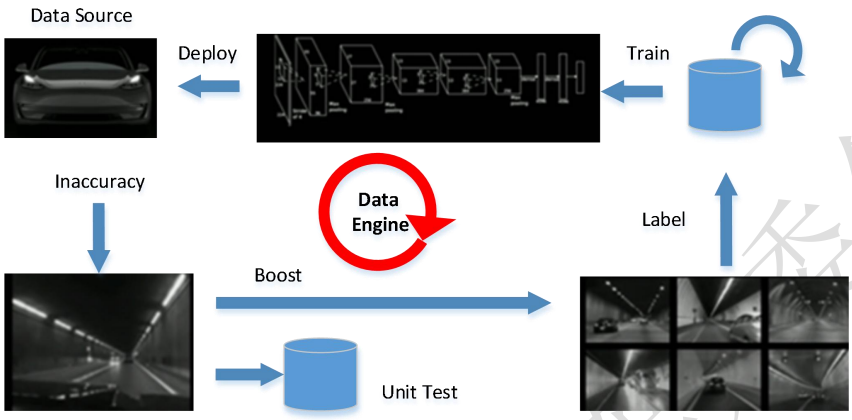


图 6-23 特斯拉的 Data Engine[128]

综上所述，数据处理框架贯穿了端到端自动驾驶系统研发的各个环节。通过高效的数据采集、标注、预处理和反馈机制，框架能够为模型提供高质量的训练数据，并不断推动模型性能的提升，为安全可靠的自动驾驶系统奠定基础。

6.2.4.3 功能开发流程

在端到端自动驾驶的研发中，功能开发流程的设计和工具链的使用是关键所在。相比传统模块化系统，端到端系统通过流程的简化和工具链的统一实现了更高的研发效率和更低的复杂度。这一部分将结合图 6-24 和图 6-25，详细阐述功能开发流程和工具链在端到端研发中的具体实现。

端到端功能开发流程涵盖从数据处理到模型训练、验证、部署、监控与优化的完整闭环。图 6-24 展示了这一流程的核心阶段，每一步都为下一阶段奠定了基础，形成了从输入到输出的清晰路径。

汽标委智能网联汽车委员会

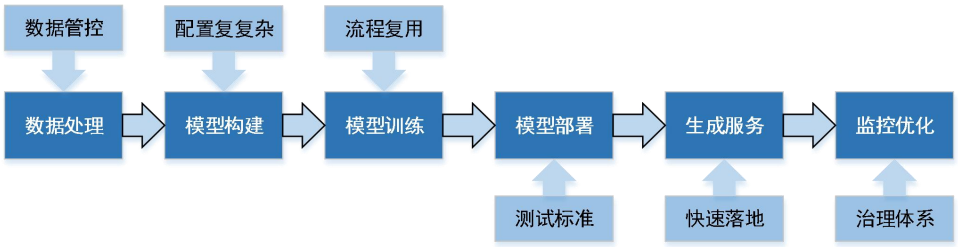


图 6-24 模型开发流程

功能开发的起点是数据处理，这是整个流程的基石。在这一阶段，开发者通过多模态传感器采集到的海量数据进行清洗、归一化、去噪和增强，以确保数据的质量和多样性。特别是对于长尾场景的数据，开发者需要通过模拟各种驾驶工况（如雨雪天气、夜间驾驶等）生成更加全面的训练数据。

接下来是模型构建阶段。基于预处理后的高质量数据，开发者设计并训练端到端模型。这一阶段通常需要配置复杂的深度学习模型架构，结合多轮实验不断优化模型性能。与传统方法不同，端到端系统省略了模块间的接口设计，直接将任务作为整体输入模型中进行学习。

在完成模型训练后，流程进入模型验证阶段。在这一阶段，通过仿真环境和真实道路场景对模型进行测试和评估。仿真平台（如 Carla 或 AirSim）可以帮助开发者在虚拟环境中验证模型的稳定性、鲁棒性以及在复杂场景下的表现。验证后的模型将进入生成服务环节，完成模型的快速落地。由于车载硬件资源有限，模型需要在部署前进行优化，例如剪枝、量化等技术，以减少计算复杂度和存储需求。随后，模型被集成到车辆系统中，并通过实时监控工具确保运行的稳定性。

最后是监控优化阶段。通过实时收集车辆运行数据并分析模型在不同场景中的表现，开发者可以不断优化模型，从而形成一个数据驱动的闭环开发流程。

功能开发流程的高效运行离不开强大的工具链支持。图 6-25 展示了 MLOps 工具链的全景结构，从数据采集到部署与监控，构建了端到端研发的完整技术体系。

汽标委智能网联汽车分委会

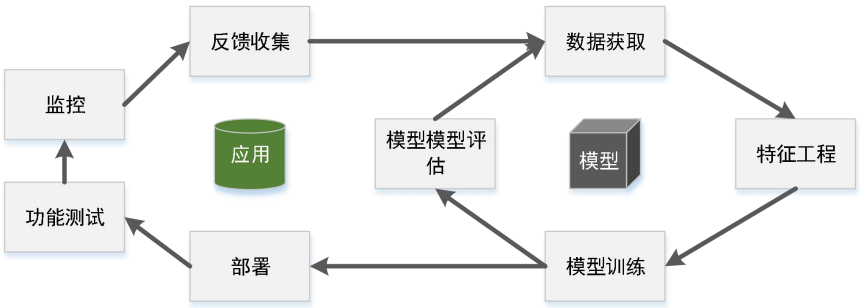


图 6-25 MLOps 工具链全景图

数据管理是整个链路的起点。通过多模态数据采集工具，开发者可以从摄像头、毫米波雷达、激光雷达等传感器获取海量环境数据。这些数据经过特征工程处理后，形成适合模型训练的高质量数据集。

在模型训练阶段，深度学习框架（如 TensorFlow 和 PyTorch）为模型的高效开发提供了核心支持。这些工具能够处理大规模数据集，并支持复杂的神经网络建模。

训练完成后，进入模型评估与验证阶段。通过仿真平台和实际场景测试工具，开发者可以验证模型的性能，评估其在安全性、鲁棒性和响应速度等方面的表现。

模型部署与监控环节则是另一个重点。部署工具能够帮助开发者快速将模型集成到车载系统中，同时实时监控模型的运行状态。结合监控工具（如 Prometheus、Grafana 等），可以对系统性能进行全面分析，识别潜在问题并提出优化方案。

整个工具链的核心优势在于其闭环特性。在车辆实际运行过程中，采集的反馈数据能够实时回流至工具链体系，驱动模型的持续优化与迭代。通过这一机制，端到端系统能够快速适应多变的驾驶场景，持续提升整体性能。

端到端开发流程和 MLOps 工具链的深度融合，为自动驾驶技术的研发构建了强有力的支持体系。功能开发流程的简化显著提升了开发效率，而工具链的自动化能力保障了数据采集、模型训练和验证部署等环节的高效协同。相较于传统的模块化架构，端到端方案不仅降低了系统开发的复杂度，更通过闭环机制实现了从数据到决策的全链路性能提升。

6.2.5 端到端系统透明度分析

(1) 透明度的概念引入与定义

透明度指的是系统能够清晰地展示其内部工作原理、决策过程和操作方式。这包括解释系统为什么以及如何做出某个决定。对于端到端自动驾驶系统来说，透明度有助于用户、开发者和监管机构理解系统的行为，提高系统的可接受性和信任度。

在端到端自动驾驶系统的标准中，可以引入类似透明度定义，并强调在自动驾驶决策过程中的透明性。这包括对算法、数据处理和系统行为的可解释性要求，以增加系统的信任度和可接受性。

(2) 利益相关者的需求分析

不同的利益相关者（如用户、开发者、监管机构等）对系统的透明度有不同的需求。用户可能希望了解系统的基本决策逻辑以便信任它，而开发者则需要深入了解系统的各个部分以便调试和优化。利益相关者需求分析帮助识别和满足这些需求，确保系统设计符合各方的期望。

(3) 系统级和模型级的透明度分类

这指的是透明度的不同层次。在系统层面，透明度涉及整个系统如何工作，包括各个子系统之间的交互和整体流程。在模型层面，透明度更细致，涉及到具体的算法、模型输入输出、参数设置等技术细节。系统级透明度提供宏观的理解，而模型级透明度则是对技术细节的深入披露。

(4) 数据集级别的透明度要求

这是指对用于训练和测试系统的数据集的披露和管理。例如，数据是如何收集的、标注的标准是什么、数据集的质量如何等。这些透明度要求确保数据的可靠性和一致性，从而支持系统的稳健性和泛化能力。

(5) 技术特征和系统分解的描述

汽标委智能网联汽车分委会

这涉及系统的详细技术描述和各个子系统的分解。透明度要求包括明确每个技术组件的功能、接口以及它们之间的关系。这种描述有助于理解系统的整体结构，确保在开发和集成过程中，各部分能够顺畅协作，并且系统的行为可以被准确预测和控制。

6.3 端到端系统测试要求

6.3.1 传统自动驾驶测试方法的局限性

传统的自动驾驶测试方法已经发展出较为成熟的流程，主要针对模块化自动驾驶系统。在感知算法的测试中，回灌数据被用于离线开环测试，系统在缺乏环境交互的情况下，通过对预先记录的数据进行分析，评估其表现。而规控算法的测试通常依赖于模拟器进行闭环测试，通过模拟器反馈系统的输出，从而评估算法在不同环境下的反应。然而，这种模块化的测试方法在端到端自动驾驶系统中遇到了显著的局限性。

端到端自动驾驶系统不再区分感知、决策和控制模块，因此测试方法必须能够有效评估系统在动态交互环境中的表现。传统的开环测试方法无法模拟端到端系统在偏离预设操作路径后的反应，而闭环测试中，现有模拟器在闭环测试中的传感器同步性和仿真准确性与实际驾驶环境存在较大差距，导致其在复杂场景下难以准确反映真实表现。此外，模拟器在处理动态交通流和复杂驾驶场景时的能力有限，无法充分评估系统的鲁棒性和适应性。

当前端到端测试方法中，在线/闭环测试可以在模拟环境中进行，通过仿真环境控制参数、交通流和传感器数据，评估系统的整体性能。然而，长尾场景和异常驾驶场景的数据较难获取，且模拟器在场景挖掘和理解能力上的不足，限制了系统在复杂场景中的性能评估。离线/开环测试基于预先记录的人类驾驶数据，通过将系统预测轨迹与数据集中的人类驾驶轨迹进行对比来衡量系统的表现。尽管这类测试易于实现，且基于真实数据集，但由于无法模拟动态交互，系统对偏离人类驾驶轨迹后的恢复能力仍然难以评估。

汽标委智能网联汽车分委会

在实车测试中，自动驾驶系统可以与真实环境进行交互，评估系统在各种场景下的实际表现。然而，开环测试无法模拟这种交互，导致其验证结果与实车测试结果之间存在差异，开环验证的离线测试结果和实车运行效果一致性较低。这说明了传统开环测试方法的局限性。端到端系统需要更具交互性和真实性的测试方法，以确保在实际驾驶中的可靠性。

6.3.2 端到端系统对仿真测试的需求

闭环仿真已经成为端到端自动驾驶系统量产过程中不可或缺的工具。由于实车测试的成本极高，仿真测试提供了一种低成本、快速迭代的方式，可以在安全的环境中优化系统性能。闭环仿真允许系统与模拟环境进行交互，通过反馈循环模拟真实驾驶中的动态变化，评估系统在复杂场景中的表现，并确保其鲁棒性和适应性。

与模块化自动驾驶系统不同，端到端系统不再区分感知、决策和控制模块，因此其测试方法必须具备高度的真实性和交互性。模块化系统中的开环测试能够有效评估感知算法的性能，但无法适用于端到端系统的复杂交互需求。闭环仿真能够弥补这一不足，通过动态反馈，评估系统在偏离人类操作后的恢复能力和系统稳定性。

闭环仿真工具的开发对于端到端自动驾驶系统上车至关重要。目前，许多研发团队正在同步推进闭环仿真工具的开发，以解决场景生成、传感器仿真和交通流仿真等技术难题。高保真度的仿真场景将显著提升自动驾驶系统的测试和验证效率。此外，闭环仿真通过强化场景数据挖掘和理解能力，能够更好地模拟长尾场景和异常驾驶场景，确保系统在各种复杂情况下的性能表现。

6.3.3 端到端系统仿真器能力需求

用于端到端自动驾驶系统的闭环仿真模拟器需要具备几项关键能力，包括场景定义和生成、传感器仿真、交通流仿真和车辆动力学仿真。其中，车辆动力学仿真技术经过多年发展，已相对成熟，能够精确模拟车辆的动力学行为，包括加

汽标委智能网联汽车分委会

速、刹车和转向等，确保仿真车辆的运动轨迹与真实车辆一致。这对于评估自动驾驶系统的控制算法尤为重要。随着计算机图形学、计算机视觉以及生成式 AI 的发展，场景生成、传感器仿真和交通流仿真也取得了显著进步。以下是对这些能力的详细介绍：

场景定义和生成是闭环仿真模拟器的基础能力之一。仿真器需要能够创建多样化的驾驶环境，包括城市道路、高速公路、乡村道路等，同时还需模拟不同的天气条件、时间段和交通状况。场景的多样性和复杂性对于全面评估端到端自动驾驶系统的性能至关重要。

传感器仿真需实现高保真度的多模态数据模拟，涵盖视觉、激光点云、射频波束及声波等全维度传感信号的精确复现。传感器仿真的真实性和一致性是确保闭环仿真有效性的关键。通过技术创新，光轮智能团队基于 3D Gaussian Splatting（3DGS）[129]技术，开发了面向开放世界街道场景的三维重建技术，显著提升了传感器仿真的精度。

交通流仿真模拟真实交通流中的各种动态行为，包括车辆行驶、变道、超车、行人过街等。光轮智能还通过基于 LLM+Diffusion 的智能体 Agent 技术，利用大量真实交通数据进行训练，为每辆交通流车辆分配不同的驾驶风格，显著提升了仿真交通流的多样性及生成复杂场景的效率。

车辆动力学仿真技术经过多年发展，已经相对成熟，主要用于模拟车辆的动力学行为，包括加速、刹车、转向等。确保虚拟车辆的运动轨迹与实际车辆一致，对于评估自动驾驶系统的控制算法尤为重要。

6.3.4 仿真试验方法

6.3.4.1 测试场景构建

(1) 场景多样性的重要性

对于端到端自动驾驶模型，仿真测试是验证其性能和安全性的重要手段。与传统模块化测试不同，端到端系统依赖于多样化、复杂的测试场景，来确保其鲁

汽标委智能网联汽车分委会

棒性和适应性。由于端到端模型是数据驱动的黑盒系统，开发者难以通过逻辑推理预判其行为，因此，丰富的测试场景覆盖成为确保系统可靠性的关键因素。具体的测试场景用例元素可参照 8.3.2 和 8.3.3 章节相关内容。

(2) 传统分段评测模式与端到端系统的差异

在传统自动驾驶系统中，分段测试模式是通过对各模块进行独立评测，构建相应场景集。例如，感知模块通常在各种天气和光照条件下进行测试，而规控模块则针对车辆操作如变道、汇流等典型场景进行验证。这种测试方式虽然能有效评估每个模块的性能，但对于端到端系统来说，不同模块之间的高度耦合使得这种方法难以全面覆盖其行为特性。

(3) 端到端测试的耦合性与复杂性

端到端自动驾驶系统中的感知、决策和控制模块相互依赖，形成一个整体的模型架构。因此，测试场景的设计不仅需要覆盖感知和规控的单独需求，还需要在不同条件下交叉测试，尽可能全面地覆盖模型可能遇到的场景。例如，在构建汇入场景时，需要同时考虑不同天气、光照、交通密度等多种变化因素。场景间的复杂交互性使得端到端测试的构造难度远高于传统分段评测方法。

(4) 部署需求

随着端到端自动驾驶技术发展，多传感器融合、大模型方案的技术演进，端到端场景类仿真模型（尤其传感器模型、交通环境模型）将趋向融合复杂化。仿真硬件需要满足当前和后续需求。支撑 MIL、SIL 场景的标准计算机应满足单传感器或模块化算法仿真测试需求下的单机部署形态，也应满足端到端自动驾驶系统执行全链条仿真需求下的集群部署形态，集群部署形态应支持资源池化和异构横向扩展的能力。

6.3.4.2 仿真测试数据生成

测试数据生成是仿真平台的核心功能。在日志仿真系统中，实车的落盘数据可以直接作为模型的输入测试在当前时间片中模型的输出结果是否符合预期。由于模型的输入数据全部来自实车采集，日志仿真结果具有最高的可信度，通常用

汽标委智能网联汽车分委会

于复现线上问题，以及验证一些无需闭环的功能。

当前的量产算法普遍还是采用模块化的算法框架，在测试的时候无论是 LogSim 还是 WorldSim，都可以针对模块或者模块之间的组合进行测试，仿真或者测试系统能够去构造显示的输入，接收显示的输出进行闭环控制以及结果评价。实车测试验证的成本高昂，针对经典的自动驾驶架构，行业已经有一套行之有效的方案进行模型上车前的测试验证，即感知算法使用回灌数据进行离线开环测试，规控算法基于模拟器进行闭环测试验证。

对于感知“端到端”、决策规划模型化，传统的测试方法依然有效，显示的输出能够被定义和构造。但对于模块化端到端、One Model/ 单一模型端到端，隐示的表达可以被捕捉，但无法评测，也无法构造，这就造成传统方法的失效。在基于数据回灌的开环测试条件下，端到端系统无法与环境交互，系统一旦出现偏离采集路径的操作，后续系统的响应将无法评估。

当然这里的传统测试方法主要指的是 SIL（软件在环），对于 HIL（硬件在环）而言，现在主流的测试已经是信号级传感器的注入，对完整的功能进行测试验证，端到端的改变并没有很大影响。

端到端仿真需要更加真实的传感器输入，尤其是视觉传感器，这是行业的共识。实现高保真模拟的技术路线包括基于光线追踪的游戏引擎、基于三维重建的仿真器以及基于世界模型的仿真器。

在世界仿真器（WorldSim）中，端到端模型直接依赖传感器的原始信息（相机图像和点云等）作为输入，而传统的世界仿真器只能构造规控模块所需的结构化仿真输入，为了解决这个问题，近年来出现了三种解决思路：基于 3D 建模的仿真器，基于世界模型的仿真器和基于游戏引擎的仿真器。

(1) 基于 3D 建模的仿真器

基于 3D 建模的仿真器通过构建车辆运行场景的 3D 模型生成传感器输入数据。此类仿真器通过 3D 环境建模，可以模拟出自车在不同方位、不同角度下的传感器信息。只要有足够精细的 3D 场景模型，理论上就能够生成高度逼真的传

汽车智能网联汽车分委会

感器输入数据，足以以假乱真。目前，学术界最常用的 Carla 仿真器就是此类方案的典型代表。

近年来，自动驾驶技术的迅猛发展不断推动着相关领域的技术革新。自 ECCV 2020 会议上 NeRF[130]（神经辐射场）技术的首次亮相，三维重建技术跨入了一个全新的发展阶段。NeRF 通过先进的深度学习算法，能够从稀疏的二维图像中重建出连续的三维场景。紧接着，SIGGRAPH 2023 会议上提出的 3DGS [129]技术，进一步加速了三维重建技术的发展。清华大学智能产业研究院团队提出的首个开源自动驾驶 NeRF 仿真工具 MARS[131]，Waabi 和多伦多大学在 CVPR 2023 上的工作汇报中介绍了 UniSim[132]，浙江和理想汽车在学术会议 ECCV 2024 上发表了一种新的显式场景表示 Street Gaussians[133]。此外还有 GaussianPro[134]、LidaRF[135]等相继被提出。尽管三维重建的热点不断，但从实际表现来看，局限性比较多，重建效果非常依赖数据，而且当视角与原车视角变换大时，效果不可控。此外，动态场景的添加、场景与车辆的交互、光影的变换，也存在一些实现难点。构建精细的 3D 模型本身耗费巨大，尤其是当需要验证复杂且多变的场景时，针对每个场景创建精细模型成本极高，限制了其在大规模测试中的应用。为了克服这些问题，近年来基于世界模型的仿真器逐渐成为研究热点。

(2) 基于世界模型的仿真器

世界模型特指一类能够自主习得现实世界基础物理法则与数学规律的智能模型，这类模型具备理解万有引力、力学作用等自然现象的能力，并能据此推演出具有时间连续性的未来场景视频。

世界模型为智能驾驶系统提供了强大的技术支撑。在自动驾驶应用中，精确预判交通场景的动态演变具有决定性意义，通过前瞻性地推测可能发生的交通事件，车辆系统能够更加从容地制定行驶策略并执行最优控制。具体到自动驾驶领域，世界模型既可用于场景合成，也能直接参与决策规划。其主要价值体现在：1）批量生成高真实度的驾驶场景，特别是那些实际采集困难的特殊工况，有效

汽标委智能网联汽车分委会

扩充训练样本库；2）作为虚拟测试环境，为算法提供闭环验证平台；3）基于多模态融合的世界模型可直接输出驾驶策略，指导自动驾驶系统做出合理决策。

(3) 基于游戏引擎的仿真器

OASIS SIM 3.0 是基于游戏引擎的仿真器，对于端到端自动驾驶系统的仿真，OASIS SIM 能够提供高保真场景模拟与传感器以及高质量、规模化的测试用例场景搭建，提升仿真的置信度以及测试覆盖度。

OASIS SIM V3.0 通过精确的物理模型，实现了传感器仿真的高精度模拟，能够实现传感器采集层、模型层、协议层的全流程仿真。此外，该仿真器能够在仿真环境中自定义交通流，重现真实世界的交通情景，帮助测试和优化自动驾驶算法。交通流场景控制功能支持基于规则的控制模型以及基于 AI 的控制模型。

OASIS SIM 3.0 支持直接导入 OpenDrive 文件，并自动识别和解析道路网络等关键信息自动化生成三维场景，无需复杂的建模和人工编辑，提高了场景构建的速度和效率。此外，该平台实现了 3DGS 场景和存量建模场景的渲染融合表达，支持对融合场景进行深度编辑，例如添加/移动车辆，改变天气光照等。

图 6-26 OASIS 重建场景演示，并且进行了车辆添加和天气改变

(4) 合成数据

自动驾驶技术发展已进入攻坚阶段，端到端架构正成为主流演进方向，世界模型的价值日益凸显。究其原因：首先，高阶自动驾驶对训练数据提出了更高要求，包括复杂交通流模拟、多样化场景覆盖、长尾场景采集以及精准 3D 标注等。然而现实数据采集面临成本高企、危险场景获取困难、长尾样本稀缺以及 3D 标

注费用昂贵等挑战。在此背景下，基于世界模型的合成数据生成方案展现出独特优势，既能高效创造丰富场景，又能准确预测交通演变。

(5) 深度强化学习

在强化学习中，智能体学习如何在环境中做出决策和行动，以最大化累积奖励信号。在自主智能体的背景下，强化学习为这些智能体提供了一种通过试错学习来学习如何在没有明确编程的情况下执行任务和做出决策的方法。智能体是强化学习框架中的学习者。它与环境相互作用，并采取行动实现某些目标。环境是代理与之交互的外部系统。

6.3.4.3 仿真场景评估指标

在评估仿真场景的效果时，需要从多个维度进行评估，以确保仿真场景的有效性、效率和可靠性。以下是几项关键评估指标：

(1) 真实性分析：收集并分析同一感知系统在仿真场景和实际道路测试中的表现数据，验证仿真场景的有效性；

(2) 仿真效率：考虑仿真场景执行的计算资源消耗、运行时间和成本效益；

(3) 标准化合规性：根据如 ASAM OpenX 系列标准（包括 OpenDRIVE、OpenCRG、OpenSCENARIO 等）评估仿真场景是否遵循行业标准，这有助于确保仿真结果的通用性和可比性；

(4) 可重复性与可追溯性：仿真场景需要能够在不同时间和条件下重复执行，并且每次执行的结果都应该是可追溯和分析的，这对于故障排查和性能优化至关重要。

6.3.4.4 自动驾驶系统性能评估指标

在自动驾驶技术的发展过程中，性能评估是至关重要的一环，它涉及到多个方面的考量，以确保自动驾驶系统在实际应用中能够达到预期的效果，同时保障乘客和周围环境的安全。具体的评估指标可以参照 8.3.3 节内容。

汽标委智能网联汽车分委会

6.4 总结与展望

本章节系统梳理了端到端自动驾驶系统的研究背景与技术框架，深入剖析了端到端自动驾驶系统的核心优势与实现路径，并提出了针对该类系统的测试要求。端到端系统架构通过消除传统自动驾驶系统模块间误差累积和信息延迟问题，进一步提高了系统的整体性能和效率，也使系统具备了更强的自适应能力和场景处理能力，为自动驾驶系统的发展提供了新的技术范式。

展望未来，端到端系统的演进将沿着几个关键方向持续深化：首先，在模型能力层面，需重点突破动态环境适应性瓶颈，发展基于在线学习的自进化网络架构，解决开放场景下的概念漂移问题；其次，在系统设计层面，应进一步探索可解释性建模与模型轻量化部署，开发面向边缘计算的模型蒸馏等技术；另外，在测试验证层面，需要构建涵盖多模态异常注入的高保真仿真平台，建立完善的模型可靠性测试体系。随着大语言模型与具身智能的融合发展，端到端系统有望突破单一任务边界，向具备跨场景迁移能力的通用智能体演进。但同时也需警惕数据隐私、算法偏见等伦理风险，需在系统设计中嵌入可信 AI 治理机制，以实现技术创新与社会价值的平衡发展。

汽标委智能网联汽车分委会

第七章 智能座舱 AI 技术应用场景与标准化研究

7.1 智能座舱 AI 技术应用现状

智能座舱 AI 技术起源于 21 世纪初的电子化座舱，早期以机械仪表盘和基础信息娱乐系统为主。21 世纪 10 年代进入智能化阶段，触控交互与语音识别技术普及，但受限于被动式交互和功能堆砌，体验碎片化。2023 年生成式 AI 技术突破，大模型开始上车，实现多模态感知与拟人化交互，重构了自然语言交互和场景化服务能力。2024 年座舱芯片被推出，支持大模型本地化部署，推动座舱向主动感知、情绪识别、个性化服务演进。

7.1.1 智能座舱 AI 技术应用场景和效果分析

随着智能网联汽车的快速发展，越来越多的车型开始搭载 AI 大模型。目前在智能座舱领域，AI 技术主要应用于车载语音交互、视觉交互、多模态交互、其他开放性任务场景，实现效率、智能、情感、舒适等多方面提升。

7.1.1.1 车载语音交互

车载语音交互是车辆内部以语音为主要方式，当前车载语音系统仍存在智能化、个性化、情感化、交互性不足等问题，通过自然语言处理类大模型可以弥补以上不足，持续理解对话的上下文，提供更加准确、个性化的响应，从而有效提升车载语音交互的体验。

7.1.1.2 视觉交互

座舱视觉交互是指在车辆内部，通过视觉显示技术与驾驶员或乘客进行信息交换和互动的过程。它通常涉及车辆内部的各种显示屏，如中控屏、仪表盘、HUD（抬头显示系统）等，以及与之相关的软件和界面设计。通过搭载 AI 算法模型，能够提升识别准确率，并根据用户习惯和喜好给出最合理的反馈，提升座舱视觉交互体验。

汽标委智能网联汽车分委会

7.1.1.3 多模态交互

多模态交互是指通过融合多种感知模态（如声音、视觉、动作、环境等）来实现的人与机器或人与人之间的交互方式。这种方式模拟了人类之间自然、多样的交流形式，旨在提供更直观、高效、自然的人机交互体验。

7.1.1.4 开放式任务

开放式任务指的是通过 AI 技术来检索完成某一任务需要的信息，并通过系统传感器、数据库等渠道获取跨领域信息来完成复杂的任务。基于先进 AI 技术开发的智能座舱产品能够实现一系列令人瞩目的功能，极大地提升驾驶的便捷性、安全性和个性化体验。

首先，智能推荐功能能够根据驾驶员的偏好和习惯，自动推荐音乐、电台、导航路线等，使驾驶过程满足用户的个性化需求。其次，行车辅助功能利用 AI 技术，通过实时监测车辆周围环境、交通状况以及驾驶员的驾驶行为，提前预警潜在危险，并自动采取紧急制动、车道保持等辅助措施，为驾驶员提供全方位的安全保障。再者，车辆设置功能允许驾驶员通过简单的语音指令或触摸屏操作，轻松调整座椅、空调、车窗等设备的设置。最后，服务支持功能通过智能座舱的联网能力，为驾驶员提供丰富的在线服务。总之，基于 AI 技术开发的智能座舱产品通过实现智能推荐、行车辅助、车辆设置、服务支持等功能，为驾驶员带来了前所未有的智能出行体验。

7.1.2 智能座舱 AI 技术应用的问题和难点

7.1.2.1 AI 大模型部署问题

(1) 云端部署难点

国产化的云端硬件与国外主流 AI 硬件平台可能存在差异，导致 AI 大模型部署时面临硬件兼容性问题，需要进行适配和优化；同时，国产云端平台可能缺乏完善的软件生态和工具链支持，增加了部署难度和复杂性；此外，国产化平台

汽标委智能网联汽车分委会

还可能面临算力不足的问题，需要设计高效的资源调度和分配策略，以充分利用有限的计算资源[136]。

(2) 端侧部署难点

端侧设备硬件性能有限，可能无法满足 AI 大模型实时性和高效性要求，需对模型压缩优化；同时，设备需兼顾能耗和散热以保障持续稳定运行，AI 大模型部署要在保证性能的同时降低能耗散热；此外，端侧设备 AI 大模型要定期更新维护保证准确性和时效性，但设备分散多样使更新维护难度大[137]。

7.1.2.2 安全和隐私问题

AI 应用于智能座舱，需要本地存储和上传大量的个人信息、视频、语音等数据，如何保证数据存储安全，不被网络攻击和窃取，也是人工智能在汽车广泛应用必须要解决的问题。在云端部署 AI 大模型时，需要确保数据的隐私保护，建立严格的数据管理机制和安全防护措施，以防止数据泄露和滥用。端侧设备通常涉及用户隐私问题，需要在 AI 大模型部署时考虑用户隐私和权限管理，建立严格的用户隐私保护机制和权限管理制度，以防止数据泄露和滥用。

7.1.2.3 数据问题

数据是 AI 大模型的核心要素之一，数据质量和数据规模很大程度决定了 AI 的整体效率和质量。高质量训练数据获取面临多个难点，主要包括：

(1) 数据多样性

AI 模型通常需要处理各种复杂和多样化的场景，因此高质量的训练数据需要具备广泛的多样性。然而，在实际获取过程中，往往难以覆盖所有可能的场景和情况，导致训练数据的多样性不足。

(2) 数据标注准确性

对于监督学习任务，训练数据需要准确地标注。然而，由于标注任务通常需要人工完成，标注质量往往受到标注人员专业能力和主观判断的影响，导致标注数据存在误差和不一致性。

汽标委智能网联专分委会

(3) 数据隐私和合规性

在获取训练数据时，需确保数据隐私性与合规性，这要求遵守法规、保护隐私、避免侵权，但实际操作中二者常难以兼得，需权衡因素并制定解决方案。

(4) 数据规模和效率

AI 模型需要大量的训练数据才能达到理想的性能。然而，在实际获取过程中，往往面临数据规模不足或获取效率低下的问题。这需要采用更高效的数据采集方法、优化数据存储和处理流程以及利用云计算等先进技术来提高数据获取效率。

(5) 数据质量评估

数据质量不仅取决于数据的准确性和多样性，还受到数据分布、噪声和异常值等多种因素的影响。如何准确评估训练数据的质量，并根据评估结果调整数据获取策略，是一个需要解决的问题。

7.1.3 智能座舱 AI 技术应用相关标准分析

目前通用人工智能领域相关标准较多，但针对车载 AI 应用的标准较少。国内，针对车载座舱 AI 技术相关的标准仍在持续建设中，在汽车信息安全、个人信息安全和智能语音交互系统测试方面，国家标准提出了部分技术要求，能够对智能座舱 AI 应用提供参考。

表 7-1 相关标准

标准名称	发布单位	主要内容
《面向行业的大规模预训练模型技术和应用评估方法 第 4 部分：汽车》	中国信息通信研究院	聚焦汽车行业高质量发展，结合汽车的研发、生产、销售、使用等全过程，形成汽车大模型应用成熟度评价方法，助推汽车大模型产品升级优化
GB/T36464.5-2018 《信	全国信息技术标准	规范了车载终端智能语音交互系统的

息技术 智能语音交互系统 第 5 部分：车载终端》	化技术委员会	术语和定义、系统框架、要求和测试方法
GB/T41797-2022 《驾驶员注意力监测系统性能要求及试验方法》	全国汽车标准化技术委员会	旨在规范驾驶员注意力监测系统的设计和性能，确保该系统能够有效地监测和评估驾驶员的注意力水平，从而提高道路安全性。推动相关技术的健康发展，并为车辆制造商和系统供应商提供明确的指导
汽车智能座舱语音分级与测评方法	中国智能网联汽车产业创新联盟（CAICV）	旨在建立一套科学、有效的测评方法，以指导智能座舱语音产品的研发设计，为企业开展智能座舱语音水平测试提供依据
汽车智能座舱智能水平测试与评价方法	国汽（北京）智能网联汽车研究院有限公司等机构	旨在解决现有智能座舱产品与用户需求之间的不匹配问题，并为汽车智能座舱的智能化水平提供一个统一的测试和评价方法
汽车智能座舱交互体验测试评价规程	中国汽车工业协会	旨在通过不同交互模态在座舱中执行相应交互任务时，评价用户体验优劣的标准
道路车辆免提通话和语音交互性能要求及试验方法	中国汽车技术研究中心有限公司	旨在探究车载语音交互系统的交互性能表现及用户体验

汽标委智能网联汽车分委会

7.2 智能座舱 AI 技术应用场景与技术路线探索

7.2.1 智能座舱语音交互场景发展趋势与技术路线

在智能座舱之中，语音交互作为最自然、最便捷的人机交互方式，其在智能座舱中的应用和发展趋势，以及所依托的技术路线，成为了推动汽车行业向智能化转型的关键力量。

7.2.1.1 发展概述

近年来，智能座舱语音交互技术经历了飞速发展。基于深度学习的语音系统不仅能准确识别多种语言和方言，还能理解复杂的语境和指令，甚至能够根据上下文进行逻辑推理和预测用户需求。例如，当用户说“我有点冷”，系统不仅能自动调高车内温度，还能贴心询问是否需要关闭车窗。

智能座舱的语音交互已广泛应用于多个场景，包括但不限于导航、娱乐、通讯、车辆控制（如空调、车窗、座椅调节）以及紧急求助服务。

为了更好地满足用户的个性化需求，现代智能座舱语音系统支持定制化设置，如语音助手的名字、音色，甚至性格特点。此外，情感识别技术的应用使得语音助手能够感知用户的情绪状态，并据此调整回应策略。

在驾驶过程中，语音交互的引入，使驾驶者无需动手即可完成大多数车内功能操作，有效减少了视线转移和手动操作频率，极大提升了行车安全性。

语音交互不仅简化了操作流程，还为用户提供了一种更加自然、直接的交流方式，使得人车互动如同与真人对话般顺畅。

智能座舱的语音交互技术还特别关注到了不同用户群体的需求，包括但不限于老年人、残障人士等。对于这些群体而言，语音控制车辆功能尤为关键，它降低了操作门槛，体现了社会包容性和技术的人文关怀。

汽标委智能网联汽车分委会

随着自动驾驶技术的逐步成熟，智能座舱将成为未来汽车的核心。在高度自动化驾驶场景下，人与车的交互将更加频繁，语音作为最自然的人机交流方式，其智能化水平将直接影响到乘客的乘车体验和对车辆品牌的认同感。

7.2.1.2 场景应用

智能座舱语音交互技术的不断演进，预示着未来汽车内部空间将转变为一个高度智能、个性化且情感丰富的交流场景。随着技术边界的拓展，几个关键的应用场景正逐渐成形。

7.2.1.2.1 情感化的语音交互场景

智能座舱语音系统已借助深度学习算法的深入应用，以及情感识别技术的精进，逐渐实现与人类对话的更高层次相似度。系统不仅能理解命令式的指令，还能领悟用户言谈间的情绪变化，并做出相应回应。这样的设计极大地增强了人车之间的情感纽带。

7.2.1.2.2 多模态融合交互场景

单一维度的语音交互正逐渐向整合视觉、听觉、触觉等多种感官通道的多模态交互转变，以更全面的方式理解乘客的意图。通过车内安装的高清摄像头捕捉乘客的面部表情和肢体动作，结合语音指令，系统能更精准地判断乘客的真实需求。这种多维度的信息整合，使智能座舱的服务更加细腻、个性化，也更加符合人类自然交流的习惯[138]。

7.2.1.2.3 主动服务场景

基于大数据分析和机器学习算法，智能座舱语音系统将能够学习每位乘客的偏好、行为模式，从而在用户提出需求之前就能预判并主动提供服务。这种从被动响应到主动服务的转变，让智能座舱更加贴合用户的生活节奏和实际需求。

7.2.1.3 技术路线

座舱语音交互技术是一种允许用户通过语音命令与车辆进行交互的技术，通常包括唤醒、识别、理解、播报四个环节。用户通常需要说出特定的唤醒词或短

汽标委智能网联汽车分委会

语（语音唤醒），系统通过内置的麦克风识别用户的语音指令（语音识别），系统将接收到的语音信号理解为可识别的指令（语义理解），系统执行完指令后，会通过语音反馈结果给用户（语音合成）。以下是座舱语音交互主要技术定义：

- 声学前端：当语音信号被各种各样的噪声干扰时，从噪声背景中提取有用的语音信号。
- 语音唤醒：语音交互系统在音频流监听模式下，一旦检测到特定特征或事件，便会从监听状态切换至命令字识别或连续语音识别等其他处理状态。
- 语音识别：将人类的语音中的词汇内容转换为计算机可读的输入。
- 语义理解：把自然语言内容，转换为有一定结构的文本数据，使应用能够理解用户的意图。
- 语音合成：合成人类语音的过程。

7.2.1.3.1 算力平台技术路线

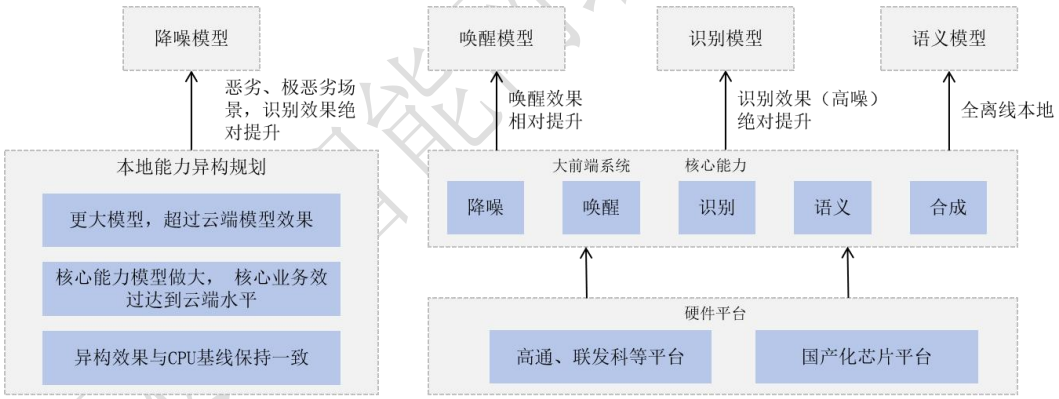


图 7-1 依托算力平台提升基础算法效果图

基于深度神经网络的语音识别技术无疑是智能座舱语音交互系统的重要基础。通过对大量语音数据的学习和训练，能够不断提升对语音的识别准确率和速度。随着芯片算力不断提升，对语音基础能力的提升有强烈的诉求，需要依托算力平台不断提升基础算法的效果。

7.2.1.3.2 计算协同技术路线

边缘计算与云计算的协同工作在智能座舱语音交互系统中发挥着重要作用。在车辆行驶过程中，很多语音交互请求需要快速得到响应，如紧急刹车预警、变道提示等。边缘计算可以在本地快速处理这些数据，从而大大减少了响应时间。

而云计算则提供了强大的计算资源和存储能力。它可以对大量的车辆数据进行分析和处理，挖掘出有价值的信息。同时，云计算还可以用于训练更加复杂的语音识别和自然语言处理模型，不断提升系统的性能。

在实际应用中，对于一些实时性要求较高的任务，可以由边缘计算来完成；而对于一些需要大量计算资源和数据存储的任务，则可以交给云计算来处理。同时，还需要建立高效的数据传输通道，确保边缘计算和云计算之间的数据能够快速、稳定地传输。

7.2.1.3.3 异构计算技术路线

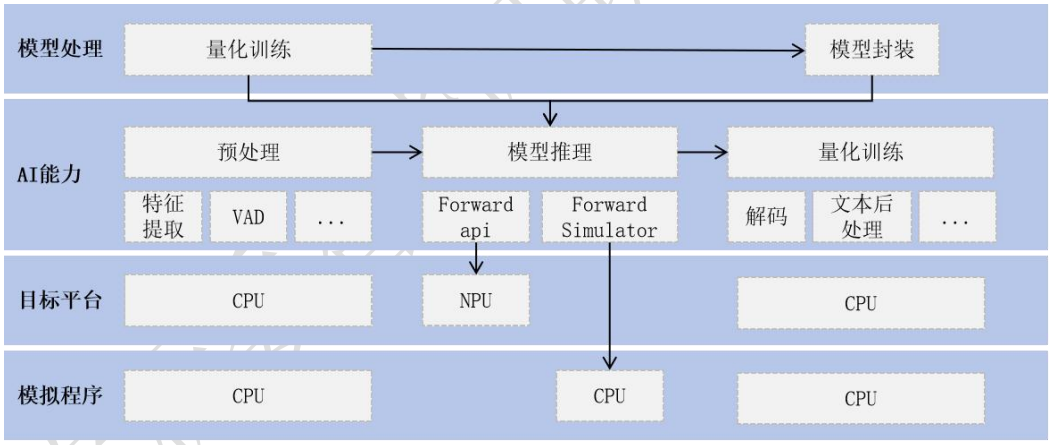


图 7-2 异构计算技术图

异构计算技术是一种创新的计算架构，通常包括 CPU、GPU、FPGA 和专用 AI 加速器等多种硬件组件，每种组件都有其独特的优势和适用场景。例如，CPU 擅长处理顺序执行的任务，而 GPU 则在并行处理大量数据时表现出色。FPGA 可以根据特定应用进行编程，以优化特定的计算任务。AI 加速器则专门为深度学习算法设计，能够在保持高能效比的同时提供强大的计算能力。通过合理分配任务到这些不同的硬件上，可以实现最佳的性能表现。

在智能座舱系统中，异构推理计算技术的实施方法涉及多个层面。首先，需要进行硬件选择和配置，确保各种组件能够满足系统的性能要求。接着，软件层面的优化也非常关键，包括算法的并行化设计、内存管理的优化以及跨硬件的通信机制。此外，还需要开发高效的调度策略，以确保任务能够在最适合的硬件上运行。异构推理计算技术的优势在于其灵活性和可扩展性，还能够根据任务的实时性要求动态调整资源分配，从而实现高性能与低功耗的平衡。

7.2.1.3.4 安全保护技术路线

在追求智能化的同时，数据安全和隐私保护是智能座舱语音交互技术发展中不可忽视的一环。为了确保数据安全，需要采用一系列的加密技术。对用户数据进行加密存储和传输，防止数据被非法窃取或篡改。匿名处理也是保护用户隐私的重要手段。同时，还需要建立严格的权限管理机制，只有经过授权的人员才能访问和使用用户数据。

此外，加强用户教育也是非常重要的。让用户了解数据安全和隐私保护的重要性，提高用户的安全意识和自我保护能力。同时，智能座舱语音交互系统的开发者和提供商也应该承担起相应的社会责任，遵守相关的法律法规，保障用户的合法权益。

7.2.2 智能座舱视觉交互场景发展趋势与技术路线

7.2.2.1 发展概述

AI 视觉交互技术源于车载视频监控系统，旨在增强驾驶安全。早期侧重环境感知，后随深度学习和 CNN 的应用实现质变，能分析驾乘人员表情、姿势、手势和眼动，检测疲劳分心并采取安全措施。随着智能座舱发展，其应用从安全监控拓展到提升乘车体验，如情绪理解、手势识别、唇动视线识别等。未来，随着技术进步和数据处理能力提升，座舱 AI 视觉交互场景将更广泛，推动智能座舱向多模态、情境感知交互模式演进，提供个性化体验，成为人车互联实践者。

汽标委智能网联汽车分委会

7.2.2.2 场景应用

AI 视觉交互技术进化已渗透汽车使用全程，并带来全方位智能体验。用户上车前，人脸识别技术可解锁车门、启动系统；驾驶时，系统实时监测驾驶员疲劳度，提高安全性，还支持手势控制车内娱乐和环境系统，减少分心操作。乘客识别、物品监测、唇动与视线追踪等新功能，丰富了智能座舱交互，提升了车辆安全配置，展现了 AI 视觉交互技术的广阔应用与潜力。

7.2.2.2.1 人脸认证

人脸认证技术利用高精度摄像头捕捉面部特征，与车辆数据库中预存数据匹配，成功后自动解锁车门并激活内部系统。

图 7-3 人脸认证技术图

7.2.2.2.2 疲劳与注意力监测

在驾驶过程中，疲劳监测系统通过分析驾驶员的眼睛活动、眨眼频率和头部位置来评估其警觉状态，并通过声音或视觉警告提醒驾驶员，以确保行车安全。

图 7-4 疲劳检测&驾驶行为识别预警系统图

7.2.2.2.3 手势识别控制

手势识别系统利用车内的 3D 摄像头捕捉解析手势动作，允许驾驶员通过手势来操作车辆的多媒体系统、空调等功能，从而减少驾驶过程中的物理分心。

图 7-5 手势识别控制图

7.2.2.2.4 危险行为监测与反馈

在监控并纠正驾驶员的危险驾驶行为的场景中，系统分析驾驶行为与道路交通规则，实时反馈给驾驶员并在必要时自动调整行车状态，以维护道路安全。

图 7-6 危险驾驶行为识别与反馈图

7.2.2.2.5 乘客识别

乘客识别系统运用高级视觉 AI 技术，捕捉车内乘客面部特征并与预存数据匹配，确认身份后提供个性化服务。

图 7-7 乘客识别系统图

7.2.2.2.6 遗留物品监测

物品监测系统专注于识别并追踪车内物品的位置和状态。利用内置的摄像头，该系统能够检测到重要物品是否遗留在车内，并监测没有儿童或宠物被无意留在车内，提高车内安全管理。

图 7-8 遗留物品检测图

7.2.2.2.7 唇动识别

唇动识别技术通过捕捉和分析驾驶员的口部动作来辅助语音命令系统，特别在车内嘈杂环境中提升命令识别的准确率。

图 7-9 唇动识别技术图

7.2.2.2.8 视线追踪

视线追踪技术通过实时分析驾驶员的视线方向和焦点，用来评估驾驶员的注意力分散程度，通过视觉或听觉警告提醒驾驶员注意驾驶安全[139]。

汽标委 智联汽车分委会

图 7-10 视线追踪技术图

7.2.2.2.9 AR-HUD

AI 增强的 AR-HUD 技术融合人工智能与 AR，通过车辆传感器收集道路信息，并以增强现实形式展示给驾驶员，如避碰撞提示和最佳路径规划等。同时，系统分析驾驶员行为，提供个性化建议并调整显示设置，提升驾驶安全性和便利性。

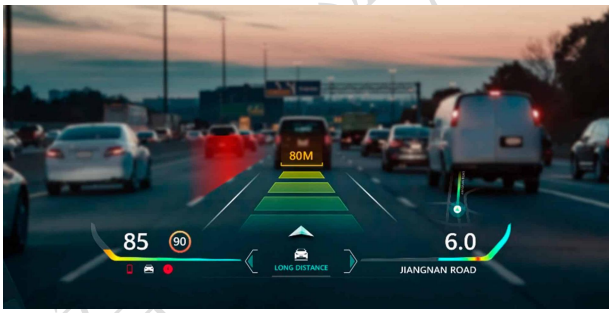


图 7-11 AR-HUD（增强现实抬头显示）技术图

7.2.2.3 技术路线

智能座舱视觉技术以机器视角的方式来观察交互者，分析判断出用户意图，再通过对应执行机构来满足交互者的使用需求。通常包括感知、识别、执行三个步骤。感知阶段机器视觉传感器实时采集舱内画面将图像输入给检测单元，检测单元通过训练好的算法模型对图像进行逐帧识别，识别到用户意图后，通过通信技术将执行结果传输给执行器执行，激活相应的车辆功能。此外，它还可以将驾驶员视线之外的区域通过车载显示屏或其他界面呈现给驾驶员，从而增强驾驶安全和便利性。

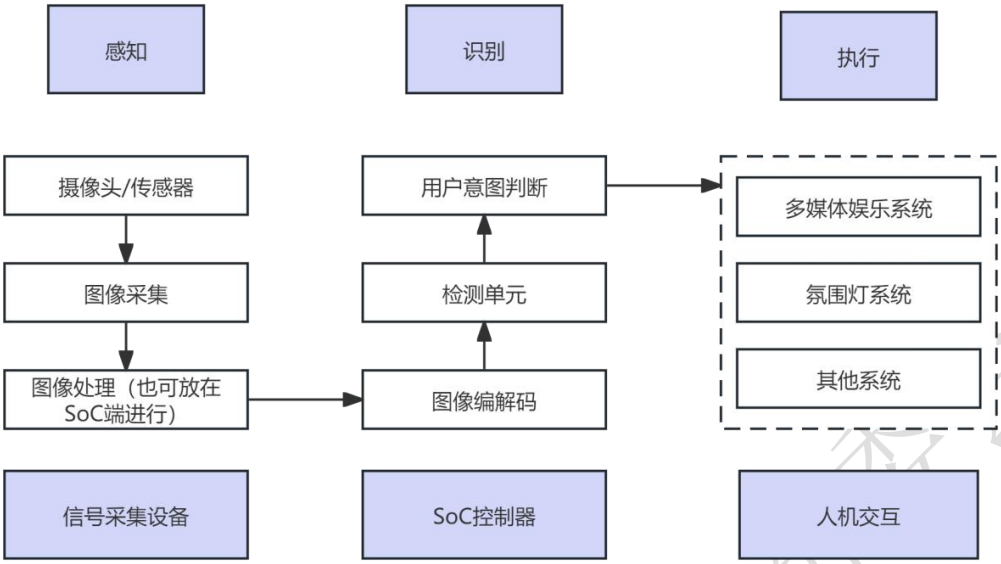


图 7-12 智能座舱视觉技术图

7.2.2.3.1 DMS 技术路线

DMS（驾驶员监控系统）是智能座舱视觉交互技术的重要领域，早期方案通过监测车辆信息或生物信号采集装置来间接判断驾驶员疲劳状态，但准确性和干扰性差。随着机器视觉技术进步，利用摄像头和近红外技术，从人眼、嘴部的开合状态来判断驾驶员疲劳状态成为主流方案。

系统通过算法模型对图像进行人脸和关键点检测，将嘴部和眼部关键位置抠图出来，根据基于眨眼行为的 EAR 和 PERCLOS 特征，和基于哈欠行为的 MAR 和 FOM 特征，判断是否出现闭眼和打哈欠状态，当达到阈值设定即可判断为疲劳状态。目前 DMS 功能不仅限于疲劳监测，还集成了吸烟、打手机等不良行为监测功能。系统识别到驾驶员有疲劳状态或不良行为后通过通信技术，将信号传输给声光报警器或仪表灯执行机构发出提醒，提醒驾驶员谨慎驾驶。

汽标委智能网联汽车分委会

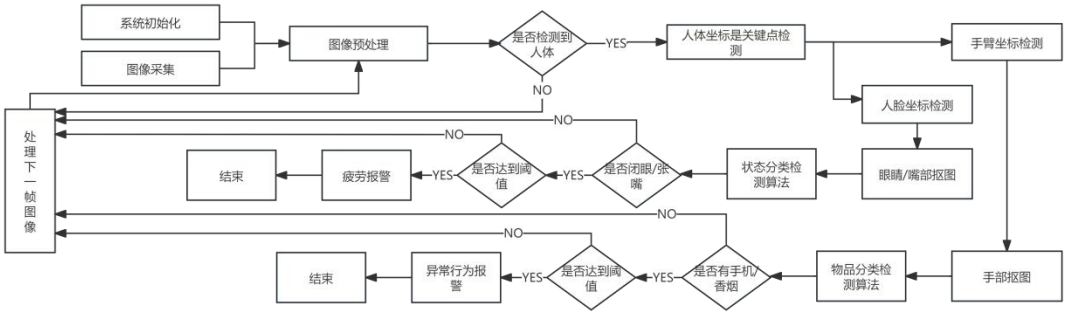


图 7-13 DMS 技术路线图

7.2.2.3.2 人脸识别技术路线

人脸识别技术分 2D 和 3D 两种，2D 人脸识别通过 RGB 相机采集 2D 人脸图片，提取特征向量并与数据库比对，但存在对光照条件依赖、易被照片欺骗等弊端。常见的 3D 成像方案有双目立体视觉、结构光和 TOF 相机方案，其中基于 TOF 相机的人脸识别技术是一种非接触式生物识别技术，更具性价比和易用性，在智能座舱领域应用更广泛。

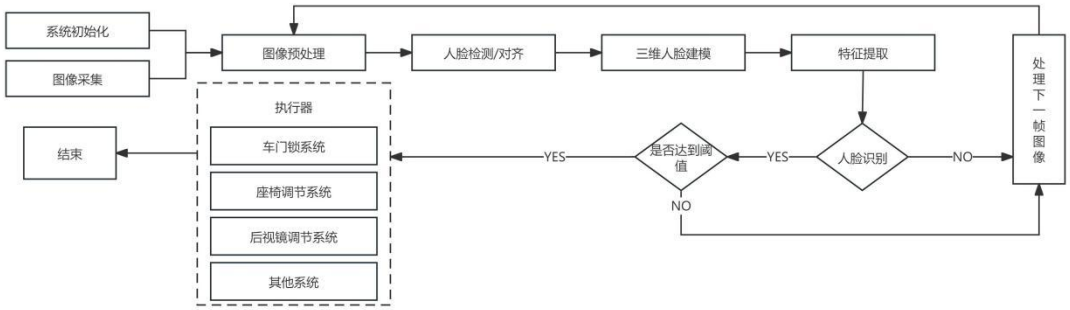


图 7-14 人脸识别技术路线图

7.2.2.3.3 手势识别技术路线

传统座舱交互采用按键方式，需要驾驶员花费时间寻找按键，增加安全隐患。手势识别交互可通过简单肢体语言实现座舱控制，增加交互的多样性、娱乐性和沉浸感，且更加安全、直观。智能座舱手势识别技术结合计算机视觉和人工智能，通过高清摄像头或深度传感器感知捕捉手部动作，经过去噪、增强和色彩校正等

预处理。对手部进行检测、定位关键点并识别特定手势，再将识别到的手势映射到预设动作类别中，明确驾驶员需求并传递信号给车辆系统执行操作。

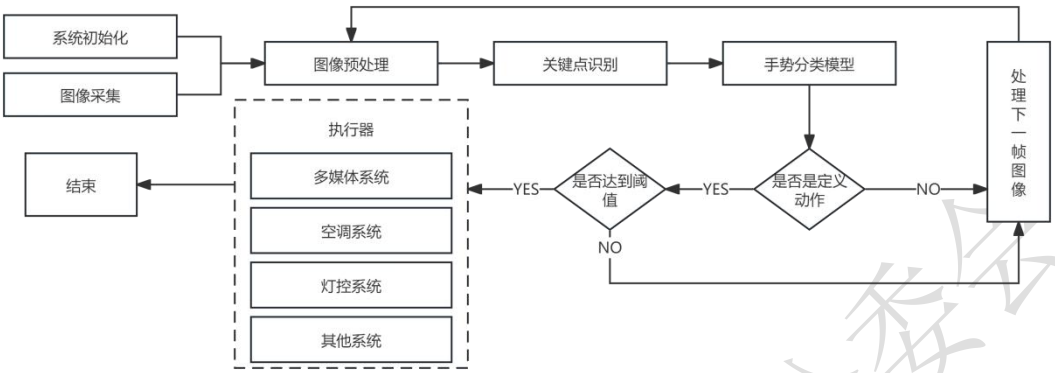


图 7-15 手势识别技术路线图

7.2.2.3.4 AR-HUD 技术路线

智能座舱视觉交互技术涵盖机器与人的视觉交互，显示设备是关键的人机交互终端，传统仪表因需转移视线存在安全隐患，故引入 HUD 技术。AR-HUD 将增强现实信息投影到驾驶员前方视线内，按投影技术分为 LCD、DLP、激光扫描投影等，LCD 成熟，DLP 优势渐显，激光投影在探索。为实现数字投影与驾驶环境融合，AR-HUD 多采用双焦面技术，近景显示车速等信息，远景显示导航等，提升驾驶安全性和信息获取效率。

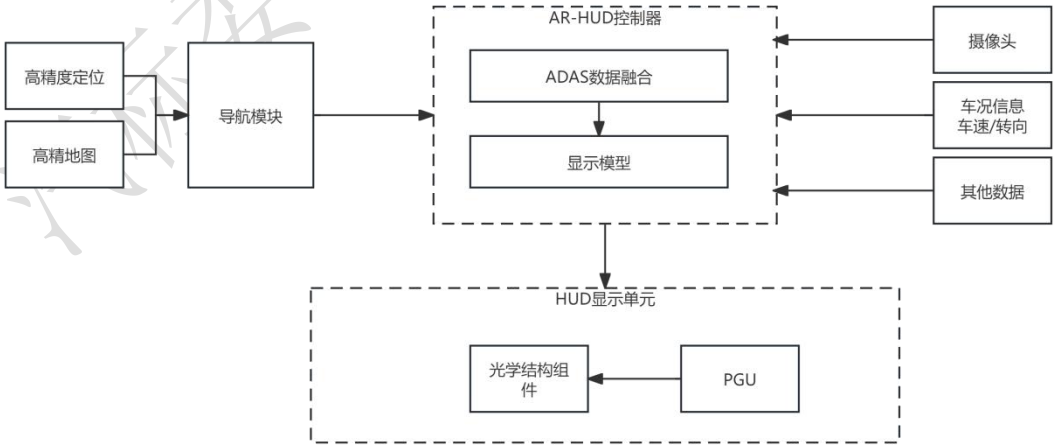


图 7-16 AR-HUD 技术路线图

7.2.3 智能座舱多模交互场景发展趋势与技术路线

7.2.3.1 发展概述

模态最初指生物体通过感知器官接收外界信息。在人机交互中，多模态交互指系统整合人类的视觉、听觉、触觉等感官输入，利用多种通信通道响应，模拟人与人之间自然的交互方式[140]。

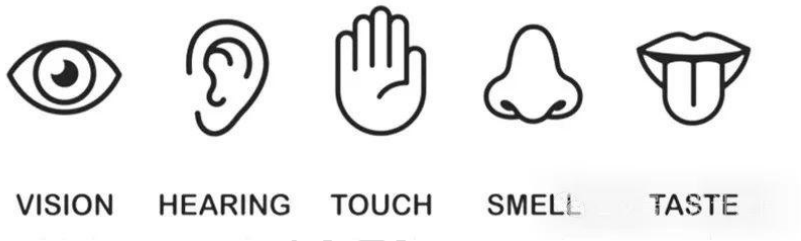


图 7-17 生物体多模态交互图

汽车制造商已实施多种多模态交互方案，如语音结合唇动、面部、手势识别等，其中语音与其他模式结合的交互形式是主流，已在多款车型中应用。这些系统丰富了交互体验，提高了可用性和直观性。随着技术进化和需求提高，未来智能座舱将通过更精细智能的多模态交互，创造更安全、舒适、个性化的驾驶环境。

7.2.3.2 场景应用

目前车企已实现的多模态融合包括但不限于语音+唇动识别、语音+面部识别、语音+手势识别、语音+头姿、面部+情绪识别、面部+眼球追踪、香氛+面部+语音识别等，其中语音多模态交互方式为当下主流。

7.2.3.2.1 听觉融合

听觉融合在智能座舱中通过结合语音识别与其他感官输入来增强交互体验和响应准确性。相关示例包括：

表 7-2 听觉融合相关示例

交互模态	功能介绍
语音+手势	驾驶员可以同时发出语音指令和手势，比如通过伸出食指并指向左右或上方，来控制车窗或天窗
语音+人脸	通过摄像头识别语音指令的发出者，自动过滤主驾位的非控制语音，如与其他乘员的对话
语音+眼球追踪	通过视线来唤醒语音识别功能，使驾驶员无需物理接触或明确命令即可激活系统
语音+唇动	采用红外光感摄像头智能识别主驾唇形，结合唇语读取和语音特征的双重识别技术，极大提升系统的唤醒和识别准确率
语音+行为追踪	允许驾驶员在车外通过语音指令“跟随我”，使卡车能在安全距离内自动跟随

7.2.3.2.2 视觉融合

在智能座舱的视觉融合技术中，通过结合人脸识别、眼球追踪以及其他生物识别技术，车辆能够提供更为精准和个性化的驾驶辅助及车辆控制功能。以下是几个具体的应用示例：

表 7-3 视觉融合相关示例

交互模态	功能介绍
人脸识别+眼球追踪	DMS 驾驶员监控系统集成了人脸识别和眼球追踪技术，支持疲劳监测和眼球追踪，实时调整 AR-HUD 投影角度
人脸识别+心率监测	通过搭载 52 个生物传感器，可以识别驾驶员的生理状态从

汽标委智能网联汽车分委会

	而自动调节车内的音乐和温度，以及危险时接管车辆
人脸识别+静脉识别	“Leap In 生物钥匙系统”，结合了人脸识别和静脉识别技术。可实现车门的快速解锁和车辆启动等

7.2.3.2.3 嗅觉融合

嗅觉融合通过结合香氛系统与语音命令、人脸识别等技术，极大地增强了驾驶体验的丰富性和个性化程度，以下是几个具体的应用示例：

表 7-4 嗅觉融合相关实例

交互模态	功能介绍
香氛+语音	通过语音命令唤醒座舱的场景模式，在醒神模式下，系统自动开启空调、香氛、氛围灯、座椅震动以及动感音乐等多种功能，帮助驱散驾驶疲劳
香氛+人脸	系统能够监测到驾驶员的疲劳状态，并自动启动车内香氛系统释放提神醒脑的香氛
香氛+语音+人脸	能够通过语音唤醒特定的场景模式，如“醒神潮汐”模式下系统会与 DMS（驾驶员监控系统）进行联动，监测到驾驶员疲劳时，系统进行语音提醒及释放醒神香氛

7.2.3.3 技术路线

智能座舱的多模态融合技术整合多种感官输入，通过高精度传感器捕捉数据，经预处理和深度学习算法分析，实现多模态数据融合与协同工作。该技术需优化算法和硬件配置以确保实时响应，并通过系统集成优化用户体验和驾驶安全性[141]。

汽标委智能网联汽车分委会

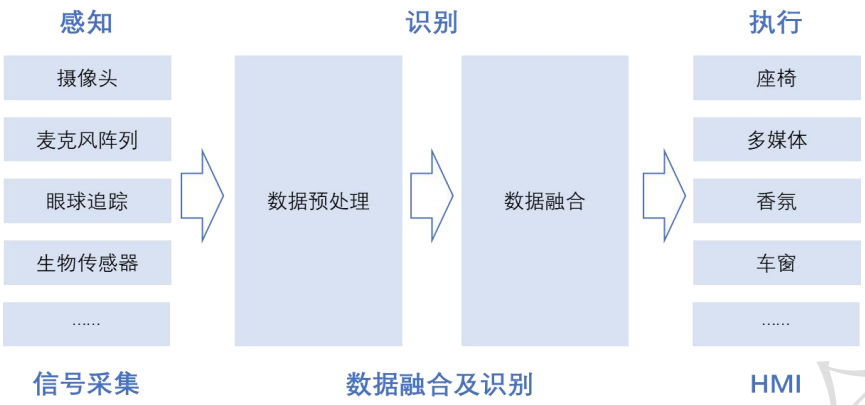


图 7-18 智能座舱多模态融合技术路线图

7.2.3.3.1 听觉融合技术路线

听觉融合技术在智能座舱中应用，通过结合语音识别与其他感官输入，实现自然高效交互。其核心在于多模态传感器集成与高级数据处理。高精度麦克风阵列和手势识别摄像头的部署，可捕捉语音和手势数据，经降噪、滤波等预处理后，利用深度学习算法解析并融合数据，生成精确控制指令。例如，驾驶员可通过语音指令结合手势（如伸出食指指向左右或上方）控制车窗、天窗或遮阳帘。未来，听觉融合技术将向多感官协同与智能化方向发展，结合环境感知和情感识别技术，实现更加个性化和自然的交互体验。

7.2.3.3.2 视觉融合技术路线

视觉融合技术通过结合人脸识别、眼球追踪等生物特征识别技术，为驾驶员提供精准个性化驾驶辅助和车辆控制。其关键在于多传感器系统部署和数据融合技术应用。高分辨率摄像头、眼球追踪设备等布置在驾驶舱关键位置，全面覆盖驾驶员面部特征和生理状态。系统集成与性能优化确保视觉融合系统与车辆中央控制系统无缝对接，性能测试和用户反馈循环不断改进系统功能和体验。未来，该技术将融合情感识别，监测驾驶员情绪状态，提供更人性化的驾驶辅助和提醒服务[142]。

汽标委智能网联汽车分委会

7.2.3.3.3 嗅觉融合技术路线

嗅觉融合技术通过香氛系统与语音、人脸识别等技术结合，提升驾驶体验丰富性和个性化。其实现依赖多感官系统集成和多模态数据处理，如香氛装置与麦克风阵列、摄像头集成，可依用户需求和车内环境精确控制香氛。该技术还可结合驾驶员监测系统，在检测到驾驶员疲劳时自动释放提神香氛，同时通过数据融合技术综合分析语音、人脸和驾驶状态数据，生成控制指令。未来，嗅觉融合技术将结合更多生物识别和环境感知技术，实现更智能、个性化的交互体验。

7.2.4 智能座舱大模型

7.2.4.1 大模型应用于语音交互场景

智能网联汽车中的大模型语音交互是指通过将大模型应用于智能网联汽车中的语音交互场景。相比于传统车内语音交互，大模型可以提供更加自然的拟人化交互，以及具有更加智能的理解能力，能够从语音中识别出情感，并作出对应响应。大模型应用于语音交互场景可以提供更加智能化，拟人化，情感化的交互以及更加丰富的场景。下面以语音识别，车内指令控制，情感化拟人化助手，语音合成等四个方面介绍大模型应用于语音交互场景下的技术路线与应用价值[143]。

7.2.4.1.1 语音识别

语音识别是将语音转化成对应文本的技术，传统的语音识别使用 DNN（Deep Neural Network），CNN（Convolutional Neural Network），RNN（Recurrent Neural Network）等方法，传统方法对语音序列进行建模，提取语音信号中的声学特征，再将特征映射到对应的单词以进行语音识别。传统方法对于语音中噪声的过滤以及一些地方方言，多语言识别效果有限。

汽标委智能网联汽车分委会

大模型可以更好地对语音序列中的时间序列信息进行特征提取，从而实现更高纯度的噪声过滤。同时根据大模型的理解能力，可以实现免唤醒，多人语音识别，语音断点检测等。相比于传统方法，大模型在语音识别中的应用优势如下：

(1) 多语言，方言识别。由于大模型具有更加复杂的架构以及更多的参数，能够捕捉到更加细微的语音信号特征，同时由于参数量的提升，大模型在语音识别泛化性上也具有明显的优势。通过在训练数据中加入多种不同语言，方言等数据，大模型可以在推理时对一句话中的多语言以及地方方言进行准确识别。

(2) 对噪声，语速变化的鲁棒性。大模型具有更强的特征提取能力，对于存在噪声以及语速突然变化的语音识别条件下，大模型通过对语音进行特征提取，依然可以保证语音识别结果的鲁棒性。

(3) 多任务学习。大模型可以同时学习多个相关任务，例如在语音识别的同时进行说话人识别，情感分析以及语音唤醒等任务。传统方法则需要对每个不同任务单独训练一个模型来完成。

(4) 持续学习。大模型可以通过后续的增量预训练或者微调技术来进行迭代，不断适应新的环境数据变化，例如网络上新出现的流行语识别等，以提高长期性能。

7.2.4.1.2 车内指令控制

车内指令控制是指通过用户输入语音转化后的文本对车内一些模块进行控制，传统的指令控制方法通常采用预定义好的语音指令来实现车内模块的控制，存在着识别指令有限且必须按照指令预定义好的规则才能触发的问题，采用大模型进行车内指令控制存在以下几点优势：

(1) 多维度分析与全面预测。大模型具备多维度分析能力和全面预测能力，通过对海量数据进行训练分析，同时结合当前车内状态以及与用户的上下文进行理解，然后在复杂场景下做出准确判断。例如根据分析出的多个指令，控制车窗的具体打开程度，车门开关，各类灯光（远光，近光，雾灯）的开关，

汽标委智能网联汽车分委会

后视镜开合，雨刷器运行等功能。根据习惯分析用户的意图，控制各类媒体播放器，播放音频，视频等多媒体信息。理解用户的复杂指令，泛化出用户的详细操作指令，控制各类多媒体应用程序播放音乐，视频和图像。控制音量大小，播放进度，显示大小等。

(2) 持续优化与自我学习。通过不断与用户交互，大模型可以持续学习用户的习惯，这是传统基于规则的指令控制所不具备的。能够识别使用者，理解使用者的语言习惯，能够更精准地读懂用户表达不够充分的语言。根据习惯分析用户的意图，调节空调温度风速，座椅展开角度，灯光类型，雨刷器模式等车辆设置。持续学习后，大模型可以更加准确地识别根据用户习惯进行的指令控制。

(3) 主动推荐。基于大模型的车内指令控制可以通过当前环境状态主动推荐用户做出某些特定的指令，在用户进行确认后执行该指令。实现更加智能化的指令控制。

7.2.4.1.3 情感化、拟人化助手

情感交互助手是指大模型通过分析使用者的情感，生成符合使用者情感的回答内容，实现共情闲聊。能够了解用户的情绪和心理状态，并通过对话完成心理解压和疏导；用户遭遇情感难题时，可以为其提供情感指导；遇到职场问题时，也可以为其答疑解惑。此外，大模型在与驾驶员对话时，大模型可以根据对话内容和情绪氛围生成更加贴合情感的回复，使对话更加生动和富有同理心。大模型在情感化、拟人化助手中有以下几种应用价值：

(1) 情感化导航指引在提供导航指引时，大模型可以根据驾驶员的情绪状态调整语言风格。如在驾驶员紧张时使用更加温和和安抚的语言。

(2) 情感化紧急响应。在发生事故或车辆故障时，系统可以生成富有同情心的语言以及语音来安慰车主，并提供紧急情况下的指导和支持。

(3) 拟人化的用车助手。基于大模型的理解能力，帮助用户在车辆参数，零部件操作，功能使用，汽车保养以及故障处理等全链路用车问题上可以随时随

汽标委智能网联汽车分委会

地获取专业的用车知识和解决方案，提高用车的便利性和安全性，成为用户贴心的用车助手。

(4) 拟人化的车内健康顾问。基于大模型，可以为车内配备一个虚拟健康顾问，能够结合与车主对话的情绪状态以及 DMS，OMS 识别到当前车主的生理状态提供健康建议与情绪支持。

(5) 拟人化角色定制。基于大模型，用户可以根据个人喜好定制一个或多个虚拟角色，并可赋予角色不同功能，比如旅游规划助手等。这些角色能够展现不同的性格特征，并在与用户的交互中展现相应的情感反应。

(6) 情感化历史解说。基于大模型，当车辆经过某个历史地标区域时，大模型可以结合驾驶员的情绪状态提供相匹配的历史解说，使历史信息更加生动化，情感化。

(7) 自由对话。实现人机自由对话，多轮对话。通过语音助手可以对话的内容包括但不限于知识问答，景点推荐，美食推荐，口语教练，信息查询等功能。

7.2.4.1.4 语音合成

语音合成是指将文本合成为语音的过程，传统的语音合成方法首先对文本进行语义分析，再将结果进行波形预测后合成得到语音。这种方法虽然可以得到较为自然的音频，但是在波形预测的过程需要大量的音频片段以供后续预测拼接，这使得语音合成存在很大的局限性。基于大模型的语音合成相比于传统方法存在以下几种优势：

(1) 自然度与灵活性。大模型通过学习到更丰富的语音特征，能够生成更加流畅的语音。同时大模型可以适应不同的语音风格和情感表达，提供更多样化的语音输出。

(2) 个性化与多语言支持。大模型可以模仿特定人的语音特征，实现高度个性化的语音合成。大模型同样支持多种语言的语音合成，而不需要为每种语言单独开发模型。

汽标委智能网联汽车分委会

(3) 上下文理解。大模型具有传统方法所不具备的长上下文理解能力，可以通过上下文生成合适的语音风格以及情感等。

大模型在上述功能或者产品体现出对于强大的理解能力，泛化能力进行语音识别或合成，以及通过推理能力解决用户的复杂需求。大模型将智能网联汽车中若干个小功能模块全部集成到一个大模型内部，不需要对独立模块进行开发。在情感化拟人化场景中，大模型作为情感助手，体现出对于上下文的理解能力，通过对用户的情绪理解来安慰或鼓励用户。

7.2.4.2 大模型应用于视觉交互场景

基于大模型的视觉交互技术在智能座舱领域展现了广泛而多样的应用场景。驾驶员监控系统（DMS）、乘员监控系统（OMS）、人脸识别 （FaceID）这些技术的日益成熟，不仅标志着汽车座舱向智能化、个性化空间的转变，也是智能出行时代加速演进的关键驱动力。在此基础上，大模型（LLM）的引入能够对智能车舱内视觉交互场景进行进一步深度优化与革新，以下是大模型应用的视觉交互场景[144]。

(1) 疲劳检测

通过摄像头监控驾驶员的眼睛和面部表情，检测驾驶员的疲劳状态，并在必要时发出警告。

(2) 注意力监控

追踪驾驶员的视线，判断其是否在关注道路，若注意力分散（如长时间未看前方），系统会发出提醒。OMS 还能够更细致地监控乘员动态，例如乘员的位置、年龄、是否系好安全带，甚至通过面部表情和身体姿态判断乘客的情绪状态。

(3) 驾驶行为分析

通过摄像头监控驾驶员的行为，如频繁转头、使用手机等，提供驾驶行为分析报告，并在检测到危险行为时发出警告。

(4) 主动响应与环境感知

汽标委智能网联汽车分委会

结合视觉和其他传感器数据，实时监测车内外情况，做出智能响应，如检测到紧急情况自动报警等。

(5) 人脸识别

基于视觉识别的乘客身份信息，系统能够不断学习特定用户的偏好，提供更贴心的服务，例如提供个性化内容推荐。

总体而言，这些融合了驾驶员监控系统（DMS）、乘员监控系统（OMS）和人脸识别（FaceID）等技术的应用，不仅显著提升了汽车的智能化和安全性，还为驾驶员和乘客带来了更智能、更便利、更个性化的驾乘体验。

7.2.4.3 大模型应用于多模态交互场景

智能网联汽车中的多模态大模型应用是指结合语音、视觉、行车信息，生物信息等多种模态的输入信息，通过大模型进行理解和泛化推理，之后输出文本，任务，语音，图像，视频等多种模态信息。大模型的多模态应用极大地改善了智能座舱交互方式。下面从精确感知乘客意图、车辆状态感知与反馈，智能体虚拟人交互体验和多模态生成能力的应用四个方面分析大模型在多模态交互场景中的应用价值[145]。

7.2.4.3.1 精确感知乘客意图

大模型可以根据用户的语音信息，图像信息，以及车内外的环境信息来精确感知用户的指令意图。避免理解不充分和误解用户的真实意图。相较于传统算法对于复杂，可变，无法预测的环境适应能力较弱，容易受到各类输入噪声和未训练情况影响，大模型可以将语音信息，视频信息以及其他传感器的信息通过编码向量化作为大模型的输入信息。再凭借大模型的泛化能力，推理能力可以更准确地感知和理解用户的意图，从而做出更加准确的判断和行动。例如，通过车内语音和视频信息判断说话人的潜在语义，肢体动作，视角朝向，情绪，从而理解对方具体想操作座舱的哪部分功能。不需要用户提供复杂精确的指令。

汽车委智能网联汽车分委会

7.2.4.3.2 车辆状态感知与反馈

大模型除了可以感知来自用户的信息之外，还可以感知车辆的信息。包括但不仅限于车内的温度，湿度，空气质量，车辆的胎压，发动机温度，车速等信息。相较于传统算法基于设定的模式和规则来判断车辆状态，大模型可以将座舱内各类传感器获得的信息，车辆的行驶信息，状态信息通过编码作为大模型的输入。再利用大模型的学习能力，泛化能力和推理能力判断车辆状态并自动调整车辆设置或者向乘客预警。大模型可以适应多变的车辆状态和车内环境以及用户的个性化需求。例如，可以综合车内的温湿度传感器，气压传感器，空气流量传感器等数据，感知座舱内实时的温湿度和气压，气流状况。参考以往用户偏好实时调节空调系统来调节座舱温度和湿度，调整通风系统和空气过滤器。也可以实时监测轮胎压力、发动机温度等机械状态，出现异常时可以及时溯源问题来源和预警提示。

7.2.4.3.3 智能体虚拟人交互体验

借助大模型可以实现具备人类丰富表情，肢体动作，情绪，语言能力的虚拟人 AI 智能体，实现接近与真人交流的感受。并且也可以通过形象复刻，声音复刻来模拟真实存在的人类，实现交互对象的数字孪生。通过虚拟人 AI 智能体，可以像老朋友一样理解乘客的需求，模拟人类的语言，行为和情绪和表情来和乘客进行交互。帮助乘客操作座舱中的各项功能，实现复杂的操作需求。与传统的人机交互方式相比，大模型凭借推理和泛化能力赋予虚拟人带有情绪的，更贴近于日常语言的语音表达。并且大模型根据表达内容和情绪来生成和控制虚拟人的面部表情和肢体动作。在虚拟人建模方面，也可以根据用户提供的素材，描述来定制个性化的虚拟人形象。例如，智能体虚拟人可以像真人小伙伴一样和后排的儿童做各种游戏，和孩子聊天。也可以体会乘客的情绪，对乘客进行心理辅导，情感交互。在交互的体验方式上，从与传统的人与机器交互的体验转变为与高度拟人化的虚拟人 AI 智能体进行交互，是交互方式的重大升级，也将重塑智能座舱交互的方式。

汽标委智能网联汽车分委会

7.2.4.3.4 多模态生成能力

大模型在生成能力方面可以实现文生图，文生视频，文生音乐，图生视频等多样的功能。实现各种模态信息之间的创新创造，为智能座舱附加新的能力。例如，可以根据描述生成车机的壁纸，孩子的童话绘本，甚至一段儿动画视频，提升乘客的幸福感和满意度；还可以根据语音，图像，行驶信息识别驾驶员和乘客的情绪和需求，生成合适风格的音乐。也可以根据行程路线与车载相机素材，生成旅行短视频。

综上所述，相较于其他现有的技术，多模态大模型可以将多种模态的数据作为大模型的输入，利用大模型适应力鲁棒性强，泛化能力和推理能力强的特点，能够理解复杂环境，处理更加复杂的任务。使得智能网联汽车中的多模态交互场景更加丰富与智能，更加贴近人类交流的方式。未来，现有的模态将进一步扩充，从现实世界中获取更多类型模态的信息作为输入。各类训练数据的质量和多样性的提升，将提升模型的训练效率以及各项能力和质量。同时，随着模型轻量化部署技术的发展，以及车机端芯片能力的提升，更多的多模态应用场景可以在车机端运行推理，在网络环境不佳的特殊情况下也能运行部分功能。

7.2.4.4 大模型应用于开放式任务场景

大模型开放式任务指的是通过大模型来检索完成某一任务需要的信息，并通过系统传感器、数据库等渠道获取跨领域信息来完成复杂的任务，逐步实现座舱智能体。下面从任务规划、生成式交互、健康监测、插件与信源扩展和检索增强生成五个方面分析大模型在开放式任务中的技术路线和应用价值[146]。

7.2.4.4.1 任务规划

大模型在复杂任务生成方面的应用体现在三个方面：首先，在对话交互方面，车机系统可以精准识别用户的语音指令，并创建符合用户需求、用户偏好及与场景匹配的复杂任务。例如，主驾可以通过语音指令在车机系统上设置“当后排乘客在下午四点后上车时，就开启后排空调到 25 度并播放英语教学视频。”其次，在任务规划方面，大模型可以将复杂的任务拆解为更小的子任务。例如，用户提

汽标委智能网联汽车分委会

出“根据日出日落开关氛围灯”，可以将任务拆解为日出关闭和日落打开氛围灯。最后，在数据管理和记忆的方面，大模型可以作为工具，保存和检索过去的数据。例如，大模型记住用户的偏好设置，并在需要时提供相应的信息。

7.2.4.4.2 生成式交互

生成式 UI 界面和交互模式可以根据用户需求和环境变化进行实时调整，从而提供个性化的体验。首先，大模型可以自动生成与用户需求相对应的脚本。这些脚本能够描述各种交互场景、功能调用和用户操作步骤。举例来说，当驾驶员通过语音指令请求导航到某一目的地时，大模型能够实时生成对应的导航脚本，包括路径规划、交通状况更新和实时语音指导等功能。一旦生成，这些脚本可以自动调用相关接口，实现无缝的功能集成。

其次，动态注册功能使得系统能够灵活应对不同的场景和需求。大模型可以通过动态注册机制，自动调整车载 UI 的布局和功能。例如，当检测到车内有儿童时，系统可以自动注册并显示相关的娱乐内容和教育应用。

此外，通过自动生成与调用接口方式的脚本，系统可以更迅速地整合新功能和应用。大模型在后台自动生成接口调用代码，简化了开发流程，大幅缩短了应用更新和功能拓展的时间。

7.2.4.4.3 健康监测

大模型具有强大的数据处理和分析能力，能够实时监测并解析驾驶员的生理状况，识别潜在的健康风险。例如，如果的心率突然升高，血压不稳定，大模型能够智能预测这些生理变化背后的原因，从而及时给予预警。其次，大模型可以通过深度学习形成对驾驶员习惯和状态的理解。这种个性化的分析能够为驾驶员提供更加准确和量身定制的建议，如提示驾驶员进行短暂休息，调整驾驶姿势，或是激活辅助驾驶模式以减轻压力。最后，当系统检测到驾驶员状态异常时，可以通过声音提示、座椅震动等多种方式提醒，使得驾驶员更容易注意到并采取相应措施。

汽标委智能网联汽车分委会

7.2.4.4.4 插件与信源扩展

通过将大模型与各种插件和信源集成，智能座舱不仅能实现功能的高度定制化和扩展性，还能提供更加详尽、实时的环境感知和响应能力。首先，大模型与插件的结合使智能座舱具备了高度的扩展性和灵活性。例如，智能座舱通过加载健康监测插件实时获取和分析驾驶员的生理数据，从而做出个性化的建议。

其次，信源的多样性和实时性是智能座舱能力提升的重要因素。大模型能够从多个信源获取数据，包括但不限于车内传感器数据、外部交通信息等。通过对这些多维度数据的整合分析，大模型能够做出智能决策[147]。例如，通过整合交通信息和天气预报，系统可以提前建议驾驶员选择更加安全的行车路线或合理的出行时间。

最后，大模型在接收信源和调用插件功能的同时，能够不断学习和适应用户的习惯和偏好，实现个性化定制。它可以通过对历史数据的分析，预判驾驶员可能的需求和行为，从而提前准备相关信息和功能。例如，大模型可以根据驾驶员的日常路线和习惯，为其推荐加油站、休息区或餐馆，实现“未询先知”的智能化服务。

7.2.4.4.5 检索增强生成

检索增强生成（RAG）是一种根据用户的查询语句搜索信息，并以搜索结果为 AI 参考从而生成回答。这项技术是多数基于 LLM 工具的重要组成部分，而多数的 RAG 都采用向量相似性作为搜索的技术。普通 RAG 技术在私有数据，如企业的专有研究、商业文档表现较差，而 GraphRAG 则基于前置的知识图谱、社区分层和语义总结以及机器学习技术可以大幅度增强此类场景的性能[148]。

7.3 智能座舱 AI 应用的关键技术

随着智能网联汽车技术的快速发展，大数据和人工智能技术在汽车行业中的应用越来越广泛。汽车正迅速从传统的交通工具转变为提供丰富体验的移动空间。

用车场景也日益多样化，用车需求的新变化推动了行业向更主动且智能化的服务方式转变。本章将针对智能座舱内 AI 应用的关键技术，从感知、认知、表达三个方面阐述 AI 算法是如何重塑座舱功能并提供个性化体验的。

7.3.1 智能座舱感知技术

7.3.1.1 声学前端技术

声学前端技术是音频信号处理的最前沿，负责对输入的原始音频信号进行预处理，如：噪声抑制、回声消除、语音增强等，以消除噪声、回声等干扰因素。且随着技术的不断发展，从传统的信号处理方法到深度学习方法，再到可学习音频前端的出现，声学前端技术正不断向更高效、更智能的方向发展。

(1) 噪声抑制（Ambient Noise Suppression，ANS）技术

噪声抑制技术基于噪声与语音的时频特性差异，通过信号处理分离目标语音与背景噪声。传统方法包括频域处理与时域处理，通过噪声估计和信号增强两阶段实现。现代技术结合深度学习，通过深度神经网络（DNN/CNN）或生成对抗网络（GAN）学习噪声与语音的复杂非线性映射，直接生成纯净语音谱或时频掩码。核心挑战在于平衡噪声抑制强度与语音失真程度，满足实时性约束。

(2) 回声消除（Acoustic Echo Cancellation，AEC）技术

回声消除技术旨在消除通话过程中由于声波反射导致的回声现象，确保语音通信的清晰性与舒适性。基于人工智能的回声消除技术通过深度学习模型替代传统自适应滤波器的线性假设，构建端到端非线性映射以解决复杂声学场景下的回声干扰问题。如图 7-19 所示，输入信号由远端参考信号（$x(n)$）与近端混合信号（$y(n)=d(n)+e(n)+v(n)$）（含回声 $d(n)$、近端语音 $e(n)$ 及环境噪声 $v(n)$）为输入，利用短时傅里叶变换（STFT）提取双通道时频域特征，并通过卷积循环网络（CRN）或 Transformer 等架构估计回声分量 $\hat{d}(n)$，计算初始残差 $r(n)=y(n)-\hat{d}(n)$，其次引入时域卷积网络（TCN）对残差信号进行谱增强，进

一步抑制残留回声与噪声。针对双讲场景，融合注意力机制动态分配近端语音与回声的权重，避免语音截断，继而输出纯净语音信号。

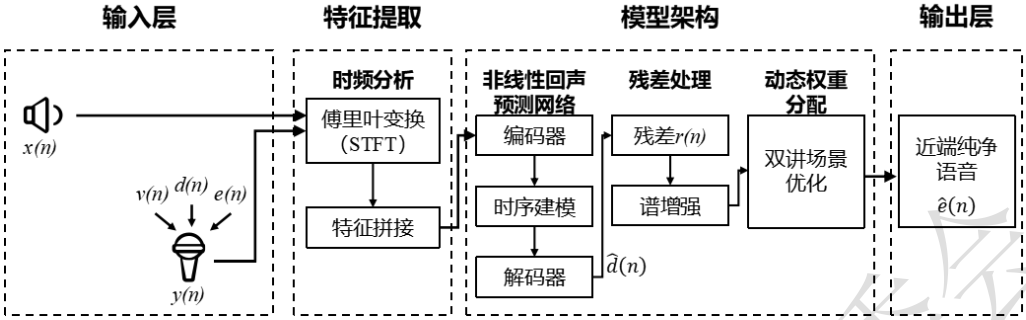


图 7-19 回声消除（Acoustic Echo Cancellation，AEC）技术原理

(3) 语音增强技术

语音增强技术旨在从复杂环境中提取纯净的语音信号，缓解噪声、混响、人声干扰及回声等问题，尽可能保留原始语音特征。语音增强系统主要包括数据读取、语音降噪、语音去混响和结果呈现与保存四个模块[149]，如图 7-20 所示。

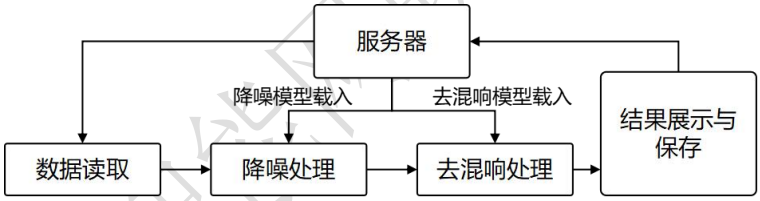


图 7-20 语音增强系统流程图

语音增强方法可分为传统数字信号处理和基于机器学习两类。在传统方法中，单通道和麦克风阵列方法是两类主要技术。单通道方法中，短时谱估计是最常用的技术，主要包含谱减法[150]、维纳滤波[151]和统计模型方法[152]。此外，自适应滤波法需依赖噪声或纯净语音数据进行优化，但受限于先验知识获取难度，通常需借助环境噪声信号实现降噪。

7.3.1.2 语音识别技术

语音识别（Automatic Speech Recognition, ASR）是将人类语言转换为机器可识别文本的重要技术，是人机交互和跨人类通信的关键桥梁。其研究涉及信号处理、概率论、模式识别、人工智能等多个实用领域。

语音识别主要有语音输入、特征提取、识别解码和文本输出这四个部分构成，如图 7-21 所示。输入的语音信号需经过预处理（如预加重、分帧加窗、傅里叶变换和对数处理），以提取 Fbank 特征[153]或 MFCC 特征[154]。此外，也可直接将原始语音信号输入神经网络模型，利用其隐含特征提取能力。语音信号的预处理旨在消除设备差异和噪声干扰，获取表征语音本质的特征。

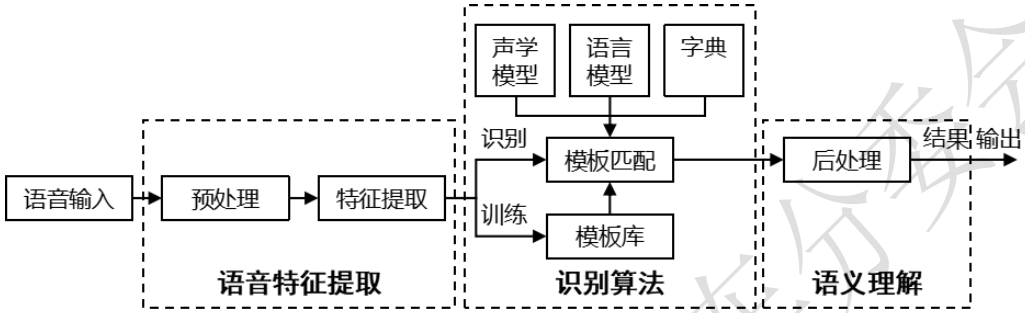


图 7-21 语音识别技术框架图

语音识别的核心是解码过程，其中声学模型（Acoustic Model, AM）[155]负责将语音特征映射为音素、字或词等最小单元，语言模型（Language Model, LM）则根据语言学规则和语义信息将其组合为完整语句。声学模型与语言模型的协同作用是语音识别实现的关键，共同构成了从声学特征到文字表达的核心技术。

7.3.1.3 语音唤醒技术

语音唤醒任务是一种轻量级的关键词检索任务，如图 7-22 所示。

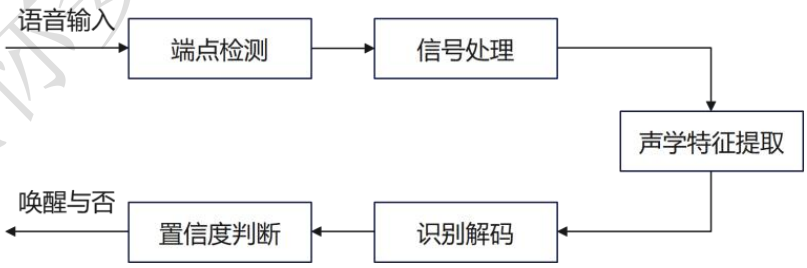


图 7-22 语音唤醒结构图

其实质是一个语音二分类任务，通过学习关键词和非关键词的音频特征，判断音频流中是否包含目标关键词，而非对音频内容进行完整转录。首先从大量训练数据中提取语音的特征模式，并通过复杂的声学模型分析语音信号的频谱特征、

汽标委智能网联汽车分委会

时序结构及上下文信息，从而在复杂噪声环境中准确区分唤醒词与其他相似词汇，有效降低误唤醒率。该技术不仅显著提升了识别的准确性和鲁棒性，还为个性化、多语言支持及安全性提供了技术保障。

7.3.1.4 人脸识别技术

人脸识别是一种基于图像采集设备获取人脸动态视频或静态图像的技术，通过提取有效特征并进行分析和归类实现身份识别，如图 7-23 所示。研究内容主要包括人脸检测与对齐、特征提取、身份识别及表情分析等方面[156]。

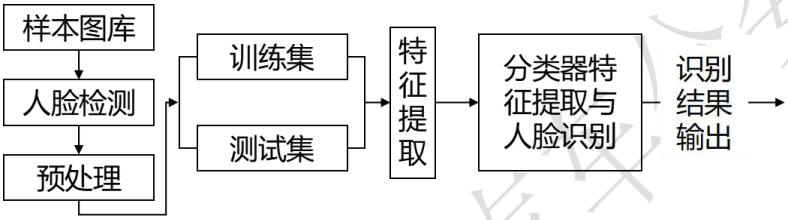


图 7-23 典型的人脸识别过程

典型人脸识别过程可分为四个关键步骤：通过人脸检测与定位，确定目标图像中的人脸区域并完成几何对齐；其次对检测到的面部图像进行预处理，包括几何校正和光照补偿等；之后从预处理后的人脸图像中提取具有辨识性的特征信息，并将其转化为可计算的特征表达；最后通过合适的分类算法对特征进行匹配与划分，完成身份识别。

7.3.1.5 动作识别技术

动作识别是对动作符号进行标签化的过程[157]，如图 7-24 所示。

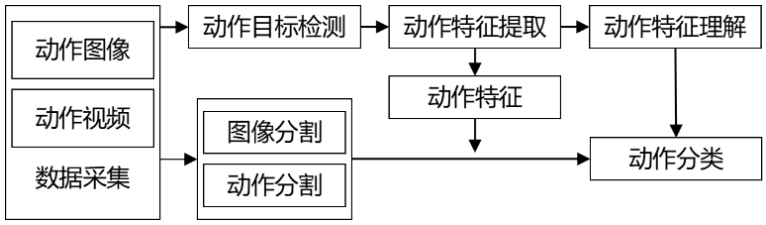


图 7-24 典型人体动作识别流程

汽标委智能网联汽车工委会

其本质是模拟人眼和人脑对动作信息进行理解区分的行为，典型的动作识别通常包含：运动目标检测、动作特征提取和动作特征理解。传统的方法多依赖于手动特征提取，主要分为全局特征方法和局部特征方法。全局特征通过定位人体位置进行动作识别，但易受噪音和遮挡影响。相比之下，局部特征方法通过提取关键点及其周围区域的图像信息，不依赖人体追踪和分割，因此更具鲁棒性。随着卷积神经网络（CNN）的发展，双流网络和 3D 卷积神经网络方法在动作识别获得广泛应用。双流网络处理图像的空间和时序信息，而 3D 卷积神经网络则在三维空间中处理视频数据，实现了更高效的动作识别。

7.3.2 智能座舱认知技术

7.3.2.1 语义理解

语义理解是自然语言理解（Natural Language Understanding, NLU）的重要研究领域，旨在通过建模使计算机能够理解和推断自然语言。语言的理解与分析是一个层次化渐进的进程，从表及里，如图 7-25 所示。

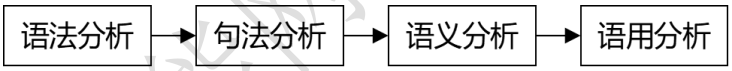


图 7-25 NLU 过程的一般步骤

词法分析旨在提取词素及其语言学信息，句法分析则通过分析词素间的线性序列，构建语法结构并揭示词素在结构中的语义和句法作用。常见的句法分析方法包括短语结构语法、功能语法等。语义分析侧重于建立词法意义、句法结构意义与任务域对象之间的映射关系，以确定语言文本所表达的任务指令或概念。语用分析则关注语言在特定环境中的实际含义，强调上下文和语境对语义解释的影响。

随着深度学习技术的发展，基于深度学习的语义理解算法逐渐成为研究主流。这类算法能够有效捕捉文本的上下文信息，显著提升了语义表示的准确性和完整性。基于深度学习的语义理解算法主要分为三类：无监督、有监督和自监督模型。无监督模型通过大规模语料库训练词向量，并通过词向量的线性组合生成文本的

汽标委智能网联汽车分委会

语义表示，典型算法包括 word2vec、GloVe 和 SIF 等。有监督模型则利用标注数据进行端到端学习，通过深度神经网络直接生成语义表示向量。为解决标注数据获取的困难，自监督模型被提出，这类模型通过自监督损失函数摆脱对标注数据的依赖，成为当前研究的热点方向。

7.3.2.2 对话引擎

对话引擎作为语音交互系统的核心组件，是实现接近甚至等同于人类交互体验的关键模块。在不同系统中，对话引擎（Conversation Engine）也可被称为“对话管理（Dialog Management，DM）”。

传统的对话引擎系统主要基于逻辑规则实现，通常采用有限状态机（Finite State Machine，FSM）或框架（Frame）原理，通过人工编撰对话系统的状态转移和决策逻辑。其优势在于规则明确、灵活可控且定制性强，适用于对话流程清晰的场景；但面对离题、歧义等复杂问题时，需依赖烦琐的规则配置。

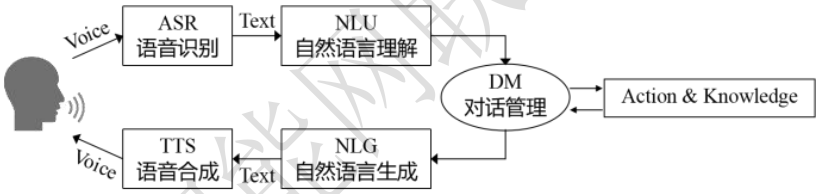


图 7-26 基于规则的对话引擎

典型的机器学习对话引擎则采用马尔可夫决策（Partially Observable Markov Decision Process，POMDP），将对话过程建模为状态－动作决策问题。系统通过对话理解的结果作为输入，在每个状态中执行动作并获得回报，最终选择期望回报最大的动作策略。该方法无需显式定义状态间的转移关系，可通过强化学习自动获取最优策略，但需要大规模标注对话训练数据作为支撑。

7.3.2.3 场景引擎

随着智能座舱功能的不断延展，功能多样化与用户需求满足度之间的矛盾日益突出，用户体验成为企业的核心竞争力。场景引擎结合人工智能引擎，深度融合车辆、用户、环境、生态和交通数据，将用户日常分解为多个智能座舱场景，

汽标委智能网联汽车分委会

以场景为中心，借助数据和算法进行主动服务，为用户提供高频、刚需、原子化、触手可及的服务，从而提升用户智能体验，驱动智行生活。

7.3.2.4 知识问答

问答系统是一种高级信息检索形式，旨在通过自然语言处理技术，从异构数据源中精准提取用户提问的准确答案。系统由五个核心组件构成：数据源分析、问题分析、知识检索、答案生成与答案评估，如图 7-27 所示。

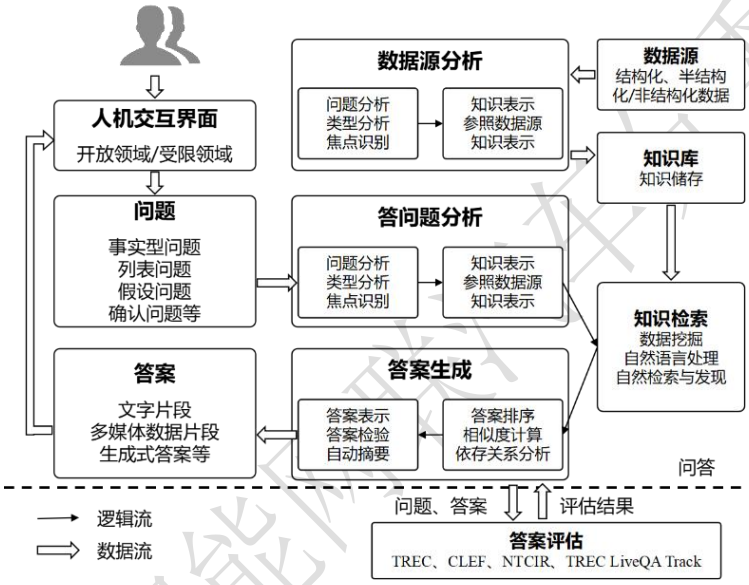


图 7-27 问答系统架构图

数据源分析模块运用知识挖掘技术处理多源数据，生成知识表示；问题分析模块解析用户查询意图，执行语义处理及查询扩展，生成查询语句；知识检索模块基于生成的查询语句搜索知识库，提取相关信息；答案生成模块利用信息抽取技术，从检索结果中选择最优答案；答案评估模块则对答案质量进行评估，其反馈用于优化系统各模块，提升系统整体性能。

7.3.2.5 大语言模型

大模型以大量参数和计算资源为特征，需大量数据以确保泛化能力，依赖高性能 GPU 或 TPU 集群增强计算能力，参数规模从数百万至数十亿不等，以学习更复杂的数据特征，提高表示能力和任务处理性能。当前，大模型多采用 NLP 领域中流行的 Transformer 架构，如图 7-28 所示。相较于传统的 RNN 和 LSTM，

Transformer 的注意力机制不仅提高了模型的理解力，还增强了并行性和扩展性，使其能处理更长的序列，在文本处理任务中表现出色[158]。

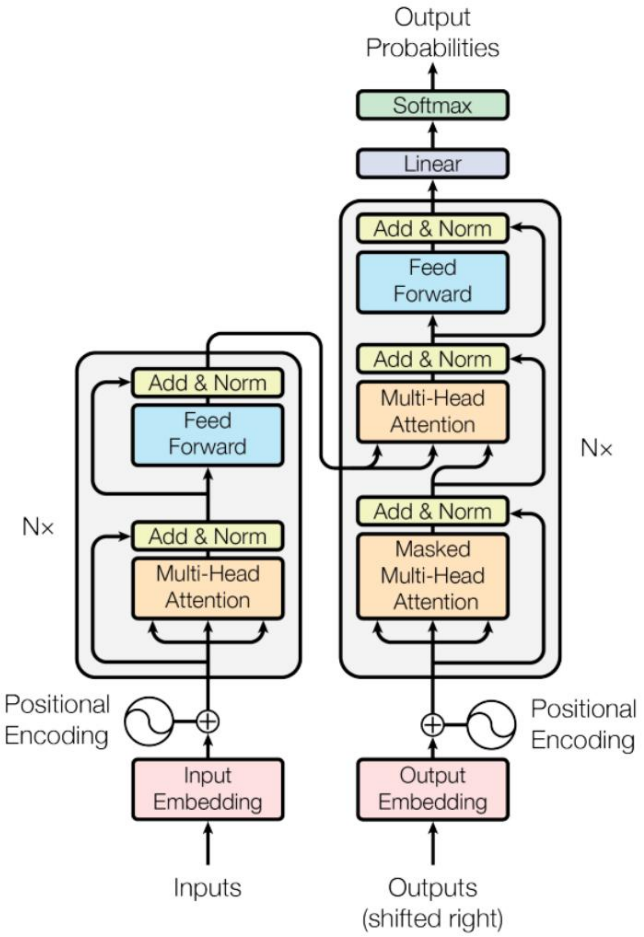


图 7-28　Transformer 结构

大语言模型作为大模型的特定子集，专注于理解和生成自然语言。通过海量文本数据训练，大语言模型能够学习语言的复杂结构和模式，实现文本生成、语义理解、机器翻译、问答系统等多种功能，特别适用于智能座舱等场景中的自然语言处理任务。

7.3.3 智能座舱表达技术

7.3.3.1 语音合成

语音合成（Speech Synthesis），即文本转语音（Text-to-Speech, TTS）技术，是人机交互领域的重要研究方向，其核心目标是生成清晰且自然的语音信号。该

技术通过解放视觉注意力，使用户能够在接收视觉信息的同时获取听觉信息，从而提升信息处理效率。语音合成可分为三个层次[159]：文本到语音（Text-To-Speech, TTS）、概念到语音（Concept-To-Speech）以及意向到语音（Intention-To-Speech）。

传统语音合成模型基于“源滤波器”框架，基于统计参数的语音合成模型（Statistical Parametric Speech Synthesis, SPSS），相较于传统方法更具灵活性，如图 7-29 所示。SPSS 通过机器学习训练模型，使其能够将文本输入映射为语音输出，而无需依赖预存语音片段。该模型利用语音数据集进行训练，输出结果为描述语音特征的统计参数，从而实现语音的高效合成。而端到端神经语音合成模型则由声学模型和声码器组成。声学模型将输入文本转换为频谱，声码器再将频谱转换为音频信号。

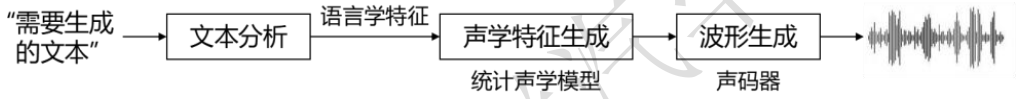


图 7-29 基于统计参数的语音合成模型

7.3.3.2 声音复刻

声音复刻（Voice Cloning）作为语音合成技术（TTS, Text-to-Speech）的个性化应用，通过大数据训练使 AI 模型能够精准学习并模仿人类语音特征。该技术弥补了传统语音合成在人声数字化方面的局限性，显著提升了语音生成的真实性和丰富度。随着计算能力的增强和大规模语音数据的积累，声音复刻能够更精细地还原音调、音色和音量的变化，使生成的语音更具自然质感，为声音塑造带来了无尽的可能性。

7.3.3.3 基础音效

基础音效技术是以声学理论为基础，结合电子与数字信号处理手段，实现声音的采集、处理与再现的技术体系。其技术原理围绕声波的物理特性（如频率、振幅、相位）及人耳感知机制（20 Hz－20 kHz 频率响应、空间定位能力）展开，核心处理模块包括混响、均衡、动态范围控制以及调制效果。

汽标委智能网联汽车分委会

7.3.3.4 独立音区

独立音区技术旨在对音源进行独立控制和定位，为汽车座舱或特定空间用户提供沉浸式和定向化的音频体验，使声源能够在三维空间中被精确放置和移动，同时确保不同区域的音频内容互不干扰。

汽车智能座舱配备的多声道音频系统为独立音区技术提供了硬件基础。通过独立控制每个扬声器或扬声器组，向特定区域播放不同的音频。基于波束形成（Beamforming）原理，结合数字信号处理（DSP）技术，将声音精确“指向”座舱中的特定区域或座位，将特定的音频流只播放给某一座位上的用户，而不会显著影响其他区域的听感。独立音区不止包含音频的分区发出，还涉及音频的独立采集。多音频源定位技术通过声音传播的物理特性和声学原理，并利用信号处理算法，从混合音频信号中分离出不同的声源信号，对这些信号进行处理和分析，利用阵列信号处理技术，以推导出音频源的具体位置，让麦克风阵列锁住某个方向的声音，精确识别语音指令发出者并为其提供个性化服务。

7.3.3.5 主动降噪

内外部嘈杂环境噪声干扰不但影响司乘人员的音频体验，引发烦恼和疲劳，还有可能影响到安全驾驶。人工智能主动降噪（AI-ANC）技术，深入分析了各种噪声 NVH 性能（包括噪声（Noise）、振动（Vibration）和声振粗糙度（Harshness）等），借助麦克风实时采集环境噪声，利用深度学习模型快速解析噪声频谱特征，基于传统的“反向声波抵消”原理，预判噪声传播路径与时间延迟，确保反向声波精准同步叠加。深度学习模型的动态优化噪声抑制策略，突破了传统自适应滤波器的线性假设限制，实现了复杂场景下的动态降噪优化。

7.3.3.6 车内交流补偿

音频交流补偿技术旨在通过技术手段改善或校正音频通信或播放过程中因环境、设备特性或传输过程导致的音频质量下降，从而提升听觉体验。该技术广泛应用于语音合成、实时通信、音频播放系统优化等多个领域。

汽标委智能网联汽车分委会

在端到端语音合成中，跨说话人情感迁移的韵律补偿技术通过引入特定的韵律补偿模块（Prosody Compensation Module, PCM），结合预训练的自动语音识别（ASR）模型提取的韵律信息，增强情感表达并保持目标说话人的音色稳定，从而在跨说话人迁移时提升合成语音的情感表现力；在智能音频管理系统中，声源补偿技术通过精确的声源定位和播放补偿，优化声音的定向传播。在特定频带的扬声器频响中，通过多速率采样和滤波器优化设计，改善扬声器在特定频率的响应特性，从而提升音频播放的均衡性和质量；此外，音频恢复技术可以通过降噪算法等手段，识别并去除背景噪声，恢复被噪声干扰的音频信号，确保在复杂环境中提供清晰的音频体验。

7.4 智能座舱 AI 技术应用测试与评价的流程和要求

7.4.1 场景交互评测的流程和要求

7.4.1.1 语音交互场景

当前智能座舱的语音交互测评规范主要围绕车载语音交互测评，涵盖语音唤醒、识别、交互等方面，涉及语音、触控及无线交互等高频交互方式。我国已有 7 个相关评价标准，如《汽车智能座舱语音分级与测评方法（起草）》《智能语音交互质量评价规范》等。这些规范覆盖了识别准确性、交互流畅性、系统响应速度等关键测试领域，确保了用户体验的高标准。然而，随着大语言模型的引入，语音交互系统的应用范围和深度不断扩展，现有测评规范未能充分考虑大语言模型的新特性和挑战。因此，构建面向大语言模型赋能的语音技术在座舱中的用户体验评价规范变得尤为关键。

（1） 生成能力

大模型应用于语音交互中，生成能力方面的标准建立应在用户体验层面和性能层面。用户体验层面包含《语音自然度评测标准》《措辞多样性评测标准》和《情感表达能力评测标准》，性能层面的标准为《生成流畅性评测标准》。

1) 用户体验：《语音自然度评测标准》旨在建立科学系统的评估体系，客

汽标委智能网联汽车分委会

观准确衡量大模型生成语音的真实感和自然程度。通过分析语音的韵律、音质、发音准确性、情感传递等，判断其与人类自然语音的接近程度；《措辞多样性评测标准》旨在评估文本或言论中措辞的多样性。该标准不仅关注内容的准确性和完整性，更注重表达形式的多样性，如词汇选择、句式结构、修辞手法等，以提升信息传达的效果和影响力；《情感表达能力评测标准》旨在评估个体或文本在情感表达方面的能力。该标准注重评估情感传递、识别和调控等方面的表现，确保其能准确、生动地传达情感信息，增强人机情感共鸣和连接。

2) 性能：《生成流畅性评测标准》从性能的角度系统性地评估生成系统或算法在输出文本流畅性方面的表现。该评测标准考量生成系统或算法在处理大规模数据时的效率、速度和准确性，以确保生成的文本能够在实时性和性能上达到要求。

（2） 学习能力

大模型应用于语音交互中，学习能力方面的标准建立应在用户体验层面和性能层面。用户体验层面包含《记忆更新能力评测标准》和《学习准确性评测标准》，性能层面的标准为《学习效率评测标准》。

1) 用户体验：《记忆更新能力评测标准》此标准专注于智能座舱系统对用户偏好的长期记忆、行为模式学习及新信息的快速更新同步，强调用户隐私保护与技术整合性，考察系统在不同使用场景下的可靠性、稳定性及用户反馈机制效率，确保用户能轻松纠正系统记忆错误；《学习准确性评测标准》此标准着重评估智能座舱系统对用户行为数据的精确收集、有效分析与模式识别能力，考察其在个性化服务提供上的准确性、对用户行为变化的快速响应与持续优化能力。

2) 性能：《学习效率评测标准》评估大模型学习速度和数据处理能力，确保系统能快速提取语音指令信息并更新知识库。还评估系统在连续交互中的性能优化能力，推动语音交互进步，实现对用户指令的即时反应和习惯预测，提供流畅、个性化交互体验。

（3） 理解能力

汽标委智能网联汽车分委会

大模型应用于语音交互中，理解能力方面的标准建立应在用户体验层面和性能层面。用户体验层面包含《情感理解评测标准》《用户意图理解评测标准》和《多语言理解评测标准》，性能层面的标准为《理解效率评测标准》。

1) 用户体验：《情感理解评测标准》衡量大模型在语音交互中识别和响应用户情感的能力，通过用户的语音语调、语速、音量以及语言表达中的情感色彩来识别情绪状态，评估系统对情感信号的敏感度、准确性及恰当反应的能力；《用户意图理解评测标准》衡量大模型在语音交互中理解和满足用户意图的能力，评估系统对用户输入意图的准确性，确保系统在用户表达不清晰或意图转换时能正确理解需求；《多语言理解评测标准》评估大模型在多语言环境下理解和处理用户意图的能力，包括对不同语言用户意图的准确识别，理解语言特有的表达方式和文化背景，测试系统在多语言环境中进行语言转换和适应的能力。

2) 性能：《理解效率评测标准》模型在实际应用中进行理解时的速度和资源消耗。

综上所述，理解和应用效率评测标准可以帮助我们全面衡量和优化大模型在实际应用中的性能表现和成本效益。

（4） 推理能力

大模型应用于语音交互中，推理能力方面的标准建立应在用户体验层面和性能层面。用户体验层面包含《泛化能力评测标准》《逻辑推理能力评测标准》《推理一致性评测标准》和《推理可解释性评测标准》，性能层面的标准为《推理效率评测标准》。

1) 用户体验：《泛化能力评测标准》大模型泛化能力至关重要，涉及其在不同数据集和任务上的稳定性能。具有良好泛化能力的模型应学习到普适性特征，而非过度拟合训练数据；《逻辑推理能力评测标准》逻辑推理是模型在处理信息、推断结论或解决问题时展现的逻辑思维能力。评测模型的逻辑推理能力，可采用设计逻辑推理任务和选择评估指标等方法。此外，还可采用开放式问题或复杂推理场景来评估模型处理真实世界复杂逻辑问题的能力；《推理一致性评测标准》

标委智能网联汽车分委会

推理一致性指模型在处理逻辑推理或问题求解时的逻辑连贯性和稳定性。评估推理一致性可采用逻辑规则一致性检查、逻辑矛盾检测等方法，同时考虑推理过程中的信息保持和传递性；《推理可解释性评测标准》关注模型推理和决策过程的解释性。评估方法包括解释性文本生成、可视化展示和推理结果一致性检查，旨在确保模型的推理过程和结果能够被人理解和信任，提高模型的可接受性，可靠性和可解释性。

2) 性能：《推理效率评测标准》关注模型推理过程中的速度和资源利用。评估方法包括分析推理时间、资源消耗和计算复杂度等指标，以定量评估模型在不同场景下的效率。

7.4.1.2 视觉交互场景

当前智能座舱视觉交互涉及人机交互设计、眼动追踪、虚拟现实技术及界面评价等，中国汽车工业协会制定了相关团体标准，但缺乏通用且权威的推荐设计标准，导致视觉交互设计因车企而异。

当前智能座舱视觉交互设计标准多为设计规范层面，缺乏通用权威的推荐标准，导致车企间视觉交互设计差异显著。随着大模型应用场景丰富，视觉交互作为核心交互模态之一，重要性凸显。为确保设计符合人体工程学，未来应结合先进生理监测技术开展人机实验与模拟，融合多学科专业进行全面优化迭代，以满足用户需求。

（1） 生成能力

大模型应用于视觉交互中，生成能力方面的标准建立应在用户体验层面和性能层面。用户体验层面包含《生成准确性评测标准》《个性化适应评测标准》和《多样性和创造性评测标准》，性能层面的标准为《生成流畅性评测标准》。

1) 用户体验：《生成准确性评测标准》评估大模型在座舱视觉交互中的生成准确性对确保驾乘安全至关重要。信息匹配度、数据处理精度和预测能力是关键指标，这些共同确保信息准确，提升系统实用性和用户体验；《个性化适应评测标准》：评价大模型在座舱视觉交互中的个性化适应能力，需考量身份识别准

汽标委智能网联汽车分委会

确性、个性化设置效果和用户反馈信息；《多样性和创造性评测标准》：在座舱视觉交互中，评估大模型的多样性和创造性是提升用户体验和吸引力的关键。

2) 性能：《生成流畅性评测标准》同语音交互的相关标准。

（2） 学习能力

大模型应用于视觉交互中，学习能力方面的标准建立应在用户体验层面和性能层面。用户体验层面包含《记忆更新能力评测标准》和《学习准确性评测标准》，性能层面的标准为《学习效率评测标准》。这些标准与语音交互的相关标准相同。

（3） 理解能力

大模型应用于视觉中，理解能力方面的标准建立应在用户体验层面和性能层面。用户体验层面包含《视觉交互－情感理解评测标准》《视觉交互-用户意图理解评测标准》和《个性化识别评测标准》，性能层面的标准为《理解效率评测标准》。

1) 用户体验：《视觉交互－情感理解评测标准》聚焦大模型对驾乘人员情感状态的识别与反馈能力。核心评估点包括情感识别的准确性，如对基础情绪的捕捉能力；反馈机制的有效性，包括通过环境调节或主动交互满足情感需求；《视觉交互－用户意图理解评测标准》重点考察多模态意图解析能力。评估框架分为多源信号处理（语音/手势/表情的跨模态融合）、显性需求响应（指令执行准确率与时效性）、隐性需求预判（基于用户画像和场景上下文的行为预测）。三维度协同可提升交互系统的主动服务能力。

2) 性能：《理解效率评测标准》同语音交互的相关标准。

（4） 推理能力

大模型应用于视觉交互中，推理能力方面的标准建立应在用户体验层面和性能层面。用户体验层面包含《泛化能力评测标准》《逻辑推理能力评测标准》《推理一致性评测标准》和《推理可解释性评测标准》，性能层面的标准为《推理效率评测标准》。这些标准与语音交互的相关标准相同。

汽标委智能网联标准分委会

7.4.1.3 多模交互场景

多模态交互技术依托人工智能、情感计算等领域发展，面临技术瓶颈突破、系统精准度优化及用户体验升级等挑战。

（1） 生成能力

大模型应用于多模态交互中，生成能力方面的标准建立应在用户体验层面和性能层面。用户体验层面包含《多模态响应生成评测标准》，性能层面的标准为《响应速度与准确性评测标准》。

1) 用户体验：《多模态响应生成评测标准》要求系统能够根据用户的语音、视觉和触觉指令，生成准确且及时地响应。系统应提供自然、直观的交互方式，减少用户的学习成本，并能够根据用户的个性化需求定制相应内容。

2) 性能：《响应速度与准确性评测标准》则侧重于系统处理多模态输入的速度和执行任务的准确度，确保系统能够在极短时间内理解并执行用户的指令。

（2） 上下文连贯能力

大模型应用于多模态交互中，上下文连贯能力方面的标准建立应在用户体验层面和性能层面。用户体验层面包含《上下文感知与应用评测标准》，性能层面的标准为《响应速度与准确性评测标准》。

1) 用户体验：《上下文感知与应用评测标准》强调系统在多模态交互中保持上下文连贯性的能力，以提供一致且流畅的用户体验。系统应能够记忆用户的操作历史，并根据当前上下文环境，提供恰当的交互响应。

2) 性能：《上下文切换平滑度标准》则评价系统在不同上下文间切换时的稳定性和平滑度，确保用户在不同场景下都能获得无间断的交互体验。

（3） 学习能力

1) 用户体验：《自适应学习与优化标准》衡量系统根据用户行为和反馈进行自我学习和调整的能力，以提升用户的满意度和忠诚度。系统应能够识别用户的偏好，并主动适应用户的变化。

汽标委智能网联汽车分委会

2) 性能：《学习效率与效果标准》则关注系统学习新行为模式的速度和学习后性能提升的幅度，确保系统能够快速吸收新知识并有效应用。

（4） 理解能力

1) 用户体验：《多模态意图理解标准》要求系统能够准确识别和理解用户的多模态指令背后的意图，即使在用户指令不明确或含糊时也能提供有效的帮助。

2) 性能：《指令解析精准度标准》则侧重于系统对用户指令的解析精准度，包括对复杂指令的拆分和执行，确保系统能够正确理解并执行用户的每一个指令。

（5） 推理能力

1) 用户体验：《决策推理与异常处理标准》评估系统在面对复杂情境时的决策推理能力，以及在遇到异常情况时的应对策略，确保用户在面对不确定性时仍能获得可靠的支持。

2) 性能：《逻辑推理准确性标准》则侧重于系统逻辑推理的准确性和效率，确保系统能够基于正确的推理提供恰当的决策。

7.4.1.4 开放式任务场景

智能座舱开放式任务的标准与现状正不断完善发展。当前国际上暂无统一标准，但国内已发布多项标准与研究报告，有力推动了智能座舱技术应用。不少企业积极开发开放平台以顺应市场与技术趋势。2024 年 1 月，《智能座舱标准体系研究报告》发布，为产业发展提供系统化支撑。不过，开放式任务作为大模型赋能下的人机交互场景，相关标准需尽早构建，以完善产业规划和打造全流程用户体验服务。

（1） 生成能力

大模型应用于开放式任务交互中，生成能力方面的标准建立应在用户体验层面和性能层面。用户体验层面包含《任务整合评测标准》《个性化内容评测标准》，性能层面的标准为《复杂指令响应能力评测标准》。

1) 用户体验：《任务整合评测标准》旨在评估系统对用户需求的深度理解

能力，包括准确识别指令意图及跨领域任务关联性，强调通过上下文和交互历史主动预判未明示需求（如纪念日提醒、乘客偏好适配）。其核心还涵盖对解决方案的创造性、实用性和个性化匹配度的评估，要求系统在技术可行基础上灵活响应用户需求变化，动态调整执行策略以适应新场景；《个性化内容评测标准》主旨在于评估系统如何深入理解用户的个性化需求和偏好，并将这些理解融入到任务生成和执行中。强调系统在处理多指令任务时，要准确执行每个指令，还要能够根据上下文信息，创造性地整合这些指令。

2) 性能：《复杂指令响应能力评测标准》聚焦系统对复杂指令的解析精度，涵盖准确捕捉条件语句、时间表达、序列动作和用户偏好。它突出系统对用户潜在需求的预测与主动服务能力，还考量系统整合多源信息、处理并发任务及保持执行连贯性的能力。同时，考察系统在指令不明确时的澄清策略，以及执行中对用户反馈的响应速度与调整能力。以此标准，智能座舱大模型技术能更精准响应复杂指令，提供更流畅高效服务。

（2） 上下文连贯能力

大模型应用于开放式任务交互中，上下文连贯能力方面的标准建立应在用户体验层面和性能层面。用户体验层面包含《环境信息识别评测标准》和《情境适应性评测标准》，性能层面的标准为《上下文连贯流畅性评测标准》。

1) 用户体验：《环境信息识别评测标准》旨在评估大模型在开放式任务中对环境信息的识别和利用能力。通过衡量系统如何准确捕捉并解析外部环境信息，包括但不限于时间、地点、气候、交通状况以及用户的活动模式。它强调系统在整合这些信息以响应复杂用户指令时的效率和准确性；《情境适应性评测标准》旨在评估系统对用户所处的具体情境的敏感度，包括用户的位置、时间、活动状态以及可能的情感需求。当系统在面对用户复杂指令时，如何灵活调整其行为以适应当前的情境。

2) 性能：《上下文连贯流畅性评测标准》同语音部分的标准。

（3） 理解能力

汽标委智能网联标准分委会

大模型应用于开放式任务交互中，上下文连贯能力方面的标准建立应在用户体验层面和性能层面。用户体验层面包含《用户意图理解评测标准》和《多指令理解评测标准》，性能层面的标准为《理解效率评测标准》。

1) 用户体验：《用户意图理解评测标准》同语音部分的标准；《学习准确性评测标准》同语音部分的标准。

2) 性能：《理解效率评测标准》同语音部分的标准。

（4） 推理能力

大模型应用于开放式任务交互中，上下文连贯能力方面的标准建立应在用户体验层面和性能层面。用户体验层面包含《因果关系推理评测标准》和《决策制定逻辑评测标准》《异常情况处理评测标准》，性能层面的标准为《学习效率评测标准》。

1) 用户体验：《因果关系推理评测标准》旨在评估大模型识别用户指令因果逻辑的能力。其核心要求包括通过因果链分析预测指令可能引发的连锁反应（如调节温度需兼顾舒适度与健康影响）；主动挖掘未明示需求，基于行为逻辑推理提供前瞻性解决方案；突破单次指令局限。该标准强调从因果关联维度提升服务精准度和场景覆盖度；《决策制定逻辑评测标准》旨在系统决策的透明度和合理性，确保用户能够理解系统做出的决策背后的原因。例如，当用户提出一系列可能相互冲突的指令时，系统需要能够根据优先级和潜在影响来做出明智的决策；《异常情况处理评测标准》用于评估系统对异常情况的识别速度与准确度，涵盖对用户指令错误、环境致服务中断或系统内部故障的响应能力。着重考察系统面对异常时的鲁棒性，以及提供替代方案或恢复计划的能力。鉴于座舱空间的有限与特殊，标准还需考量系统处理用户紧急请求时的优先级排序和快速响应机制，保障关键任务及时处理。

2) 性能：《推理效率评测标准》；同语音部分的标准。

汽标委智能网联汽车分委会

7.4.2 内容安全评测的流程和要求

国内外对生成式人工智能的关注日益增加，各大企业陆续在各国和地区正在积极探索制定相关的标准和法规，以应对这一新兴技术的快速发展和广泛应用。

欧盟《人工智能法案（AI Act）》是欧盟针对人工智能技术发展和应用提出的全面法规框架。根据 AI 系统对用户构成的风险，将 AI 系统进行分类，并根据风险等级决定监管的严格程度。美国则通过一系列行政命令和政策文件，强调了在促进人工智能技术发展的同时，需要考虑其对社会的影响，并在必要时制定相应的治理方法[160]。

在中国，国家网信办联合国家发展改革委、教育部、科技部、工业和信息化部、公安部、广电总局公布了《生成式人工智能服务管理暂行办法》。《办法》提出了国家坚持发展和安全并重、促进创新和依法治理相结合的原则，并采取有效措施鼓励生成式人工智能的创新发展。此外，还明确了技术发展与治理的相关鼓励措施，以及对服务提供者和使用者的基本要求。

在技术标准方面，全国网络安全标准化技术委员会（TC260）也正在积极推进《网络安全技术 生成式人工智能服务安全基本要求》的国家标准计划。

7.4.2.1 现有内容安全场景评测方案分析

针对生成内容安全的评测，在 CCSA TC1 WG1 中已有生成式人工智能技术及产品系列标准。技术能力标准中规定了生成式 AI 技术及基于技术形成的相关工具、平台的评估指标，从主观和客观两个维度进行评测。能力标准中规定了具备生成式 AI 技术能力的互联网系统、网页、平台等的评价指标，从基础能力、服务能力和可控能力等三个维度进行评测。有助于降低生成结果不可控、版权争议、恶意滥用等风险，促进生成式人工智能技术的健康发展。

7.4.2.2 内容生成安全评测

随着生成式人工智能（AIGC）和大模型技术的普及，未来的智能座舱将可能集成更高级的内容生成应用。例如，智能座舱能够根据驾驶员和乘客的偏好、行

为模式生成定制化的内容推荐，或根据车内环境和驾驶状况提供即时提醒与警示信息。这些新功能虽然提升了用户体验，但也带来了新的安全风险，需要进行针对性的安全评测。

在目前出台的指导文件中，网信办《生成式人工智能服务管理办法》和信安标委的《生成式人工智能服务安全基本要求》[161][162]对生成式人工智能具体的安全风险进行了定义和分类划分。这些指导原则有助于制定具体的安全评测标准和方法，以确保评测内容全面覆盖各类潜在风险，提升评测的针对性和有效性。

7.4.2.3 内容拒答安全评测

“内容拒答”指大型语言模型面对无法或不宜回答的问题时，能正确判断并拒绝作答或给出无意义、不相关响应，对保障模型安全可靠意义重大。其实现依赖模型对上下文的深度理解分析，需构建拒绝回答机制。评估时，设计一系列无法或不宜回答的问题做测试数据集，观察模型响应，以此判断其面对不确定或信息不足情况时能否正确拒答，确保实际应用中的可靠与准确[163]。

评测方面，中国科学院软件研究所中文信息处理实验室提出重点分析大型语言模型在 RAG（检索增强生成）所需的噪声鲁棒性、拒答、信息整合和反事实鲁棒性四项能力，要求模型对无法回答的问题进行拒答，并将拒答比例作为评价指标。清华大学的 SuperBench 测评平台也将拒答行为纳入安全测评框架中，为了精准评估模型在拒答情况下的性能，基准特别设置了拒答分数和非拒答分数：拒答分数将拒答题目视为回答错误，反映模型在面对不确定性时的保守态度或知识盲区；非拒答分数则排除拒答题目，更纯粹地评估模型在已作答题目上的表现。

7.5 智能座舱 AI 技术标准化需求与关键挑战

7.5.1 标准化需求

智能座舱 AI 技术的快速发展对标准化提出了迫切需求，主要涵盖技术规范、数据管理、安全隐私、测试评价等多个维度。

汽标委智能网联汽车分委会

技术规范标准化：需统一多模态交互接口协议，统一语音、视觉、触觉等交互接口协议，确保不同模态数据的协同处理与响应一致性。例如，制定语音+手势、语音+唇动等多模态融合的交互标准；制定大模型轻量化部署标准，制定云端与端侧模型部署的硬件适配、压缩优化及能耗管理标准，支持本地化轻量级大模型的高效运行；并规范异构计算平台协同，明确 CPU、GPU、FPGA 等异构硬件的任务分配策略与资源调度标准，提升算力利用效率。

数据管理标准化：需建立高质量、多场景训练数据的采集规范，统一标注流程与准确性验证标准，解决方言、噪声等数据多样性问题；需制定数据匿名化处理、用户授权管理及跨境数据传输的合规标准，确保符合《中华人民共和国个人信息保护法》等法规要求；需定义多模态数据（语音、图像、传感器数据）的格式与融合规则，支持跨领域信息整合。

安全与隐私保护标准化：需规范数据传输加密算法、存储加密机制及访问权限控制，防止网络攻击与数据泄露；需建立生成式 AI（如语音合成、虚拟人）的伦理审查标准，限制恶意内容生成与滥用；需明确生物特征数据（人脸、声纹）的采集边界与销毁机制，避免过度收集与非法使用。

测试与评价标准化：需制定语音识别准确率、视觉交互响应延迟、多模态融合鲁棒性等量化指标；可建立自然度（语音合成）、个性化（推荐系统）、情感理解（情绪识别）等主观体验的评测体系。

7.5.2 标准化关键挑战

智能座舱 AI 技术的标准化推进在技术实现上面临着技术复杂性、数据治理难题、安全风险及跨领域协调等多重挑战。

技术复杂性：大模型在车载端侧部署需兼顾低延迟与低功耗，现有硬件算力与散热设计难以满足复杂场景需求；不同感知模态（如视觉与语音）的数据融合存在时序对齐与语义一致性挑战，需突破跨模态特征提取技术；CPU、GPU、AI 加速器的任务分配策略动态调整困难，难以实现资源利用率最大化。

汽标委智能网联汽车分委会

数据管理挑战：覆盖极端天气、复杂路况等长尾场景的训练数据稀缺，标注成本高且易引入人工误差；数据匿名化可能降低模型性能，如何在隐私保护与算法精度间取得平衡成为难题；车企、互联网平台、政府间的数据孤岛现象严重，缺乏统一的数据共享机制与信任框架。

安全与隐私挑战：恶意攻击者可能通过噪声注入欺骗视觉或语音识别系统，需制定抗干扰认证标准；生成式 AI 可能输出误导性信息（如错误导航指令），需强化内容过滤与实时监控机制；人脸、声纹等数据一旦泄露可能被用于身份伪造，亟须生物识别信息的生命周期管理标准。

标准化协调挑战：欧盟、美国与中国在 AI 伦理、数据主权等方面存在政策分歧，企业需应对多重合规压力；AI 技术更新速度远超标准制定周期，现有标准难以覆盖新兴应用；汽车、通信、互联网等领域的标准体系碎片化，缺乏统一的顶层设计框架。

7.6 总结与展望

智能座舱 AI 技术正面临多方面的挑战与机遇。语音交互需突破复杂环境下的噪声抑制、多语言识别及跨文化障碍，而 AI 视觉交互则依托域控制器和高性能处理器提升数据处理效率，可结合大模型增强视觉理解能力，推动个性化服务发展。而未来视觉系统将深化情绪与行为识别，整合触摸、语音、视线等多模态交互，并借助 AR 技术提供实时导航与信息增强。技术架构的演进正逐步替代传统 ECU，通过集中式处理优化响应速度与整车性能，为精准化、沉浸式体验奠定基础。

多模态交互融合了视、听、触、嗅多感官输入输出，极大地提高了驾乘人员的沉浸式体验，是座舱智能化的未来演进路线。作为智能座舱的核心方向，需通过数据融合、上下文感知及优化算法实现动态适配，缓解信息过载并提升安全性。逐步实现预测性服务，基于大数据分析用户习惯，主动规划路线或推荐内容，并实时性能优化确保交互流畅性。未来需持续突破异构计算、隐私保护等关键技术，推动标准化与跨领域协作，最终实现安全可靠、自然智能的交互革命。

汽标委智能网联汽车分委会

参考文献

[1] Jiang P, Ergu D, Liu F, Cai Y, Ma B. A Review of Yolo algorithm developments[J]. Procedia computer science. 2022 Jan 1;199:1066-73.

[2] Duan K, Bai S, Xie L, Qi H, Huang Q, Tian Q. Centernet: Keypoint triplets for object detection[C]. In Proceedings of the IEEE/CVF international conference on computer vision 2019 (pp. 6569-6578).

[3] Wang T, Zhu X, Pang J, Lin D. Fcos3d: Fully convolutional one-stage monocular 3d object detection[C]. In Proceedings of the IEEE/CVF International Conference on Computer Vision 2021 (pp. 913-922).

[4] Liu L, Chen X, Zhu S, Tan P. Condlanenet: a top-to-down lane detection framework based on conditional convolution. InProceedings of the IEEE/CVF international conference on computer vision 2021 (pp. 3773-3782).

[5] Wang J, Ma Y, Huang S, Hui T, Wang F, Qian C, Zhang T. A keypoint-based global association network for lane detection[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2022 (pp. 1392-1401).

[6] Bai Y, Chen Z, Fu Z, Peng L, Liang P, Cheng E. Curveformer: 3d lane detection by curve propagation with curve queries and attention[C]. 2023 IEEE International Conference on Robotics and Automation (ICRA) 2023 May 29 (pp. 7062-7068). IEEE.

[7] Brust CA, Sickert S, Simon M, Rodner E, Denzler J. Convolutional patch networks with spatial prior for road detection and urban scene understanding. arXiv preprint arXiv:1502.06344. 2015 Feb 23.

[8] Bayón-Gutiérrez M, García-Ordás MT, Alaiz Moretón H, Aveleira-Mata J, Rubio-Martín S, Benítez-Andrades JA. TEDNet: Twin Encoder Decoder Neural Network for 2D Camera and LiDAR Road Detection[J]. Logic Journal of the IGPL. 2024 May 10:jzae048.

[9] Caltagirone L, Scheidegger S, Svensson L, Wahde M. Fast LIDAR-based road detection using fully convolutional neural networks[C]. 2017 ieee intelligent vehicles symposium (iv) 2017 Jun 11 (pp. 1019-1024).

汽标委智能网联汽车分委会

[10] Wu X, Jiang L, Wang PS, Liu Z, Liu X, Qiao Y, Ouyang W, He T, Zhao H. Point Transformer V3: Simpler Faster Stronger[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2024 (pp. 4840-4851).

[11] Fan R, Wang H, Cai P, Liu M. Sne-roadseg: Incorporating surface normal information into semantic segmentation for accurate freespace detection[C]. European Conference on Computer Vision 2020 Aug 23 (pp. 340-356).

[12] Li Z, Wang W, Li H, Xie E, Sima C, Lu T, Qiao Y, Dai J. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers[C]. European conference on computer vision 2022 Oct 23 (pp. 1-18).

[13] Liu Z, Tang H, Amini A, Yang X, Mao H, Rus DL, Han S. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation[C]. 2023 IEEE international conference on robotics and automation (ICRA) 2023 May 29 (pp. 2774-2781).

[14] Wei Y, Zhao L, Zheng W, Zhu Z, Zhou J, Lu J. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision 2023 (pp. 21729-21740).

[15] Wang S, Clark R, Wen H, et al. Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks[C]//2017 IEEE international conference on robotics and automation (ICRA). IEEE, 2017: 2043-2050.

[16] Cho Y, Kim G, Kim A. Unsupervised geometry-aware deep lidar odometry[C]//2020 IEEE international conference on robotics and automation (ICRA). IEEE, 2020: 2145-2152.

[17] Sindagi VA, Zhou Y, Tuzel O. Mvx-net: Multimodal voxelnet for 3d object detection[C]. 2019 International Conference on Robotics and Automation (ICRA) 2019 May 20 (pp. 7276-7282).

[18] Vora S, Lang AH, Helou B, Beijbom O. Pointpainting: Sequential fusion for 3d object detection[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 2020 (pp. 4604-4612).

[19] Yoo JH, Kim Y, Kim J, Choi JW. 3d-cvf: Generating joint camera and lidar features using cross-view spatial feature fusion for 3d object detection[C]. Computer vision – ECCV 2020:

浙江省智能网联汽车分委会

16th European conference, Glasgow, UK, August 23 – 28, 2020, proceedings, part XXVII 16 2020 (pp. 720-736).

[20] Liang M, Yang B, Wang S, Urtasun R. Deep continuous fusion for multi-sensor 3d object detection[C]. Proceedings of the European conference on computer vision (ECCV) 2018 (pp. 641-656).

[21] Zhao X, Liu Z, Hu R, Huang K. 3D object detection using scale invariant and feature reweighting networks[C]. Proceedings of the AAAI Conference on Artificial Intelligence 2019 Jul 17 (Vol. 33, No. 01, pp. 9267-9274).

[22] Lin H, Parsi A, Mullins D, et al. A Study on Data Selection for Object Detection in Various Lighting Conditions for Autonomous Vehicles[J]. Journal of Imaging, 2024, 10(7): 153.

[23] Li H, Bamminger N, Magosi Z F, et al. The Effect of Rainfall and Illumination on Automotive Sensors Detection Performance[J]. Sustainability, 2023, 15(9): 7260.

[24] Li J, Li B, Tu Z, et al. Light the Night: A Multi-Condition Diffusion Framework for Unpaired Low-Light Enhancement in Autonomous Driving[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 15205-15215.

[25] Leung H K, Chen X Z, Yu C W, et al. A Deep-learning-based Vehicle Detection Approach for Insufficient and Nighttime Illumination Conditions[J]. Applied Sciences, 2019, 9(22): 4769.

[26] Wang K, Zhou T, Li X, et al. Performance and Challenges of 3D Object Detection Methods in Complex Scenes for Autonomous Driving[J]. IEEE Transactions on Intelligent Vehicles, 2022, 8(2): 1699-1716.

[27] Almalioglu Y, Turan M, Trigoni N, et al. Deep learning-based robust positioning for all-weather autonomous driving[J]. Nature machine intelligence, 2022, 4(9): 749-760.

[28] Min X, Zhai G, Gu K, et al. Objective quality evaluation of dehazed images[J]. IEEE Transactions on Intelligent Transportation Systems, 2018, 20(8): 2879-2892.]

[29] Zhang Y, Carballo A, Yang H, et al. Perception and sensing for autonomous vehicles under adverse weather conditions: A survey[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2023, 196: 146-177.

汽标委智能网联汽车分委会

[30] Hnewa M, Radha H. Object detection under rainy conditions for autonomous vehicles: A review of state-of-the-art and emerging techniques[J]. IEEE Signal Processing Magazine, 2020, 38(1): 53-67.

[31] J. M. Jose Valanarasu, R. Yasarla and V. M. Patel, "TransWeather: Transformer-based Restoration of Images Degraded by Adverse Weather Conditions," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 2343-2353.

[32] Saleh K, Szénási S, Vámossy Z. Occlusion handling in generic object detection: A review[C]. 2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI). IEEE, 2021: 000477-000484.

[33] Solawetz J. Occlusion techniques in computer vision[J]. 2020.

[34] Sinha S N. Occlusion[J]. Computer Vision: A Reference Guide, 2020: 1-4.

[35] Wang K, Peng X, Yang J, et al. Suppressing uncertainties for large-scale facial expression recognition[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 6897-6906.

[36] Zhu Z, Wang Q, Li B, et al. Distractor-aware siamese networks for visual object tracking[C]. Proceedings of the European conference on computer vision (ECCV). 2018: 101-117.

[37] Saleh K, Szénási S, Vámossy Z. Generative adversarial network for overcoming occlusion in images: A survey[J]. Algorithms, 2023, 16(3): 175.

[38] Carlini N , Wagner D .Towards Evaluating the Robustness of Neural Networks[C]//2017 IEEE Symposium on Security and Privacy (SP).IEEE, 2017.DOI:10.1109/SP.2017.49.

[39] Borkar T S, Karam L J. DeepCorrect: Correcting DNN Models Against Image Distortions[J]. IEEE Transactions on Image Processing, 2017, PP(99).

[40] Zhang, Yi, et al. Rethinking noise synthesis and modeling in raw denoising[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[41] Yue, Z., Zhao, Q., Zhang, L. and Meng, D., 2020. Dual adversarial network: Toward real-world noise removal and noise generation[C]. In Computer Vision – ECCV 2020: 16th European Conference, 2020, Proceedings, Part X 16 (pp. 41-58).

清华智能网联汽车分委会

[42] Shyam P, Yoo H J. PAIR: Perception Aided Image Restoration for Natural Driving Conditions[C]. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2024: 7459-7470.

[43] Yang X, Sang F, Wang T, et al. Research on the influence of camera velocity on image blur and a method to improve object detection precision[C]. 2021 International Conference on Cyber-Physical Social Intelligence (ICCSI). IEEE, 2021: 1-6.

[44] Zhang Z, Zhang L, Meng D, et al. Vehicle Kinematics-Based Image Augmentation against Motion Blur for Object Detectors[R]. SAE Technical Paper, 2023.

[45] Gong D, Yang J, Liu L, et al. From motion blur to motion flow: A deep learning solution for removing heterogeneous motion blur[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2319-2328.

[46] Uřičář M, Rashed H, Ranga A, et al. VisibilityNet: Camera visibility detection and image restoration for autonomous driving[J]. Electronic Imaging, 2020, 32: 1-8.

[47] Ezaki Y, Moko Y, Hayakawa T, et al. Angle of View Switching Method at High-Speed Using Motion Blur Compensation for Infrastructure Inspection[J]. Journal of Robotics and Mechatronics, 2022, 34(5): 985-996.

[48] Sayed M, Brostow G. Improved handling of motion blur in online object detection[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 1706-1716.

[49] Balunovic M , Baader M , Singh G ,et al.Certifying Geometric Robustness of Neural Networks[J]. Advances in Neural Information Processing Systems. 2019;32.

[50] Wang, F., Xu, P., Ruan, W., & Huang, X. (2023, June). Towards verifying the geometric robustness of large-scale neural networks. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 12, pp. 15197-15205).

[51] Fayyad J, Jaradat M A, Gruyer D, et al. Deep learning sensor fusion for autonomous vehicle perception and localization: A review[J]. Sensors, 2020, 20(15): 4220.

[52] Woodward B, Kliestik T. Intelligent transportation applications, autonomous vehicle perception sensor data, and decision-making self-driving car control algorithms in smart

汽标委智能网联汽车分委会

sustainable urban mobility systems[J]. Contemporary Readings in Law and Social Justice, 2021, 13(2): 51-64.

[53] Huang X, Kroening D, Ruan W, Sharp J, Sun Y, Thamo E, Wu M, Yi X. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability[J]. Computer Science Review. 2020 Aug 1;37:100270.

[54] Li R, Yang P, Huang C C, et al. Towards practical robustness analysis for DNNs based on PAC-model learning[C]//Proceedings of the 44th International Conference on Software Engineering. 2022: 2189-2201.

[55] ISO 34501:2022. Road vehicles — Test scenarios for automated driving systems — Vocabulary.

[56] ISO 34502:2022. Road vehicles — Test scenarios for automated driving systems — Scenario based safety evaluation framework.

[57] ISO 34503:2023. Road Vehicles — Test scenarios for automated driving systems — Specification for operational design domain.

[58] ISO 34504:2024. Road vehicles — Test scenarios for automated driving systems — Scenario categorization.

[59] 杜丽. 面向自动驾驶的场景理解关键技术研究[J].

[60] Werling M, Kammel S, Ziegler J, et al. Optimal trajectories for time-critical street scenarios using discretized terminal manifolds[J], 2012, 31: 346 - 359.

[61] Wei J, Dolan J M, Litkouhi B. Autonomous vehicle social behavior for highway entrance ramp management[C]. 2013 IEEE Intelligent Vehicles Symposium (IV), 2013: 201-207.

[62] Trautman P, Ma J, Murray R M, et al. Robot navigation in dense human crowds: Statistical models and experimental studies of human – robot cooperation[J], 2015, 34: 335 - 356.

[63] Willners J S, Toohey L, Petillot Y. Sampling-Based Path Planning for Cooperative Autonomous Maritime Vehicles to Reduce Uncertainty in Range-Only Localization[J]. IEEE Robotics and Automation Letters, 2019, 4(4): 3987-3994.

[64] Xue Y, Liu C, Zhang X. State Bounding Description and Reachable Set Estimation for Discrete-Time Genetic Regulatory Networks With Time-Varying Delays and Bounded

汽标委智能网联汽车分委会

Disturbances[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2022, 52(10): 6652-6661.

[65] Pilania V, Gupta K. A localization aware sampling strategy for motion planning under uncertainty[C]. 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2015: 6093-6099.

[66] Shamsfakhr F, Antonucci A, Palopoli L, et al. Indoor Localization Uncertainty Control Based on Wireless Ranging for Robots Path Planning[J]. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 1-11.

[67] Indelman V, Carlone L, Dellaert F. Planning in the continuous domain: A generalized belief space approach for autonomous navigation in unknown environments[J], 2015, 34(7): 849-882.

[68] Althoff M, Dolan J M. Online Verification of Automated Road Vehicles Using Reachability Analysis[J]. IEEE Transactions on Robotics, 2014, 30(4): 903-918.

[69] Althoff M, Guernic C L, Krogh B H. Reachable set computation for uncertain time-varying linear systems[C]. Proceedings of the 14th international conference on Hybrid systems: computation and control, 2011: 93 – 102.

[70] Althoff M. Reachability Analysis and its Application to the Safety Assessment of Autonomous Cars[C], 2010.

[71] Scarinci R, Heydecker B, Hegyi A. Analysis of Traffic Performance of a Merging Assistant Strategy Using Cooperative Vehicles[J]. IEEE Transactions on Intelligent Transportation Systems, 2015, 16(4): 2094-2103.

[72] Chen D, Srivastava A, Ahn S. Harnessing connected and automated vehicle technologies to control lane changes at freeway merge bottlenecks in mixed traffic[J]. Transportation Research Part C: Emerging Technologies, 2021, 123: 102950.

[73] Hu X, Sun J. Trajectory optimization of connected and autonomous vehicles at a multilane freeway merging area[J]. Transportation Research Part C: Emerging Technologies, 2019, 101: 111-125.

汽标委智能网联汽车分委会

[74] Ding J, Li L, Peng H, et al. A Rule-Based Cooperative Merging Strategy for Connected and Automated Vehicles[J]. IEEE Transactions on Intelligent Transportation Systems, 2020, 21(8): 3436-3446.

[75] Tian R, Li N, Kolmanovsky I, et al. Game-Theoretic Modeling of Traffic in Unsignalized Intersection Network for Autonomous Vehicle Control Verification and Validation[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(3): 2211-2226.

[76] Min H, Fang Y, Wu X, et al.On-ramp merging strategy for connected and automated vehicles based on complete information static game[J]. Journal of Traffic and Transportation Engineering (English Edition), 2021, 8(4): 582-595.

[77] Wang K, Wang Y, Du H, et al. Game-Theory-Inspired Hierarchical Distributed Control Strategy for Cooperative Intersection Considering Priority Negotiation[J]. IEEE Transactions on Vehicular Technology, 2021, 70(7): 6438-6449.

[78] Ferreira J C, Fonseca C M, Gaspar-Cunha A. Methodology to select solutions from the pareto-optimal set: a comparative study[C]. Proceedings of the 9th annual conference on Genetic and evolutionary computation, 2007: 789 – 796.

[79] Wu Q, Zhao Y, Fan Q. Time-Dependent Performance Modeling for Platooning Communications at Intersection[J]. IEEE Internet of Things Journal, 2022, 9(19): 18500-18513.

[80] Xu B, Li S E, Bian Y, et al. Distributed conflict-free cooperation for multiple connected vehicles at unsignalized intersections[J]. Transportation Research Part C: Emerging Technologies, 2018, 93: 322-334.

[81] Hu Z, Huang J, Yang D, et al. Constraint-tree-driven modeling and distributed robust control for multi-vehicle cooperation at unsignalized intersections[J]. Transportation Research Part C: Emerging Technologies, 2021, 131: 103353.

[82] Liu M, Lu G, Wang Y, et al. Preempt or yield? An analysis of driver’s dynamic decision making at unsignalized intersections by classification tree[J]. Safety Science, 2014, 65: 36-44.

汽标委智能网联汽车分委会

[83] Liu M, Lu G, Wang Y, et al. Analyzing drivers' crossing decisions at unsignalized intersections in China[J]. Transportation Research Part F: Traffic Psychology and Behaviour, 2014, 24: 244-255.

[84] Hafner M R, Cunningham D, Caminiti L, et al. Cooperative Collision Avoidance at Intersections: Algorithms and Experiments[J]. IEEE Transactions on Intelligent Transportation Systems, 2013, 14(3): 1162-1175.

[85] Sun G, Zhang Y, Yu H, et al. Intersection Fog-Based Distributed Routing for V2V Communication in Urban Vehicular Ad Hoc Networks[J]. IEEE Transactions on Intelligent Transportation Systems, 2020, 21(6): 2409-2426.

[86] Nguyen V L, Lin P C, Hwang R H. Enhancing Misbehavior Detection in 5G Vehicle-to-Vehicle Communications[J]. IEEE Transactions on Vehicular Technology, 2020, 69(9): 9417-9430.

[87] Castiglione L M, Falcone P, Petrillo A, et al. Cooperative Intersection Crossing Over 5G[J]. IEEE/ACM Transactions on Networking, 2021, 29(1): 303-317.

[88] Zhao B, Xu N, Chen H, et al. Design and Experimental Evaluations on Energy-Efficient Control for 4WIMD-EVs Considering Tire Slip Energy[J]. IEEE Transactions on Vehicular Technology, 2020, 69(12): 14631-14644.

[89] Gao J, Sun C, Zhao H, et al. VectorNet: Encoding HD Maps and Agent Dynamics From Vectorized Representation[J], 2020: 11522-11530.

[90] Zhou Z, Wang J, Li Y, et al. Query-Centric Trajectory Prediction[C]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023: 17863-17873.

[91] Zhao H, Gao J, Lan T, et al. TNT: Target-driveN Trajectory Prediction[C]. Conference on Robot Learning, 2020.

[92] Zhou Z, Ye L, Wang J, et al. HiVT: Hierarchical Vector Transformer for Multi-Agent Motion Prediction[C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022: 8813-8823.

[93] Liang M, Yang B, Hu R, et al. Learning Lane Graph Representations for Motion Forecasting[J], 2020.

汽标委智能网联汽车分委会

[94] Alahi A, Goel K, Ramanathan V, et al. Social LSTM: Human Trajectory Prediction in Crowded Spaces[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 961-971.

[95] 高铭. 基于深度学习的复杂交通环境下目标跟踪与轨迹预测研究[D]. 吉林大学.

[96] Wang Z, Guan C, Zhao Z, et al. Expressway lane change strategy of autonomous driving based on prior knowledge and data-driven[J]. Physica A: Statistical Mechanics and its Applications, 2024, 640: 129672.

[97] Jiang J, Rui Y, Ran B,等. Design of an Intelligent Vehicle Behavior Decision Algorithm Based on DGAIL[J], 2023.

[98] 覃熊艳, 张雄飞, 汽车与驾驶维修 张 J. 关于 AI 大模型在自动驾驶中的应用研究[J], 2024(003): 000.

[99] Yu L, Shao X, Wei Y, et al. Intelligent Land-Vehicle Model Transfer Trajectory Planning Method Based on Deep Reinforcement Learning[J], 2018, 18.

[100] Pérez-Gil Ó, Barea R, López-Guillén E, et al. Deep reinforcement learning based control for Autonomous Vehicles in CARLA[J], 2022, 81: 3553 - 3576.

[101] Gening Y, Sethi I K. Road-following with continuous learning[C]. Proceedings of the Intelligent Vehicles '95. Symposium, 1995: 412-417.

[102] Moriarty D E, Handley S, Langley P. Learning distributed strategies for traffic control[C], 1998.

[103] Malik S, Khan M A, El-Sayed H, et al.How Do Autonomous Vehicles Decide?[J], 2022, 23.

[104] Kuutti S, Bowden R, Jin Y, et al. A Survey of Deep Learning Applications to Autonomous Vehicle Control[J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 22(2): 712-733.

[105] Wang P, Zhu M, Lu H, et al. BEVGPT: Generative Pre-trained Large Model for Autonomous Driving Prediction, Decision-Making, and Planning[J]. CoRR, 2023, abs/2310.10357.

汽标委智能网联汽车分委会

[106] Xu Z, Zhang Y, Xie E, et al. DriveGPT4: Interpretable End-to-end Autonomous Driving via Large Language Model[J], 2023, abs/2310.01412.

[107] Apollo 3.0 软件架构，https://apolloauto.readthedocs.io/en/latest/specs/Apollo_3.0_Software_Architecture_cn/.

[108] 辰韬资本，南京大学上海校友会自动驾驶分会，九章智驾.《端到端自动驾驶行业研究报告》. 2024 年 6 月.

[109] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017.

[110] Chen, L., Wu, P., Chitta, K., Jaeger, B., Geiger, A., & Li, H. End-to-end autonomous driving: Challenges and frontiers. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2024.

[111]Pomerleau D A. Alvinn: An autonomous land vehicle in a neural network[C]. NeurIPS, 1988.

[112] Kendall A, Hawke J, Janz D, et al. Learning to drive in a day[C]. ICRA, 2019.

[113] Chitta K, Prakash A, Jaeger B, et al. Transfuser: Imitation with transformer-based sensor fusion for autonomous driving[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022.

[114] Chitta K, Prakash A, Geiger A. Neat: Neural attention fields for end-to-end autonomous driving[C]. ICCV, 2021.

[115] Wu P, Chen L, Li H, et al. Policy pre-training for autonomous driving via self-supervised geometric modeling[C]. ICLR, 2023.

[116] Hu Y, Yang J, Chen L, et al. Planning-oriented autonomous driving[C]. CVPR, 2023.

[117] Tian X, Gu J, Li B, et al. DriveVLM: The convergence of autonomous driving and large vision-language models[EB/OL]. 2024.

[118] Xing S, Qian C, Wang Y, et al. Openemma: Open-source multimodal model for end-to-end autonomous driving[C]. Proceedings of the Winter Conference on Applications of Computer Vision, 2025.

[119] Zhang D, Wang G, Zhu R, et al. SparseAD: Sparse query-centric paradigm for efficient end-to-end autonomous driving[EB/OL]. 2024.

汽标委智能网联汽车分委会

[120] Chen L, Wu P, Chitta K, et al. End-to-end autonomous driving: Challenges and frontiers[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024.

[121] Hu A, Russell L, Yeo H, et al. Gaia-1: A generative world model for autonomous driving[J]. arXiv preprint arXiv:2309.17080. 2023.

[122] Su F C, Chen C M, Chen Y C, et al. LINGO-2 polymorphism and the risk of Parkinson’s disease in Taiwan[J]. Parkinsonism & Related Disorders. 2012.

[123] Bojarski M, Del Testa D, Dworakowski D, et al. End to end learning for self-driving cars[J]. arXiv preprint arXiv:1604.07316. 2016.

[124] Jiang B, Chen S, Xu Q, et al. Vad: Vectorized scene representation for efficient autonomous driving[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

[125] Xu Z, Zhang Y, Xie E, et al. Drivegpt4: Interpretable end-to-end autonomous driving via large language model[J]. IEEE Robotics and Automation Letters. 2024.

[126] Shao H, Hu Y, Wang L, et al. Lmdrive: Closed-loop end-to-end driving with large language models[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

[127] Wang T, Xie E, Chu R, et al. Drivecot: Integrating chain-of-thought reasoning with end-to-end driving[J]. arXiv preprint arXiv:2403.16996. 2024.

[128] Mazurowski M A, Dong H, Gu H, et al. Segment anything model for medical image analysis: an experimental study[J]. Medical Image Analysis. 2023.

[129] Kerbl B, Kopanas G, Leimkühler T, Drettakis G. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. 2023.

[130] Ben M, Pratul P. S, Matthew T, et al. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. 2020.

[131] Wu Z, Liu T, Luo L, et al. MARS: An Instance-Aware, Modular and Realistic Simulator for Autonomous Driving. 2024.

[132] Ze Y, Yun C, Jingkang W, et al. UniSim: A Neural Closed-Loop Sensor Simulator. 2023.

汽标委智能网联汽车分委会

[133] Yunzhi Y, Haotong L, Chenxu Z, et al. Street Gaussians: Modeling Dynamic Urban Scenes with Gaussian Splatting. 2024.

[134] Kai C, Xiaoxiao L, Kaizhi Y, et al. GaussianPro: 3D Gaussian Splatting with Progressive Propagation. 2024.

[135] Shanlin S, Bingbing Z, Ziyu J, et al. LidaRF: Delving into Lidar for Neural Radiance Field on Street Scenes. 2024.

[136] 张悦,陈亮.国产云端平台 AI 大模型部署的硬件适配策略研究 [J]. 计算机科学前沿，2024, 18 (6): 135-144.

[137] Li X, Zhang Y. Efficient Model Compression for Edge Devices: A Survey[J]. Journal of Edge Computing, 2024, 6(2): 45-62.

[138] 覃京燕,郝泽宇.无人驾驶车多种空间人类移动性交互设计研究 [J]. 包装工程，2018

[139] Kellnhofer P, et al. Gaze360: Physically Unconstrained Gaze Estimation in the Wild[C]. ICCV,2019.

[140] 高欣.汽车智能座舱的生命健康状态监测系统研究 [J]. 汽车维护与修理，2019

[141] 朱红月,王逸,储水晶,刘健.无人驾驶汽车智能座舱的场景交互研究 [J]. 汽车维修技师，2024 (10):10 - 11.

[142] 陈建清,陆建明,褚玉峰.基于多模态融合的汽车智能座舱人机交互系统优化 [J]. 汽车测试报告，2024 (16):17 - 19.

[143] 大模型专题报告：AI 大模型如何赋能智能座舱_车载大模型 [EB/OL]. 2025

[144] 智能座舱技术在自动驾驶汽车中的应用 [J]. 汽车知识，2024,24 (5):218 - 220.

[145] 邓建明，龚循飞，于勤，等。基于 AI 大模型的新能源汽车智能座舱多模态交互技术研究综述 [J]. 客车技术与研究，2025 (1).

[146] 汽车智能座舱基于多模态交互与 AI 融合的范式创新 [EB/OL]. 电子工程世界,2025

[147] 数智网编.解锁大模型和 AI Agent 在车端的革新应用和落地方法 [EB/OL].2024

汽标委智能网联汽车分委会

[148] Dong, L., et al. "GraphRAG: Graph-based Retrieval-Augmented Generation for Structured Knowledge." Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), 2024

[149] 朱训谕.基于深度学习的小尺度麦克风阵的语音增强研究[D].浙江大学,2021.DOI:10.27461/d.cnki.gzjdx.2021.001990.

[150] Berouti M, Schwartz R, Makhoul J. Enhancement of speech corrupted by acoustic noise[C]//ICASSP'79. IEEE International Conference on Acoustics, Speech, and Signal Processing. IEEE, 1979, 4: 208-211.

[151] Lim J, Oppenheim A. All-pole modeling of degraded speech[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1978, 26(3): 197-210.

[152] Ephraim Y. Statistical-model-based speech enhancement systems[J]. Proceedings of the IEEE, 2002, 80(10): 1526-1555.

[153] Vetterli M. Filter banks allowing perfect reconstruction[J]. Signal processing, 1986, 10(3):219-244.

[154] Logan B. Mel frequency cepstral coefficients for music modeling[C]. In International SymposiumonMusicInformation Retrieval. 2000.

[155] Canyon IP Holdings;LLC,Timely speech recognition[P],海外专利：US9099090, 2015-08-04.

[156] Weinland D, Ronfard R, Boyer E. A survey of vision-based methods for action representation, segmentation and recognition[J]. Computer vision and image understanding, 2011, 115(2): 224-241.

[157] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[158] 胡亚军.基于神经网络的统计参数语音合成方法研究[D].中国科学技术大学,2018.

[159] 覃京燕,郝泽宇.无人驾驶车多种空间人类移动性交互设计研究 [J]. 包装工程，2018

[160] "Artificial Intelligence Act (AI Act) | GFMD Policy & Advocacy Center." GFMD Policy & Advocacy Center, 10 Sept. 2024

汽标委智能网联汽车分委会

[161] 国家互联网信息办公室，国家发展和改革委员会，教育部，等。生成式人工智能服务管理暂行办法 [EB/OL].2023

[162] 全国网络安全标准化技术委员会。网络安全技术 生成式人工智能服务安全基本要求（征求意见稿）[EB/OL].2024

[163] Abujabal A, Minchev P, Hofstätter P, et al. Benchmarking Large Language Models in Retrieval-Augmented Generation[J]. arXiv preprint arXiv:2305.15017, 2023

汽标委智能网联汽车分委会

三、车用 AI 安全及标准化

第八章 车用 AI 测试与评估标准化研究

8.1 车用 AI 测试评价技术概述

随着 AI 系统在智能网联汽车中的广泛渗透，模型、数据、隐私与伦理等多方面的安全问题已引起社会各界的高度关注。标准化测试作为保障 AI 系统安全、可靠、可控的核心手段，能够为技术研发与产品迭代提供明确目标和评估基准，有效评估其在不同场景下的准确性、合理性与控制精度，避免潜在系统缺陷引发事故，保障安全，助力技术进步与产业升级。因此，车用 AI 系统测试与评估的标准化研究，旨在构建统一的评价体系，在保障交通安全、引导技术演进、加强法规监管与促进国际互认等方面发挥关键作用，推动自动驾驶技术健康有序发展。本章通过对国内外车用人工智能测试评价方法、流程、工具及标准体系的系统梳理，探讨车用 AI 系统测试评价技术的发展态势与标准化路径，了解当前技术实施与标准制定中的关键难点与和技术挑战。

8.2 车用 AI 测试评价方法

8.2.1 通用测试方法

8.2.1.1 仿真和物理测试

按照测试手段分，测试评估方法可以包括模拟仿真和物理测试，模拟仿真可分为软件在环测试 SIL、硬件在环测试 HIL 和车辆在环测试 VIL 等；物理测试包括模拟环境下的测试以及实际环境下的测试[1]。

8.2.1.2 开环测试和闭环测试

开环测试是在系统不考虑输出反馈的情况下，对自动驾驶系统（ADS）的各组件进行独立或串联测试，便于快速定位和修复问题。由于感知算法不涉及系统

汽标委智能网联汽车分委会

与环境的交互，开环测试适用于模块化自动驾驶算法。决策规划算法的测试可基于独立环境输入，模拟器与物理世界的一致性影响较小。

闭环仿真模拟系统与环境的相互作用，通过反馈循环评估系统表现，确保鲁棒性和适应性。随着计算机图形学、计算机视觉和生成式 AI 的发展，闭环仿真工具将显著提升端到端自动驾驶系统的测试和验证效率。

8.2.1.3 白盒测试、黑盒测试、灰盒测试

白盒测试基于程序内部结构设计测试用例，关注内部控制流程，不考虑整体功能实现，属于逻辑驱动测试。

黑盒测试视被测对象为黑盒，仅考虑整体特性，不涉及内部实现，适用于系统、子系统或模块等。

灰盒测试结合白盒和黑盒测试，既利用整体特性信息，又考虑内部实现信息，适用于集成测试阶段，关注输入输出正确性及程序内部状态。两种信息的比例决定了灰度，灰盒测试在集成测试中尤为常见。

8.2.1.4 离线测试和在线测试

AI 系统模型常需迭代更新，A/B 测试用于比较新旧模型的性能差异，以确定优化或衰退，从而选择合适模型上线。离线测试在模型部署前使用预设数据集评估性能，是判断模型迭代效果的重要手段。在线测试，即闭环评估，在真实环境中进行，实时接收传感器输入并生成控制策略，提供真实多样数据以优化模型。新旧模型可同时部署，通过决策对比来验证新模型性能。在线测试还能监控用户行为和环境变化，确保 AI 系统适应性和有效性。图 8-1 展示了基于 AI 的智驾系统部署流程。

汽标委智能网联汽车分委会

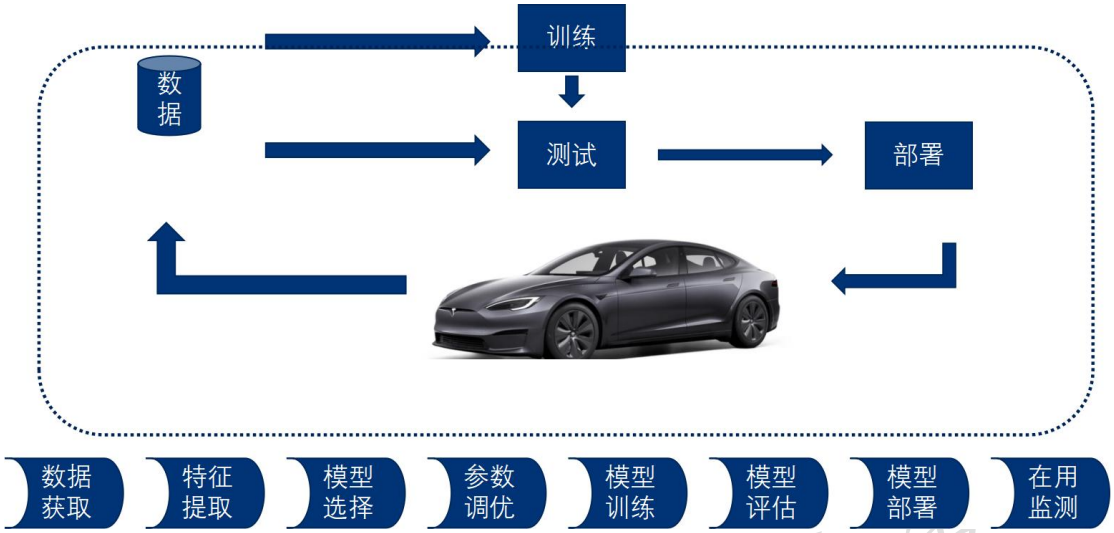


图 8-1 基于 AI 的智驾系统部署图

8.2.1.5 试验方法推荐附表

表 8-1 主要感知系统指标

	模拟仿真			封闭道路	实际道路
	SIL	HIL	VIL		
测试方法	●	○	○（纯视觉）	○或 ●（纯视觉）	×或 ○（纯视觉）

注：● 建议的必要测试，○建议的选做测试，×建议的可不做的测试

8.2.2 感知系统测试与评估方法

感知系统的测试目标是验证传感器及感知算法获取环境信息的准确性和稳定性。重点评估感知模块在各种环境条件下的目标检测、识别与定位能力，确保输出高精度环境模型，为后续预测决策提供可靠依据。

8.2.2.1 主要性能指标

智能驾驶系统感知系统测试的性能指标可参考 4.3.1 章节，实际测试时需根据需求选择。表 8-2 列出了推荐的感知系统测试主要指标供参考。

表 8-2 感知系统测试主要性能指标

指标分类	识别类指标	精度类指标
感知系统	准确率 Accuracy、精确率 Precision	平均精度（mAP）、平均平移误差（ATE） 平均尺度误差（ASE）、平均方向误差 (AOE)、平均速度误差 (AVE)、平均属性误差（AAE）

8.2.2.2 干扰要素生成方法

图 8-2 天气对传感器的影响

在真实道路环境中，智能驾驶系统需面对雨雾、光照、遮挡等多种复杂且不确定的环境挑战，这些干扰要素会显著影响传感器性能，进而削弱系统的整体识别与决策能力。为系统性评估感知模块在异常条件下的鲁棒性与稳定性，干扰要素生成方法通过构建典型干扰变量，复现传感器在不同场景下的干扰特征，为仿真测试和实车验证提供了可控、可复现的输入，可以有效支撑 AI 模型在边界条件下的安全性验证[2]。

(1) 基于实体构建的干扰要素生成

基于实物构建通过考虑传感器干扰、场景设计、实体模型和道具使用，模拟真实复杂环境。场景设计包括复制城市街道、高速公路、乡村小路等；设计极端场景如恶劣天气、低光照、施工、事故现场等，并用环境模拟设备实现。利用灯光控制光照强度和角度，使用实体模型或虚拟仿真模拟动态交通行为，测试自动驾驶系统的感知和决策能力。实体模型和道具可通过静态障碍物模型（如路牌、路灯）、动态模型和天气设备干扰场景，测试识别、避障和感知性能。

(2) 基于生成模型的干扰要素生成

基于生成模型的干扰要素生成主要分为两种：一是通过干扰模型在图像/视频上添加干扰生成新测试数据；二是利用生成式神经网络学习数据集，生成含干扰要素的测试场景。干扰模型常用于雨、雪、雾等噪声的建模，通过调整图像的像素参数构建噪声模型，生成测试感知系统的图像或点云样本。该过程通常包括噪声特性分析、模型设计、参数设定、验证调整和应用测试等步骤。例如，雨雪噪声表现为图像中的随机亮点或暗点，需考虑其形状、大小、密度、透明度及动态变化；雾噪声则主要降低图像对比度，使颜色偏淡并产生模糊，需理解雾对光线的散射和吸收作用。模型结构设计上可采用基于物理特性的数学模型、统计分布，或将两者融合构建合成模型；参数设定需根据天气条件设定如透射率、大气光值等噪声参数，并结合图像分辨率、色彩空间等图像参数；模型验证可通过视觉检查、PSNR、SSIM 等指标评估其准确性与实用性，并据此优化。

图 8-3 干扰模型的应用

相比于在原始数据上添加干扰的方式，生成式神经网络方法则依赖神经网络模型直接生成带有干扰要素的图像、点云或视频，不再依赖已有干净图像作为输入，适用于构建更加复杂和真实的测试样本。

汽标委智能网联汽车分委会

表 8-3 主要干扰要素及实现方式

分类	干扰要素	主要指标	描述	模拟仿真实现方式	推荐	封闭场地实现方式	推荐	实际道路实现方式	推荐
环境	雨	雨量	微雨、小雨、中雨、大雨和暴雨	场景生成软件	推荐	环境模环设备	推荐	随机天气	不推荐
	雪	雪量	零星小雪，小雪，中雪，大			雪地试验场			
	雾	影响能见度	几乎不影响			极端天气试验场 环境模拟设备			
	霾		轻微影响						
	沙尘暴		中度影响				不推		
	烟雾		重度影响				推荐		
	光照	强度、方向	顺光、逆光			试验场			推荐
	温度	高温、极寒	高温（高于 60 摄氏度）	NA	不推荐	极端天气试验场			不推荐
			极寒(低于-40 摄氏度）						
	时间	黎明、白天、黄昏、黑夜	NA	场景生成软件	推荐	试验场		选取适合时间	推荐

汽标委智能网联汽车分委

	能见度	距离		NA	不推荐			随机天气	不推荐
	湿度	湿度区间				NA			
道路要素	路面	干燥；积水；积雪；结冰；	NA	场景生成软件	推荐	试验场		选取适合路段	不推荐
	摩擦力	摩擦系数		NA	不推	试验场 实体模拟构建			
	交通标识	交通标志牌、路面标线	指引线、限速标线等	场景生成软件	推荐				推荐
	护栏设施	位置，状态	NA	场景生成软件	推荐				
	信号灯								
	道路涂鸦								
	广告牌								
	行道设施	行道树、路灯等							
	车道	车道线	颜色，模糊、清晰、中断、						
交通参与者	行人	姿态	站立，半蹲，蹲坐，躺卧	场景生成软件	推荐		推荐	选取适合路段	推荐
		携带物	伞，行李箱，货物，轮椅，						
		衣着	反光背心，连体衣，人偶服						

汽标委智能网联汽车分委会

	动物	姿态	站立，躺卧					NA	不推
		衣着	反光背心，连体衣，人偶服						推荐
	障碍物	体积、位置、姿态	NA	场景生成软件				选取适合路段	推荐
	机动车	车辆类型、车辆行为							
	非机动车	自行车							
传感器干扰	自身	传感器性能局限	NA					选取适合路段	不推
	屏蔽物	遮挡	遮挡程度						推荐
运动状	自车状态	速度	NA						

汽标委智能网联汽车分委会

8.2.2.3 测试场景生成方法

(1) 基础测试场景的构建

基础测试场景的构建主要依赖于智能驾驶系统的设计运行条件以及功能，从大的方面来说，可以分为高速、城市环境以及行车和泊车功能。通过分析实际感知系统的特点，可以进行设定若干用于测试的基础场景，表 8-4 为场景要素示例。

表 8-4 典型场景要素示例

序号	区域分类	道路环境	备注
1	高速/城区/停车场	直道	高速/城区，叠加干扰要素
2	高速/城区	弯道	
3	高速	匝道	
4	城区	十字交叉路口	
5	城区	丁字路口	
6	城区	环岛	
7	城区	路面停车场	
8	城区	地下停车场	

(2) 扩展测试场景的构建

通过对基础场景增加干扰要素构建扩展测试场景，比如更改测试环境或者替换目标物测试开展试验。

汽标委智能网联汽车分委会

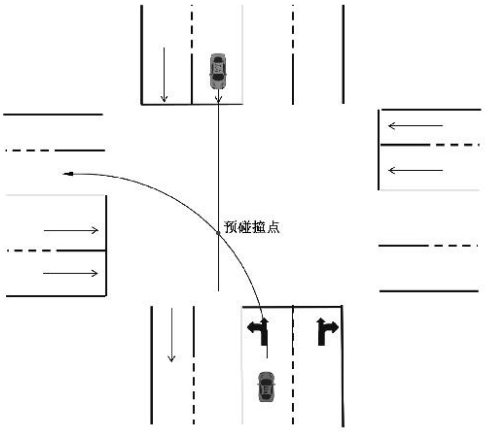


图 8-4 无保护左转路口对向存在直行车辆试验场景示意图

表 8-5 无保护左转路口对向存在直行车辆试验场景基础及干扰要素

基础要素	城市道路：进口道车道线、停车线、限速
干扰要素	雨天 照明 传感器遮挡 车辆速度 道路标志涂鸦 异形目标物

注：参考国标 GB/T 41798-2022 智能网联汽车 自动驾驶功能场地试验方法及要求给出部分实例，实际需要根据产品设计运行条件进行构建。

8.2.3 预测与决策系统测试

8.2.3.1 预测与决策系统测试主要指标

智能驾驶系统预测决策系统测试的相关性能指标可以参考 5.4 章节的主要内容，但在开展测试时一般会根据实际需要进行选取，表 8-6 推荐了一些预测决策系统测试时的主要指标供参考。

指标分类	指标	含义
安全指标	MPCI	两个安全接管之间自动驾驶 AI 系统能安全行驶的平均里程数。该指标数值越大，则系统在 AI 决策规划方面的安全性方面大概率越高。
	跟车纵向安全距离	为避免碰撞，自车需与前车保持的理论安全距离下界。即使出现前车突然急刹停情况，本车在及时反应后，一定可以在纵向上避免本次碰撞。自动驾驶决策规划系统应该能长期保持其跟车距离在该指标指示的范围之外，如果 AI 系统有较高频率或长时间不满足该距离，那么决策规划系统的安全度可能较低。
	横向安全距离	为避免横向碰撞，自车需与并行车辆保持的横向距离。即使双方同时并线情况下，本车即使反应后，一定可以在横向上避免本次碰撞。自动驾驶决策规划系统应该能长期保持其跟车距离在该指标指示的范围之外，如果 AI 系统有较高频率或长时间不满足该距离，那么决策规划系统的安全度可能较低。
	碰撞事件 TTC	该指标表示车辆在当前速度下达到前车/前方障碍物位置的时间。TTC 值越小，表明车辆与前方障碍物之间的碰撞风险越高。如果 TTC 达到一个临界值，系统通常会触发警告或采取自动制动措施以避免碰撞。TTC 值为负，表示车辆正在远离障碍物，通常这种情况下不会触发任何警告。
	碰撞率 CR	碰撞率表示为每百万英里行驶里程内发生的碰撞次数。较低的碰撞率表示自动驾驶系统的安全性较高。
	车道偏离事件	车道偏离事件是指自动驾驶车辆在没有合法原因情况下偏离车道的事件次数。该指标用于评估自动驾驶车辆的车道保持能力，较低的车道偏离

		表示自动驾驶系统的车道保持能力较强，安全性较高。
	紧急制动频率	该指标表示自动驾驶车辆每百万英里行驶里程内发生紧急制动的次数。较低的 EBF 值表示自动驾驶系统在预测与避障决策方面表现更好。
	事故严重度	事故严重度通常根据事故造成的损失和伤亡程度进行打分。事故严重度指数越高，表明事故的平均严重程度越高。
	法规遵从性	法规遵从性反映自动驾驶系统是否严格遵守交通法规，如速度限制、信号灯响应等。值越高，说明自动驾驶系统越难以遵从交通法规。
效率指标	平均通行时间	指自动驾驶系统多次通过相同或相近的路况条件下的平均每次完成通行的时间。可通过实车选择具有代表性的路况或搭建特定仿真场景后测试的方法计算指标。在相同或相近路况下，平均通行时间越短，代表自动驾驶系统通行效率越高。
	换道回退率	指自动驾驶系统执行换道行为过程中回退到换道前车道的比例，一般可通过实车测试或仿真测试后统计行为次数进行计算。换道回退率越低，代表自动驾驶系统执行的有效换道效果越好。

表 8-6 主要预测决策系统指标

表 8-7 不同测试方法下的安全性指标适用性

安全性指标	基于场景片段	连续道路测试
接管里程	×	√
跟车纵向安全距离	√	√
横向安全距离	√	√
碰撞时间	√	√

碰撞率	×	√
车道偏离事件	×	√
紧急制动频率	×	√
事故严重度	√	√
法规遵从性	√	√

8.2.3.2 挑战测试场景要素及生成方法

挑战测试场景是自动驾驶系统验证中至关重要的部分，主要聚焦在低频高风险的长尾情形，这类情形虽发生概率低，但对 AI 模型构成极大挑战。其挑战要素主要包括动态干扰要素与静态干扰要素[3]。

动态干扰主要指交通参与者的复杂行为，如机动车急刹、变道、无保护左转，非机动车突入行驶，以及行人闯红灯、密集分布等。生成高风险周车运动轨迹指的是控制一辆或多辆车辆，令其沿着特定的对抗性轨迹行驶，从而干扰自动驾驶系统的预测和决策模块，如图 8-5。常见的对抗轨迹生成方法包括数据驱动方法、强化学习算法、启发式搜索算法等，以模拟对抗性运动干扰，检验系统鲁棒性。

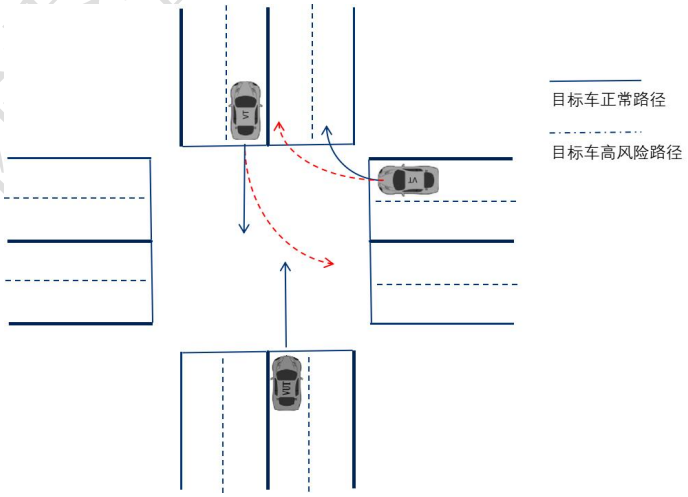


图 8-5 高风险周车运动轨迹

静态干扰则包括模糊的车道线、矛盾的路牌标识、突发障碍物（如静止车辆、施工物料、倒地树木）等。这类干扰虽不具备动态演化能力，但对路径规划、障碍判断等模块影响显著。通过系统性整理动态行为特征与静态结构干扰形式，可形成标准化挑战要素库，为后续测试场景设计提供支持。

典型动静态干扰要素汇总如下：

表 8-8 动静态干扰要素汇总表

要素类别	要素	要素特征属性
动态要素	机动车	交叉口掉头 切入/切出 急刹（自动驾驶跟车） 无保护左转冲突 变道/超车 急转弯
	非机动车	误入机动车道 突然闯入 交叉口内随意行驶
	行人	闯红灯 突然闯入 随机走动 人流密集度
静态要素	道路	道路类型：城市道路/高快速路 车道数/车道宽度

汽标委智能网联汽车分委会

		几何特征：曲率、坡度 路面摩擦系数、材质
	交叉口	交叉口类型：环岛、信控、无信控 进出口道数量 交叉口几何特征：环岛半径、进出口道夹角等
	标志线	标志标线类型 清晰度
	信号灯	信号配时以及相位
	基础设施（收费站、车站、立交桥、施工段等）	设施布设的几何特点：位置、角度、规模尺寸、与道路的拓扑关系等

8.2.3.3 挑战测试场景生成方法

测试场景构建是确保自动驾驶系统性能评估准确性的关键步骤。测试场景需要模拟真实世界中的各种驾驶条件，包括但不限于道路类型、交通流量、环境因素等。

(1) 片段场景的构建

片段式场景构建是挑战测试的重要基础方法，强调对特定交通任务和典型功能进行结构化测试。其基本思路是基于智能驾驶系统的设计运行条件构建标准化基础场景，如城市路口通行、高速变道、泊车入库等，再叠加环境干扰与交通参与者异常行为形成挑战性扩展场景。标准构建可参考国家推荐标准 GB/T 41798-2022，确保场景的标准化和一致性。

汽标委智能网联汽车分委会

表 8-9 典型预测决策场景

序号	预测决策场景列表		
1	交通信号灯控制场景	绿灯起步/红灯停车	基础功能场景
2		无保护交叉口穿行	
3		左转等待绿箭头	
4	行人和非机动车辆穿行	城市交叉口行人穿越	
5		非机动车并行与切换车道	
6	复杂交叉路口通行	多车道交叉口穿越	
7		转弯车道选择	
8	车流密集通行	高峰时段城市高速公路	
9		城市交通堵塞	
10		紧急情况下的车流穿插	
11		停车场和商业区驾驶	
12	高速跟车	高速变道	
13		上下匝道	
14		紧急停车	
15		交通堵塞绕行	
16		隧道行驶	

汽标委智能网联汽车分委会

17	带有路权博弈的场景	匝道汇入	
28		十字路口或 T 字路口，直行遇到右转汇入	
19		无交通灯的十字路口直行遇到左转	
20		十字路口车辆违反交通灯行驶	

此类方法具有参数可控、测试可重复等优势，适用于对单项功能的鲁棒性、安全性和容错能力进行精准验证。通过预定义的剧本化输入，可实现对模型边界条件下响应策略的细粒度评估。例如，针对图 8-6 无保护左转场景中叠加对向车辆突入、标线模糊与路权冲突等多种干扰组合，有助于系统在复杂博弈条件下进行压力测试，从而检验其感知准确性、预测稳健性与决策合理性。

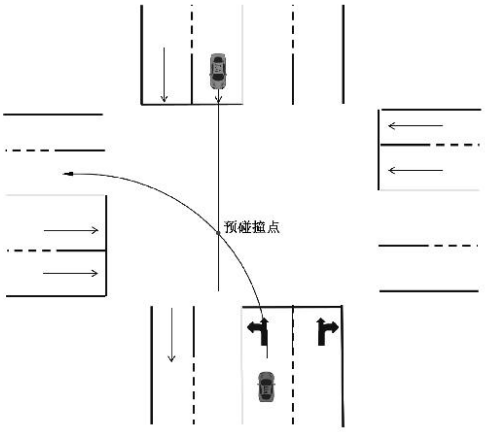


图 8-6 无保护左转路口对向存在直行车辆试验场景示意图

表 8-10 无保护左转路口对向存在直行车辆场景挑战测试基础及干扰要素

基础要素	城市道路：进口道车道线、停车线、限速
干扰要素	路牌遮挡，对抗因素 道路标识线覆盖，缺失 路径重叠，碰撞风险 路权博弈

(2) 连续仿真测试场景

相比于片段场景测试“参数化、剧本化”的特征，连续场景测试具有场景闭环连续、交通参与者随机可交互的特点，支持构建连续化场景集、可进行无限时长仿真、且每次仿真均有不同的“剧本”，可真正实现面向自动驾驶的多场景加速测试。同时，若在仿真平台中嵌入合适的驾驶员模型，则连续仿真场景测试可支持交通参与者与自动驾驶的交互博弈，且场景概率化触发，参与者非固定，从而实现对复杂交通环境的真实模拟。

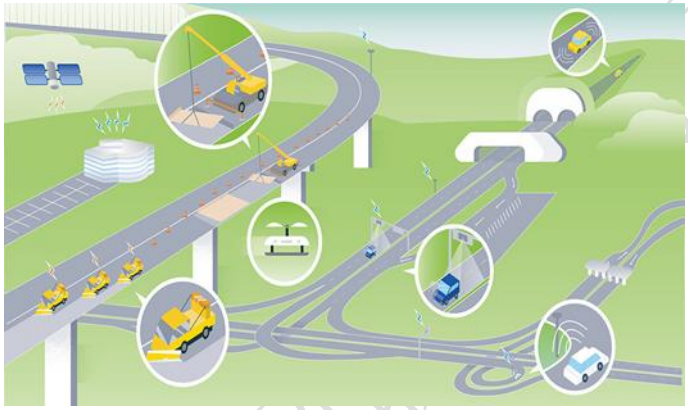


图 8-7 连续仿真测试场景路线示意图

该方法适用于大规模仿真验证与泛化能力测试，典型的评价体系如 CARLA 排行榜与 Bench2Drive，分别设计了包含城市、高速、居民区等多类型道路与多种天气条件下的闭环路线任务，嵌入行人、自行车、对抗车辆等动态元素。系统需在每次测试中自主适应不同交互路径，实现端到端的复杂任务处理。该方法优势在于测试规模大、数据覆盖全面，尤其适用于评估 AI 模型在动态博弈、多目标冲突与长尾风险场景下的稳健性与适应性。

8.2.4 智驾系统测试

8.2.4.1 主要指标

对于车用人工智能的自动驾驶系统，一般基于整车进行系统测试，主要验证系统的泛化能力和鲁棒性水平，泛化能力指的是自动驾驶系统在面对未知情况或未在训练数据中见过的场景时，仍能做出正确决策的能力，一个具有高泛化能力

的系统能够识别和适应新环境、新交通模式或不常见的障碍物，从而在各种不同的驾驶条件下保持稳定和安全的操作。鲁棒性指自动驾驶系统在面对极端条件、意外情况或系统内部错误时，仍能保持功能或安全地降级其功能的能力。鲁棒的系统能够抵抗外部攻击、硬件故障、传感器噪声和其他非预期的干扰。鲁棒性测试可以通过进行压力测试、模拟极端情况、增加外部干扰等方式来实现。

现阶段通过模拟仿真、场地测试、实际道路的“三支柱”测试方法进行车用人工智能的自动驾驶系统验证，在进行场地和仿真测试时，对于单一测试用例，采用通过和不通过的方式判定试验结果，如在 GB/T 41798-2022 《智能网联汽车自动驾驶功能场地试验方法及要求》的 5.2 章节给定了试验通过条件，各个测试用例又定义了通过要求。目前，对于通过率的阈值，目前的“三支柱”的国家标准中为 100%，即不允许有不通过的情形，但是由于干扰要素的原因，是否保持这一要求，还需要进一步的研究。

通过率指标可以分为单项测试的通过率，比如单个测试用例进行 N 次，则单项的通过率可以计算

$$\text{pass_rate}=\frac{N_{\text{pass}}}{N_{total}}$$

对于实际道路测试，除了通过率标准之外，还可以设置一些参考性指标，用于评判智能驾驶的安全性能。

最小安全距离保持：模型遵守安全跟随距离的能力，是否能够在保持体感优秀的情况下达到安全跟车距离。

紧急制动响应时间：遇到危险情况时，从检测到做出制动反应的时间。

遵守交通规则的情况，如红绿灯识别、限速遵守等。

对于效率性、舒适性指标，以及参考性指标不作为本章的主要内容，可以参考预测与决策系统的测试方法进行，区别在于感知输入使用外部输入和传感器模型代替真值输入。

汽标委智能网联汽车分委会

8.2.4.2 干扰要素生成方法

相对传统基于规则的自动驾驶系统，基于人工智能的自动驾驶系统的验证，在进行场景设计时主要考虑。

(1) 测试干扰场景

自动驾驶系统测试中的干扰场景指的是在测试过程中故意引入的变量或条件，用以评估自动驾驶系统在面对非典型或极端情况时的性能和鲁棒性。这些场景可能包括但不限于：

传感器干扰：模拟传感器失效被遮挡，被污染的目标物体、异形障碍物等情况，以测试系统的冗余、感知算法的能力。

环境干扰：包括恶劣天气条件如大雨、雾、雪等，这些条件可能影响车辆的感知系统。

交通干扰：在测试中引入意外的交通情况，如突然切入的车辆、行人或其他障碍物。

以上场景可以通过对车辆自动驾驶系统的功能安全、预期功能安全的安全分析进行后进行设计，可以是通过经验的方法进行也可以通过生成式人工网络生成挑战场景。

(2) 模拟极端情况

极端场景主要来源，主要是对于逻辑场景中可选择参数的边界，以及潜在碰撞的场景，比如：

ODC 边界：车辆在即将接近 ODC 边界或者刚刚超出 ODC 边界时的场景。

车辆性能的边界：可能涉及车辆的加速、制动、转向等性能边界的场景。

潜在碰撞的危险场景：在进行设计时，寻找到可能发生碰撞事故的最大化参数组合，这个可以采用算法筛选的方式，也可以通过生成式神经网络模型进行危险场景的生成。

汽标委智能网联汽车分委会

冲突场景：路权博弈场景、交通规则符合性冲突场景、伦理冲突场景等。

8.2.4.3 测试场景生成方法

与感知决策和预测决策系统的测试不同在于，对于本章的智能驾驶系统的测试，是将整个系统看成一个测试对象，是不同外部环境条件下车辆的整体行为表现，因此是将系统看成一个黑盒的测试，而对于感知系统测试和预测决策系统测试是基于产品开发的白盒、黑盒或者是灰盒测试，因此测试指标的要求不同，其测试场景也需要做一定的调整。对于智能驾驶系统的研发测试，可以在本章节的基础上，参考 3.3 预测与决策系统测试的内容进行。

根据车辆的设计运行条件和功能，构建基础测试场景，并根据 8.2.2.2 相关干扰和挑战要素构建扩展测试场景，方法和步骤可以参考 8.2.2.3。对于基础场景可以参考 GB/T 41798-2022《智能网联汽车自动驾驶功能场地试验方法及要求》等标准。

8.2.4.4 试验方法及要求

在智能驾驶系统验证的过程中，模拟仿真、封闭场地测试以及实际道路测试都是必不可少的环节，各自具有独特的优势和用途。由于基于人工智能驾驶系统的验证包含干扰场景和交通挑战场景，对于试验人员和道路使用者存在着一定的安全风险，因此模拟仿真测试相对而言最为安全，且场景构建相对方便，成本低、效率高，但是需要保证仿真工具链的可信度水平；封闭场地测试提供了一个受控的环境，可以在没有真实交通风险的情况下测试车辆，可以实现一定的高风险或极端条件下的测试，但是受限于昂贵的环境模拟设备和较低的场景搭建效率；对于实际道路测试，能够验证系统在复杂多变的真实交通状况下的表现，由于考虑到安全性和测试效率，对于环境干扰以及交通环境挑战，并不做特意的规定，在设计运行条件下，尽量覆盖不同的地点和时段进行测试。

汽标委智能网联汽车分委会

8.3 车用 AI 测试评价流程

8.3.1 车用 AI 智能驾驶领域测试需求

智能驾驶系统由感知、预测、决策与控制四个模块构成，协同实现自动驾驶功能。AI 测试评价的核心目标是验证系统在设计运行条件（ODC）下的安全性、适应性与合规性，其主要需求体现在三个层面：

功能安全性：通过 MPCI、TTC 等指标评估关键时刻系统的响应能力；

极端场景鲁棒性：模拟雨雾、光照突变等干扰条件，验证感知与融合算法的抗干扰能力；

法规与伦理合规性：评估系统在复杂规则与道德冲突情景下的行为合理性。

此外，由于传统基于规则的测试方法难以覆盖全部风险，测试体系需构建“场景生成—测试执行—数据闭环”的迭代流程，借助仿真平台、生成模型等技术手段扩展测试边界，提升覆盖率与工程适应性。

8.3.2 测试评估基本原则

车用 AI 系统测试与评估的基本原则包括：

(1) 安全性至上：车用 AI 系统必须确保行车安全，具备强大的鲁棒性，应对输入错误、传感器噪声及外部干扰。同时，系统的实时性能至关重要，任何延迟可能导致严重后果，且需要验证系统在不同驾驶场景下的泛化能力。

(2) 系统级视角：测试不仅要在单独的 AI 组件上进行，还要在软硬件集成后的 AI 系统上进行，以更真实地反映系统的实际表现。

(3) 全面性覆盖：测试场景应涵盖所有可能的使用情形，包括正常操作、故障模式、极端环境条件和用户交互等，确保系统在各种实际应用中表现良好，同时要符合相关法规、标准和行业规范。还需检验系统应对对抗场景与恶意攻击的能力。

清华大学智能网联汽车分委会

(4) 定量与定性结合：测试既要关注定量性能指标，如识别率、反应时间等，也需进行定性评估，如决策合理性和用户体验，全面评估系统的性能。

(5) 持续性评估：由于 AI 系统具备持续学习和适应能力，需要进行长期、持续的性能监测与评估，确保系统在实际使用中能够不断优化并保持安全、可靠的性能。

8.3.3 测试评估流程

测试评估流程以及测试规范的编制，是一个包括以下六个主要步骤的复杂过程，每个步骤都需要详细规划与执行，确保 AI 系统的质量满足预定标准，如图 8-8 所示。

图 8-8 AI 测试流程

定义测试范围：明确测试对象和生命周期，包括从单元/组件测试到集成测试、系统测试、系统集成测试、验证测试及维护性测试等各阶段。

分析测试要求：仔细分析 AI 系统的功能、性能、可靠性等要求，并进行详细拆解，为后续测试做好准备。

确认测试方法：依据相关标准、技术及流程，确认具体的测试步骤和方法。

定义测试用例：编写功能、性能、安全和可靠性测试用例，确保覆盖 AI 系统的各个方面，保证系统质量。

设计测试环境：设计适合 AI 系统的测试环境，确保支持测试所需的资源。

进行 AI 测试：根据测试用例执行功能、性能、安全和可靠性测试，评估系统的正确性和可用性。

汽标委智能网联汽车分委会

8.4 车用 AI 测试评价工具

8.4.1 模拟仿真工具

针对基于人工智能的智驾系统的测试，模拟仿真测试工具分为场景设计工具、软件在环测试工具、硬件在环测试工具、车辆在环测试工具。

8.4.1.1 仿真平台

仿真平台通常分为两种：日志仿真（LogSim）和世界仿真（WorldSim）[4]。

日志仿真：日志仿真能够根据实车的落盘日志记录精准复现实车场景，并利用日志数据离线重新执行特定的线上模块，如感知、预测、 规控等，从而达到评估算法模块效果，复现线上问题的目的。对于自动驾驶系统，该方法具有逻辑简单、稳定可靠、仿真精度高等优势，是不可缺少的仿真评估工具。然而，日志仿真的核心局限在于仿真平台和模型的交互数据是固定不变的，场景变化完全依赖日志内容，无法根据模型输出实时调整，因而难以实现算法在连续时间片下的闭环响应验证。若需开展可交互、动态响应的闭环仿真，则需借助具备行为建模与反馈能力的世界仿真系统以实现更真实的系统级验证。

世界仿真：和日志仿真相比，世界仿真的核心优势就是仿真系统能够根据仿真算法的输出自由构建仿真器中的场景。仿真系统能够根据模型输出动态调整自车状态，并据此生成新的输入，实现针对特定模块的闭环验证流程。一些平台还支持基于车辆几何与动力学特征的建模，使仿真结果更贴近实车行为。然而，世界仿真的输入需由仿真系统主动构造，其真实性直接影响模拟精度。目前多数平台仍以结构化信息为主，难以还原原始传感器数据，限制了对端到端模型的评估能力。

为满足高阶智能驾驶与端到端模型的发展需求，越来越多仿真平台引入世界模型、强化学习与扩散模型等新兴技术。典型的开源平台如 CARLA，提供城市级场景建模与 Python 接口，适用于决策系统与强化学习训练；AirSim 基于虚幻引擎，支持无人机与自动驾驶算法的高保真训练；Waymax 能复现真实交通交互

汽标委智能网联专业分委会

过程，适用于多 Agent 交互场景仿真。在企业级方面，NVIDIA Drive Constellation 支持全栈闭环仿真并仿真多种天气与光照；Waymo Carcraft 每日支撑大规模虚拟车队运行，用于持续学习与技能测试；CarSim 及其拓展软件广泛应用于整车动力学仿真。在国内，百度 Apollo 仿真平台和腾讯 TAD Sim 分别聚焦于云端决策测试和高精度场景建模，支撑大规模 AI 算法验证与回归测试。

8.4.1.2 场景设计工具

场景设计工具通过模拟真实世界的道路条件、交通情况、天气变化等，能够在没有实际上路测试的风险和成本下，对车用人工智能系统进行全面和深入的测试。常见的模拟环境测试工具有 CARLA、SUMO、51Sim-One 等，这些工具虽然在具体的实现、功能和专注点上存在差异，但它们也存在一些共性特征。能够提供多样化的环境和场景模拟能力，包括不同的天气条件、时间（白天/夜晚）、交通流量等，以模拟现实世界中自动驾驶汽车可能遇到的各种情形。同时支持多种车载传感器（如摄像头、雷达、激光雷达）的模拟，生成虚拟的传感器数据供算法处理和分析。目前主流的场景生成方法包括基于神经网络渲染的场景生成和基于世界模型的自动驾驶场景生成两类。

(1) 基于神经网络渲染的场景生成

基于神经网络渲染（Neural Rendering）的场景生成技术，通过深度学习模型直接从数据中学习场景的视觉与几何特征，能够高效合成高度逼真且可扩展的虚拟环境，为自动驾驶算法的训练与测试提供全新解决方案。

(2) 基于世界模型的自动驾驶场景

在自动驾驶数据获取方面面临着重大挑战，包括高昂的数据收集和标注成本、法律限制以及安全考量。世界模型通过自监督学习范式，提供了一种有前途的解决方案，可以从大量未标记数据中提取有价值的见解，从而以成本效益提升模型性能。世界模型在驾驶场景生成中的应用尤为值得注意，因为它能够创建多样化和逼真的驾驶环境。这种能力显著丰富了训练数据集，使自主系统具备在稀有和复杂驾驶场景中导航的稳健性。

北京智能网联汽车分会

8.4.1.3 软件在环测试工具

软件在环测试工具允许开发人员在没有物理硬件的情况下执行软件组件，这是通过在计算机上模拟软件与其交互的硬件和外部环境来实现的。这种方法不仅可以降低测试成本，还能提早发现和修复软件缺陷。常见的 SIL 测试工具有 CARLA、AirSim、VIRES VTD、Prescan、PanoSim 等。SIL 工具的使用促进了更高效的迭代开发过程，支持模拟传感器数据和执行器响应，为开发嵌入式系统和自动驾驶技术等复杂应用提供了强大的支持。

8.4.1.4 硬件在环测试工具

硬件在环测试工具允许车用人工智能开发人员在实际的硬件上测试和验证其控制策略和算法，而不需要在真实环境中进行成本高昂且可能危险的实车测试。HIL 测试通过模拟外部环境（如车辆传感器输入、车辆动态、通信网络等）与实际的硬件（如电子控制单元 ECU 或整个车辆系统）交互，以评估系统的响应和性能。车用人工智能领域常用的 HIL 测试工具有 dSPACE SCALEXIO、NI VeriStand、Vector VT System 等。这些工具通过模拟外部物理系统如车辆动力学、环境条件和传感器反馈，并与真实硬件进行交互，以评估系统的响应和性能。通常采用模块化、可扩展的结构，支持多种汽车通信协议和接口，如 CAN、LIN、FlexRay、Ethernet 等，以适应不同的测试需求和技术发展。HIL 测试平台还具备故障注入功能和测试自动化能力，使得可以模拟各种故障场景并自动执行大量测试案例，大大提高了测试效率。

8.4.1.5 车辆在环测试工具

车辆在环测试工具测试人员创建可准确代表真实情境的模拟环境，包括交通、道路标志、道路标记等。测试设备将模拟环境呈现给真实车辆，传感器的输入来自模拟环境。该方法降低实际道路测试的风险性和难度，同时减少测试成本和时间。验证停车辅助、变道、紧急制动、规避转向等功能，以及自动驾驶功能。并提供了安全的环境执行高风险操作，而不必担心发生碰撞。填补了软件或硬件模拟与完整道路测试之间的空白。在一个封闭场地内实现任意开放道路的集成测试，

汽标委智能网联汽车分委会

满足了智能驾驶车辆在环测试系统的需求。

整车在环测试系统用于对自动驾驶规划决策控制器与底盘系统，在真车环境进行特定工况、极端工况下的测试验证和评价。通过构建基于整车的在环仿真测试系统，充分结合虚拟仿真测试与实车测试的优势，通过仿真软件生成虚拟测试场景，将虚拟传感器探测到的场景信息注入给算法控制器中，进而测试实际车辆在仿真场景中的运动状态。一般而言整车在环测试系统可分为数字孪生整车在环测试系统与轴耦合式整车在环测试系统。

(1) 数字孪生整车在环测试系统

基于仿真软件生成所需开发测试的虚拟场景，并在仿真软件中建立多种虚拟传感器模型，将虚拟摄像头检测到的仿真视频信息通过视频注入板卡注入、虚拟毫米波传感器探测到的目标列表信息通过汽车 CAN 总线的形式注入到实车上的 ADAS/AD 控制器中进行信息融合与控制决策，ADAS/AD 控制器将决策后的加速、制动以及转向指令发送给实车执行器，再通过组合惯导将实车车身姿态和车辆位置信息反馈给虚拟场景来完成实车在仿真场景中的姿态位置映射，进而实现整个系统的闭环实时仿真。

图 8-9 数字孪生整车在环测试实现效果图

此系统不仅具备虚拟场景多样化、安全和可重复性高的优势，还通过在虚拟交通环境中加入真实汽车的方式来加强仿真技术的真实性。系统架构如图 8-10 所示。

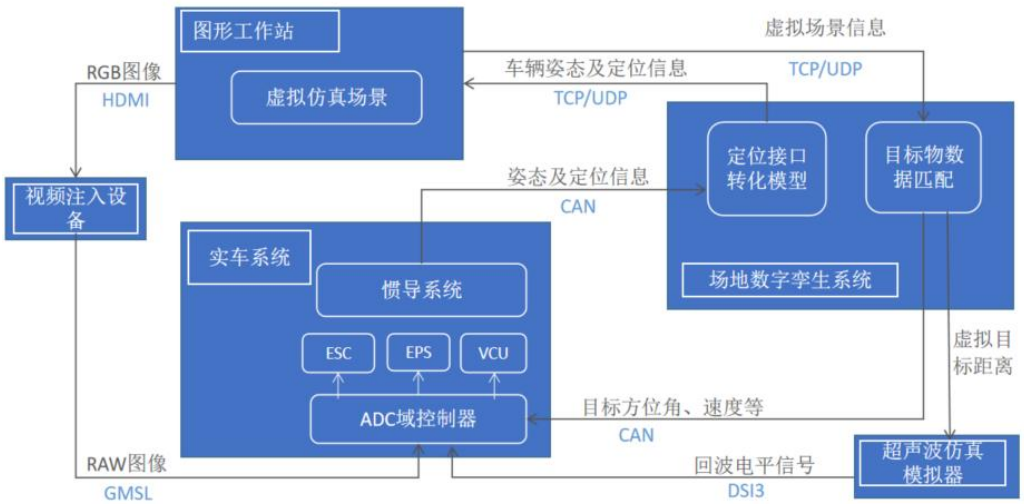


图 8-10 数字孪生整车在环测试系统架构

(2) 轴耦合式整车在环测试系统

类似于数字孪生整车在环测试系统，基于轴耦合式的整车在环测试系统同样借助仿真软件生成开发和测试所需的虚拟场景，并在仿真软件中建立多种虚拟传感器模型，通过总线注入、目标注入和视频注入等方式将传感器数据发送至 ADAS/ADS 控制器，控制器经过决策后，将加速、制动以及转向等指令发送给实车执行器，再通过测功机将得到车辆的实际扭矩、车轮转角和角速度实时反馈给虚拟场景，在仿真场景中完成车辆状态映射，进而实现整个系统的闭环实时仿真。

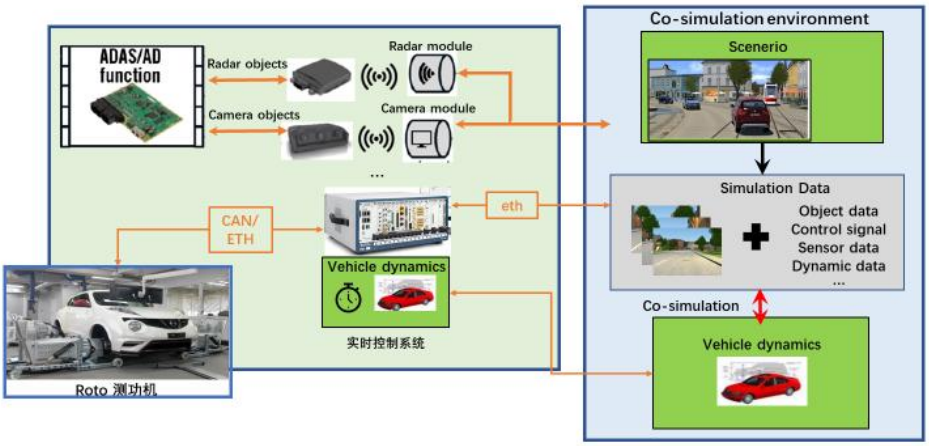


图 8-11 轴耦合式整车在环测试系统总体架构

这项方案不仅具备了虚拟场景多样化、安全和可重复性高的优势，同时还通

过在虚拟交通环境中加入真实汽车的方式来加强了仿真技术的真实性。该系统可实现智能网联汽车 ADAS 系统的高精度相关仿真测试，车辆动力加载、转向模拟等功能的仿真测试，以及应用于智能网联实车环境下的动力部件测试等。根据所采用传感器仿真方案的不同，注入原始级/信号级/目标级等多层级感知数据，可实现不同层级的感知端测试，以及规控决策算法的测试验证，最终实现自动驾驶功能优化工作的效率提升。

通过整车在环，还可以添加外部干扰要素，如图 8-12 所示将 VUT 安装在测试转毂上运行时，激活自动驾驶功能。然后进行电磁干扰测试，记录和观察 VUT 的自动驾驶功能异常或故障。VUT 的响应观察可以通过测试轮毂控制器上的车轮速度，以及通过摄像头对仪表盘上的刹车灯及其它任何指示灯的监控。

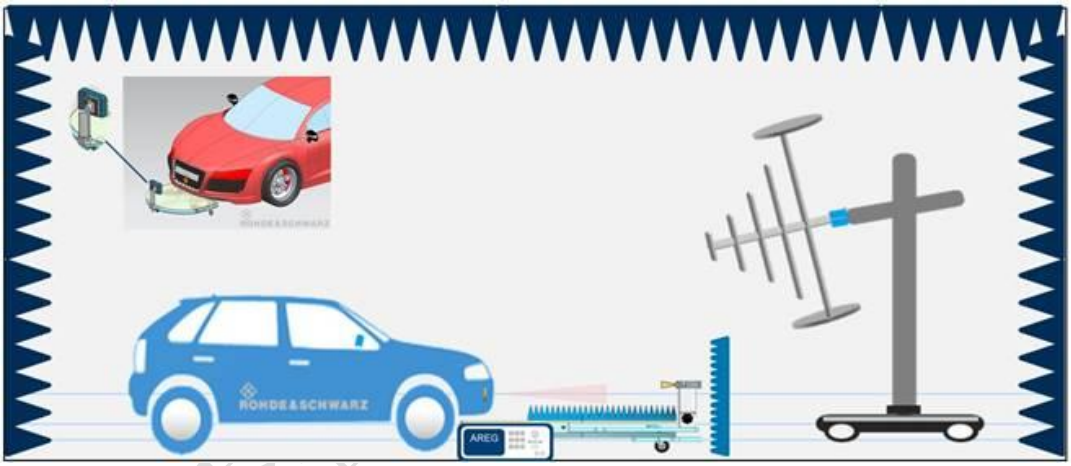


图 8-12 电磁干扰环境下验证基于毫米波雷达的自动驾驶功能

8.4.2 封闭场地测试工具

封闭场地测试需要构建各种模拟真实交通环境的场景，包括城市道路、高速公路、乡村道路、交叉口、坡道、隧道等。这些场景应覆盖不同的交通状况和环境条件，如不同的天气状况（雨、雪、雾等）。在日常环境下，恶劣气象条件发生的概率有限，在封闭环境中可以模拟各种天气条件下的路面情况、交通状况等，可使得测试更加全面和细致。为了进行雨雾等特殊天气下的自动驾驶场景测试，通过在综合试验场内搭气象模拟通道，可以研究自然界气象环境（如雨、雾等）的物理表征形式以及不同气象环境对智能汽车自动驾驶功能和性能的影响机制。

如图 8-13 所示为中汽研汽车试验场股份有限公司封闭场地气象环境模拟封闭场地测试环境，该试验场基于雨雾阵列技术，建立一种封闭场地环境下的气象多级可调模拟方法。雨雾通道建设规模总长 200 米，宽 8 米（双车道），支持降雨强度 10~100mm/h 三级可调，降雾能见度 15~80 米。

a) 气象模拟通道

b) 雨天遇行人测试

c）雨天遇静止车辆测试

d）雾天遇静止车辆测试

图 8-13 中汽研汽车试验场股份有限公司封闭场地气象环境模拟

8.4.3 实际道路测试工具

现阶段，道路测试主要是通过软件结合硬件的方式，按照设计好的路线在全国范围内进行被测系统可靠性和稳定性的开放道路测试。通过主观和客观两种方式，记录系统误报、漏报等问题事件。离线回放中，同步还原并分析事件发生时间段内的车辆总线、视频图像、音频信息、GPS 定位等数据，为系统开发人员进行问题定位和系统升级提供依据。

系统主要组成如下图所示：

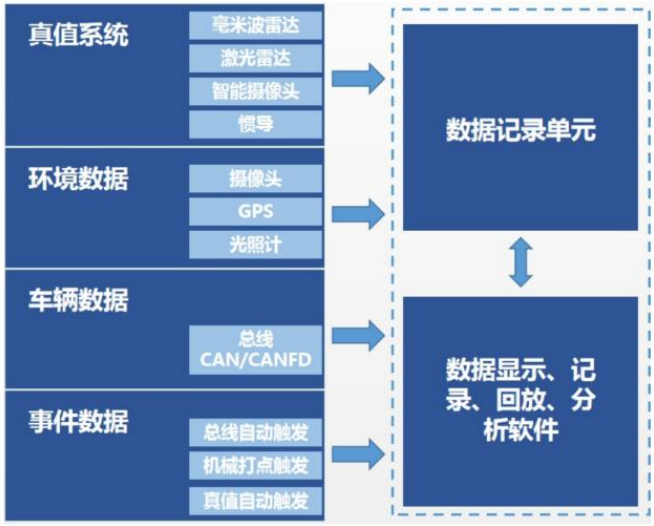


图 8-14 道路测试系统组成

从测试设备上讲，实际道路测试设备大体上可以分为以下几大类：数据采集系统，采集与分析软件，组合导航系统，真值系统，驾驶员监控系统等。

8.5 总结与展望

8.5.1 国内外法规概况

8.5.1.1 国际法规与政策导向

全球范围内，车用 AI 测试评价标准正呈现“欧盟引领、多极协同”的发展态势。欧盟于 2024 年正式通过《人工智能法案》[5]，将 AI 系统按风险等级分为四类（不可接受风险、高风险、有限风险、最小风险），其中自动驾驶汽车被归类为“高风险 AI 系统”，适用更严格的开发、测试与透明性要求。与此同时，欧盟明确提出将 AI 监管要求纳入整车法规框架，与《欧盟整车型式核准框架法规》《通用产品安全法规》[6][7]等联动实施，对车载 AI 系统的安全与可控性提出统一技术要求；联合国世界车辆法规协调论坛（WP.29）陆续颁布《UN R155》《UN R156》，要求整车企业建立覆盖产品全生命周期的网络安全管理体系（CSMS）和软件升级管理体系（SUMS），并明确 OTA 流程的认证与风险管控机制。2024 年，WP.29 新发布《UN R171 驾驶员控制辅助系统》法规，对 L2 级辅助驾驶功能提出驾驶员状态监测和警示机制要求，强化其安全边界识别与人

汽标委智能网联汽车分委会

机责任划分。此外，WP.29 已启动 L3 及以上 ADS 法规制定计划，并发布 VMAD/FRAV 工作组起草的 ADS 安全评估与测试方法框架，为后续法规统一奠定技术基础。

美国方面自 2020 年以来，已先后出台《国家人工智能倡议法案》[8]等多项指导性政策，并于 2024 年签署《AB 2655》《AB 2839》《AB 2355》《AB 2602》《AB 1836》五项 AI 相关法律，以提升人工智能应用透明度与公众信任。同时，美国交通部和国家公路交通安全管理局（NHTSA）发布的《联邦自动驾驶汽车政策》，作为美国联邦层面首个系统指导自动驾驶汽车发展的政策文件，提出自动驾驶汽车的系统安全目标、测试评估要求与数据共享机制，强调保障公众安全、技术可靠性和责任清晰度。政策在明确监管方向、推动技术规范统一的同时，也引发了关于数据共享、企业隐私保护与联邦—州权力分配的广泛讨论[9]。

日本在 2019 年提出修订《道路交通法》，正式允许 L3 级自动驾驶汽车在特定条件下上路，并对其运行速度、传感器探测范围、数据记录装置及驾驶员责任等提出明确法律要求，为自动驾驶车辆的法定使用设定了基本门槛。同时，《道路运输车辆法》修正案引入了自动驾驶系统的更新许可制度，允许通过 OTA 对系统功能进行升级，并规范了相关硬件安装与维护责任。此外，2018 年国土交通省发布的《自动驾驶汽车安全技术指南》作为法规性技术指导文件，从 ODD 设定、人机交互、网络安全、数据记录、运行安全评估等方面提出了系统性安全要求，成为车用 AI 功能合规性测试的法律依据。这一系列法规为日本推动车用 AI 的实际应用和安全管理奠定了基础[10]。

8.5.1.2 国内法规与政策导向

中国智能网联汽车产业正经历从“野蛮生长”向“高质量发展”的关键转型期，国家通过多层次法规政策构建起涵盖测试验证、准入管理、数据安全和责任认定的完整监管框架。2025 年 2 月，工业和信息化部、市场监管总局联合发布《关于进一步加强智能网联汽车产品准入、召回及软件在线升级管理的通知》[11]，对搭载 L2 级辅助驾驶系统及具备 OTA 功能的车型提出更严格的合规要求。同

汽标委智能网联汽车分委会

时，建立“安全沙盒”机制，要求企业提交辅助驾驶功能深入测试方案和风险评估报告；强化事故通报机制，要求厂商及时上报与智能驾驶相关的故障和事故数据；并将 OTA 行为分为改/不改主要参数两类实施分类管理，防止逃避监管。

2024 年 1 月，工信部、公安部、交通运输部等五部委联合印发《关于开展智能网联汽车“车路云一体化”应用试点工作的通知》[12]，明确在全国开展车路协同与智能网联汽车融合应用试点，提出“构建统一测试评价体系”的任务要求，包括场景数据库建设、信息交互标准制定与仿真验证能力提升等内容，推动形成车用 AI 测试标准、评估流程与应用数据闭环的联动机制。

法规制度建设方面，《道路交通安全法》拟将自动驾驶汽车纳入法定交通管理体系，补充道路测试许可、事故处理、交通违法责任等关键条款，为自动驾驶上路和责任界定提供法律依据[13]。深圳正式施行《深圳经济特区智能网联汽车管理条例》[14]，成为国内首个涵盖自动驾驶测试、登记、运行、车路协同及交通管理的综合性法规，首次提出“高度自动驾驶”与“完全自动驾驶”可在限定区域开展商业化运营，明确责任主体与数据安全要求。北京市人大则通过《北京市自动驾驶汽车条例》[15]提出建立“包容审慎监管”机制，明确测试运营准入流程、安全责任划分与数据接入要求，为无人化示范应用提供法治支撑。此外，上海、重庆、广州、武汉等地已发布配套道路测试管理办法或开展准入试点，探索基于不同自动驾驶等级的差异化测试和运营监管策略，逐步形成多地联动、中央统筹的政策执行格局。

8.5.2 国内外标准概况

8.5.2.1 国际标准

全球车用 AI 测试评价标准呈现“核心框架趋同、区域特色并存”的发展态势，由 ISO、SAE、IEEE、UNECE 等组织主导，涵盖功能安全、预期功能安全、AI 特有风险、以及自动驾驶测试与评估等关键领域。

在 AI 安全标准方面，2024 年 12 月，ISO 正式发布 ISO/PAS 8800:2024《道路车辆—安全和人工智能》[16]，由 ISO/TC22/SC32/WG14 工作组牵头制定。该

汽标委智能网联汽车分委会

标准是首个专门面向汽车 AI 系统的国际安全框架，重点提出 AI 安全生命周期管理、数据驱动的风险识别与验证机制，补充了 ISO 26262 和 ISO 21448 无法覆盖的 AI“黑箱性”问题[17]，是当前 AI 车载化测试的重要指导标准。

在管理体系方面，2023 年 12 月发布的 ISO/IEC 42001《人工智能管理体系》[18]为 AI 系统的开发、部署与使用提供了组织层级的管理框架，强调风险评估、责任分配与技术伦理，适用于构建 AI 产品全生命周期合规体系。

在自动驾驶系统分级方面，SAE J3016 标准定义了 L0 至 L5 六级自动驾驶系统，被全球广泛采纳为分级基准[19]。SAE J3018 规范自动驾驶实路测试流程，包括测试人员资质、应急机制等内容，是测试安全保障的重要补充。

在自动驾驶测试评价标准方面，ISO 3450x 系列标准正构建覆盖术语、场景分类、安全评估与测试用例生成的完整体系[20]。其中，中国主导编制的 ISO 34501:2022 统一了自动驾驶系统测试场景中的核心术语，成为国际自动驾驶测试语言的基础标准；已发布的 ISO 34502 提出了基于场景库的安全评估框架，ISO 34504 则定义了自动驾驶场景的分类方法[21]；ISO 34503 和 ISO 34505 将进一步补全标准链条，共同构成覆盖自动驾驶系统全生命周期的测试评价方法体系。此外，UL 4600（美国保险实验室标准）提供自动驾驶系统完整安全性评估指南[22]，IEEE 2846 规范自动驾驶决策逻辑的安全假设验证，IEEE 2851/2852 定义了自动驾驶系统性能指标和验证框架[23]。

这些标准体系共同构成车用 AI 从系统开发到测试评价的完整规范路径，既支撑产品合规，也为各国法规落地提供技术基础。

8.5.2.2 国内标准

中国在车用 AI 测试评价标准化领域已实现从“国际对标”到“自主引领”的跨越式发展，逐步形成基础术语、管理要求、功能安全、测试方法与产品规范多层次协同的标准生态。

在基础标准方面，2021 年发布的 GB/T 40429-2021《汽车驾驶自动化分级》明确了 L0－L5 级别划分，是与 SAE J3016 对齐的国家推荐性标准，为功能分类

汽标委智能网联汽车分委会

和测试分层提供术语支撑[24]。

在安全与管理标准方面，GB/T 34590《道路车辆 功能安全》[25]对标 ISO 26262，构建了功能安全管理要求体系；工业和信息化部组织制定的《汽车整车信息安全技术要求》《汽车软件升级通用技术要求》《智能网联汽车 自动驾驶数据记录系统》三项强制性国家标准由国家市场监督管理总局、国家标准化管理委员会批准发布，将于 2026 年 1 月 1 日起开始实施。这些强制标准参考了《UN R155》《UN R156》核心内容，并结合我国监管与技术实践，首次对车载 AI 系统的运行安全、软件更新与数据采集提出国家级刚性要求，对提升智能网联汽车安全水平、保障产业健康持续发展具有重要意义[26]。

在自动驾驶测试方法标准方面，在全国汽车标准化技术委员会的归口管理下，联合相关领域单位，已编制形成一系列自动驾驶测试方法标准。其中，国家标准包括《智能网联汽车模拟仿真测试规范》《智能网联汽车自动驾驶功能场地试验方法及要求》《智能网联汽车运行安全测试技术要求》等，覆盖仿真、场地和运行测试等关键环节[27][28][29]；团体标准则包括《智能网联汽车仿真测试场景库场景分类》《智能网联汽车城市道路场景无人化测试场地试验方法及要求》《智能网联汽车自动驾驶系统测试场景自然语言描述方法标准化需求》等，进一步完善了测试场景构建与表达方式的技术体系。同时，中国积极主导国际标准制定并加快国内转化进程，在 ISO 34501 的基础上积极开展本地化适配与扩展，立项制定《道路车辆 自动驾驶系统测试场景 术语》，为构建统一、规范的自动驾驶测试语言体系奠定基础。

此外，为推动标准落地实施，相关监管文件已明确将部分关键标准纳入产品认证与准入审查范围，推动标准在实际应用中的强制执行，逐步建立起覆盖研发验证、准入审查与事后监管的“标准—测试—监管”闭环体系[30]。

汽标委智能网联汽车分委会

第九章 车用 AI 数据安全测试与评估标准化研究

9.1 车用 AI 数据安全测试与评估现状

9.1.1 研究背景

智能网联汽车基于 AI 技术提升驾驶安全与体验，其核心在于通过多类传感器（摄像头 / 雷达 / 激光雷达等）和车载计算平台实现环境感知与决策。随着 AIGC 技术成为生成高质量训练数据的常用手段，车用 AI 系统对多源数据的依赖程度持续提升，涵盖车内外影像、定位信息、V2X 通信及用户行为数据等。

当前技术发展面临双重挑战：既要保障实时数据处理能力以支撑自动驾驶功能，又需构建多维安全防护体系。数据安全风险涉及隐私泄露、数据篡改及系统攻击等，直接影响行车安全与用户权益。研究车用 AI 数据安全需从技术创新、隐私保护和法律合规三个维度协同推进，在确保算法效能的同时构建可信安全边界，这对智能交通系统的可持续发展具有重要现实意义。

9.1.2 研究目的及意义

车用 AI 数据安全测试与评估的研究目的及意义主要体现在以下几个方面：

(1) 确保数据安全与隐私保护：随着汽车行业向智能化、网络化发展，车辆收集和处理的个人数据量急剧增加。这些数据往往涉及用户的个人隐私，通过对智能网联汽车人工智能进行数据安全测试与评估，可以确保这些数据的安全，防止数据泄露、滥用等，保护用户隐私。

(2) 确保车辆系统可靠性与安全性：随着人工智能技术的发展，攻击者可能利用人工智能的弱点进行攻击。研究数据安全测试方法、增强智能网联汽车人工智能的安全性，可以防止恶意攻击导致的系统故障或误操作，确保行车安全。

(3) 促进人工智能技术的健康发展：对人工智能技术在汽车领域进行有效的数据安全测试与评估，有助于推动人工智能技术在汽车领域的健康、可持续发展。

汽标委智能网联汽车分委会

(4) 应对法律法规要求：通过研究并实施严格的数据安全测试与评估，有助于车企符合相关法律法规，避免因数据安全问题而引发的法律风险。

9.1.3 车用 AI 场景及数据安全需求

AI 在汽车行业已拥有丰富的应用场景，涵盖自动驾驶、智能交互、智能出行等场景。

9.1.3.1 自动驾驶场景

AI 在自动驾驶中的应用涵盖了从感知到决策、从控制到通信的各个方面。基于大模型，可实现汽车自主感知周围环境、识别道路和交通标志，实现自动巡航、自动变道、停车和避障等功能：

(1) 环境感知与理解：基于 AI 深度学习算法和模型的视觉感知方案、激光雷达感知方案，通过传感器感知周围环境实时动态环境信息，通过构建三维环境模型，自动驾驶汽车能够准确理解周围环境的布局和动态变化。

(2) 智能决策与规划：在获得环境信息后，自动驾驶汽车需要基于这些信息做出决策，AI 系统通过结合交通规则、道路状况以及实时交通信息，为车辆规划出安全、高效的行驶路径。当面对复杂的交通状况，AI 系统能够实时分析并做出准确的驾驶决策。

9.1.3.2 智能交互场景

车用 AI 可实现与驾驶员的智能对话，提供准确的控车、娱乐和问答服务。传统的单指令、多意图语义理解基于特定语义结构实现，无法识别口语化、带语气助词等区别于特定语义结构的自由语音输入。智能座舱依托大模型技术实现多模态交互升级：语音系统突破传统指令限制，通过语义理解增强技术精准执行控车、导航等操作。视觉感知模块通过车内摄像头实时监测驾驶员状态，当检测到疲劳或分心时自动触发警示并启动陪聊模式，结合实时资讯进行互动唤醒。同时基于微表情分析识别乘客情绪，动态调整音乐、氛围灯等配置。个性化服务系统

汽标委智能网联汽车分委会

通过行为数据构建用户画像，自动推荐音乐、影视等内容，并利用生成式 AI 能力提将车内空间拓展为融合娱乐、办公、生活服务的智能终端。

基于生成式 AI 大模型，使智能座舱具备了多种生成式 AI 能力，可实现生成式语音回复基础上叠加音色、方言、人设等设置，为用户提供更多情感化交互感知，并能提供包括情绪感知、智能助手、深度对话、创意内容生成等多维度能力。

9.1.3.3 智能出行场景

基于大模型，可提供行程推荐、出行问题解决、任务提醒及固定任务安排等服务，能够极大地提升驾驶者的出行效率和生活便利。

行程规划与推荐：驾驶者输入目的地、旅行天数等信息后，可自动规划包含热门景点、特色美食、住宿等内容的详细行程；并可根据驾驶者的喜欢历史、自然风光等兴趣爱好，推荐更符合用户兴趣的行程安排。

出行问题解决：针对驾驶者身份证丢失等紧急情况，可基于当地情况推荐最近的补办地点及详细流程指导。

任务提醒与固定任务安排：根据驾驶者设定的会议提醒、生日祝福等任务，自动在指定时间进行相关内容提醒；当用户进入上班开车等特定场景时，可自动触发开启导航、播放音乐、调整车内温度等相关任务，提升用户体验。

9.1.3.4 数据安全需求

AI 系统高度依赖数据获取的正确性，数据集质量、算法训练模型鲁棒性、生成式 AI 数据安全都可能影响 AI 的安全。数据集的质量取决于数据集的规模、均衡性、准确性等，有多种原因使得获取的数据质量低下，都会对 AI 系统形成严重的干扰，直接影响到 AI 算法的执行效果；数据投毒是指人为地对数据进行修改来改变 AI 算法的执行效果，即通过输入不正确的样本数据， AI 算法训练得到不正确的模型参数，从而引发算法错误；对抗样本是指在不改变模型参数的情况下，对 AI 算法需要识别的数据加以修改，让算法失效，是对 AI 算法识别

汽车智能网联专分委会

过程的攻击。因此采取措施防止数据被恶意修改或破坏，采取措施防止数据泄露和滥用，对 AI 在智能网联汽车上应用至关重要。

9.1.4 现状和发展

9.1.4.1 法规政策方面

在全球范围内，各国政府正逐步制定和完善相关法规政策，以确保智能网联汽车人工智能的数据安全。2024 年 3 月 13 日，欧盟议会通过了《人工智能法案》，这是全球首部人工智能（AI）法案，对“基础模型”或经过大量数据训练的 AI 系统施加了披露生成式 AI 训练数据版权和风险评估要求。美国的监管重点在于推动自动驾驶技术的安全测试和部署，NHTSA 的准则强调了系统的可控性和数据安全。在 AI 方面，美国颁布了《关于安全、可靠和可信的 AI 行政命令》该行政令强调了人工智能系统的安全性和可信度，美国务院发布的《人工智能战略》中也强调保护数据安全的重要性。日本政府颁布《人工智能指导方针》，该方针旨在为人工智能企业的研发与应用提供参考，强调在实现“以人类为中心的人工智能社会”愿景的过程中，必须重视数据安全与隐私保护。这些政策为未来车用 AI 的开发、部署和使用提供了重要的法律保障，同时也推动了相关企业在技术层面的改进。以保障智能网联汽车人工智能的数据安全。

我国政府也在积极统筹智能网联汽车产业发展与数据安全保护。2020 年 8 月，国家标准委、中央网信办等五部委联合印发《国家新一代人工智能标准体系建设指南》，旨在加强人工智能领域的顶层设计，推动技术研发和标准制定，促进产业的健康可持续发展。2021 年 7 月，国家发展和改革委员会、工业和信息化部等五部门联合发布《汽车数据安全管理若干规定（试行）》明确了汽车重要数据范围，为汽车领域个人信息保护工作明确了基本原则，以及倡导汽车数据处理者在开展汽车数据处理活动中坚持车内处理、默认不收集、精度适用、脱敏处理的原则。2024 年 1 月，工信部发布《国家人工智能产业综合标准化体系建设指南》（征求意见稿），进一步加强人工智能标准化工作系统谋划，加快构建满

汽标委智能网联汽车分委会

足人工智能产业高质量发展需求的标准体系，更好发挥标准对推动技术进步、促进企业发展、引领产业升级、保障产业安全的支撑作用。

9.1.4.2 技术研发方面

智能网联汽车 AI 数据的引入带来了新的数据安全挑战，高级持续性威胁（APT）和零日攻击等复杂攻击手段难以检测和防御。各大汽车制造商、科技公司和科研院所都在持续加大对智能网联汽车人工智能数据安全的投入。新的数据安全测试方法和技术逐渐被引入，以便更全面地检测和发现潜在的安全风险。

静态应用程序安全测试（SAST）：通过代码走读的方式对源代码的安全性进行测试的方法。通过数据流、控制流、信息流等测试方法与安全规则库进行匹配，进而发现潜在的安全漏洞。该方法主要在编码阶段进行，旨在尽可能早地发现安全性问题。

动态应用程序安全测试（DAST）：借助工具或手工模拟黑客的输入，对应用程序进行安全性测试，以发现系统中的安全性问题。该方法一般在系统测试阶段进行，虽然覆盖率较低，但可以通过提供更多的测试数据来达到较高的覆盖率。

扫描程序中的数据：在程序运行过程中对数据进行测试，以发现诸如缓冲区溢出之类的漏洞。这种方法主要对内存进行测试，是静态应用程序安全测试和动态应用程序安全测试很难测试到的部分。

网络安全性测试：涉及防护措施的正确装配、非授权攻击检查、网络漏洞工具使用等，以确保网络系统的安全防护策略正确实施。

数据库安全性测试：包括数据库备份和恢复功能、数据加密、安全日志文件记录等，以确保数据库的安全性和可靠性。

软件组成分析（SCA）：对生成式 AI 模型的源代码和依赖关系进行分析，确定其中包含的组件和第三方库信息，检测模型是否不存在侵犯知识产权的情况。

汽标委智能网联汽车分委会

9.1.4.3 标准制定方面

为了规范能车用 AI 数据安全测试与评估，国际电信联盟（ITU）、国际标准化组织（ISO）和国际电工委员会（IEC）等组织正在制定相关标准，主要集中在基础术语、参考架构、数据治理、系统设计、性能评估等方面。国际标准组织（ISO）中的 ISO/IEC JTC1 SC42 人工智能分技术委员会开展人工智能领域的标准化工作，主要聚焦在人工智能管理、可信赖、安全与隐私保护等方面，详细如下表所示。

表 9-1 ISO 数据相关标准列表

序号	标准名称	主要内容
1	ISO/IEC 8183:2023 《信息技术 人工智能 数据生命周期框架》	描述了一个全面的框架，涵盖数据管理的各个方面，从数据采集和创建到开发、部署、维护和最终退出，通过强调数据处理在人工智能系统生命周期各个阶段的关键作用，该国际标准将有助于加强系统管理、数据质量、数据安全和系统实用性，确保企业与行业最佳实践保持一致它还能帮助企业通过提高数据的可访问性、可用性和安全性，最大限度地发挥数据的价值。
2	ISO/AWI PAS 8800：2023《道路车辆 人工智能安全》	明确了数据（集）在人工智能系统开发的整个生命周期起到至关重要的作用，尤其是数据（集）安全分析阶段；确定了与数据相关的安全属性包括完整性、可追溯性、可验证性、可用性、时间性、相关性和独立性；约束了基于过程失效模式和影响分析（PFMEA）等方法，研究关于数据安全风险的风险决策，以及可用于减轻/预防数

汽标委智能网联汽车分委会

		据安全风险的决策类型、程度和措施，最终满足数据相关的安全特性和/或智能网联汽车人工智能系统的安全要求。
3	ISO／IEC 24668-2022《信息技术 人工智能 大数据分析的流程管理框架》	提供了一个开发流程的框架，以便在整个组织中有效利用大数据分析，流程类别包括组织利益相关者流程、能力开发流程、数据管理流程、分析开发流程和技术集成流程。
4	ISO／IEC 20547-3：2020《信息技术 大数据参考架构 第 3 部分 参考架构》	规定了大数据参考架构(BDRA)，参考体系结构包括概念和体系结构视图。
5	ISO／IEC 42001：2023《人工智能管理体系》	采用类似于 ISO 9001 质量管理体系标准和 ISO/IEC 27001 信息安全管理体系标准的方法,提供了人工智能管理风险和操作方面的最佳实践、规则、定义和指导。
6	ISO/IEC 5259-1《人工智能 分析和机器学习（ML）的数据质量 第 1 部分：概述、术语和示例》	涵盖了数据质量管理的基本概念、关键术语和具体示例，定义了数据质量的维度和指标，提供了评估和提升数据质量的方法。
7	ISO/IEC 5259-2《人工智能 分析和机器学习（ML）的数据质量 第 2 部分：数据质量措施》	介绍了数据质量的测量方法和指标，包括数据完整性、一致性、准确性、及时性、相关性和可解释性等方面。
8	ISO/IEC 5259-3：2024《人工智能分析和机器学习（ML）的数据质量第 3 部分：数据质量管理要求和指南》	规定了数据质量管理的要求和指南，涵盖数据治理框架、 质量控制流程、数据评估方法、质量改进策略和监控机制。

汽标委智能网联汽车分委会

9	ISO/IEC 5259-4：2024《人工智能分析和机器学习（ML）的数据质量第 4 部分：数据质量过程框架》	定义了数据质量管理的过程框架，包括数据质量规划、数据采集、数据处理、 数据存储和数据验证的流程。它详细描述了每个阶段的活动、角色和责任， 以及质量控制和改进措施。

欧洲电信标准化协会（ETSI）发布了人工智能安全研究报告，阐述了人工智能安全问题和缓解措施，美国电气电子工程师学会（IEEE）也在汽车人工智能领域制定了多项标准，涵盖自动驾驶决策系统、安全验证、通信架构及数据安全等方面。

国内针对汽车数据安全也陆续出台相关标准，20214342-T-469《信息安全技术 汽车采集数据的安全要求》定义了车外数据、座舱数据、运行数据和位置轨迹数据四类数据，对上述数据在传输、存储、出境方面提出了具体要求； GB/T 44464-2024《汽车数据通用要求》对汽车数据的分类分级提出了具体要求。此外车联网数据安全评测与评估相关标准包括《车联网数据跨境流动安全评估规范》（待制定）、2018-0183T-YD《基于移动互联网的汽车用户数据应用与保护评估方法》、《车联网用户个人信息合规检测要求》（待制定）、2020-1317T-YD《车联网信息服务 数据安全保护能力评估规范》。在国家人工智能标准化总体组全国信标委人工智能分委会发布的《人工智能伦理治理标准化指南》中，人工智能伦理治理标准体系包含了基础共性、治理技术、管理以及行业应用四个部分，在行业应用中自动驾驶作为重要的一个行业应用，其中规划了《人工智能 自动驾驶系统仿真测试平台技术要求》、《人工智能 可信赖规范 第 6 部分：自动驾驶》两项标准。

当前针对 AI 数据安全测试与评估的标准尚少，而随着人工智能在智能网联汽车领域应用的逐步深入，智能网联汽车 AI 数据安全测试与评估相关标准的提出与讨论，有助于提升智能网联汽车人工智能数据安全的水平，提升人工智能数据安全管理水平和技术能力，加速智能网联汽车产业智能化转型。

汽标委智能网联汽车分委会

9.1.4.4 产业合作方面

智能网联汽车 AI 技术的创新，尤其是大模型的应用，逐渐成为企业产品力、生命力的重要组成部分，同时也带来了新的数据安全风险，成为智能网联汽车发展道路上的核心问题。为了共同应对智能网联汽车 AI 数据安全挑战，各大汽车制造商、科技公司和研究机构正积极开展多层面，多参与者，涉及技术、管理、法规等多个方面的产业合作，例如建立健全企业数据管理制度、搭建汽车安全监测平台，建立覆盖车端、云端，通信管道，以及移动端等全链条的信息安全体系，加大数据安全方面投入，逐步研究数据合规的技术方案等。通过企业间的资源共享、技术交流和合作研发，提前进行防范和布局，实现高效的数据流通与保护，共同提高智能网联汽车 AI 数据安全水平，构建起强大的数据安全防线，推动汽车行业的健康可持续发展[1]。

9.2 车用 AI 数据安全总体要求

9.2.1 车用 AI 数据安全测试评估基本原则

车用 AI 数据安全测试与评估的过程需要遵循全面性、合规性、专业性、透明性、客观性、持续性六大原则。这些原则是确保能够有效识别和管理数据安全风险的基础，同时也是满足法律法规要求、保护数据主体权益和提升数据安全管理水平的重要保障。在进行数据安全测试与评估时，应当严格遵循上述原则，以确保评估工作的质量和效果。需要遵循的原则包括但不限于以下几点：

(1) 全面性：确保评估涵盖车用 AI 数据的全生命周期，包括收集、存储、使用、加工、传输、提供和公开等各个环节。通过全面评估，掌握车用 AI 数据安全的总体状况，发现潜在的安全隐患。基于评估结果，提出车用 AI 数据安全管理和技术防护建议，以提升数据的安全性。

(2) 合规性：遵守相关法律法规的要求，如《中华人民共和国数据安全法》中对数据处理活动的规定。此外，应遵循国家或行业标准进行评估，如工业领域数据安全风险评估标准等，以保证评估的统一性和比较性。在发现车用 AI

汽标委智能网联汽车分委会

数据安全缺陷或漏洞时，应立即采取补救措施；发生数据安全事件时，应立即处置并按规定报告。

(3) 专业性：车用 AI 数据安全的测试和评估应由具备相应专业知识和经验的人员或机构进行，全面地覆盖所有相关的风险点，准确地识别安全问题，提出有效的解决方案，以确保评估的专业性和有效性。

(4) 透明性：评估过程和结果应公开透明，让利益相关者了解车用 AI 数据安全状况和采取相应措施。透明性有助于明确数据的所有权和使用权，促进数据资源的合理配置和市场化运作。

(5) 客观性：评估应基于事实和数据，避免主观偏见，确保评估结果的准确性和可靠性，增加组织内部和外部利益相关者（如客户、合作伙伴、监管机构）信任，增加评估公信力。

(6) 持续性：车用 AI 数据安全风险评估是一个持续的过程，需要定期开展，以适应数据环境和威胁情况的变化。

综上，遵循车用 AI 数据安全测试与评估的原则对于组织的运营、法律遵从性、风险管理以及持续改进等方面都至关重要。这些原则为组织提供了一个框架，以确保智能网联汽车 AI 数据安全，同时促进了数据的有效利用和保护。

9.2.2 车用 AI 数据安全测试评估对象和范围

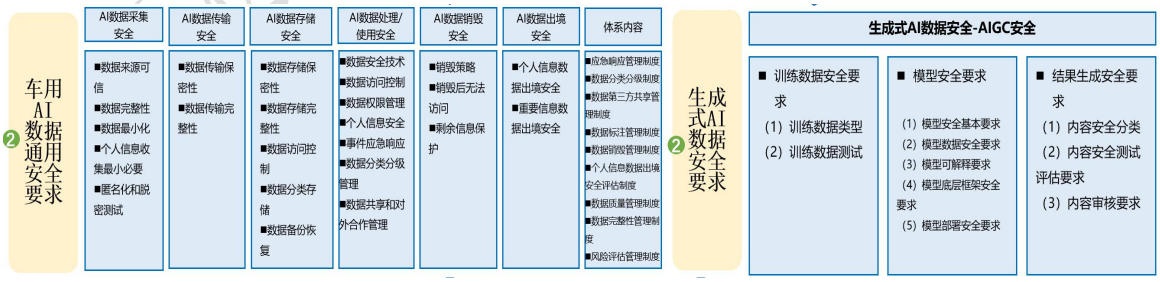


图 9-2 智能网联汽车 AI 数据安全测试评估对象和范围

AI数据通用安全测试评估要求包括AI数据全生命周期以及体系评估两个部分，具体包括AI数据采集、AI数据传输、AI数据存储、AI数据处理、AI数据销毁、AI数据出境和体系评估内容。

AI数据采集：包括智能网联汽车在运行过程中产生的各种可以被用于车用AI系统的数据，如车辆行踪轨迹、音频，图像和生物识别特征等个人信息，以及车辆流量、交通等数据。确保AI数据在采集过程中的安全性。

AI数据传输：在智能网联汽车中，在使用部署在云端的智能网联汽车人工智能模块时，AI数据会在云端与车辆之间进行数据传输，在AI数据传输时应注意数据传输安全，确保车辆网络中的AI数据安全。

AI数据存储：在智能网联汽车中，AI数据可能存储在本地车辆的存储设备上，或者上传到云端服务器进行存储。存储阶段需要考虑AI数据的安全性、可访问性和备份策略。

AI数据处理：智能网联汽车在处理采集到的AI数据时，应当通过AI数据安全技术以及数据访问控制，权限管理等技术确保AI数据处理的安全性。AI数据处理过程应当遵守《汽车数据安全管理若干规定》，《中华人民共和国数据安全法》、《汽车采集数据处理安全指南》（TC260-001）等法律、行政法规规定。

AI数据销毁：当AI数据不再需要或达到一定的保留期限时，应当被安全地销毁。AI数据销毁包括彻底删除数据，以防止未经授权的恢复和访问。

AI数据出境安全：AI数据在跨境传输以及使用时需确保AI数据的出境安全，需对AI数据进行分类，明确哪些数据属于个人信息，重要数据。并对不同类型的AI数据采用个人信息保护以及重要数据的保护，确保AI数据出境满足《汽车数据安全管理若干规定（试行）》《数据出境安全评估办法》等法律法规要求。

AI数据安全体系评估内容：包括应急响应管理制度、数据分类分级制度、数据第三方共享管理制度、数据标注管理制度、数据销毁管理制度、个人信息数

清华大学智能网联汽车分学会

据出境安全评估制度、数据质量管理制度、数据完整性管理制度、风险评估管理制度。

生成式 AI 数据安全要求：从训练数据安全、模型安全、结果生成安全三个类别考虑，并针对不同的类别明确给出安全要求，从而确保 AIGC 的数据安全。

训练数据安全：针对生成式车用 AI 训练数据类型和训练数据测试提出要求，用于规范训练数据的选择和使用安全。

模型安全：针对生成式车用 AI 算法模型提出安全要求，包括模型安全基本要求、模型数据安全要求、模型可解释要求、模型底层框架安全要求、模型部署安全要求。

结果生成安全要求：针对生成式车用 AI 结果生成内容提出安全要求，包括内容安全测试和评估要求。

9.2.3 车用 AI 数据通用安全测试评估要求

根据第 9.2.2 章节介绍的测试与评估的对象和范围, AI 数据通用安全测试评估要求从 AI 数据全生命周期以及体系评估两个维度考虑，主要分为以下几个方面：

9.2.3.1 车用 AI 数据采集安全测试评估要求

(1) AI 数据采集匿名化和脱敏测试：确保 AI 数据采集活动遵守相关法律法规，如《数据通用技术要求》、《汽车采集数据处理安全指南》所提及的要求，以及《个人信息保护法》等相关规定。

(2) AI 数据最小化：确保采集的 AI 数据应当限于实现处理目的的最小范围，不得过度收集个人信息。

(3) AI 数据来源可信：通过采集授权、身份验证等方式，确保 AI 数据来源是可信赖和可验证的

(4) AI 数据完整性：采用哈希函数、数字签名等技术保证 AI 数据采集过程中的

汽标委智能网联汽车分委会

完整性，确保 AI 数据未被篡改。

9.2.3.2 车用 AI 数据传输安全测试评估要求

(1) AI 数据传输保密性：通过进行身份验证，端到端加密，安全的通信协议，使用 VPN 建立安全的网络隧道等手段，确保 AI 数据传输时的保密性。

(2) AI 数据传输完整性：对传输的 AI 数据包进行完整性检查，确保 AI 数据传输的完整性，如采用哈希算法加签名的方法。

9.2.3.3 车用 AI 数据存储安全测试评估要求

(1) 物理和逻辑安全：确保存储设施具备足够的物理和逻辑安全措施，防止 AI 数据泄露、篡改或丢失。

(2) AI 数据备份与恢复：建立数据备份和恢复机制，确保 AI 数据的可恢复性。

(3) AI 数据存储保密性：对存储的 AI 数据使用加密算法进行加密处理，即使 AI 数据被未授权访问，也无法被读取或理解。

(4) AI 数据存储完整性：使用哈希算法、数字签名等方法，确保 AI 数据存储的完整性。

(5) AI 数据访问控制：实施严格的访问控制机制，确保只有授权用户才能修改 AI 数据。

(6) AI 数据分类存储：基于分类分级结果，采取不同的存储措施，如将重要数据和敏感数据存储在加密或隔离的环境中。

9.2.3.4 车用 AI 数据处理安全测试评估要求

(1) AI 数据安全技术：采用算法安全、匿名化技术、隐私增强技术、个人信息脱密等技术手段保护 AI 数据在处理过程中的安全。其中包括：算法安全性测试，实施模型逆向工程测试，使用代码审计工具和深度学习模型测试框架如 CleverHans 进行安全测试，以确认没有安全漏洞；隐私增强技术测试，通过代码审查等方式验证在智能网联汽车人工智能数据

汽标委智能网联汽车分委会

处理过程中是否采用了隐私增强技术（PETs），如同态加密、差分隐私等。

(2) AI 数据访问控制：确保用户和系统只拥有完成其职能所必需的最小数据访问权限。根据用户的角色来分配访问权限，不同角色拥有不同的 AI 数据访问和处理权限

(3) 个人信息安全：对 AI 数据进行分类分级并加密存储，未经用户同意不将云端 AI 数据分享给任何第三方。保障用户对自己个人信息的控制权，包括访问、更正、删除和撤回同意的权利。遵守相关的个人信息保护法律法规，如《个人信息保护法》和《数据安全法》等。

(4) 事件应急响应：根据事件的严重程度、影响范围和紧急性对事件进行分类。制定详细的应急响应计划，包括事件确认、影响评估、隔离措施、数据恢复、责任分配、沟通策略和事后分析。

(5) AI 数据分类分级管理：根据 AI 数据的类型和用途，将 AI 数据分为不同的类别。例如，可以分为环境的 AI 数据、车辆自身的 AI 数据（如车辆状态和性能数据）、用户的 AI 数据（如个人信息和驾驶行为数据）等。对每个类别的 AI 数据进行敏感性和重要性评估，划分为不同的级别。

(6) AI 数据共享和对外合作管理：制定明确的 AI 数据共享政策，规定哪些 AI 数据可以共享，共享的条件和限制。对外合作时应对合作伙伴进行严格的审查，确保具备足够的 AI 数据保护能力，并遵守相关的法律法规。与合作伙伴签订 AI 数据保护协议，明确 AI 数据的使用范围、处理方式、保密义务和 AI 数据泄露时的责任。

9.2.3.5 智能网联汽车 AI 数据销毁安全测试评估要求

(1) AI 数据安全销毁：制定明确的 AI 数据销毁策略和流程，确保不再需要的 AI 数据能够被彻底且安全地销毁。

(2) AI 数据销毁后无法访问：销毁方法应确保数据无法被访问，防止 AI 数据泄

汽标委智能网联汽车分委会

露风险。

(3) AI 数据剩余信息保护：在 AI 数据销毁时，保证需要销毁的 AI 数据没有可恢复的数据残留。并对剩余的信息进行保护避免被销毁。

9.2.3.6 车用 AI 数据出境安全测试评估要求

智能网联汽车 AI 数据数据出境安全：根据相关法律法规，如《数据出境安全评估办法》，对 AI 数据出境活动进行安全评估，确保 AI 数据跨境传输的安全性。确保 AI 数据出境后，AI 数据处理者具备足够的 AI 数据保护能力。AI 数据出境应当对 AI 数据进行采取加密保护等措施。重要信息数据出境前，应开展数据出境风险自评估，重点评估 AI 数据出境的目的、规模、范围、种类、敏感程度以及可能带来的风险。

9.2.3.7 车用 AI 数据制度体系完整性要求

(1) 数据质量管理体系：确认企业是否具备相应管理体系，管理体系是否能覆盖 AI 产品的要求。

(2) 风险评估管理制度体系：确认企业是否具备相应管理体系，管理体系是否能覆盖 AI 产品的要求。

(3) 应急响应管理制度：确认企业是否具备相应管理体系，管理体系是否能覆盖 AI 产品的要求。

(4) 分类分级制度：确认企业是否具备相应管理体系，管理体系是否能覆盖 AI 产品的要求。

(5) 数据共享和对外合作管理制度：确认企业是否具备相应管理体系，管理体系是否能覆盖 AI 产品的要求。

(6) 数据全生命周期管理制度、风险评估管理制度：确认企业是否具备相应管理体系，管理体系是否能覆盖 AI 产品的要求。

(7) 数据销毁管理制度：确认企业是否具备相应管理体系，管理体系是否能覆盖 AI 产品的要求。

汽标委智能网联汽车分委会

(8) 数据出境管理制度：确认企业是否具备相应管理体系，管理体系是否能覆盖 AI 产品的要求。

(9) 用户权益响应管理制度：确认企业是否具备相应管理体系，管理体系是否能覆盖 AI 产品的要求。

9.2.4 车用 AI 生成式数据安全测试要求

9.2.4.1 训练数据安全要求

(1) 训练数据类型

在智能网联汽车（ICV）领域，基于真实数据和合成数据的闭环系统是一个至关重要的概念，它不仅贯穿于车辆从设计、开发到实际运营的全过程，还涉及到数据的云端存储与管理。高效地利用大数据资源对于推动自动驾驶技术的进步具有重要意义。

1) 真实数据：

A. 真实数据的获取方式

真实数据通常来源于实际道路测试中的车辆传感器（如摄像头、激光雷达、毫米波雷达等）。这些传感器能够捕捉车辆周围环境的信息，包括但不限于交通参与者的行为、路面状况、天气条件等。此外，真实数据还包括用户反馈、事故报告等非结构化信息。

B. 真实数据的高效挖掘

特征提取：通过对原始传感器数据进行预处理，提取出有助于自动驾驶决策的关键特征，比如物体检测、分类及跟踪。

异常检测：识别并标记出异常行为或事件，比如紧急制动、碰撞预警等，这对于提升自动驾驶系统的安全性和可靠性至关重要。

场景重建：利用多源传感器数据融合技术，重建复杂交通场景，为后续的仿真闭环提供基础。

汽标委智能网联汽车分委会

C. 真实数据的高效利用

模型训练：将自动化标注或人工标注的高质量真实数据用于训练深度学习模型，提高其在特定场景下的识别精度和泛化能力。

性能评估：通过对比测试集上的表现，评估自动驾驶系统在不同条件下（如雨雪天气、夜间行驶）的表现，从而指导系统优化方向。

规则制定：依据实际驾驶经验总结出的安全规范和操作准则，不断完善自动驾驶算法逻辑框架。

D. 真实数据的存储与管理

在云平台上存储真实数据时，需要考虑以下几个方面：

数据安全：采用加密技术保护敏感信息，防止未经授权访问。

数据清洗：去除无效或重复的数据记录，确保数据质量。

数据分发：根据不同应用场景的需求，将数据分门别类地存储于不同的数据库中，方便后续检索和使用。

备份与恢复：定期备份重要数据，并建立完善的灾难恢复计划，确保在发生意外情况时能够迅速恢复正常运作。

隐私保护：遵守相关法律法规，对涉及个人隐私的数据进行匿名化处理。

2) 合成数据：

A. 合成数据的获取方式

合成数据则一般是通过计算机图形学手段生成的虚拟环境中的数据。这包括使用游戏引擎（如 Unity、Unreal Engine）构建逼真的三维场景，以及借助物理引擎模拟真实世界的物理现象。

B. 合成数据的多样性泛化

参数调整：设定各种环境变量（如光照强度、风速等），观察其对传感器输出的影响，从而获得更广泛的训练样本。

汽标委智能网联汽车分委会

多样性扩展：利用随机噪声添加机制增加数据集的多样性，使模型能够在更多未知情况下做出合理判断。

跨域适应：通过改变虚拟世界与现实世界之间的差异程度，训练模型更好地适应不同环境的变化。

C. 合成数据的高效利用

快速迭代：相较于收集昂贵且耗时的真实数据，合成数据允许研究人员快速生成大量用于实验的数据集，加速算法迭代周期。

极端情况模拟：在虚拟环境中可以轻松复现现实中难以遇到的极端事件，如突发障碍物出现、行人突然横穿马路等，帮助提升自动驾驶系统应对复杂情况的能力。

成本节约：减少对真实道路数据采集和标注的依赖，降低了研发成本，并提高了开发效率。

D. 合成数据的存储与管理

合成数据同样需要在云平台上进行有效的管理和维护：

版本控制：合成数据的生产往往伴随着模型更新而变化，因此需要实施严格的数据版本控制系统，确保每次更改都能被追踪。

数据共享：鼓励跨团队间的数据共享，促进知识交流，加速整个行业的发展。

计算资源调度：合理分配计算资源，保证数据生成和处理过程中的高效运行。

自动化流程：实现数据生成、标注、验证等环节的自动化，减轻人工负担，提升工作效率。

无论是真实数据还是合成数据，在智能网联汽车数据闭环体系中都发挥着不可替代的作用。真实数据提供了最直接的反馈，合成数据则极大地丰富了训练样本空间。未来，随着技术进步，两者之间会实现强有力的相互补充，共同推动自动驾驶技术向着更加安全、智能的方向发展。而合理的云平台存储与管理策略，

汽标委智能网联汽车分委会

则为这一进程提供了坚实的基础。

(2) 训练数据测试

1) 数据投毒

训练数据是生成式 AI 模型学习的基础，其质量和安全性直接影响模型的性能和输出的可靠性。因此，选择高质量、安全、合规的训练数据对于车用 AI 系统至关重要。数据投毒攻击 (Data Poisoning) 是一种通过在训练数据中植入恶意样本或修改数据以欺骗模型的方法。攻击者可能会植入具有误导性标签或特征的数据，扭曲生成式模型的学习过程，导致模型在未来的模型生成中产生错误结果。因此，在训练数据选择过程中，必须对数据投毒进行严格的分析和判断。为了防止数据投毒攻击的产生，智能网联汽车生成式人工智能的训练数据安全测试评估需要从以下几个方面进行：

A. 数据源审查

确保训练数据来源于可信的数据源，并对数据采集过程进行严格的审查。原则上来说，所有原始数据的来源都应进行标注，并对来源进行审查。对曾发现可疑或恶意的数据源进行备案和预警。

B. 数据清洗

在数据输入模型前，进行彻底的数据清洗。特别的，应重点识别并剔除异常值、内容与标签不一致的数据、和内容（图片、文本等）可能被篡改的数据，这些数据存在较高的被投毒风险。

C. 数据验证

实施数据验证机制，包括统计测试、可视化分析等传统验证与自动化投毒检测等基于深度学习的先进验证，以识别潜在的数据投毒行为。

D. 访问控制

限制对训练数据的访问，确保只有授权人员能够进行数据的添加、修改或删除操作，防止内部人员进行数据投毒。

汽标委智能网联汽车分委会

E. 数据加密

在训练数据存储和传输过程中使用加密技术，确保数据的机密性和完整性，防止在传输或传输过程中被数据投毒。

F. 数据增强

通过数据增强技术，如随机旋转、平移和噪声添加，来扩展数据集并保持其固有特征。这种方法可以增加样本量，从而降低数据投毒的影响。

G. 数据的全面性与多样性

训练数据应涵盖智能网联汽车可能遇到的各种情况和环境，以提高模型的泛化能力和适应性。同时，应避免数据偏见，确保数据集的多样性和平衡性。通过这些措施，在提升模型整体性能与泛化性的同时，降低数据投毒的不良影响。

H. 风险评估与应急预案

定期开展训练数据安全风险评估，识别和分析数据安全风险。同时，制定数据安全应急预案，明确应急响应流程和措施，提高应对数据安全事件的能力。

I. 持续监控与审计

建立监控审计机制，对数据处理活动进行持续监控和审计。通过日志管理平台收集系统和访问者操作日志，实现对数据处理活动的全面监控。

J. 人员培训

加强对员工的数据安全意识培训，提高员工的安全防范意识和技能。鼓励员工积极参与数据安全管理工作，形成全员参与的数据安全文化。

1) 隐私泄露

在智能网联汽车领域，生成式 AI 技术的应用日益广泛，其训练数据的选择直接关系到数据安全和个人隐私保护。例如，智能网联汽车在运行过程中，会收集大量的个人和环境数据，包括但不限于位置信息、行驶习惯、个人偏好等。这些数据在未经妥善处理的情况下，可能存在隐私泄露风险。例如，数据滥用：未

汽标委智能网联汽车分委会

经授权的第三方可能通过非法手段获取数据，用于商业广告、身份盗窃等；数据误用：在数据共享过程中，可能因缺乏适当的数据脱敏处理，导致敏感信息被不当使用。训练数据的选择是确保智能网联汽车生成式 AI 系统数据安全的关键环节。智能网联汽车生成式 AI 数据安全测试应通过训练数据选择和处理的方式降低隐私泄露风险，保护用户隐私权益，同时提升公众对智能网联汽车的信任度，促进行业的健康发展。智能网联汽车企业应将数据安全和隐私保护作为核心竞争力之一，不断加强技术防护能力，完善管理体系，确保数据处理活动的安全性和合规性。为了防止隐私泄露，智能网联汽车生成式 AI 的训练数据隐私泄露测试评估需要从数据泄露和非法使用测试、数据脱敏和匿名化处理测试、法律遵从性评估、隐私影响评估等方面考虑。

2) 数据确权

数据确权问题源自于数据交易中存在的数据拥有权与数据交易本身的矛盾性。具体是指：通过数据交易获得指定训练数据拥有权的买家，可以通过直接对训练数据进行篡改后将该数据放在市场上继续流通，从而使自己从数据的拥有者变成了数据的生产者。这种做法将会破坏数据交易的流程，严重打击乐数据生产方的积极性，使得各方越来越倾向于自用训练数据或者拒绝生产训练数据。这将会对数据的使用效率和流通性产生不良影响。

智能网联汽车生成式 AI 的训练数据确权测试评估包括数据水印测试和知识产权评估等方面。下图所示汽车训练数据在面对多样化的数据攻击手段下，通过确权手段完成数据的确权的方法。针对训练数据的攻击可以分为数据修改、数据删减、数据洗牌和数据转化等。其中，数据修改是指通过加入随机或者非随机的噪声，对训练数据的值进行篡改，从而使得在训练数据上的添加的确权信息被人为破坏从而无法完成数据确权；数据删减则是直接对训练数据中的一些词条进行删减和销毁，从而对潜在的数据水印进行破坏；数据洗牌和转换都是针对训练数据分布的改变，只改变数据的排列，但会在一定程度上保留训练数据的结构。面对多样化的攻击，要求训练数据的确权具有无害性、鲁棒性、隐蔽性的特征。

汽标委智能网联汽车分委会

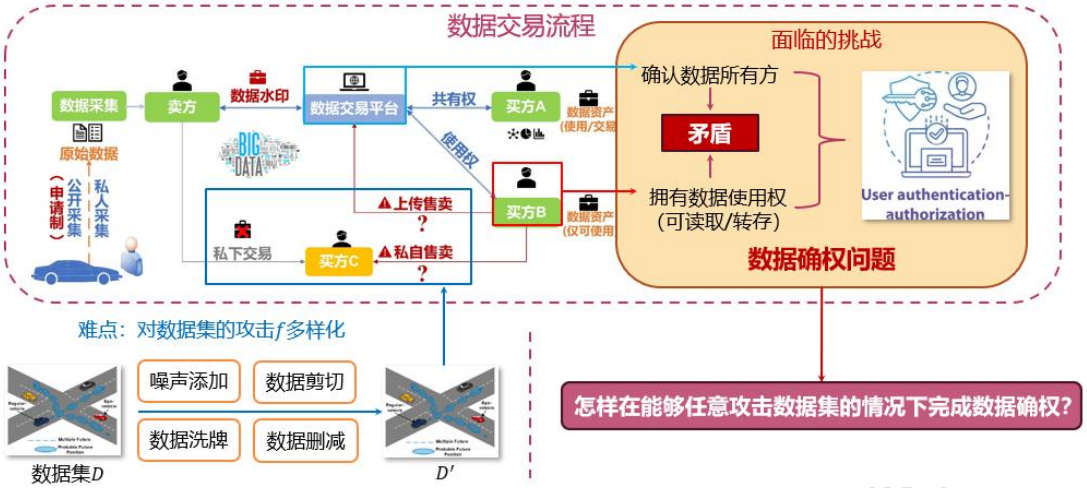


图 9-3 汽车训练数据交易结构与数据确权

9.2.4.2 生成式 AI 模型安全要求

(1) 模型安全基本要求：

汽车行业大模型应按照《生成式人工智能服务管理暂行办法》进行备案，如使用第三方基础模型提供服务，应使用已经主管部门备案的基础模型[2]。

(2) 模型数据安全要求：

模型应采用差分隐私、数据清洗、同态加密等数据安全防护技术，保护模型参数、重要数据、隐私数据等不易遭受非授权窃取、篡改、毁坏等，应对模型数据进行分类分级，并根据不同的数据级别实施不同的数据访问控制策略。

(3) 模型可解释要求：

模型应对模型本身内置可解释性和模型训练后的可解释性进行评估，包括算法的透明性、表达的解构性、算法适用边界、黑盒模型的事后解释以及因果分析等方面。

(4) 模型底层框架安全要求：

模型应对开源框架和依赖库的软件版本、软件组件、API 接口等进行安全检测，开源框架及依赖库应不存在公开的安全漏洞。

汽标委智能网联汽车分委

生成式 AI 如涉及车辆控制，其算法框架应具备高容错能力。当算法面临传感器故障、数据异常或其他突发情况时，仍能保持一定的工作能力或采取适当的应急措施，确保车辆的基本安全和功能。例如，当某个传感器出现故障时，算法能利用其他传感器的数据进行补偿和判断。

(5) 模型部署安全要求：

模型部署环境要至少满足网络安全等级保护三级要求，针对模型访问控制、安全审计、入侵防范、数据完整性、数据保密性、数据备份与恢复、等方面进行制定安全防护策略。

9.2.4.3 结果生成安全要求

(1) 车用 AI 生成式内容安全分类

智能网联汽车生成式 AI（AIGC）内容安全可分为智能网联汽车 AIGC 生成内容的隐私安全、合规风险、伦理安全三个类别，每个类别的要求建议如下：

1) 智能网联汽车 AIGC 生成内容的隐私安全

数据保护：确保训练生成式 AI 模型的数据来源合法，尊重数据主体的隐私权利，针对个人敏感信息进行加密传输和存储或采用匿名化技术去除可识别的个人信息。

访问控制：实施严格的访问权限管理，仅授权必要的人员访问涉及个人数据的系统和模型，采用身份验证和权限分离机制，防止未授权访问。

数据水印：利用暗水印等技术为生成内容添加不可见标识，便于追踪数据泄漏源头，保护原创内容和用户隐私。

2) 智能网联汽车 AIGC 生成内容的合规风险

遵守《生成式人工智能服务内容标识办法》等法律法规，对生成内容进行明确标识，确保透明度和可追溯性。

遵循 ISO/IEC 27001、ISO/IEC 27701、ISO27017、ISO27018 等国际安全与

汽标委智能网联汽车分委会

隐私保护标准，建立全面的信息安全管理机制。

A. 生成内容不含色情内容或暴力内容。

B. 生成内容不含影响社会或国家安全的内容。

C. 生成内容不含危害个人隐私的内容。

D. 生成内容不含歧视或仇恨的内容。

E. 生成内容不含不宜传播或不良价值观的内容。

3) 智能网联汽车 AIGC 生成内容的伦理安全

公平性与无偏见：确保生成式 AI 模型不加剧社会偏见，通过算法审计和调整，消除性别、种族、宗教等方面的歧视性内容。

透明度与可解释性：提高模型决策过程的透明度，使用户能理解生成内容背后的逻辑，增强公众信任。

责任归属：明确生成内容的责任主体，当生成内容引发问题时，能够追溯到相应的开发者或使用者，确保责任可追究。

用户教育与意识提升：提升公众对生成式 AI 潜在风险的认识，教育用户合理、负责任地使用此类技术，避免滥用。

(2) 车用 AI 内容安全测试评估要求

样本和训练管理：包括样本人工审核、隐私信息过滤、违规内容过滤、伦理问题内容过滤等。

用户输入环节的车用 AI 生成式内容安全检测：包括文本个人隐私信息识别、文本涉黄识别、文本暴恐识别、文本违禁识别、文本辱骂识别、文本伦理问题识别等风险识别。

内容生产环节的车用 AI 生成式内容安全检测：包括文本的隐私信息、涉黄、暴恐、违禁、辱骂、伦理问题内容识别，图片的人脸车牌等隐私信息、涉黄、暴恐、伦理问题内容等的识别，视频的人脸车牌等隐私信息、

汽标委智能网联汽车分委会

涉黄、暴恐、伦理问题内容等的识别，语音的隐私信息、涉黄、辱骂、伦理问题内容识别等。同时，还需要具备异常反馈和用户实名认证等功能。

举报与处置环节：主要包括投诉举报、账号处罚和内容处罚等。

审核优化环节：主要包括词库管理、检测范围管理和人工审核等。

(3) 车用 AI 生成内容审核的具体要求

文本审核：支持昵称、聊天和生成式人工智能指令审核，识别隐私信息、涉黄、暴恐、违禁敏感内容、伦理内容等，能够识别内容变形变异，关联用户上下文，兼容中英文指令。

图像审核：支持图片和 AI 绘图审核，识别人脸车牌等隐私信息、涉黄、暴恐、伦理问题等内容，能够区分动漫软色情尺度，检测图片和 AI 绘图画面中文字，提供丰富的标签管理。

视频审核：支持视频画面审核，识别人脸车牌等隐私信息、涉黄、暴恐、伦理问题等内容，能够区分动漫软色情尺度，检测视频画面中文字，提供丰富的标签管理。

语音审核：支持音频文件和语音流审核，识别语音隐私信息、涉黄、辱骂、伦理问题等内容，检测静默语音和呻吟声，提供转写后的文字和功能辅助人工审核。

9.3 车用 AI 数据安全测试技术研究

9.3.1 测评流程标准化研究

9.3.1.1 流程标准化目的

车用 AI 数据安全测试流程标准化的目的是为了确保 AI 系统在使用过程中能够保障数据的安全性和完整性。测试流程标准化可以规范测试的步骤和流程，明确责任和义务，可以帮助企业和组织发现和修复潜在的数据安全问题，预防数据泄露、篡改、丢失、AIGC 生成违规内容等问题的发生，从而保障用户的隐私

汽标委智能网联汽车分委会

和权益。此外，标准化的测试流程还可以提高 AI 系统的稳定性和可靠性，降低系统故障和事故的风险，保障企业和组织的业务正常运行。

9.3.1.2 流程及要求

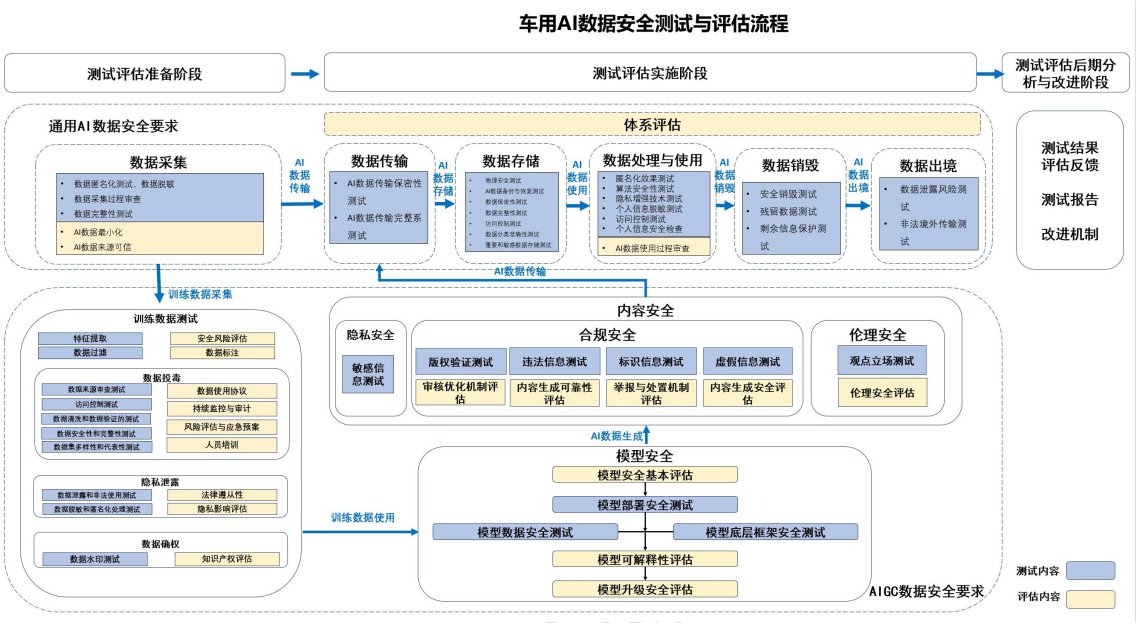


图 9-4 智能网联汽车人工智能数据安全测试与评估流程

(1) 前期准备

数据采集：采集数据用于通用 AI 应用场景以及生成式 AI 应用场景。

训练数据安全测试：用于生成式 AI 应用场景的数据，开展训练数据安全测试及风险评估，识别潜在威胁并评估其影响程度。之后进行训练数据的选择测试，包括数据投毒、隐私泄露、数据确权测试，识别 AIGC 模型训练使用的安全数据。

制定策略：依据风险评估结果，制定相应的数据安全策略，明确测试范围、方法及预期目标。

组建团队：成立专门负责此次测试工作的项目小组，成员应具备相关专业知识背景。

资源准备：确保测试所需软硬件设施齐备，搭建适合开展测试的实验环境。

清华智能网联

(2) 测试实施阶段

1) 数据生成阶段

模型安全：针对模型参数数据、模型底层框架安全和模型部署安全进行测试。

内容安全：针对 AI 数据生成内容合规、伦理、隐私方面进行测试。

2) 数据传输过程

传输通道安全：分析数据在传输过程中使用的协议安全性，如 HTTPS 等。

完整性校验：实施数据包完整性检查，防止中途篡改。

流量监控：建立实时流量监控体系，及时发现异常通信行为。

3) 数据存储与管理阶段

加密技术应用：考察是否采用先进的加密手段保护存储中的数据，防止未经授权的访问。

访问权限控制：测试系统对于不同级别用户的权限分配是否合理，能否有效防止越权操作。

备份恢复机制：验证数据备份方案的有效性，确保在发生意外情况时能够快速恢复。

4) 数据使用与销毁阶段

生命周期管理：制定详细的数据生命周期管理制度，规定数据何时何地可以被访问及使用期限。

销毁标准制定：确立科学合理的数据销毁流程，确保过期或不再需要的数据彻底清除。

遗留问题处理：针对测试中发现的问题，提出具体的整改措施，并跟踪落实情况。

(3) 后期分析与改进阶段

汽标委智能网联汽车分委会

编制报告：整理测试期间收集的所有信息，形成书面报告，客观反映测试成果。

整改建议：针对测试中暴露出来的不足之处，给出建设性的改进意见。

持续改进：鼓励企业和研究机构根据测试反馈不断完善自身数据安全管理机制，形成良性循环。

9.3.1.3 流程可行性分析

随着 AI 在智能网联汽车自动驾驶、智能交互等场景中的应用日益广泛，AI 技术的应用也带来了数据泄露、恶意攻击等数据安全问题，对汽车的安全性、隐私性和合规性造成了潜在威胁。因此，建立统一的智能网联汽车 AI 数据安全测试流程标准，明确测试范围、测试方法、测试指标等内容，对于保障汽车行业的健康发展具有重要意义。

在行业需求方面，AI 广泛应用于智能网联汽车自动驾驶、智能交互等场景中，面临数据泄露、车辆失控等安全威胁，AI 数据安全相关法律法规也在逐渐完善中，汽车厂商和供应商对数据安全测试流程标准化的需求日益迫切，通过 AI 数据安全流程标准化可以确保符合相关法规和标准要求，降低潜在的数据安全风险。

在数据安全测试基础方面，随着 AI 和数据安全技术的不断发展，已具备测试原则、测试阶段及测试内容、测试方法及测试的工具等可统一测试流程标准的基础。在测试原则方面，AI 数据安全测试原则主要包含数据完整性、保密性、可用性、合规性等方面在测试阶段及测试内容方面，AI 数据安全测试阶段分为数据采集、数据传输、数据存储、数据处理/使用安全、数据销毁、数据出境及 AIGC 安全，包含数据来源可信、数据完整性等测试内容。在测试工具和方法方面，针对智能网联汽车 AI 的数据安全测试工具越来越多样化，能够全面覆盖从算法模型到系统部署的各个环节。这些工具能够模拟各种攻击场景，对智能网联汽车 AI 进行全面的安全评估，发现潜在的安全漏洞和风险。同时，测试方法也

汽标委智能网联汽车分委会

在不断创新，结合静态分析、动态检测、模糊测试等多种技术，提高了测试的准确性和效率。

在国际合作趋势方面，国际社会对数据安全问题的关注度不断提高，各国通过分享经验、交流技术、共同研究等方式，可以不断提升对智能网联汽车 AI 数据安全问题的认识和理解；还可以推动各国在智能网联汽车 AI 数据安全领域的法律法规、技术标准等方面实现统一。

9.3.2 测评方法研究

9.3.2.1 通用 AI 数据安全测试方法及结果评价

AI 技术在车辆中的应用涉及到大量数据的处理，其中包括车辆的行驶数据、环境感知数据、用户行为数据等。为了确保这些数据的安全性，需要对智能网联汽车 AI 中的数据进行全面的安全测试，保障数据在整个生命周期中的安全，每个场景的数据安全测试方法都是为了在 AI 生命周期的各个阶段确保数据的安全性、隐私性和合规性，从而保护个人和企业免受数据泄露和滥用的风险。

(1) 数据采集安全

1) AI 数据采集匿名化和脱敏测试：

标识去除检查：审查数据集中是否存在明确的个人标识符，如姓名、身份证号、电话号码等。可以通过自动化脚本或人工检查的方式进行。对数据集中的每条记录进行遍历，确保没有直接可识别个人身份的信息存在。

间接标识符分析：检查数据中是否存在可能间接识别个人身份的信息，如特定的地理位置、职业、爱好组合等。可以通过统计分析和关联分析的方法，查看是否存在某些特征组合能够高概率地指向特定个人。

匿名化算法验证：了解数据采集过程中使用的匿名化算法，并对其进行验证。检查算法的参数设置是否合理，是否能够有效地去除个人标识信息。

可以通过对比原始数据和匿名化后的数据，评估算法的性能。评价是否数据中不包含常见的个人敏感信息，如电话，姓名，家庭住址等以及车辆的敏感信息，

汽标委智能网联汽车分委会

如车辆 VIN 码，车辆历史行驶轨迹等。如果评价结果为数据中不包含任何敏感信息，则评价通过，不存在隐私风险。

2) AI 数据最小化测试：

确定处理目的，详细分析 AI 系统的具体处理目的，将处理目的以文档形式记录下来；根据处理目的，列出实现该目的所必需的数据特征，对每个数据特征进行必要性评估，确保其确实与处理目的直接相关。通过审查数据采集的来源的方式，确定数据源是否与处理目的相关，检查数据源的合法性和合规性，确保数据采集是在合法的框架内进行。

3) AI 数据完整性测试：

验证是否使用签名算法或其他 hash 算法，并验证算法的安全性；对采集的数据进行篡改，评估系统针对不完整数据的处理方式。

(2) 数据传输安全

AI 数据传输完整性测试：验证智能网联汽车人工智能数据传输的可靠性，通过验证数字签名等方式验证智能网联汽车人工智能数据的完整性和一致性，防止智能网联汽车人工智能数据在传输过程中被窃取或篡改。价数据传输是否完整，如果数据完整，则评价通过，不存在数据丢失风险。

AI 数据传输保密性测试：验证智能网联汽车人工智能数据传输过程中是否采用了合适的加密算法和加密强度，检查是否使用了如 SSL/TLS 等协议，所使用的加密算法符合当前安全标准，通过使用网络抓包工具如 Wireshark 来抓取数据包，对数据包进行评价是否对数据进行加密。如果已加密，则评价通过，不存在数据泄露风险。

(3) 数据存储安全

物理和逻辑安全测试：检查物理接口和逻辑区域是否做了安全防护策略，防止非授权进行存储设施数据访问。

AI 数据备份与恢复测试：检查是否有明确的数据备份功能入口和设置

汽标委智能网联汽车分委会

选项。确认可以手动或自动启动备份操作，测试不同类型的数据是否都能被纳入备份范围，如行驶数据、导航设置、车辆诊断信息等。检查是否有明确的数据恢复功能入口和操作流程。确认可以选择特定的备份版本进行恢复，测试在不同情况下的恢复操作，如系统故障、数据丢失、软件升级失败等。

AI 数据存储保密性测试：依据车用 AI 数据存储区域和地址范围说明，通过零部件调试接口，使用软件分析工具提取存储的智能网联汽车人工智能数据（如敏感个人信息），评价是否存储的敏感隐私数据都是用了强加密标准，如 AES 加密等，如果对应存储数据已进行加密，则评价通过，不存在隐私风险。

AI 数据存储完整性测试：检查是否有机制保证智能网联汽车人工智能数据在存储过程中不被篡改，确保存储的智能网联汽车人工智能数据未被未经授权修改。可使用哈希算法生成数据的摘要，并存储这些摘要作为数据完整性的校验标签。在数据存储后，可以通过重新计算哈希值并与原始哈希标签比较来验证数据是否被篡改。通过重新计算哈希值并与原始哈希值进行对比，评价对比结果，如对比结果相同，则评价通过，不存在数据丢失风险。

AI 数据访问控制测试：基于用户账号及分配权限，访问及操作智能网联汽车人工智能数据，验证是否能访问授权智能网联汽车人工智能数据资源，是否能正确执行增加、删除或修改等操作，是否能阻止未授权用户对智能网联汽车人工智能数据资源的访问。对测试结果进行评价，如果可以阻止访问未被授权使用的数据，则评价通过，不存在隐私风险。

(4) 数据处理/使用安全

算法安全性测试：实施模型逆向工程测试，使用代码审计工具和深度学习模型测试框架如 CleverHans 进行安全测试，以确认没有安全漏洞。

隐私增强技术测试：通过代码审查等方式验证在智能网联汽车人工智能数据

处理过程中是否采用了隐私增强技术（PETs），如同态加密、差分隐私等；

访问控制测试：通过获取授权的智能网联汽车人工智能应用，使用工具篡改其代码，并安装、执行篡改后的授权第三方应用，测试其是否被限制直接访问系统资源、个人信息等。对测试结果进行评价，如果可以阻止访问未被授权使用的数据，则评价通过，不存在隐私风险。

AI 数据个人信息安全测试：检查汽车系统中与个人信息相关的设置选项。确认是否有明确的开关或选项允许用户控制数据。验证是否有同意提示，明确说明数据使用处理的对象、目的和范围，以便用户做出明智的决策。

(5) 数据销毁安全

AI 数据安全销毁测试：使用数据销毁测试工具，如 DBAN 等工具，来验证智能网联汽车人工智能数彻底销毁，测试智能网联汽车人工智能数是否可以被彻底销毁，无法恢复；如果数据被安全销毁，则评价通过，不存在数据销毁安全问题。

AI 数据销毁后无法访问测试：在数据声明销毁后，检查所有系统和备份中是否还有智能网联汽车人工智能数残留。通过关键字搜索、技术检测、日志审查等方式，确认销毁后的智能网联汽车人工智能数无法访问。评价是否被销毁的数据存在残留数据，如果不存在残留数据，则评价通过，数据被完全销毁。

AI 数据剩余信息保护测试：依据剩余智能网联汽车人工智能数存储区域、地址范围说明，测试剩余智能网联汽车人工智能数是否为按要求进行密文存储及访问控制防护。评价是否剩余信息对不同权限的访问控制是否正常，如果正常，则评价通过。

(6) 数据出境安全

AI 数据出境安全测试：通过模拟跨境数据传输过程，开启车辆全部移动蜂

清华大学智能网联汽车分学会

窝通信通道和 WLAN 通信通道，在各项智能网联汽车人工智能功能正常启用状态下，使用网络数据抓包工具对通信网络通道进行抓包，解析通信报文数据，检查目的 IP 地址中是否包含非法境外 IP 地址。如果评价不是非法境外传输，则评价通过，不存在数据境外传输合规风险。

9.3.2.2 生成式 AI 数据安全

(1) 训练数据测试方法

1) 训练数据特征提取

特征提取：从数据集中提炼出与安全性相关的特征，这些特征可能包括数据安全事件类型、发生频率、异常行为等，利用这些特征训练 AIGC 模型。

2) 训练数据过滤

数据过滤：通过语义分析、关键词检索等方式，确认 AIGC 模型训练数据中是否充分过滤违法不良信息。

3) 数据投毒测试

A. 数据来源审查测试

该测试旨在确保训练数据来源的合法性和可靠性，防止使用非法或不可靠数据源。

数据提供方资质审核测试：测试数据提供方的合法资质，如营业执照、数据服务许可证等。

数据采集合法性验证测试：测试数据采集过程是否符合相关法律法规，如是否获得了数据主体的明确同意。

版权信息检查测试：测试数据是否涉及版权问题，检查数据是否获得了适当的授权或处于公有领域。

B. 数据清洗和数据验证的测试

该测试旨在评估数据清洗和数据验证的有效性。数据清洗和数据验证测试可

汽标委智能网联汽车分委会

以采用多种方法和技术，根据数据的特点和问题进行选择。以下是一些常见的测试方法。

数据完整性检查：通过比较处理前后的数据数量和特征分布，检查处理后的数据是否完整，是否存在缺失值、空白值或不一致的数据。

数据一致性检查：使用条件语句、数据验证工具或基于模型的方法进行测试，比较不同数据源或数据字段之间的差异，检查数据的一致性。

逻辑错误检查：通过对数据进行逻辑性验证，检查数据中的逻辑错误和矛盾。例如，检查日期的合理性、数值之间的关系是否符合逻辑等。

异常值处理测试：检查是否剔除异常值、替换为合理值或使用统计方法进行处理，并验证处理后的数据是否合理。

缺失值处理测试：检查数据中的缺失值是否通过填充默认值、删除含有缺失值的行或使用插值等方法进行了处理。

重复值处理测试：检查数据是否存在重复值，确认数据中的重复记录已被正确移除或标记。

C. 访问控制测试

该测试旨在检查数据管理是否存在访问控制机制，是否只有授权人员能够进行数据的访问和处理操作。

用户身份验证安全测试：测试用户身份验证过程的安全性。需要对系统采用的认证机制进行全面的测试，检验是否存在可以被利用的弱点。检查密码策略的严格性与会话管理的安全性，以及多因素认证的安全性（如有）。

访问权限的测试：检查系统如何为不同用户分配和管理权限。验证最小权限原则是否得到了遵循，并检查是否有过度的权限授予。

访问控制策略的有效实施监控：检查访问控制策略是否有效执行，确保策略安全预期工作。检查系统是否具备行为监控和异常检测的能力，以识别出不正常的访问模式。

汽标委智能网联汽车分委会

D. 数据安全性和完整性测试

该测试旨在确保训练数据在采集、存储和传输过程中的安全性和完整性。

数据加密验证测试：测试数据在传输和存储时是否使用了强加密算法。

完整性校验测试：使用哈希算法对数据进行完整性校验测试，确保数据未被篡改。

安全协议测试：测试数据传输是否采用了安全的通信协议，如 SSL/TLS。

E. 数据集多样性和代表性测试

该测试旨在评估训练数据是否全面覆盖智能网联汽车的应用场景，确保模型的泛化能力。

场景覆盖分析测试：测试数据集是否包含各种使用场景，例如各种驾驶环境、天气条件、交通状况等。

用户行为多样性评估测试：测试数据集中是否包含不同驾驶者或用户的行为模式。

统计分析测试：使用统计工具对数据集进行分析，测试数据分布是否具有均衡性和多样性。

F. 隐私泄露测试

数据泄露和非法使用测试：通过使用数据泄露防护系统(DLP)设置策略来实时监控数据是否泄露以及在数据中嵌入水印，如果数据被非法复制和使用，可以通过检测水印来识别和追踪数据泄露的来源。

数据脱敏和匿名化处理测试：该测试旨在验证训练数据中个人隐私和敏感信息是否得到有效保护。

敏感信息识别测试：使用自动化工具测试数据中是否包含个人身份信息、财务信息等敏感数据。

脱敏处理验证测试：测试是否对敏感信息进行了脱敏处理，如替换、加密或

汽标委智能网联汽车分委会

泛化。

匿名化效果评估测试：测试处理后的数据是否无法追溯到具体的个人。

G. 数据确权测试

数据确权水印测试：通过数据驱动的数据水印技术，将数据生产者的知识产权信息隐蔽的嵌入至 AIGC 模型训练数据中并保证训练数据的分布和数值不发生改变，同时对抗性训练水印识别模型，保证该产权信息能够被特定模型给识别。

(2) 模型安全测试方法

1) 模型数据安全测试方法

模型参数测试：模型架构参数包括神经元数量、层类型、激活参数、隐藏层大小和宽度等模型重要参数数据安全进行测试，创建模型参数数据安全测试用例，包括输入敏感信息对话内容和预期系统处理结果。执行测试用例，评估系统在处理模型参数数据变化时的结果。

2) 模型底层框架安全测试方法

模型底层框架安全测试方法包括四类安全测试方法，比如随机攻击、盲盒攻击、黑盒攻击、白盒攻击。

随机攻击：采用常见的提示和问题，通过文本增强和扩展技术批量生成，用于大语言模型评估。

盲盒攻击：利用特定的攻击知识并引入恶意输入来创建对抗性样本，采用诸如指令劫持和提示屏蔽等策略。

黑盒攻击：攻击者无需了解模型的参数或训练细节即可与大语言模型进行对话式交互的能力，通过黑盒方式进行攻击测试。

白盒攻击：假定攻击者可以访问模型的结构和权重从而使他们能够使用梯度上升等技术针对模型制作对抗性样本，通过白盒方式进行攻击测试[3]。

3) 模型部署安全测试方法

环境配置检查：对被大模型的部署配置进行检查，确保其符合安全标准和要

汽标委智能网联汽车分委会

求。

安全测评：使用专业工具对被大模型部署环境各个方面进行测试，以发现潜在的安全问题和漏洞，包括物理环境、通信网络、区域边界，安全计算环境、安全管理中心进行实地检查，确保其安全性和符合等级保护的要求。

人员访谈：测评人员与被测系统相关人员进行交流、讨论，获取相关证据，了解有关信息。

文档审查：根据技术和管理标准，对模型部署安全管理制度、执行过程文档、系统设计方案等技术资料进行审查。

(3) 内容生成安全测试方法

1) 版权验证测试

测试车用 AIGC 内容是否侵犯了他人的版权或知识产权，使用内容比对算法，扫描文本、图片及视频等多种内容格式，与广泛的版权数据库进行比对。

2) 违法信息测试

如使用 NLP（自然语言处理）技术进行内容分析，确保车用 AI 生成的内容不含有违法违规、误导性或不当信息。

3) 敏感信息测试

对车用 AIGC 的生成内容进行敏感识别，测试车用 AIGC 的生成内容是否存在保密信息、已销毁内容以及个人信息。

4) 虚假信息测试

使用数据来源验证和内容分析方法，对车用 AIGC 的生成内容进行虚假信息测试，测试车用 AIGC 的生成内容是否存在虚假信息，特别是汽车行业中的交通法规和违法处罚内容的。

5) 观点立场测试

使用情感分析和内容检测方法，对车用 AIGC 生成的内容进行情感测试，测试车用 AIGC 的生成内容是否存在观点和立场错误的问题。

汽标委智能网联汽车分委会

6) 标识信息测试

测试车用 AIGC 的生成结果是否进行标识，能够和非生成结果进行区分。

9.3.3 测试工具

随着 AI 技术在汽车领域的深入应用，数据安全问题日益凸显。为确保汽车数据的安全性，开发高效、专业的数据安全测试工具显得尤为关键。测试工具分为智能网联汽车数据通用安全测试工具、智能网联汽车 AI 数据安全测试工具、AI 数据安全测试工具。

9.3.3.1 智能网联汽车数据通用安全测试工具

(1) 数据抓包检测工具：

Wireshark 能够实时抓取网络流量，并提供详细的协议解码功能，支持数百种网络协议。实时抓包与分析。可视化展示各种网络协议。支持过滤器，可以针对特定协议、IP 地址、端口等进行筛选。支持捕获和分析无线网络流量。

Tcpdump 能够在指定的接口上捕获网络流量并输出到标准输出或文件中。验证智能网联汽车人工智能数据传输过程中是否采用了合适的加密算法和加密强度，检查是否使用了如 SSL/TLS 等协议，所使用的加密算法符合当前安全标准。

(2) MD5/SHA-256 校验和工具：

WinMD5：用于 Windows 的简单 MD5 校验工具。

HashCalc：支持多种哈希算法（MD5、SHA-1、SHA-256 等）的工具，适用于文件和字符串。

(3) 数据存储漏洞扫描工具：

OpenVAS：一个开源的漏洞扫描工具，能够扫描存储系统的操作系统和网络服务中的安全漏洞，帮助检测数据存储的逻辑安全问题。

Nessus：一个专业的漏洞扫描工具，广泛用于检测存储服务器或设备中的安全漏洞，并提供补救措施。它支持对存储系统的操作系统、应用程序以及网络协

汽标委智能网联汽车分委会

议进行安全检查。

(4) 加密算法验证工具：

OpenSSL：是一个开源工具，支持多种加密算法和协议。可以用于生成、验证和测试加密密钥以及加密数据。可以用来检查加密密钥和算法的正确性，验证数据是否按照预期进行了加密。

(5) 数据完整性检测工具：

用于检测数据在传输、存储或备份过程中是否发生了变化，特别是在存储算法和压缩算法应用后验证数据的一致性。

MD5SUM：用于生成和校验文件的 MD5 哈希值，适合检测文件的一致性。

SHA256SUM：用于生成和校验文件的 SHA-256 哈希值，比 MD5 更加安全。

(6) 访问控制测试检测工具

ADB、SSH、putty：基于用户账号及分配权限，访问及操作智能网联汽车人工智能数据，验证是否能访问授权智能网联汽车人工智能数据资源，是否能正确执行增加、删除或修改等操作，是否能阻止未授权用户对智能网联汽车人工智能数据资源的访问。

nmap、putty：扫描车辆上的开放端口，识别正在运行的服务。发现主机上哪些端口是开放的，了解服务运行情况。判断是否存在访问控制授权。

(7) 代码分析工具：

SonarQube：供静态代码分析，支持多种编程语言（如 Java、C#、JavaScript、Python 等），可以检测代码中的安全漏洞、代码质量问题和技术债务。

(8) 同态加密工具、差分隐私检测工具：

Benchmarking Tools：用于测试同态加密算法的性能，如计算速度和资源消耗。评估加密算法的效率，优化性能。

汽标委智能网联汽车分委会

PrivateSetIntersection (PSI) Tools：用于隐私保护集合交集计算的工具，支持差分隐私应用场景。测试隐私保护集合交集计算的实现。

(9) 数据还原检测工具：

TestDisk：开源的数据恢复工具，支持从丢失的分区或损坏的存储设备中恢复数据。可以用它来验证数据销毁后的恢复可能性。

(10) 数据销毁评估工具：

Recuva：一款用于恢复已删除文件的轻量级工具，适合检测文件删除或销毁后的恢复情况。

(11) 剩余信息保护测试工具：

R-Studio：R-Studio 是一款专业的数据恢复工具，支持多个操作系统，包括 Windows、Mac 和 Linux。通过扫描设备并尝试恢复文件，检测销毁数据的残留情况。

9.3.3.2 智能网联汽车 AI 数据安全测试工具

(1) 生成式对抗网络（GAN）测试工具

GANFuzz：利用生成式对抗网络（GAN）来生成测试数据，以发现汽车系统中潜在的安全漏洞。GANFuzz 通过生成高度逼真的数据包来测试汽车网络的鲁棒性和安全性。

AI Fuzzing：结合生成式 AI 和模糊测试技术，自动生成多样化的测试用例，用于测试汽车软件和硬件系统的安全性。

(2) AI 驱动的渗透测试：

DeepExploit：使用深度学习技术自动化渗透测试过程，适用于复杂的汽车系统。DeepExploit 可以模拟黑客行为，生成特定的攻击向量来测试系统的安全性。

OpenAI Gym for Cybersecurity：虽然不是专门为汽车设计，但可以通过自定义训练环境，将汽车安全测试场景嵌入到 AI 驱动的渗透测试框架中。

汽标委智能网联汽车分委会

(3) AI 生成数据集与模拟：

CARLA + AI Data Generator：CARLA 是一个开源的自动驾驶模拟器，结合 AI 技术，可以生成大量用于测试汽车感知系统的数据集。这些数据可以用来测试汽车系统的应对能力，包括安全漏洞的暴露。

DeepDrive：一个基于深度学习的自动驾驶平台，能够生成模拟数据和场景，用于测试自动驾驶汽车在不同情况下的安全性和反应能力。

(4) 智能入侵检测与响应：

AI-enhanced IDS/IPS：基于 AI 的入侵检测系统（IDS）和入侵防御系统（IPS），如 Darktrace，利用生成式 AI 模型分析汽车网络中的异常行为，并自动生成安全响应策略。

Generative Adversarial IDS：使用生成式对抗网络来模拟攻击和生成恶意流量，以训练和测试汽车网络中的入侵检测系统。

(5) AI 增强的车载系统漏洞扫描：

AI-powered Vulnerability Scanners：将生成式 AI 模型集成到传统的漏洞扫描工具中，生成可能的攻击路径和变种漏洞，从而提高扫描的全面性和有效性。

这些生成式人工智能工具和技术代表了数据安全测试的新前沿，能够模拟复杂、多样化的攻击场景，帮助识别和修复汽车系统中的潜在漏洞。通过结合生成式 AI 技术，汽车制造商和安全研究人员能够更深入地了解系统在现实世界中的表现，并加强其安全性。

针对生成式人工智能的数据安全测试工具，应具备以下几项关键能力：

(6) 匿名化效果测试工具：

Privacy Analytics Risk Assessment 提供全面的数据匿名化和风险评估工具。能够评估去标识化的数据集是否仍然存在重新识别的风险。

人工智能数据安全测试工具在当前的网络安全领域中扮演着至关重要的角

汽标委智能网联汽车分委会

色。它们帮助企业和组织识别、评估并修复人工智能中的潜在安全漏洞，确保数据的安全性和完整性。以下是一些常见的能进行人工智能数据安全测试工具：

9.3.3.3 AI 数据安全测试工具

(1) PaddleSleeve：

基于百度飞桨开源深度学习平台的安全与隐私工具 PaddleSleeve 已在多个场景中实现对飞桨自定义及预训练模型，包括 ResNet、YOLO 等通用产业级模型的支持，为人工智能模型安全与隐私保护提供更好的能力支撑。PaddleSleeve 主要功能有鲁棒性测评工具 Robustness、模型隐私风险测试与评估 RrivBox、隐私增强 PrivacyGuard[4]。

(2) 人工智能安全检测平台“蚁鉴”：

蚂蚁基于自身多样化的业务场景，同时结合对相关法律法规、指南要求等的研究，总结了一套覆盖内容安全、数据安全、科技伦理三大类共 200 多个子类的测评依据标准，已有百万量级测试数据集，并提供了一系列的隐私保护测试工具和方法，用于评估人工智能模型在数据处理和存储过程中的隐私泄露风险。

(3) 人工智能安全评测平台 AIcert:

浙江大学区块链与数据安全全国重点实验室人工智能安全团队聚焦人工智能系统全栈安全，牵头研发了人工智能安全评测平台 AIcert。相较其他人工智能安全评测平台，AIcert 解决了从数据到系统、从算法开发到系统部署的多层面全方位安全检测技术覆盖[5]。

(4) AI Safety Benchmark：

AI Safety Benchmark 是一个专注于大模型安全基准测试的平台，它通过生成大量攻击样本，评估 AI 模型在多个测试维度（如底线红线、信息泄露和社会伦理）上的表现，揭示模型可能存在的安全风险。

(5) Inspect：

Inspect 是由英国人工智能安全研究所发布的开源 AI 模型安全检测工具箱，

汽车零智能网联汽车分委会

它通过灵活的组件设计，旨在评估 AI 模型在核心知识、推理能力和自主能力等关键领域的安全风险。

(6) LLM Guard：

LLM Guard 是一个开源工具包，专为保护和强化大型语言模型在生产环境中的安全性而设计。它提供全面的评估和保护功能，包括清理有害语言、检测数据泄漏、防止提示注入和越狱攻击等。

(7) 蚁天鉴 2.0：

蚁天鉴 2.0 是蚂蚁集团联合清华大学发布的 AI 大模型安全检测工具，它新增了 AI 鉴真功能，支持多模态内容真实性及深度伪造检测，并构建了双重防御护栏，提供全流程自动化的安全测评工具，为通用大模型及垂直领域行业大模型应用安全保驾护航。

(8) 绿盟 AI 大模型风险评估工具：

绿盟科技发布的 AI 大模型风险评估工具，为企业 AI 大模型的应用实践提供全面、深入的安全风险评估防线。它基于精心筛选和校准的测试用例库，能够迅速且高效地识别出内容安全和对抗安全的潜在威胁，并配备专业的风险处理建议。

9.4 车用 AI 数据安全评估技术研究

9.4.1 评估流程标准化研究

9.4.1.1 流程标准化目的

车用 AI 数据安全评估流程标准化旨在规范车用 AI 数据安全评估的流程及要求、开展车用 AI 数据安全评估方法和工具的研究、细化通用数据专项、体系、生成式 AI 数据安全评估指标，为相关单位开展车用 AI 数据安全评估提供参考。

9.4.1.2 流程及要求

车用 AI 数据安全评估涵盖前期准备、评估实施和后期分析与改进三大阶段。

汽标委智能网联汽车分委会

前期准备：此阶段为评估奠定基础，需梳理智能网联汽车人工智能数据状况，明确数据处理全流程评估目标，确定评估对象、范围和边界。组织业务、安全等多部门协同，制定评估方案，确定依据、内容并建立文档。同时，收集数据处理者相关信息，形成业务、系统、资产和活动清单，条件允许时绘制数据流图，并调研已有安全措施，汇总成调研结果表。

评估实施阶段：基于前期信息收集成果，结合自动化与人工评估方法，围绕被评估企业数据资产和处理情况，验证数据安全生命周期各环节安全措施有效性，以及 AIGC 生成内容的隐私安全与合规性。

后期分析与改进阶段：从数据保密性、完整性、可用性和处理合理性等角度，分析数据安全保护有效性，得出总体评估结果。编制详细评估报告，记录评估过程、问题及改进建议。

9.4.1.3 流程可行性分析

本评估方法依据了 GB/T 35273—2020《信息安全技术 个人信息安全规范》、GB/T 41871—2022《信息安全技术 汽车数据处理安全要求》、TC260-00X《生成式人工智能服务安全基本要求》[31]、GB/T 44464—2024《汽车数据通用要求》等法规标准，对智能网联汽车人工智能数据全生命周期中的数据安全和个人信息保护能力提出要求，故具备严格坚实的理论基础。

本评估方法能够覆盖车端与企业端多场景智能网联汽车 AI 产品，采用统一标准，具备普遍适用性；评估过程不涉及核心算法及商业秘密，企业仅需提供体系文件、功能逻辑展示及有限技术支持，无需开放底层代码或权限接口，操作便捷；其贯穿车用 AI 数据全生命周期（含告知同意、收集、传输、存储、使用、加工、公开、删除等环节），可助力企业梳理风险点、强化数据安全防护，降低泄露滥用风险。综合来看，该方法兼具市场需求、理论支撑、普适性、可操作性及社会效益，具备可行性。

通过上述因素的综合考量，本评估方法具备广泛的市场需求、坚实的理论基

浙标智能网联汽车分委会

础、普遍的适用性、可实施的路径以及积极的社会效益，故本评估方法具有可行性。

9.4.2 评估方法研究

9.4.2.1 评估内容及评价

智能网联汽车 AI 数据安全评估方法是一个系统化、面向智能网联汽车 AI 全生命周期的一个过程，该过程的核心目的在于确保数据的保密性、完整性、可用性和数据处理的合理性不受威胁，同时保障个人隐私和合规性要求得到满足。以下是围绕智能网联汽车人工智能数据全生命周期的数据安全评估方法，包括通用数据专项评估、体系评估、生成式 AI 数据安全评估。车用 AI 的数据安全评估可参照 9.3.1.2 中流程图评估部分。

(1) 通用数据专项评估

AI数据采集最小化评估包括数据采集过程审查和精度适用性评估，其中：

数据采集过程审查：确定处理目的，详细分析 AI 系统的具体处理目的，根据处理目的，列出实现该目的所必需的数据特征，对每个数据特征进行必要性评估，确保其确实与处理目的直接相关。通过审查 AI 数据的处理过程，确定处理数据是否与处理目的相关，检查数据处理的合法性和合规性，确保数据处理是在合法的框架内进行。

精度适用性评估：通过查看企业产品手册、开发文档、企业设备参数信息表等方式，确认数据采集设备精度为实现功能所必须的。

AI 数据采集评估还包括数据来源可信度评估，通过使用查看软件许可证、合作协议等方式，确认 AI 数据来源可信；通过使用 SCA 分析、语义分析、正则表达式匹配等方法，确认 AI 数据来源不包含黑名单数据。

(2) 体系评估

数据质量管理体系内容完整性：评估企业是否具备数据质量管理体系、数据质量监控措施、数据质量改进措施等机制。

汽标委智能网联汽车分委会

风险评估管理制度体系内容完整性：评估企业是否具备风险评估管理制度，处理重要数据时是否每年开展至少一次风险评估，是否具备相应的工作记录

应急响应管理制度内容完整性：评估企业是否具备智能网联汽车人工智能应急响应管理制度、应急预案等文件以及是否组建应急指挥小组；通过查看应急演练记录、工作表单等文件确认企业是否落实应急响应工作；

分类分级制度内容完整性：评估企业是否具备分类分级制度，是否能够识别数据处理过程中涉及的重要数据，并具备重点保护的能力；

数据共享和对外合作管理制度内容完整性：评估企业是否具备数据第三方共享管理制度，是否具有合同、保密协议等数据共享管理的工作记录；

数据全生命周期管理制度、风险评估管理制度内容完整性、产品数据安全管理制度完整性：评估制度文件中是否包含数据全生命周期各个阶段的安全措施要求及风险识别、溯源等要求；

数据销毁管理制度内容完整性：评估企业是销毁制度中是否具有明确合理的销毁策略；

数据出境管理制度完整性：评估企业是否在数据出境管理制度中，明确数据出境的管理要求；

用户权益响应管理制度：评估企业是否明确了用户权益响应负责人和联系方式，在用户权益响应管理制度中，是否明确了用户权益响应的流程和工作方法。

(3) 生成式 AI 数据安全评估

1) 训练数据评估

安全风险评估: 通过 AIGC 数据集安全风险评估，识别潜在威胁并评估其影响程度，依据风险评估结果，制定相应的数据安全策略，明确测试范围、方法及预期目标。

数据标注: 通过查看数据标注任务流、审核记录等工作记录，确认数据标注具备审核机制，标注规则至少包括标注目标、数据格式、标注方法、质量指标等

汽标委智能网联汽车分委会

内容，数据标注严格依照标注规则进行。

数据使用协议: 与数据提供方签订数据使用协议，明确数据的使用范围和保密义务。

法律遵从性: 严格遵守《中华人民共和国个人信息保护法》等相关法律法规，确保数据处理活动合法合规

风险评估与应急预案：定期开展训练数据安全风险评估，识别和分析数据安全风险。同时，制定数据安全应急预案，明确应急响应流程和措施，提高应对数据安全事件的能力

持续监控与审计：建立监控审计机制，对数据处理活动进行持续监控和审计。通过日志管理平台收集系统和访问者操作日志，实现对数据处理活动的全面监控 。

人员培训：加强对员工的数据安全意识培训，提高员工的安全防范意识和技能。鼓励员工积极参与数据安全管理工作，形成全员参与的数据安全文化 。

隐私影响评估：定期进行隐私影响评估，识别和评估数据处理活动中的隐私风险。

知识产权评估：通过开源协议、合作协议查看等方式，确认 AIGC 模型训练数据不存在侵犯知识产权的情况。

2) 模型安全评估

模型安全基本要求评估：审查汽车行业大模型备案情况，已上线的生成式 AI 应用或功能，应在显著位置或产品详情页面公示所使用已备案生成式 AI 服务情况，注明模型名称及备案号。

模型可解释评估：针对模型的解释结果与预测结果之间的一致性评估；用户的主观性评价；算法执行性能和效率评价；以及模型稳定性评估

模型升级安全：通过开发文档、管理文档、OTA 备案信息查阅等方式，确认车端 AIGC 模型更新、升级时采取了管理策略

汽标委智能网联汽车分委会

伦理安全评估：通过诱导、伦理题库等方式评估 AIGC 的生成内容是否存在有悖于道德伦理有关的内容。

3) 生成内容安全评估

文本审核：支持昵称、聊天和生成式 AI 指令审核，识别隐私信息、涉黄、暴恐、违禁敏感内容、伦理内容等，能够识别内容变形变异，关联用户上下文。

图像审核：支持图片和 AI 绘图审核，识别人脸车牌等隐私信息、涉黄、暴恐、伦理问题等内容，能够区分动漫软色情尺度，检测图片和 AI 绘图画面中文字，提供丰富的标签管理。

视频审核：支持视频画面审核，识别人脸车牌等隐私信息、涉黄、暴恐、伦理问题等内容，能够区分动漫软色情尺度，检测视频画面中文字，提供丰富的标签管理。

语音审核：支持音频文件和语音流审核，识别语音隐私信息、涉黄、辱骂、伦理问题等内容，检测静默语音和呻吟声，提供转写后的文字和功能辅助人工审核。

生成内容可靠性评估：通过语义分析、题库测试、人工审核等方式，确认 AIGC 生成的内容具备可靠性；能准确响应输入者的输入意图；同时符合科学常识或主流认知。

举报与处置机制评估：主要包括投诉举报、账号处罚和内容处罚等。

审核优化机制评估：主要包括词库管理、检测范围管理和人工审核等。

9.4.3 评估工具

(1) 体系评估工具：常用文档工具内嵌体系评估要求，通过文档审查、人员访谈等方式确认体系内容是否具备完整性。

(2) 数据质量评估工具：评估工具应具备对数据集进行多维度质量指标分析的能力。这意味着工具需要能够量化并评估数据的规范性、完整性、准

汽标委智能网联汽车分委会

确性、一致性。评估工具应能够分析数据集的多样性和平衡性，确保数据能够全面代表目标群体或真实世界的情况。评估工具需要提供评估数据标注质量的功能，包括通过阈值判断和真值比对等方法来验证标注的准确性。

9.5 总结与展望

当前，车用 AI 数据安全测试与评估在法规、标准、技术及产业协同方面已取得显著进展，并且随着智能网联汽车渗透率提升，AI 技术驱动下的自动驾驶、智能交互与出行服务场景不断深化，数据安全需求日益迫切。在法规层面，以欧盟《人工智能法案》、中国《汽车数据安全管理若干规定》为代表的法规政策逐步完善，从数据分类、隐私保护到跨境流动等领域加强管理；在标准层面，国际标准化组织（如 ISO、ITU）与国内机构加快制定 AI 数据质量、安全治理相关标准，推动行业规范化发展；在技术研发层面，基于数据机密性、数据完整性、数据可用性等数据安全防护技术，继续在人工智能领域数据安全对抗样本生成、数据投毒和模型抗干扰等技术方向进行探索；在产业协同方面，通过 AI 赋能车-云-管-端全链条产业的参与主体，形成可持续发展、相互协作的数据安全生态。

然而，智能网联汽车的车用 AI 数据安全当前存在诸多待解问题，一是基于 AI 模型的数据安全测试，在传统数据安全的完整性、保密性、个人隐私及合规性等方面，缺乏统一测试标准，二是在测试和评估技术上，智能网联汽车人工智能面临测试场景多样化生成困难、相关工具缺失、软硬件集成测试棘手等问题，三是在 AI 模型安全方面，缺乏有效的评估方法。可以基于这些问题进行生态搭建与技术迭代，将车用 AI 的数据安全防护从被动防御转向主动免疫，为智能交通系统的可持续发展构筑可信底座，最终实现安全、高效、人性化的智慧出行愿景。

汽标委智能网联标准分委会

第十章 车用 AI 信息安全测试与评估标准化研究

10.1 车用 AI 信息安全测试与评估现状

10.1.1 车用 AI 信息安全测试与评估的意义和目的

人工智能在智能网联汽车中的应用迅速发展，涵盖自动驾驶、智能车载助手、智能安全系统等领域。然而，AI 技术的发展伴随信息安全风险，训练过程中需采集大量数据，若数据遭受攻击，将影响行驶安全[36]。

AI 信息安全研究旨在防范攻击，确保系统稳定运行。信息安全测试与评估结合检测与测试，对基础设施、核心技术和应用进行安全评估，以防深度伪造、对抗样本、算法后门等风险，保障数据和系统的完整性、保密性、可用性、透明性及隐私性[42]。此外，测试有助于发现漏洞、确保数据安全、验证安全策略、提升用户信任并符合法规。

近年来，汽车厂商和零部件供应商加大对 AI 信息安全的研究，以降低经济损失和社会影响。风险评估测试在研发阶段识别潜在风险，指导安全防范与监管[37]，从而提升安全管理能力，推动技术创新。

10.1.2 车用 AI 信息安全测试与评估的定义与发展

AI 信息安全测试与评估是对智能网联汽车 AI 系统的全面检查，确保其安全运行，降低黑客攻击、恶意软件和数据泄露的风险。该过程涵盖算法、模型、数据集的安全评估，防止数据篡改或窃取，保障信息安全不影响车辆与乘客安全。

通过测试与评估，可识别并修复安全漏洞，增强系统对威胁的防御能力，确保智能网联汽车的整体安全性。

10.1.3 车用 AI 信息安全测试与评估的问题和难点

(1) 敏感数据信息安全：车辆涉及用户数据、VIN 码、轨迹等敏感信息。AI 组件在训练时可能泄露用户隐私，如何确保信息安全成为挑战。

汽标委智能网联汽车分委会

(2) 对抗攻击信息安全：新型攻击手段层出不穷，对抗样本可误导 AI 组件，导致感知、决策错误。需通过安全测试识别并防御此类攻击，确保 AI 系统稳定性。

(3) AIGC 信息安全：大模型技术推动 AIGC 在车载应用落地，如智能语音助手、图像生成等。如何测试并确保 AIGC 内容的合规性是关键挑战。

(4) 存储传输信息安全：AI 模型及数据在云端存储或传输时，易被篡改或窃取。需确保云端数据安全及传输过程不受攻击，以保障 AI 系统的完整性。

10.1.4 车用 AI 信息安全测试与评估的应用情况

10.1.4.1 概述

随着汽车智能化发展，AI 技术广泛应用于智能网联汽车，如辅助驾驶功能，但也带来新的信息安全问题。作为智能网联汽车信息安全的子领域，AI 信息安全正逐步受到关注，但尚未得到充分重视。目前，无论是政策法规制定，还是整车厂与供应商的研究实践，对 AI 信息安全的测试和评估仍处于初期阶段。

10.1.4.2 法规进展

信息安全对于智能网联汽车至关重要，国内外均积极开展智能网联汽车信息安全相关的法规和标准制定工作，涵盖了智能网联汽车的管理体系和产品体系。

国际已发布的相关法规包括 UNECE-R155（信息安全）和 UNECE-R156（软件升级）。R155 强调建立网络安全管理体系（CSMS），R156 要求建立标准化软件更新管理体系（SUMS），确保车辆生命周期内的安全。我国已发布 GB 44495-2024《汽车整车信息安全技术要求》，为智能网联汽车的信息安全研发提供指导。

10.1.4.3 标准进展

智能网联汽车 AI 信息安全尚无专门测试和评估标准。由于人工智能安全与传统信息安全相关，现有方法仍可提供参考。

汽标委智能网联汽车分委会

当前，汽车信息安全测试主要依据 ISO/SAE 21434 Road vehicles—Cybersecurity engineering，该标准提供系统化框架，涵盖汽车生命周期各阶段的安全保障。国内已正在制定基于该标准转化的《道路车辆 信息安全工程》。

目前国内外围绕人工智能正在积极开展标准研制工作，推动人工智能应用与行业的进一步发展，涉及人工智能信息安全的主要标准如下[43,44]。

(1) 国际标准

1) AISO/IEC JTC 1

国际标准组织/国际电工委员会的第一联合技术委员会(ISO/IEC JTC1) 依托 SC42（人工智能分委会）和 SC27（信息安全分委会）开展人工智能安全标准化工作， 目前涉及人工智能安全的标准项目如下：

表 10-1 ISO/IEC JTC 1 总结表

负责/归口	标准编号	标准名称	阶段
ISO/IEC JTC1 SC42	ISO/IEC AWI TS 29119-11	《Software and systems engineering — Software testing Part 11: Testing of AI systems》 《软件和系统工程——软件测试 第 11 部分：人工智能系统测试》	在研
ISO/IEC JTC1 SC27	ISO/IEC TR 27563:2023	《Security and privacy in artificial intelligence use cases — Best practices》 《人工智能用例中的安全和隐私——最佳实践》	发布
ISO/IEC JTC1 SC27	ISO/IEC CD 27090	《Cybersecurity — Artificial Intelligence — Guidance for addressing security threats and failures in artificial intelligence systems》	在研

汽标委智能网联汽车分委会

		《网络安全 - 人工智能 - 解决人工智能系统安全威胁和故障的指南》	
ISO/IEC JTC1 SC27	ISO/IEC WD 27091.2	《Cybersecurity and Privacy — Artificial Intelligence — Privacy protection》 《网络安全与隐私 — 人工智能 — 隐私保护》	在研

2) ETSI

欧洲电信标准化协会（ETSI）于 2019 年成立了行业规范组-保护人工智能（Industry Specification Group on Securing Artificial Intelligence，ISG SAI），致力于人工智能安全标准化，目前已发布了多项人工智能安全报告，具体如下：

表 10-2 ETSI 总结表

负责/归口	标准编号	标准名称	阶段
ETSI	ETSI GR SAI 001	《AI Threat Ontology》 《人工智能安全：人工智能本体威胁》	发布
ETSI	ETSI GR SAI 002	《Data Supply Chain Security》 《人工智能安全：数据供应链安全》	发布
ETSI	ETSI GR SAI 003	《Security testing of AI》 《人工智能安全：人工智能安全测试》	草案
ETSI	ETSI GR SAI 004	《Problem Statement》 《人工智能安全：问题陈述》	发布
ETSI	ETSI GR SAI 005	《Mitigation Strategy Report》 《人工智能安全：缓解策略报告》	发布
ETSI	ETSI GR SAI 006	《The role of hardware in security of AI》 《人工智能安全：硬件在人工智能安全中的作用》	发布

3) CEN-CENELEC

欧洲标准化委员会（CEN）、欧洲电工标准化委员会（CENELEC） 成立了

汽标委智能网联汽车分委会

新的CEN-CENELEC联合技术委员会JTC 21“人工智能”，负责制定和采纳人工智能和相关数据的标准，旨在解决人工智能网络安全的扩展问题，包括可信特性、数据质量、人工智能治理和人工智能管理系统等。为支撑《人工智能法案》的落地，JTC 21 开展了多项标准提案[45]，具体如下：

表 10-3 CEN-CENELEC 总结表

负责/归口	标准名称	阶段
CEN-CENELEC	《AI risk catalogue and risk management》 《人工智能系统风险目录和风险管理》	在研
CEN-CENELEC	《AI trustworthiness characterization》 《人工智能可信特性》	在研
CEN-CENELEC	《 Accuracy of NLP (Natural Language Model) systems》 《NLP（自然语言模型）系统的准确性》	在研

4) ENISA

ENISA （European Union Agency for Cybersecurity欧洲联盟网络安全局）已发布了多份AI网络安全相关的技术报告，具体如下：

表 10-4 ENISA 总结表

负责/归口	标准名称	阶段
ENISA	《Artificial Intelligence Cybersecurity Challenges: Threat Landscape for Artificial Intelligence》 《人工智能的网络安全挑战：人工智能威胁图谱》	发布
ENISA	《Cybersecurity Challenges in the Uptake of Artificial	发布

汽标委智能网联汽车分委会

	Intelligence in Autonomous Driving》 《自动驾驶人工智能应用面临的网络安全挑战》	
ENISA	《Securing Machine Learning Algorithms》 《保护机器学习算法》	发布
ENISA	《Artificial Intelligence and Cybersecurity Research》 《人工智能与网络安全研究》	发布
ENISA	《Multilayer Framework for Good Cybersecurity Practices for AI》 《人工智能良好网络安全实践的多层框架》	发布

5) NIST

美国国家标准与技术研究院（NIST）已发布的人工智能安全标准如下：

表 10-5 NIST 总结表

负责/归口	标准编号	标准名称	阶段
NIST	NIST SP 1270	《Towards a Standard for Identifying and Managing Bias in Artificial Intelligence》 《建立识别和管理人工智能偏差的标准》	发布
NIST	NIST IR 8312	《Four Principles of Explainable Artificial Intelligence》 《可解释人工智能的四大原则》	发布

(2) 国内标准

2020 年 7 月，国家标准化管理委员会、中央网信办、发展改革委、科技部、工业和信息化部联合印发了《国家新一代人工智能标准体系建设指南》，推动人

汽标委智能网联汽车分委会

工智能产业技术研发和标准制定，标准研制范围涵盖了人工智能的安全和隐私保护、伦理等部分[48]。

2023 年，我国发布了首个人工智能安全国家标准 GB/T 42888-2023《信息安全技术 机器学习算法安全评估规范》。该标准明确了机器学习算法技术和服务在整个生命周期内各阶段的安全要求、评估方法及安全评估流程。

2024 年 9 月，《生成式人工智能服务管理暂行办法》正式生效，旨在促进生成式人工智能的健康发展和规范应用。该办法对在中国境内向公众提供生成文本、图片等内容的服务进行管理，强调发展与安全并重，并实施包容审慎、分类分级的监管。其内容包括鼓励技术创新与基础技术自主研发，同时要求服务提供者履行相应责任，如使用合法来源数据进行训练、与用户签订协议、保护个人信息、标识生成内容等。各相关部门将依据职责加强监督检查，并对违规行为依法处罚。

此外，为规范人工智能赋能各行业的技术要求，行业标准也在积极制定和发布中，如已发布的JR/T 0221-2021 《人工智能算法金融应用评价规范》，规定了人工智能算法在金融领域应用的基本要求、评价方法、判定准则[49]。

(3) 标准进展小结

现有人工智能安全标准仍处于初级阶段，主要侧重基础性和通用性，尚缺乏针对智能网联汽车人工智能信息安全的规范。因此，应加快研究相关测试与评估标准，以明确人工智能产品和服务的安全水平，并为智能网联汽车人工智能安全技术的发展提供指导。

10.2 车用 AI 信息安全测试与评估通用要求标准化研究

10.2.1 车用 AI 信息安全测试与评估通用要求标准化目的

人工智能在智能网联汽车中的应用日益广泛，数据处理和交互规模激增，带来更高的信息安全风险。然而，目前汽车行业尚缺乏 AI 信息安全测评标准。建立相关通用要求的标准化，有助于指导产业链企业研发、规范测试流程，并提升 AI 技术的信息安全防护与风险控制能力。

标标委智能网联汽车分委会

10.2.2 车用 AI 信息安全测试与评估评价基本原则

为确保车辆安全性、用户隐私保护及产业健康发展，AI 信息安全测试需遵循以下原则：

(1) 安全优先原则：在设计、开发、测试与部署全生命周期中，安全需求优先。

(2) 全面性原则：覆盖数据保护、网络安全、软件完整性、硬件可靠性及用户交互安全。

(3) 系统性原则：考虑车辆内部组件交互及外部环境（如其他车辆、基础设施、行人等）的影响。

(4) 动态性和实时性原则：安全测试应持续进行，实时监控并响应新威胁。

(5) 合规性原则：遵循数据保护法规、隐私法律及行业标准。

(6) 透明性和可追溯性原则：测试过程与结果应可追溯，保障利益相关方知情权。

10.2.3 车用 AI 信息安全测试与评估对象及要求

10.2.3.1 模型开发环境要求

确保安全开发环境和工具、保证第三方组件可信赖、系统权限和访问控制、定期进行漏洞扫描与安全审计、及时更新和维护、建立灾难恢复与备份策略。

10.2.3.2 训练数据来源要求

模型在应用过程中防范数据投毒、数据还原等攻击，要求：

(1) 公开数据集要求：应进行安全审查、数据标注、多样性及平衡性管理。

(2) 自有数据集要求：除公开数据集要求外，还需满足合规性、隐私保护、访问控制与更新迭代要求。

(3) 预训练模型要求：确保来源可信，进行安全审查并提升模型可解释性。

北京市智能网联汽车分委会

10.2.3.3 模型部署要求

(1) 车辆部署AI模型的系统（含AI模型关键参数）进行完整性验证。

(2) 车辆部署AI模型的系统应具备安全机制防止非授权的特权访问。

(3) 车辆部署AI模型的系统（包含但不限于操作系统、固件、依赖库）不应存在由权威漏洞平台 6 个月前公布且未经处置的高危及以上的安全漏洞。

10.2.3.4 模型更新要求

(1) 模型参数更新系统需具备篡改防护机制。

(2) 参数更新时，系统与云平台通信时应验证通信对象身份真实性。

(3) 车辆应对传输的AI模型参数进行真实性和完整性验证。

(4) 应具备模型参数更新日志，存储时长不少于 6 个月。

10.2.3.5 模型安全要求

(1) AI可解释性要求

AI 的可解释性直接关系到安全性。尽管人工智能在诸多任务中表现优异，但其工作原理仍不透明，影响用户信任。对于智能网联汽车而言，增强 AI 模型的可解释性有助于提升用户接受度，并帮助研发人员优化模型。

可解释性需求源于问题的形式化不完全性，仅有预测结果往往不足，还需解释其推理过程 [50]。高度可解释的模型更易被理解和信任，同时有助于识别潜在错误，提高系统可靠性。研究表明，神经网络可能出现意想不到的错误，如对抗攻击，影响高可靠性系统的安全。

可解释性方法可从以下维度分类：

1) 解释阶段：

固有:改变模型网络或者训练过程使模型本质上有可解释性。

事后:开发可解释性技术解释已训练好的模型。

汽标委智能网联汽车分委会

2) 解释范围：

局部:解释每一个输入样本的决策过程和决策依据（针对单个样本）。

半局部：解释模型部分模块的决策特征（针对特定类别）。

全局:解释整体网络架构的逻辑以及内部的工作机制。

3) 解释方式：

敏感性分析:衡量输入样本的特征，解释模型对不同特征的敏感性。

样本原型:提供近似的原型或者样本，表示模型对特定类别的决策依据。

局部近似:对模型的局部信息进行近似，表示模型在局部的决策依据。

特征分解:使某些神经元/网络层表达的内容更直观化从而被我们理解。

概念分析:提取神经网络学习到的抽象语义。

规则提取:抽取模型的逻辑规则，表示模型的决策过程。

(2) AIGC安全要求

AIGC 安全评测涉及文本、图像等生成内容的风险评估，通过多轮对话、问答等方式训练评价模型，输出合规性和准确性测评报告。测评系统应具备对抗演练、自动化、全面扫描和标准化衡量能力。

AIGC 安全检测原则包括：

1) 数据可控：包括训练数据去毒、用户输入去毒、模型输出去毒等；

2) 训练可控：包括大模型解构、RLHF/RRHF、风险对抗训练等方面；

3) 推理可控：基于逻辑图谱的知识融合、模型可解释、可靠性等方面。

检测内容涵盖 NLP 和多模态大模型，包括：

1) 标准识别：建立内容安全、数据安全、伦理安全检测标准，如人工标注、AIGC 识别、分类识别等。

2) 测试用例：测试用例可通过人工、模板、诱导、二次改写等方式生

汽标委智能网联汽车分委会

成，分为普通对话、敏感话题、单轮和多轮诱导。

AIGC安全检测形式包括：

1) 饱和式攻击，支持万量级的攻击；

2) 准自动化，自动化率过半；

3) 多攻击手段的题库。

(3) AI模型应用安全要求

本节重点关注模型在应用阶段的安全性和鲁棒性，确保自动驾驶指令的可靠性，防止 AI 模型被攻击或误导。

主要防护措施包括：

1) 功能正确性测试：通过压缩、噪声等方式测试模型对有损数据的适应性。

2) 对抗攻击测试：黑盒攻击下模型准确率需满足特定阈值。

3) 数据投毒与后门攻击测试：攻击成功率需低于特定阈值。

自动驾驶系统受攻击可能导致功能失灵、误判甚至远程操控，危及人身和财产安全。因此，车用 AI 模型的鲁棒性和安全性至关重要。然而，训练数据覆盖不足易导致模型在恶意样本下失效。

车用 AI 面临的安全风险包括：

1) 对抗样本攻击：攻击者通过微小修改输入，使 AI 误判。例如，有研究者通过涂改路标，使 AI 误将“禁止通行”识别为“限速 45” [51]。目前建议重点关注黑盒物理攻击的测试与防御。

2) 数据投毒：攻击者可注入精心设计的样本，污染训练数据，影响 AI 系统正常运行。

3) 后门攻击：攻击者可植入隐蔽后门，使模型在特定触发条件下被控制。由于神经网络仅由参数构成，难以通过代码审查发现后门，隐蔽性极

浙江委智能网联汽车分委会

高。

10.2.3.6 数据传输安全要求

数据传输安全涵盖训练、验证、测试数据以及模型部署数据的安全性，确保机密性、完整性和可用性，降低数据泄露和篡改风险。

要求包括：

(1) 机密性与访问限制：要求确保在数据传输过程中的机密性，限制对数据传输的访问权限，即只有授权的用户可以访问和解读传输的数据。

(2) 完整性：要求确保数据传输过程中的完整性，防止数据在传输中出现任何篡改、修改或损坏。

(3) 隐私保护：要求确保数据在流转过程中，不出现数据窃取等问题。

(4) 传输可用性：确保数据在传输过程中不会受中断、延迟或其他问题影响，导致数据传输失败。

此外，车辆AI模型与云平台通信应验证其通信对象的身份真实性，车辆AI模型不应直接向境外传输数据。

10.3 车用 AI 信息安全测试工具标准化研究

10.3.1 车用 AI 信息安全测试工具标准化目的

标准化有助于提升车用AI信息安全测试工具的可靠性和有效性，提高识别和防御潜在威胁的能力。主要体现在以下方面：：

(1) 保证安全性：标准化测试工具可评估汽车 AI 系统的安全性，如对抗样本攻击、数据劫持等，确保系统鲁棒性。

(2) 促进技术发展：通过统一的测试标准，加速车用 AI 安全技术的研发和产业化。

(3) 增强用户信任：透明的测试流程提高消费者信心，促进智能网联汽车的

汽标委智能网联汽车分委会

普及。

(4) 响应技术发展与市场需求：信息安全测试标准需随技术进步和市场需求变化不断更新。

10.3.1.1 对抗样本攻击测试工具标准化目的

当前对抗样本攻击工具众多，但普遍存在性能不足、开发不规范、接口不统一、功能重叠等问题，影响车用人工智能信息安全测试体系的构建。标准化可整合各子模块，明确工具测试内容，统一接口，提高协调性，规范流程，推进安全评估体系的建立。同时，特色标准能体现工具集在车用AI中的应用特点。

10.3.1.2 AIGC 信息安全测试工具标准化目的

2024 年 4 月，世界数字技术院（WDTA）发布《生成式人工智能应用安全测试标准》，但完整的测试工具链仍未形成。现有工具协调性差，测试标准缺失，专门针对AIGC的安全测试工具较少。标准化可确保评估一致性和可信度，为监管提供基准，增强用户信任，促进技术发展与应用。

10.3.2 车用 AI 信息安全测试工具分类

10.3.2.1 对抗样本攻击测试工具分类

对抗样本攻击测试是智能网联汽车AI信息安全中的关键部分，主要评估智能模型在面对旨在误导或欺骗AI系统的输入数据时的反应。通过攻击算法生成对抗样本，测试智能网联汽车AI系统的安全性，帮助开发者发现潜在漏洞并增强防御能力，确保系统在实际应用中能抵御各种攻击。对抗样本测试不仅提升智能网联汽车安全性，还推动AI安全技术的发展，构建更安全、可靠的智能交通基础。

学术界对对抗样本生成主要有几种方式，如基于梯度优化攻击、基于查询的攻击、生成模型训练攻击及传统遗传算法攻击。考虑到攻击效果与计算开销，基于梯度优化的攻击是主流方式。

基于优化的攻击大致有两种思路：一种是直接在原始输入上施加扰动，优化目标函数生成对抗样本；另一种是通过间接方式对目标网格或纹理进行扰动，再

汽标委智能网联汽车分委会

通过优化目标函数生成对抗样本。两者的差异体现在目标函数、扰动限制、扰动模态及数据流。

(1) 按目标函数分类的优化攻击分类：优化攻击本质上是求解优化问题。例如，PGD 攻击在扰动范围内通过步长调整寻求损失函数最大化的扰动方向。调整损失函数可引入先验条件，如在 2D 图像补丁优化中加入不可打印损失和光滑损失，以生成易于物理实现的平滑对抗补丁。此外，在目标检测任务中，最大化预测框回归损失或分类误差均可导致严重影响。

(2) 按扰动限制的优化攻击分类：优化攻击通常需满足特定约束。例如，全局扰动需在肉眼不可察觉范围内，而局部扰动则可集中于特定区域，如通过打印方式实现。不同的先验限制决定了攻击算法的分类。

(3) 按扰动模态的优化攻击分类：优化问题由目标函数和约束条件构成，变量的不同定义适用于不同场景。例如，PGD 算法可用于 RGB 图像或二维点云，若优化问题对模态无关，则可视为不同攻击类别。

(4) 按数据流的优化攻击分类：对抗样本的生成依赖特定数据流，从正常样本转换为对抗样本的方式决定了攻击类别。按数据流分类可确保所有攻击方法均被考虑。

10.3.2.2 AIGC 信息安全测试工具分类

机器学习测试（Machine Learning Testing）旨在评估和验证机器学习模型及其开发、部署和运行环境，以识别、缓解或消除安全威胁，包括模型透明性、稳健性、公平性和安全性，如对抗样本抵御能力、偏见检测及敏感信息泄露防护。其不同于软件测试或模型构建过程中准确率的测试。

在人工智能，尤其是生成式人工智能（AIGC）中，安全测试涉及人工智能响应处理（AI Response Handling），即对人工智能生成内容进行准确性、相关性、无害性、隐私性和伦理对齐的评估。AIGC安全测试工具需评估生成内容的可靠性、法律与伦理合规性，防止虚假或有害信息的产生。

汽标委智能网联汽车分委会

AIGC安全测试工具主要具备以下能力：

(1) 模型和数据验证能力：工具应能检查数据的质量和安全性，检测数据投毒、后门等，并验证模型参数和结构是否符合预期，避免安全漏洞。

(2) 攻击检测和防护能力：工具应识别和防护常见攻击，如对抗样本攻击，并提供机制发现潜在安全隐患。

(3) 内容审查机制：工具应检测和评估生成内容中的有害、敏感或不当内容，支持定制化规则，并提供纠正或减轻方法。

工具可按以下方式分类：

(1) 按测试对象：包括数据测试（数据完整性、代表性、投毒检测）、算法测试（安全性评估）、部署框架测试（TensorFlow 等的安全性）。

(2) 按测试指标：涵盖鲁棒性（对抗攻击抵御能力）、安全性（隐私泄露）、公平性（偏见检测）及伦理对齐（法律合规性）。

(3) 按联机状态：分为离线测试（基于历史数据分析）与在线测试（实时交互监测）。

10.3.3 车用 AI 信息安全测试工具现状

10.3.3.1 对抗样本攻击测试工具现状

当前，对抗样本攻击测试软件涵盖多种AI应用场景，如TextAttack、AdvBox、Counterfit、Cleverhans、OpenAttack、DeepRobust、Torchattacks、Foolbox和ART。

TextAttack是一个用于NLP对抗攻击、训练和数据增强的Python框架，简化了对NLP模型鲁棒性测试，支持任务如句子编码、语法检查和单词替换。

Counterfit是一个通用自动化工具，集成多个对抗框架，允许用户创建自定义框架，专注于机器学习系统的安全性评估。

Cleverhans是由Ian Goodfellow团队开发并开源的攻防平台，模块化的攻防算法支持全球研究者开发对抗样本生成和防御算法，但版本较老，长时间未更新。

汽标委智能网联汽车分委会

OpenAttack是一个Python工具，涵盖文本对抗攻击全过程，包括模型访问、对抗样本生成、攻击评测及训练，能暴露模型弱点，提高其鲁棒性和可解释性。

DeepRobust是一个基于PyTorch的对抗库，专注于图像和图形的攻击与防御。

Torchattacks是一个PyTorch库，提供多种对抗攻击功能，简化对抗示例生成。

Foolbox是一个Python库，便于对深度神经网络等模型进行对抗性攻击，支持PyTorch、TensorFlow和JAX中的模型。

ART（Adversarial Robustness Toolbox）是一个多功能AI模型对抗测试工具，支持多个主流框架，如TensorFlow、PyTorch等，提供攻击、防御与评估工具，适用于多种数据类型和任务。但作为综合工具，ART难以完美适配车用AI的特定需求，尤其是自动驾驶场景下的智能模型攻击与加固。

10.3.3.2AIGC 信息安全测试工具现状

针对生成式人工智能（AIGC）及其集成的机器学习算法的安全测试工具，主要包括鲁棒性测试、对抗性测试和合规性检查。这些工具既采用传统软件测试方法（如模糊测试、变异测试），也结合机器学习测试方法（如神经元覆盖率）。部分工具可广泛应用于机器学习模型测试，部分则专门针对AIGC和大模型。

(1) 鲁棒性测试工具

目前，人工智能鲁棒性测试工具功能多样，方法各异。代表性的工具包括DeepXplore、DeepMutation、DeepConcolic和TensorFuzz等。

DeepXplore采用神经元覆盖率和差分测试方法，生成特定输入以检测和修复神经网络缺陷，适用于图像分类、语音识别和自然语言处理。

DeepMutation通过变异测试评估模型质量，改进版本如DeepMutation++和MuNN引入更多模型级变异算子。

DeepConcolic结合具体执行与符号分析，生成对抗性输入，增强模型鲁棒性，并已更新至基于贝叶斯网络和模糊测试引擎的版本。

汽标委智能网联汽车分委会

TensorFuzz由Google Brain Research开发，基于覆盖率导向模糊测试（CGF），迭代优化输入以发现漏洞，现阶段支持TensorFlow框架。

(2) 模型合规性检查工具

人工智能模型合规性检查工具关注公平性、可解释性和合法性，主要包括AIF360、AIX360 和AIGC-Safe。

AIF360（AI Fairness 360）开源工具库，评测并减轻模型中的不公平性，适用于经济、医疗等领域。

AIX360（AI Explainability 360）提供DIP-VAE、ProtoDash、CEM等算法，增强黑盒AI模型的透明性和可解释性，适用于高风险决策场景。

AIGC-Safe[52] 由新华网与中科院计算所联合开发，支持文本、图像、音视频的AIGC内容合规性检测，未来可用于虚假新闻、换脸技术和学术诚信评估。

(3) 综合性工具

综合性测试工具涵盖AIGC系统全生命周期安全问题，代表性工具包括BIG-bench、Evals、Responsible AI ToolBox和HELM。

BIG-bench评估大型语言模型推理能力，涵盖自然语言处理测试及社会行为评估，如对齐测试、偏见和毒性检测。其轻量版BBL提供低成本测试选项。

Evals 由OpenAI开发，可评估问答逻辑、幻觉、毒性等功能，并支持自定义评测框架，确保生成内容的安全性与合规性。

Responsible AI ToolBox由微软推出，包含模型表现统计、可解释性评估、错误分析等组件，助力负责任AI开发。

HELM评估语言模型的精度、鲁棒性、公平性，并支持问答、信息检索等任务。其衍生框架HEIM专注于文生图模型的对齐性、原创性及偏差评估。

标委智能网联汽车分委会

10.3.4 车用 AI 信息安全测试工具软件要求

10.3.4.1 对抗样本攻击测试工具软件要求

对抗样本攻击测试工具应涵盖主流攻击算法，如 FGSM（快速梯度符号法）、PGD（投影梯度下降法）和 CW（卡林尼-瓦格纳攻击）。FGSM 通过调整输入数据梯度生成对抗样本，PGD 采用小步长梯度下降并投影至合法数据空间，CW 则基于模型置信度生成更具隐蔽性和挑战性的攻击样本。工具需在主流数据集上验证可靠性，并应至少包含一种创新攻击方法，在车用 AI 等特定场景下，攻击后性能下降至少 80% 以验证有效性。此外，工具应支持安全风险评估，检测 AI 系统在已知与未知威胁下的安全性。

在可靠性与稳定性方面，工具需详细记录测试过程和结果，便于分析与改进，并提供故障恢复机制，确保数据安全，异常事件可追溯，保障测试连续性。设计上应提升可靠性，提供安全、可控的测试环境。

工具界面应简洁直观，支持 GUI 和 CLI，并提供完整的用户文档、操作指南及在线帮助，确保用户理解和使用。同时，工具应兼容主流开发与测试环境，实现与现有工具链和测试平台的无缝集成。此外，工具需符合智能网联汽车行业标准协议和数据格式（如 CAN、LIN、Ethernet、JSON、XML），确保行业适配性，并遵循《汽车数据安全管理若干规定》、ISO 21434、ISO 26262 等信息安全与功能安全标准。同时，工具应定期更新，确保符合最新法规与行业标准。

10.3.4.2AIGC 信息安全测试工具软件要求

AIGC 信息安全测试工具需评估 AIGC 模型的安全性，涵盖多个算法功能。工具应从鲁棒性角度评估数据质量和安全性，验证模型参数与结构的合理性；从对抗安全角度检测攻防能力，评估模型抗攻击能力；从合规性角度检测生成内容中的敏感或不当元素，支持定制化评估，并提供减轻有害内容影响的解决方案，确保社会信息安全。

汽标委智能网联汽车分委会

在可靠性与稳定性方面，工具应确保测试的高准确性和一致性，定期优化算法应对最新威胁，采用标准化流程和自动化框架，并结合语义分析、情感分析、图像识别等技术，实现多维度检测，适配现实场景。同时，工具应在高负载和长时间运行时保持稳定，具备故障恢复机制，避免因非安全因素导致系统崩溃。

在兼容性方面，工具应支持多种操作系统和硬件平台，确保测试工具链的可靠对接与仿真。

在易用性方面，工具需提供直观的 GUI，清晰的导航、帮助提示及错误反馈，并配备详尽的用户手册，涵盖安装、配置、使用及故障排除。联网功能应支持在线客服等多渠道技术支持。

最后，工具需符合智能网联汽车相关行业标准与法规，兼容车载网络通信标准（如 CAN、FlexRay），支持常见协议的数据流解析、检测与分析，保障车载 AIGC 应用的安全。同时，工具需遵循 ISO 26262《道路车辆-功能安全》、ISO/SAE 21434《道路车辆-网络安全工程》等标准，并定期更新以符合最新法规。

10.3.5 车用 AI 信息安全测试工具硬件要求

10.3.5.1 对抗样本攻击测试工具硬件要求

在硬件方面，为了确保所有的攻击算法均能成功应用，部署机器应至少满足以下要求：

(1) 至少包含一张 RTX 30 系列的显卡，显存容量不低于 24GB；

(2) 配备至少 2TB 的硬盘存储空间，以满足数据存储和处理的需求。

(3) 需要满足至少 64GB 内存。

(4) 不低于 256B 固态硬盘作为操作系统盘。

若工具部署于仿真环境，还需满足仿真平台的额外硬件要求。总体而言，标准化的环境配置是确保测试工具性能和可用性的关键，有助于提升自动驾驶系统安全评估的可靠性。

汽标委智能网联汽车分委会

10.3.5.2AIGC 信息安全测试工具硬件要求

生成式人工智能安全测试工具需具备强大运算、存储与传输能力，并尽可能匹配被测模型的真实运行环境。由于测试工具对计算资源需求取决于模型规模、测试目的和任务复杂度，硬件要求可能有所差异，但一般需满足以下标准：

(1) 配备至少 8 核、主频 3.5GHz 的中央处理器，如 Intel Core i9。

(2) 具备至少 8GB 显存的高性能显卡，如 RTX 3080。

(3) 需要至少 64GB 内存，以保证测试能加载程序以及所需数据。

(4) 配备至少 1TB 的硬盘存储空间，以满足数据存储和处理的要求。

若测试工具涉及云计算或分布式计算，还需确保稳定、高速的网络连接。

10.3.6 车用 AI 信息安全测试工具标准化可行性

汽汽车标准化创新基地正在主导构建《汽车信息安全检测实践库》，作为《汽车整车信息安全技术要求》国家标准的配套实施用例方法库。该库将根据智能网联汽车技术发展持续扩充测试方法，并计划建立车用人工智能信息安全对抗攻击算法库。尽管当前学术界已提出多种对抗攻击算法，完全覆盖仍有难度，但核心研究方向较为固定，具备代表性的算法已较为明确，因此可优先推进对抗算法库及车用AI攻击路径库的构建。

目前测试工具需求差异较大，其性能受被测模型结构和数据量影响，难以形成系统性评价。此外，行业智能模型推理接口尚未统一，测试工具接口标准化依赖模型接口的发展进程。因此，建议暂缓测试工具性能与接口标准化工作的推进。

汽标委智能网联汽车分委会

10.4 车用 AI 信息安全测试流程标准化研究

10.4.1 车用 AI 信息安全测试流程标准化目的

智能网联汽车的 AI 信息安全测试流程标准化至关重要，旨在建立严密、系统的测试体系，确保智能化与联网化背景下 AI 系统的安全性，维护公众出行安全、个人信息保护及社会稳定。

10.4.1.1 全面识别安全隐患

智能网联汽车集成自动驾驶、智能导航、V2X 等技术，依赖复杂 AI 系统。然而，软件漏洞、恶意代码、硬件故障等安全隐患可能导致车辆失控或功能异常。标准化测试流程可精准识别未授权访问、数据篡改、系统漏洞等问题，提前采取防范措施，保障车辆安全。

10.4.1.2 保障数据安全

智能网联汽车处理海量数据，包括车辆状态、用户隐私、路况等，其完整性和真实性至关重要。标准化测试流程对 AI 系统的数据处理机制进行严格校验，确保数据在采集、传输、存储和使用过程中不丢失、不被篡改，提升智能决策的可靠性，保障用户权益。

10.4.1.3 构建网络安全防护体系

智能网联汽车面临网络威胁，如黑客攻击、拒绝服务攻击等。标准化测试流程可识别系统防护薄弱点，并通过强化防火墙、加密技术、访问控制策略等手段，提高抗攻击能力，确保车联网稳定运行，维护网络安全。

汽标委智能网联汽车分委会

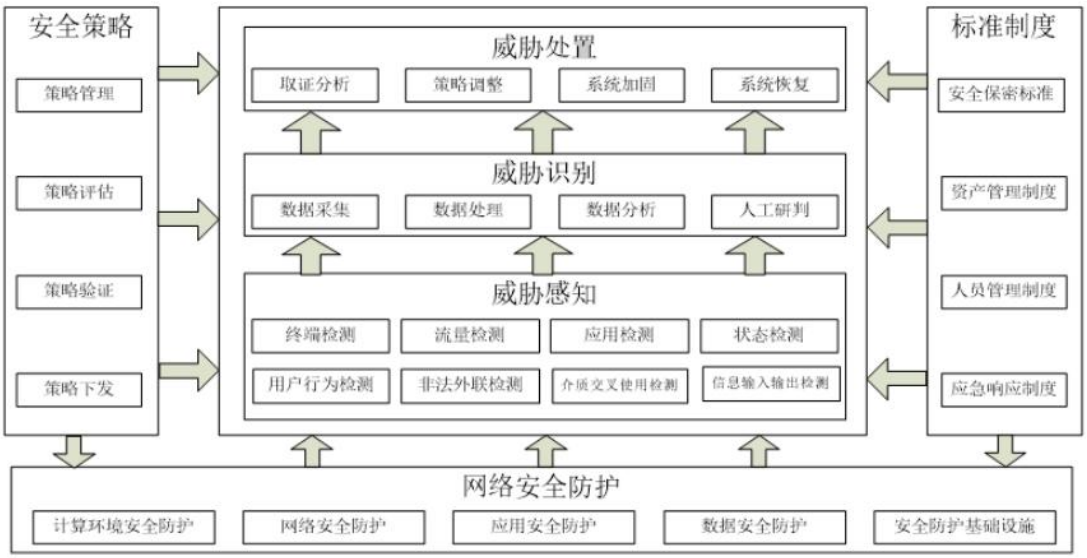


图 10-1 高安全等级网络安全防护体系架构

10.4.1.4 促进技术创新与产业升级

标准化 AI 信息安全测试流程为产业提供技术规范，推动企业加大研发投入，促进技术创新和产业优化，提高生产效率，避免低水平重复建设。同时，有助于监管机构实施有效监督，引导行业健康发展。

10.4.1.5 提升社会整体安全意识

标准化测试流程及其结果的公开，有助于提升公众对智能网联汽车信息安全的认知，增强消费者自我保护能力，营造全社会关注和维护信息安全的良好氛围，推动智能网联汽车的安全发展。

综上，智能网联汽车 AI 信息安全测试流程标准化从安全隐患识别、数据安全、网络防护、产业升级及社会意识提升等多方面入手，构建安全可靠的环境，助力智能、绿色、安全的现代交通体系。

10.4.2 车用 AI 信息安全测试流程及要求

智能网联汽车作为交通领域的前沿创新成果，其 AI 信息安全至关重要。明确科学合理的测试流程及要求，是保障智能网联汽车信息安全的关键。

汽标委智能网联汽车分

10.4.2.1 确定安全目标和风险

首先，应明确智能网联汽车 AI 系统的安全目标，并识别潜在风险。数据隐私是核心，需确保用户敏感信息在采集、存储、传输和使用过程中不被非法获取、篡改或滥用。模型完整性同样重要，应防范恶意篡改或破坏，确保系统稳定运行与精准决策。此外，对抗性攻击也是关键威胁，恶意输入可能误导 AI 产生错误决策，因此须强化 AI 系统抗攻击能力，确保在复杂环境下仍能保持稳定可靠的性能。

10.4.2.2 制定整体测试策略

在明确安全目标和风险的基础上，制定全面的测试策略，涵盖测试方法与环境规划。遵循“研发阶段和公告阶段均应以测试为主，文档检测为辅”的原则。

(1) 测试方案制定策略

智能网联汽车 AI 信息安全测试涵盖多种方法，以全面评估系统安全性：

1) 认证和授权测试：验证用户身份管理及权限分配的准确性。

2) 数据保护测试：确保加解密功能、数据传输和存储的安全性。

3) 安全配置测试：检测默认配置及系统安全性。

4) 输入验证测试：测试用户输入的正确处理，确保系统稳健性。

5) 安全审计测试：验证安全日志及审计功能的完整性和正确性。

6) 安全漏洞测试：发现系统漏洞，如栈溢出、提权攻击等。

7) 安全性能测试：确保系统在面临安全威胁时维持正常性能，测试系统在遭受恶意攻击及进行安全处理时的表现。

(2) 测试环境搭建策略

测试环境的搭建至关重要。测试环境应确保测试车辆能够安全运行。对于可能影响车辆状态的测试，应在多运行工况的整车转鼓环境下进行，以便真实模拟各种复杂的行车工况，全面检测系统在不同条件下的性能和安全性。

汽标委智能网联汽车分委会

同时，测试环境还应保证车辆通信的稳定可靠，且测试过程不会对公共网络环境造成不良影响。对于可能影响公网环境的测试，应在具备通信功能的整车暗室或类似的专业环境中进行，有效隔离测试对公共网络的潜在干扰，确保测试的安全性和合法性。

10.4.2.3 编制详细测试计划

(1) 测试用例细化

基于整体测试策略，可采用以下方法细化 AI 信息安全测试用例：

1) 需求分析法：从系统的功能需求出发，深入挖掘可能存在的安全隐患，并据此设计测试用例。

2) 等价类生成和分析：通过将输入数据划分为等价类，选取代表性的数据进行测试，提高测试的效率和覆盖度。

3) 边界值分析：专注于输入数据的边界情况，通过测试边界值来发现可能存在的异常情况。

4) 基于知识和经验的错误猜测：充分发挥测试人员的专业知识和经验，对可能出现的错误和漏洞进行推测，并针对性地设计测试用例。

(2) 明确通过准则

基于确定的安全目标和细化的测试用例，编制客观、可比的待测样件通过测试的准则。通过准则应明确具体，能够准确衡量测试样件是否达到了预期的安全标准，为评估测试结果提供清晰的依据。

(3) 确认测试周期

基于细化的 AI 信息安全测试用例，明确各测试用例执行的时间顺序和相互之间的时序关联性，从而准确确认整体的测试周期。合理安排测试时间，确保测试过程高效有序，及时发现和解决问题，保障智能网联汽车的信息安全。

汽标委智能网联汽车分委会

10.4.2.4 测试策略的评估

为了确保所制定的测试策略的有效性和合理性，需要采用以下评估方法对测试策略进行评估：

(1) 覆盖度评估：检查测试策略是否涵盖了所有已确定的安全目标和潜在风险。通过对比安全目标和风险清单，确认测试方法和环境规划能够充分覆盖各个方面。

(2) 可行性评估：考察测试策略在实际操作中的可行性。包括评估测试所需的资源（如人力、时间、设备等）是否可获取，测试环境的搭建是否具有现实可操作性。

(3) 有效性评估：分析过往类似测试策略的实施效果，或者进行小规模的预测试，以判断当前策略是否能够有效地发现潜在的安全问题。

(4) 适应性评估：考虑测试策略是否能够适应智能网联汽车技术的不断发展和变化，以及是否能够应对未来可能出现的新的安全风险和挑战。

通过以上评估方法，可以及时发现测试策略中存在的不足，并进行相应的调整和优化，以确保测试策略能够有效地保障智能网联汽车 AI 信息安全。

10.4.3 车用 AI 信息安全测试数据安全要求

人工智能实现的基础离不开海量的数据支撑，AI信息安全测试同样需要大量的特殊数据对AI模型进行安全测试。这些测试数据来源于车辆运行期间能够采集获得海量高精度的关于车辆周边行人、车辆、周围建筑物、实时路况、车内语音等目标的详细信息，也包括在用户使用过程中采集用户账户账号密码、脸部特征、声纹特征、驾驶员操作习惯、应用使用记录、语音图像、行驶轨迹、地理位置、车辆状态等特征[38]。AI信息安全测试需要基于这些数据生成对抗样本，如何保障测试数据安全是AI信息安全体系的重要环节。从数据流上划分，测试数据管理可划分为测试数据采集安全、测试数据存储安全、测试数据处理安全、测试数据

汽标委智能网联汽车分委会

使用和测试数据销毁安全，从管理维度上需要数据使用组织具备完善的数据管理组织建设、制度流程、技术工具、人员能力等[39]。

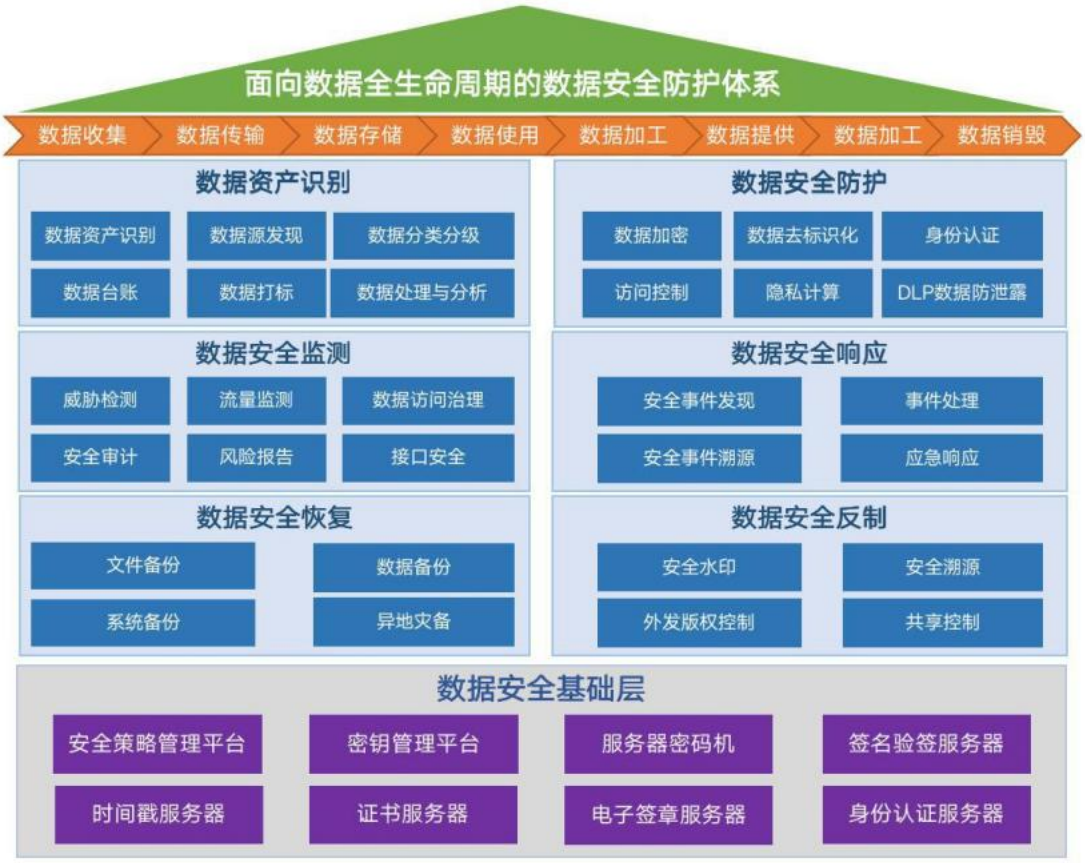


图 10-2 面向数据全生命周期的数据安全防护体系

10.4.4 车用 AI 信息安全测试方法

车用AI系统是现代汽车智能化的核心，包含多个关键部分。首先，传感器系统（如摄像头、雷达、激光雷达等）充当汽车的“眼睛”，实时采集周围环境信息。其次，数据计算平台负责快速处理来自传感器的海量数据。AI算法和模型（如深度学习技术）则分析数据并做出决策，如自动驾驶的路径规划和速度控制。此外，通信模块确保车辆与外部设施及其他车辆的信息互通，提高行车安全与效率。

在国内，汽车信息安全的强制性标准主要是GB44495-2024《汽车整车信息安全技术要求》。该标准第七章提出了外部链接安全、通信安全、软件升级安全及数据安全等基线要求，第八章则给出了相应的测试方法。然而，该标准仅覆盖了

部署环境、依赖和数据传输的安全要求，未涉及传感器系统及AI模型的安全问题。其他推荐性标准如《道路车辆信息安全工程》、《汽车信息安全应急响应管理规范》等，虽然涵盖了车辆研发阶段的信息安全及传统信息安全关键零部件的系统与通讯安全要求，但同样忽视了传感器系统与AI模型的信息安全[53]。

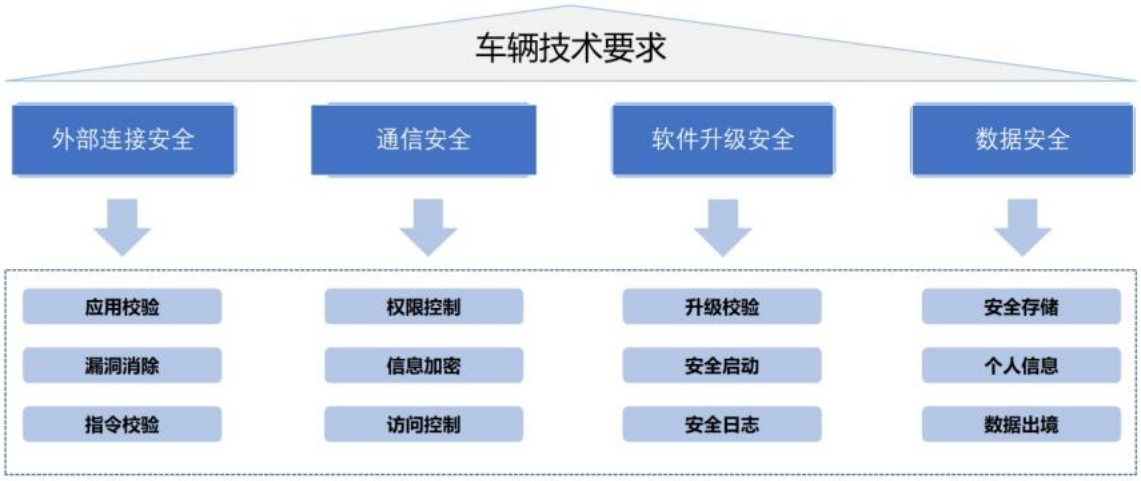


图 10-3 《汽车整车信息安全技术要求》中车辆技术要求的基线范围

在汽车领域，深度学习模型正面临严峻挑战，其中对抗性攻击尤为突出。然而，现行汽车信息安全标准对此尚无明确要求。例如，在图像识别模型中，攻击者可通过微小扰动干扰输入，导致模型误判，进而危及自动驾驶安全。但现有标准对此类攻击的防护与测试缺乏规范。同样，在语音识别模型中，特定频率的声音干扰可诱导系统误解指令，而当前标准尚未提供有效的防范和检测指导。

10.4.4.1 传感器攻击

(1) 传感器干扰攻击

传感器干扰攻击旨在降低数据质量，使传感器无法准确获取信息。例如，强光可干扰激光雷达和超声波传感器，导致测距和探测物体产生误差，影响自动驾驶避障及驾驶员检测功能，严重时可引发安全事故和经济损失。

汽标委智能网联

(a)相机干扰攻击　　(b)激光雷达干扰攻击　　(c)毫米级超声波雷达干扰攻击

图 10-4 传感器干扰攻击

(2) 传感器欺骗攻击

传感器欺骗攻击通过伪造环境物体误导传感器。例如，伪造 GPS 信号会导致定位偏差，影响依赖位置信息的应用；无人机投射欺骗性交通标志可能使自动驾驶系统做出错误决策。这类攻击不仅威胁个人安全，也可能扰乱社会秩序。

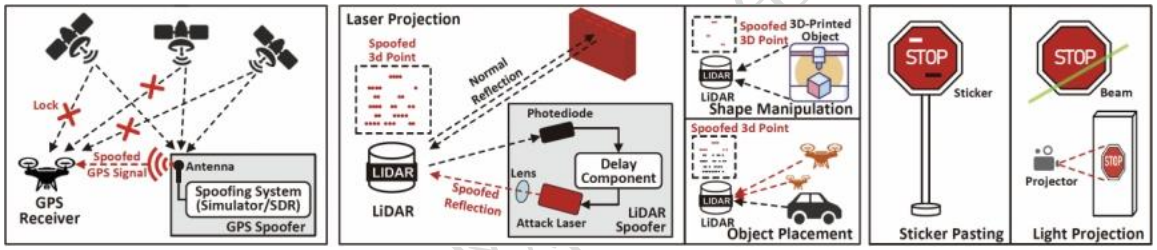


图 10-5 GPS/激光雷达/相机欺骗攻击

(3) 测试评价体系

1) 信号质量评价

评估传感器在干扰条件下的信号质量，包括图像失真、数据误差等。在信号伪造测试中，考察激光、超声波和毫米波传感器的误判率及响应时间。GPS 欺骗测试计算定位偏差，无人机欺骗测试统计自动驾驶系统错误决策比例。

2) 损坏程度评价

分析传感器在激光或高温损伤测试后的外观及功能变化，以及信号欺骗对硬件或软件的损害情况。例如，GPS 欺骗可能导致系统崩溃，无人机欺骗可能影响自动驾驶系统的决策能力。

3) 恢复能力评价

测量传感器在攻击后恢复正常工作的时间及性能下降程度，如相机强光干扰后的图像恢复速度、GPS 信号欺骗后的定位精度恢复。

4) 报警能力评价

评估系统检测攻击并发出警报的能力，如强光干扰、激光损伤、信号伪造或 GPS 欺骗时的报警及时性及准确性，以便用户及时采取措施降低风险。

10.4.4.2AI 模型对抗攻击

(1) 相机感知系统攻击

1) 目标分类攻击

目标分类攻击包括多种复杂且有针对性的手段，其中物理对抗样本攻击较为常见。这种攻击通过在图像中添加精心设计的细微扰动，巧妙地改变图像特征，进而影响模型输出。具体而言，攻击者在物理环境中的正常标志和广告上加入对抗扰动，使自动驾驶汽车误将其错误分类为高可信度的交通标志。

图 10-6 物理对抗样本攻击

对抗补丁攻击则是将与输入无关和与场景无关的对抗补丁放置在现实物体上。比如基于生成模型的对抗补丁攻击，利用先进的生成技术创建出具有高度迷

惑性的补丁，这些补丁具有很强的隐蔽性，可以伪装成涂鸦或贴纸等形式，无论物体处于何种环境或输入条件下，都能有效地干扰模型的识别。

图 10-7 对抗补丁攻击

对抗物体攻击则是构建对抗性的三维物体来欺骗神经网络。像是通过三维渲染嵌入视觉神经网络进行攻击，精心设计的三维物体在经过渲染处理后，能够对神经网络的判断产生误导，使其无法准确识别物体的真实类别。

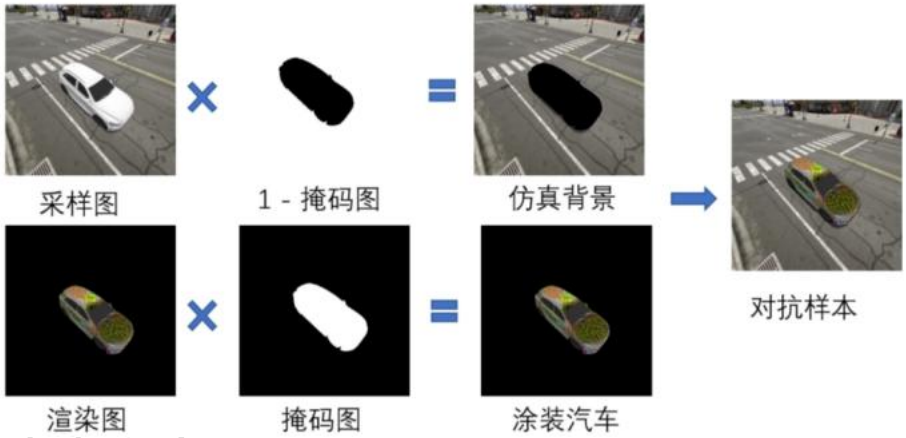


图 10-8 对抗物体攻击

2) 目标检测攻击

目标检测攻击是自动驾驶感知层面临的一种重要安全威胁，主要包括对抗补丁攻击、对抗伪装攻击和投影仪攻击等手段。

对抗补丁攻击通过生成针对目标检测的物理对抗性补丁来实现。这些对抗性补丁可以使停车标志在目标检测器的视野中“隐身”，导致自动驾驶汽车无法及时检测到障碍物，增加了碰撞的风险。

对抗伪装攻击则采用贴纸或喷漆等对抗性伪装来装饰障碍物。CAMOU方法通过近似模仿模拟器将伪装应用于车辆的方式，寻找最佳伪装来最小化近似检测分数，从而使自动驾驶视觉感知系统无法检测到障碍物。如UPC攻击、DAS攻击。

图 10-9 对抗伪装攻击

投影攻击是使用投影仪将对抗性扰动投射在物体上。一种方式是将物体或道路标志投影到自动驾驶车辆行驶的路面上，目的是让自动驾驶系统将这些幻影物体误认为真实，并做出错误反应。另一种方式是通过物理对抗补丁和对抗性样本的攻击形式进行投影，解决投影图案与汽车相机捕获图案之间的颜色差异等问题，生成在现实中能成功进行目标检测攻击的鲁棒对抗扰动。

图 10-10 投影攻击

目标检测攻击的这些手段都具有很强的隐蔽性和危害性，对自动驾驶系统的安全构成了严重威胁，需要采取有效的防御措施来应对。

3) 目标跟踪攻击

目标跟踪攻击旨在劫持跟踪器，影响自动驾驶系统对障碍物的准确跟踪。攻击手段包括生成对抗性纹理或改变目标边界框的位置和形状。例如，在线增量攻击通过利用连续帧在空间和时间上的一致性，生成细微扰动以欺骗跟踪器；3D对抗物体攻击通过构建特定三维物体在仿真场景中干扰跟踪器。此外，AttrackZone方法通过点云和图像信息确定可投影区域，并在该区域投影精心设计的动态扰动，成功将路边车辆移入驾驶车道或将前方车辆移出车道。这些攻击可能导致自动驾驶系统无法准确跟踪障碍物，从而引发安全事故。

原始图像　　攻击后图像

图 10-11 目标跟踪攻击

4) 车道线检测攻击

车道线检测攻击主要分为绘制车道线攻击和脏路补丁攻击。绘制车道线攻击通过在路面上添加小型标记，诱使车道检测模块误判为车道线，进而引导车辆驶向错误方向。脏路补丁攻击则利用路面污渍或修复补丁等隐蔽物，干扰车道检测。路面上的污渍或修复补丁较车道线更为隐蔽，且通常不易被察觉。攻击者可利用车辆运动模型和透视变换，更新摄像头图像，使自动驾驶系统误判车道线，进而

改变车辆行驶路径。这两种攻击均可能导致自动驾驶车辆驶入人行道或非机动车道，严重威胁交通安全。

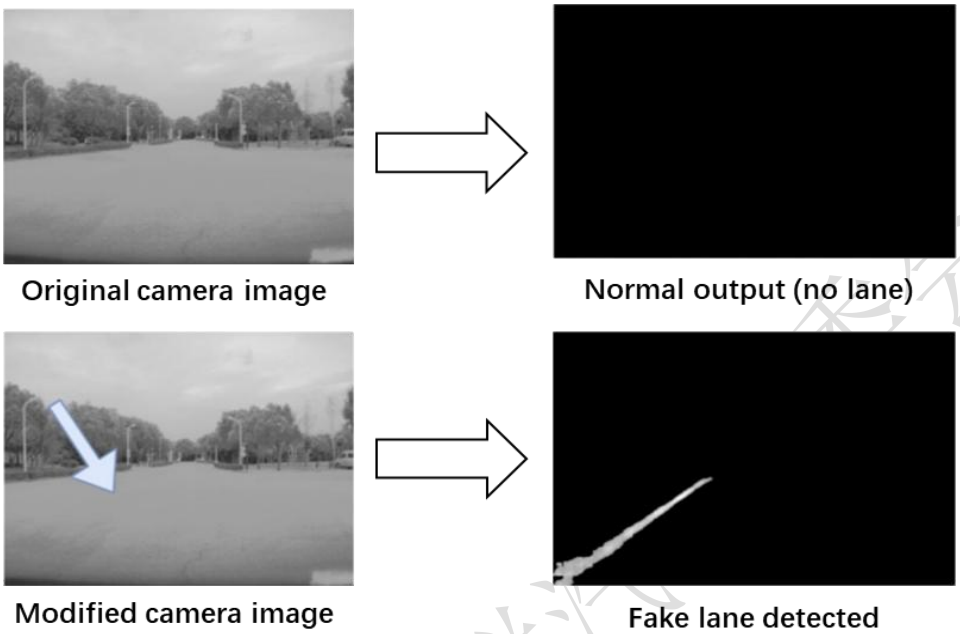


图 10-12 车道线检测攻击

(1) 激光雷达感知系统攻击

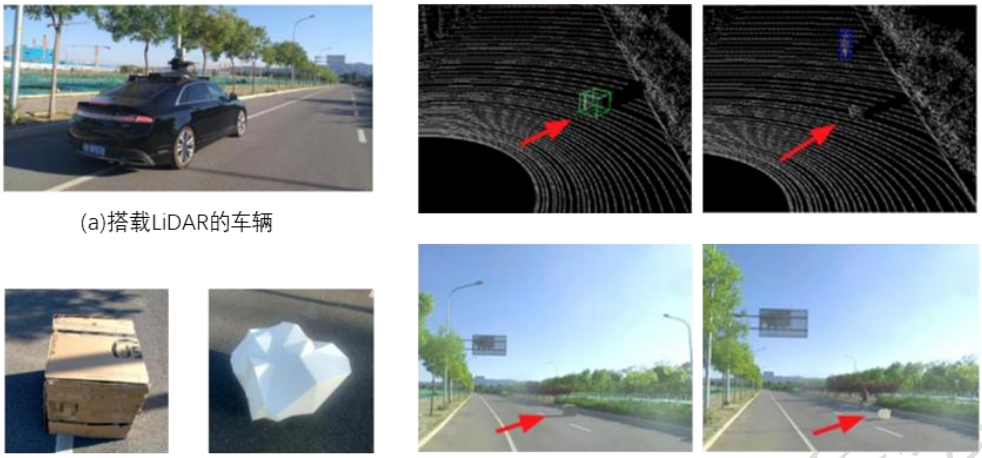
1) 目标检测攻击

目标检测攻击主要包括激光攻击和对抗性物体攻击。

激光攻击通过激光发射器向激光雷达点云图注入欺骗点，导致漏检或误检。优化方法可以确定欺骗点的位置和数量，进而影响目标检测结果，可能导致错误决策和增加碰撞风险。

对抗性物体攻击则通过特定物体欺骗激光雷达系统，例如，使用可微分的激光雷达渲染器模拟扰动，生成对抗性点云。另一方式是将对抗物体放置于车辆顶部或使用反射物体，如无人机，来隐蔽前方障碍物。这些攻击使激光雷达目标检测失效，威胁自动驾驶安全。

汽标委智能网联汽车分会



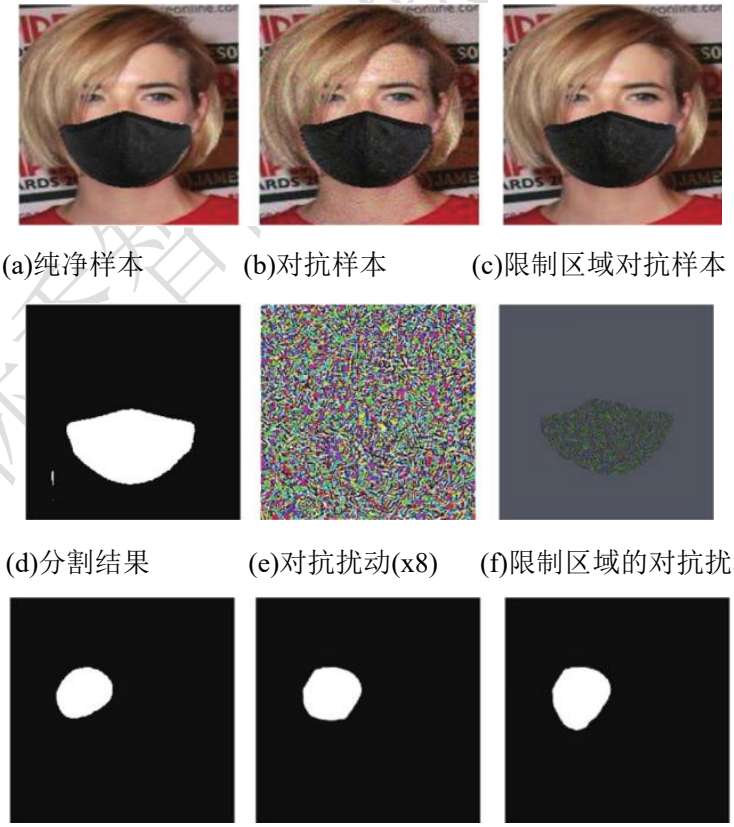
(a)搭载LiDAR的车辆

(b)良性物体　(c)对抗性物体　(d)正常检测良性物体　(e)无法检测对抗性物体

图 10-13 激光攻击

2) 语义分割攻击

语义分割攻击旨在干扰激光雷达对环境的语义理解。攻击者通过特定形状的物体或选择对抗性位置进行攻击，影响物体分类和识别，导致误判物体属性。



(a)纯净样本　(b)对抗样本　(c)限制区域对抗样本

(d)分割结果　(e)对抗扰动(x8)　(f)限制区域的对抗扰动(x8)

(g)目标分割结果　(e)对抗扰动的分割结果　(f)限制区域的对抗样本分割结果

图 10-14 针对遮挡人脸语义分割的对抗样本

(2) AI 模型预测与决策层攻击

1) 对抗轨迹攻击

对抗轨迹攻击通过操控周围车辆沿对抗轨迹行驶，误导轨迹预测模块，导致被攻击车辆做出危险行为，如偏离路线或发生碰撞，威胁自动驾驶安全。

2) 物理对抗样本攻击

物理对抗样本攻击通过制造伪造车道线或扰动图像空间，误导路径规划模块。使用贝叶斯优化生成的物理对抗性样本能够有效干扰感知和决策系统，影响自动驾驶车辆的判断和行为。

(3) 测试评价体系

1) 方法一

采用高精度仿真测试对测试车辆进行评估，通过构建多样化、真实的数据集，全面测试激光雷达、雷达、摄像头等传感器的表现。该测试评估车辆在虚拟环境中应对复杂驾驶场景的能力，特别是在高风险场景中的感知与决策表现。仿真测试的优势在于能够快速模拟复杂场景并精准控制变量，从而全面测试车辆系统的感知与决策能力。

测试评估包括正确识别率、正确响应率、误报率、关键场景通过率和反应时间。反应时间指从接收传感器输入到决策的时间，若反应时间小于 200 毫秒，则测试通过；200-500 毫秒为部分合格；超过 500 毫秒则不通过。正确识别率要求达到 90%以上，正确响应率也需达到 90%。关键场景通过率要求在高风险场景下 100%成功应对，误报率需控制在 10%以下。

该测试方法能够高效、低成本地进行多样化场景的评估，发现系统潜在弱点并提供改进反馈，相较传统物理测试具有明显优势。

2) 方法二

初步评估与仿真测试类似，重点在于验证系统在仿真环境中应对复杂场景的能力，而非统计错误率。测试旨在确认系统的鲁棒性及决策能力。初步评估后，

汽标委智能网联汽车分委会

进行实体场景测试，利用 3D 打印技术重建未通过的仿真场景，并在现实环境中验证系统表现。成功应对所有场景则通过测试。

此外，冗余性测试评估系统在传感器失效时的切换能力。若切换时间小于 5 秒则通过，5-10 秒为部分合格，超过 10 秒则不通过。

此方法通过真实环境测试验证系统的性能，确保仿真结果与实际应用一致，提升测试结果的可靠性与可重复性。

3) 方法三

采用黑盒测试评估自动驾驶系统在复杂场景中的安全性，不依赖系统内部设计，而是模拟真实世界的挑战和对抗性威胁。测试通过对抗攻击样本数据库重现攻击场景，并使用 3D 打印技术在实际环境中进行模拟。

重点评估传感器数据一致性（SDC）和系统恢复能力（SRC）。SDC 要求系统在 95%以上的场景中展示数据一致性，确保正确识别并应对多传感器数据融合后的威胁。SRC 评估系统在攻击后恢复的速度，5 秒内恢复为测试通过，5-10 秒内为部分合格，超过 10 秒则不通过。

该测试方法通过外部压力测试发现系统潜在薄弱点，为后续优化提供依据，确保自动驾驶系统在现实世界中的稳健性和安全性。

10.4.4.3AIGC 文本对抗攻击

(1) 基于受害者模型可访问性的攻击

1) 基于梯度的攻击

基于梯度的攻击假设攻击者可以访问模型的所有参数、结构和训练数据，利用这些信息计算输出对输入的梯度，确定如何修改输入以产生最大影响。尽管这种方法能提供精确的攻击方向，但它高度依赖模型的内部信息，在黑盒模型中应用较难。

2) 基于分数的攻击

基于分数的攻击通过获取模型的预测分数来推测如何调整输入。它适用于更广泛的场景，但效率较低，通常需要大量查询，消耗时间和资源，尤其在实时性

汽标委智能网联汽车分委会

要求较高的场景中，可能影响攻击效果。

3) 基于决策的攻击(Decision-Based Attacks)

该攻击方法只获取模型的分类决策，而无法得知预测分数。尽管适用于信息极度有限的情况，但效率低，攻击者需进行大量尝试，且在高效性要求较高的场景中不太实用。

4) 盲攻击

盲攻击在完全不了解模型的情况下进行，无法获取任何模型信息。此方法适用于黑盒环境，但成功率较低，难以有效实施。

(2) 基于扰动程度的攻击

1) 句级攻击

句子级攻击涉及对整个句子进行修改，以误导 NLP 模型，同时保持原文的流畅性。常见方法包括：

A. 句子简化攻击(SEA)：通过简化句子结构，移除修饰性或非核心成分误导模型。

B. 句法控制的对抗生成(SCPN)：基于深度学习的模型，通过控制句法结构生成语义相近但表达不同的句子。

C. 生成对抗网络(GAN)：通过对抗训练生成高质量的文本来误导 NLP 模型。

2) 词级攻击

词汇级攻击通过替换、删除或修改单个词汇来生成对抗样本。常见方法包括：

A. TextFooler

TextFooler 通过替换对模型预测影响最大的词汇来生成对抗样本，具有高效且易于实施的特点。

B. PWWS(Probability Weighted Word Saliency)

PWWS 基于词汇重要性和替换词选择的概率，PWWS 通过计算词汇重要性排序来确定替换词，并选取概率权重高的同义词。

C. Genetic Algorithm

汽标委智能网联汽车分委会

Genetic Algorithm 基于遗传算法，随机生成词汇替换候选集并通过选择、交叉、变异等操作优化替换策略，最终选择最优解。

D. SememePSO

SememePSO 结合语义微元和粒子群优化，选择最优替换词，同时保持语义一致性并优化模型误导效果。

E. BERT-ATTACK

BERT-ATTACK 利用 BERT 模型预测替换词，选择与上下文最匹配的词汇进行替换。

F. BAE(BERT-based Adversarial Examples)

BAE 同样基于 BERT 模型，通过在句子中插入、替换或混合这些操作生成对抗样本。

3) 字符级攻击

字符级攻击通过修改单个字符，如插入、删除或替换，生成对抗样本。此方法在对字符细节敏感的模型中尤其有效。

A. VIPER

VIPER 通过对字符进行视觉扰动来误导 OCR 系统或其他视觉文本分析模型。

B. DeepWordBug

DeepWordBug 通过小范围字符扰动，如替换、删除或插入，DeepWordBug 快速生成对抗样本，使用梯度信息识别关键字符进行攻击。

C. TextBugger

TextBugger 通过细微的字符级修改来生成对抗样本，针对影响较大的关键字进行修改。

4) 多层级攻击

多层级攻击结合句子级、词汇级和字符级攻击技术，通过多个层面的修改增强攻击效果，增加对抗样本的自然性和不可预测性。

(3) 测试评价体系

汽标委智能网联汽车分委会

1) 道德伦理合规性评估

确保 AIGC 生成内容符合信息安全与道德标准，实现全面合规。

A. 诱导测试：设定特定场景引导 AIGC 生成敏感内容，关注公正性、隐私保护与言辞文明性，揭示伦理弱点并提出优化建议。

B. 伦理审查：根据不同内容类别制定审查标准，重点防范虚假信息、确保平等呈现，动态调整标准以符合最新伦理要求。

2) 内容审核机制的有效性评价

通过技术和人工审核结合，确保 AIGC 内容合规，及时识别和处理违规信息。

A. 技术审核：利用自然语言处理和机器学习技术，自动化筛查 AIGC 生成的内容，根据不同领域的需求，优化审核技术以减少误报和漏报。

B. 人工审核：人工审核补充技术审核，在复杂情境下提供细致判断，确保内容合法性与适当性。

3) 可靠性验证与响应能力评估

系统评估 AIGC 内容的准确性和一致性，确保信息安全和用户预期。

A. 语义分析：语义分析评估 AIGC 内容的准确性与一致性，防止语义偏差导致错误信息传递。

B. 人工审核：高风险场景下，人工审核结合专业知识，确保 AIGC 内容的合法性与可靠性。

10.4.5 车用 AI 信息安全测试流程标准化可行性

10.4.5.1 技术层面的可行性

智能网联汽车的快速发展带来了对AI信息安全测试流程与评价方法标准化的迫切需求。在技术层面，标准化的可行性取决于多个因素，包括AI传感器安全、AI模型安全和AIGC安全的测试与评价方法。

(1) AI 传感器安全

汽标委智能网联汽车分委会

为了保障AI传感器安全，需采取多种测试与评价方法，包括模拟物理干扰、漏洞扫描、渗透测试及应用加密技术保护数据传输。此外，机器学习算法可用于传感器数据的异常检测。这些方法为AI传感器安全标准化提供技术基础，确保智能网联汽车的稳定与信息安全。

(2) AI 模型安全

从技术角度看，AI模型鲁棒性标准化是可行的。鲁棒性评价有助于识别和修复模型的潜在弱点，提高智能网联汽车的安全性和稳定性。

(3) AIGC 安全

智能网联汽车AI信息安全测试流程的标准化是可行的。现有技术提供了基础，但还需结合法规、经济与社会因素，并加强国际合作。跨学科合作与持续的技术研究是推动标准化的重要途径。

10.4.5.2 法规与政策支持

智能网联汽车AI信息安全测试流程与评价方法的标准化，得到国家层面的高度重视与推动。法规与政策支持是标准化可行性的关键因素。

(1) 顶层设计和政策引导

国家已明确智能网联汽车的战略意义，并通过政策文件提供法规支持。例如，《国家车联网产业标准体系建设指南（智能网联汽车）（2023 版）》明确了统一的国家标准体系，涵盖智能网联汽车的基础、技术、产品和试验标准，表明国家对AI信息安全标准化的积极态度。以下从AI模型鲁棒性、可解释性、AIGC安全和AI模型安全四个方面探讨法规与政策支持。

A. AI 传感器安全

AI 传感器是智能网联汽车获取外部环境信息的核心，其安全性至关重要。国家出台了传感器精度、可靠性和抗干扰能力的标准，确保其在复杂环境中稳定运行。同时，政策要求企业对传感器数据的采集、传输与存储采取加密与访问控制措施，以防止数据被非法获取、篡改或泄露。

汽标委智能网联汽车分委会

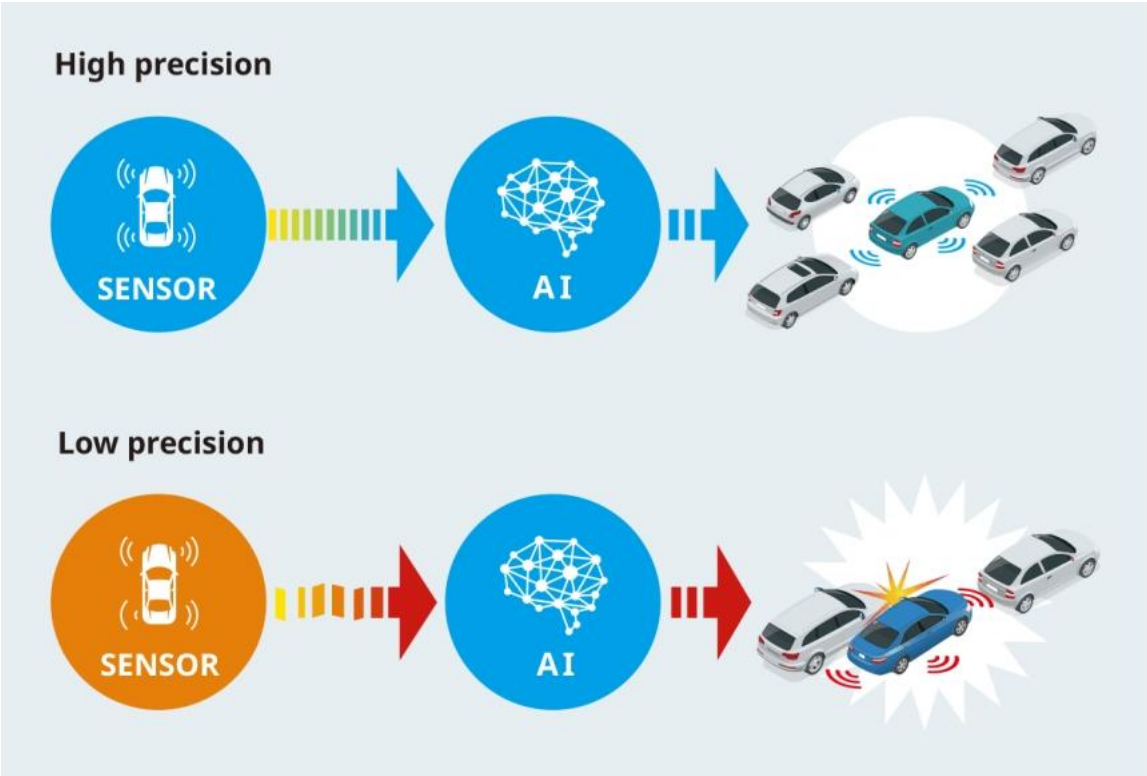


图 10-15 传感器安全

B. AI 模型安全

AI 模型安全是指模型在面对输入数据的微小变动或攻击时，能否保持准确性。在智能网联汽车中，AI 模型安全的测试和评价至关重要，直接影响车辆的安全运行。各国企业在智能网联汽车网络安全保障方面逐渐形成统一思路，并在相关政策法规和标准指导下，落实安全设计、制造和测试实践。例如，ISO/SAE 21434 标准从风险评估、产品开发、运行维护和流程审核等四个方面保障汽车网络安全。

C. AIGC 安全

AIGC 安全关注 AI 生成内容的真实性与合规性。智能网联汽车中，AIGC 技术可能用于生成导航提示、安全警告等。《中华人民共和国数据安全法》和《个人信息保护法》规定企业在数据收集与管理过程中应遵守法律，确保数据安全。

(2) 技术标准与规范

智能网联汽车信息安全测试与评估的标准化涵盖网络安全、数据安全、人机交互等多个技术领域。国家已制定相关技术标准，如智能网联汽车网络安全标准和汽车数据安全标准，明确企业的管理体系和技术要求。

(3) 数据安全管理规范

智能网联汽车在运行过程中会产生大量数据，数据安全至关重要。《汽车数据安全管理若干规定（试行）》明确了数据采集、存储、传输和使用等方面的要求，为AI信息安全测试中的数据安全提供了法律依据。

(4) 道路测试与示范应用管理规范

《智能网联汽车道路测试与示范应用管理规范（试行）》规定了道路测试主体、驾驶人及车辆的条件，以及网络安全保障能力要求，提供了AI信息安全测试实施的操作指南。

(5) 国际合作与标准协调

中国积极参与联合国WP.29 和ISO等国际标准法规的协调，推动智能网联汽车国际标准的制定。这有助于与国际接轨，提升中国在AI信息安全测试领域的影响力。ISO/SAE 21434 标准涉及智能网联汽车信息安全要求，体现了国际合作成果，为全球测试提供统一参考框架。

10.4.5.3 行业合作与共识

智能网联汽车AI信息安全测试流程的标准化不仅依赖技术支持，还需要行业合作与共识。在第九届智能网联汽车技术交流会上，中汽中心与 60 余家机构发布了汽车行业人工智能技术“5+1”标准化发展倡议，为人工智能在汽车领域的应用和创新发展提供了方向。智能网联汽车行业已形成多元化生态系统，通过合作实现资源共享和优势互补，推动AI信息安全测试的标准化。

10.4.5.4 需求与推动可行性

智能网联汽车的快速发展带来了信息安全的新挑战，尤其是在AI技术的广泛应用下，如何确保车辆系统的安全性和数据的保密性成为了行业关注的焦点。在这样的背景下，制定一套标准化的AI信息安全测试流程与评价方法显得尤为重要。

(1) 行业需求

汽车零智能网联汽车分委会

随着智能网联汽车的智能化和网联化，信息安全问题日益突出，车辆成为潜在的网络攻击目标。汽车功能日益复杂，安全风险点增多，迫切需要标准化测试流程来识别和评估安全威胁。行业亟需统一的标准和流程，以确保不同厂商生产的智能网联汽车在信息安全方面的可靠性和一致性。

(2) 推动可行性

国国家层面对智能网联汽车信息安全的重视，以及相关政策法规的出台，为标准化工作提供了政策支持和法律基础。同时，行业内已有的安全测试实践和研究为标准化提供了技术支持和经验积累。标准化测试流程和评价方法将有助于提升行业安全水平，增强消费者信心，推动市场接受度和产业发展。第三方安全测评机构的参与和专业工具的开发将进一步促进标准化工作的实施。

智能网联汽车AI信息安全测试流程与评价方法的标准化，既是行业需求，也得到了政策支持。标准化工作能有效提升信息安全水平，为行业健康发展提供技术保障。推动标准化需要政府部门、汽车制造商、安全研究机构和第三方测评机构的共同努力，共同构建完善的智能网联汽车信息安全标准体系。

10.5 车用 AI 信息安全评估流程标准化研究

10.5.1 车用 AI 信息安全评估流程标准化目的

智能网联汽车 AI 信息安全评估流程的标准化至关重要，旨在构建严谨、规范、普适的评估体系，保障 AI 系统安全，为生命财产安全、数据隐私保护及社会秩序稳定奠定基础。

10.5.1.1 确保行车安全

智能网联汽车依赖 AI 算法实现自动驾驶、远程控制、车路协同等功能，AI 系统若遭攻击或故障，可能引发交通事故。标准化评估流程可全面审查 AI 系统安全性，识别并预防算法漏洞、对抗样本攻击、系统失效等风险，确保车辆在复杂环境和紧急情况下作出正确决策，保障行车安全。

汽标委智能网联汽车分委会

10.5.1.2 保护用户数据隐私

智能网联汽车运行过程中收集、处理和传输大量敏感数据，如驾驶行为、位置信息、生物特征等，若遭恶意获取或滥用，将严重侵犯隐私。标准化评估流程要求严格审计 AI 系统数据处理全流程，确保数据采集合规、存储安全、传输加密、使用透明，防止数据泄露与非法利用。

10.5.1.3 维护网络安全秩序

作为互联网生态系统中的重要节点，智能网联汽车 AI 系统安全不仅影响单车，还关乎整个车联网乃至更广泛的网络安全。标准化评估流程有助于识别 AI 系统潜在的网络攻击风险，通过增强防护、规范应急响应、提升威胁情报共享等手段，降低大规模网络攻击的概率，维护健康的网络空间秩序。

10.5.1.4 推动产业健康发展

标准化评估流程为智能网联汽车研发、生产、销售及运维提供安全基准和操作指南，优化企业资源配置，避免重复投入。同时，统一标准利于监管部门高效执法，打击安全违规行为，促进公平竞争，推动产业健康发展。

10.5.1.5 增强公众信任与接受度

公众对智能网联汽车的信任影响其市场推广与应用。标准化评估流程通过公开透明的安全评价结果，提高消费者对车辆安全性的认知和信心，提升购买意愿与使用满意度，推动智能网联汽车的普及与应用。

综上，智能网联汽车 AI 信息安全评估流程标准化构建全面、动态的安全防线，从行车安全、数据隐私、网络安全、产业发展及公众信任等维度，为智能网联汽车广泛应用保驾护航，助力构建安全、高效的智慧交通体系。

10.5.2 车用 AI 信息安全评估流程及要求

随着车辆互联和自动驾驶技术的发展，AI技术在汽车智能化领域的应用日益广泛，同时也带来了信息安全风险。ISO/SAE 21434 汽车网络安全管理体系提供了结构化流程，以提升汽车行业的信息安全水平，确保车辆在数字化时代的安全

汽标委智能网联汽车分委会

性和可靠性。该标准对保护车辆免受网络攻击、提高安全性及增强消费者信任具有重要作用，并可用于智能网联汽车AI信息安全评估。

图 10-16 产品的生命周期过程

在产品生命周期的各阶段均需评估AI信息安全要求。概念阶段进行安全分析；需求阶段明确AI信息安全需求，包括算法模型、训练数据、数据传输、模型部署、内容合规及业务运营等安全要求；方案设计阶段需充分考虑并应对这些需求；开发阶段关注开发环境与依赖组件的安全性，进行组件安全和代码安全扫描；集成验证阶段关注模型部署安全，AI模型承载大量知识与技能，若管理不当，可能被盗取、复制或篡改，导致性能下降或功能失效，并面临对抗攻击威胁；生成与运行阶段需验证内容合规性，生成式内容涉及个人隐私、虚假信息、民族仇恨、歧视、暴力色情、政治军事敏感及版权等安全问题；运行阶段面临业务安全挑战，如恶意攻击、数据盗取等。

汽标委智能网联汽

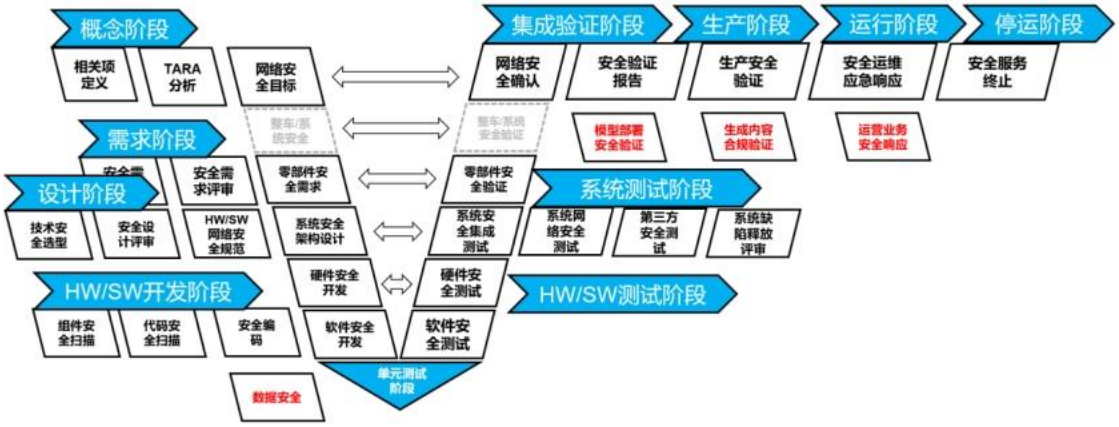


图 10-17 生命周期的不同阶段的信息安全评估

按照上文分析，可以得出在智能网联汽车开发生产过程中，需要按照生命周期的不同阶段完成AI信息安全评估，不同阶段评估的内容和要求如下表：

表 10-7 ISO/SAE 21434 汽车全生命周期网络安全要求管理规范

ISO/SAE 21434 汽车全生命周期网络安全要求管理规范				
汽车生命周期阶段	网络安全目标	阶段安全活动	产出物	AI 信息安全评估要求
概念阶段	1.根据风险评估结果，明确网络安全目标和网络安全声明； 2.明确网络安全需求，并分配到相关系统组件；	1.威胁分析与风险评估； 2.网络安全目标制定； 3.网络安全需求制定；	1.TARA 分析报告 2.网络安全目标/声明 3.网络安全需求 4.网络安全计划	1.输出 AI 信息安全需求清单
开发阶段	1.网络安全需求细化与架构设计； 2.验证细化的网络安全需求和架构设计符合更高抽象层级的网络安全需求； 3.识别设计中	1.细化网络安全需求和架构设计	1.网络安全需求细化与管控相关资料； 2.后开发阶段的网络安全需求； 3.脆弱性分析报告；	1..出 AI 信息安全需求和架构设计文档
		2.集成和验证测试	1.集成与测试安全规范 2.网络安全相关的软件信息清单 3.安全编码规范 4.安全规范扫描报告	1.输出 AI 信息安全测试验证报告

	的网络安全漏洞并进行相应的管理； 4.提供证据证明该组件符合网络安全规范且不包含网络安全相关的非预期功能。		5.开源组件安全扫描报告 6.代码安全问题扫描报告 7.渗透测试方案/测试用例 8.安全渗透测试报告 9.网络安全漏洞处置流程	
生产阶段	1.形成生产阶段安全需求并确保在生产过程中被落实； 2.确保生产过程不会引入额外的安全风险。	生产	1.生产控制计划	1.输出 AI 部署环境检测报告 2.输出 AI 内容安全审核报告
运维阶段	1.网络安全运营与监控； 2.网络安全事件分析与安全响应； 3.软件升级的安全维护；	网络安全监控	1.安全巡检 2.漏洞预警 3.安全监控 4.威胁情报	1.输出 AI 信息安全运营与监控报告
		网络事件评估与响应	1.网络安全事件评估与响应流程/规范	
		软件升级	1.软件升级风险评估报告	
报废阶段	确保报废阶段数据安全销毁	网络安全支持和退役结束	1.服务停用风险评估及影响说明 2.安全停服流程	

10.5.3 车用 AI 信息安全评估数据管理

AI在智能网联汽车中的应用涵盖自动驾驶、高精导航、智能交互等领域，涉及用户个人信息、行为数据等敏感信息。若数据被滥用或泄露，将威胁用户隐私甚至国家安全，因此，AI信息安全评估数据管理至关重要。

10.5.3.1 建立完善的数据管理体系

全生命周期的数据管理包括数据来源、传输、存储、使用、共享、销毁等环节，以确保数据安全、完整性和可靠性。

数据来源：需确保数据来源合规，避免隐私泄露风险

数据存储与传输：要关注数据的存储安全，符合法规要求，保障数据安全。

数据共享与使用：明确用途和范围，建立审计机制，遵循最小原则，确保用户知情同意。

数据销毁：车辆报废或转让时，须依规销毁相关AI模型数据。

10.5.3.2 评估有效的数据安全措施

需评估大模型应用方是否采取了加密存储、访问控制等数据安全措施，以降低安全风险。随着网络攻击和新漏洞的增加，传统安全技术已难以满足智能网联汽车AI需求，需推动技术创新，提高数据安全性。

10.5.3.3 制定明确的数据处理标准

数据质量直接影响模型性能，需去除低质量数据、重复数据、有害信息及个人隐私。

低质过滤：主要采用分类器或启发式方法去除低质量数据[54]。

冗余去除：复数重据影响模型稳定性和多样性，应进行文本去重[54]。

隐私消除：绝预训练数据可能包含敏感信息，应删除涉及个人身份的内容。此外，为确保数据处理一致性，需制定标准化流程和指标要求。

10.5.4 车用 AI 信息安全评估评价方法

对智能网联汽车AI信息安全评估的评价主要从AI模型开发评估、AI模型安全评估、AIGC评估要求及数据传输安全评估四个方面进行全面考虑。

AI模型开发评估：考察模型在不同环境下的适应性及抗攻击能力。

AI模型安全评估：分析模型决策过程，提高可解释性和用户信任度。

AIGC评估，涵盖内容审核、伦理评估及安全性要求。

汽标委智能网联汽车分委会

数据传输安全评估：包括模型安全漏洞、后门攻击防范，以及通信过程中的加密保护，以防止未经授权的访问、篡改或数据窃取。

10.5.4.1AI 模型开发要求评估示例

(1) AI 模型开发环境与依赖评估示例

1) AI模型开发环境评估

可解释性工具（如LIME、SHAP）帮助开发者分析模型决策过程，识别对抗样本。这些工具通过展示模型如何处理输入数据，揭示其在面对扰动时的反应，帮助开发者识别受影响的特征和攻击机制，从而为防御策略提供支持。

2) 增强模型的鲁棒性

可解释性工具有助于发现模型的脆弱点，如过度依赖某些特征（如颜色信息），并指导开发者通过对抗训练增强模型抵抗攻击的能力。此外，这些工具还可以优化特征提取，减少对脆弱特征的依赖，从而提升模型的鲁棒性。

3) 可解释性工具在实时监控中发挥关键作用

能够分析模型决策并及时识别异常行为。例如，自动驾驶汽车的交通标志识别可能出错，通过可解释性分析，开发者可以迅速发现并应对潜在攻击或异常情况，确保系统稳定性。

4) 这些工具提升了AI系统的透明度

帮助开发者向用户和监管机构展示决策过程，增强信任。尤其在汽车行业，透明的决策过程有助于确保模型符合严格的安全标准，增强系统在对抗攻击中的稳健性和可靠性。

5) 可解释性工具为开发更有效的防御策略提供了基础

通过识别模型的脆弱点，开发者可以设计针对性检测算法、优化模型架构，或使用多模态输入增强鲁棒性。此外，开发者可以根据反馈持续优化模型，增强其对新型攻击的防御能力。

汽标委智能网联汽车分委会

(2) AI 模型安全评估

1) 攻击方式

对抗攻击主要分为以下几种方式：

A. 对抗样本攻击（Adversarial Examples）：通过对输入数据进行微小的扰动，导致 AI 模型做出错误的决策。

B. 数据投毒攻击（Data Poisoning）：在训练数据中加入恶意数据，使得训练出来的模型性能下降。

C. 中间人攻击（Man-in-the-Middle Attack）：攻击者在 AI 通信的过程中截获和篡改数据包，影响 AI 决策。

D. 后门攻击（Backdoor Attack）：在模型训练时加入后门，使得在特定条件下模型输出攻击者期望的结果。

2) 攻击位置

在汽车 AI 系统中，可能的攻击位置包括：

A. 传感器层：激光雷达、摄像头、雷达等传感器输入的原始数据。

B. 数据传输层：传感器到 AI 处理模块之间的数据传输过程。

C. AI 处理层：AI 模型处理和决策的阶段。

D. 通信模块：车与车（V2V）、车与基础设施（V2I）之间的通信数据。

3) 不同行车场景的攻击

本节沿用 ST8 中提出的典型场景，并对相关场景进行攻击对抗，评估鲁棒性。

A. 场景一：高速公路行驶

目标：干扰车辆的巡航控制和车道保持功能。

具体操作：攻击者可使用激光设备干扰摄像头，导致车道线识别错误，或通过恶意数据包干扰自动巡航控制系统，使其接收到错误的速度或距离信息，造成加速或减速异常。

B. 场景二：城市交通

目标：干扰车辆的红绿灯识别和行人检测。

汽标委智能网联汽车分委会

具体操作：攻击者注入恶意图像，导致车辆无法正确识别红绿灯，或通过无线通信篡改车与车、车与基础设施的通信数据，令车辆接收到错误的交通信号信息。对于行人检测，攻击者可利用对抗样本干扰摄像头，使检测算法误判行人位置和动作。

C. 场景三：停车场环境

目标：干扰自动泊车系统。

具体操作：攻击者在训练数据中植入特定后门触发条件，导致系统在特定标志或信号出现时作出错误决策。还可通过物理手段（如贴标志、设置假停车线）触发后门，使系统误判停车位，导致碰撞或无法正常停车。

(3) AIGC 安全评估要求

AIGC安全评估是评估生成内容的安全性和可信度的过程，旨在确保其合法性和可靠性，因为人工智能生成的内容可能被滥用或误解。

1) 内容审核

内容审核是评估生成内容合法性和适当性的关键步骤，旨在检查内容是否含有违规、不当或敏感信息，并确保其符合法律法规和道德标准。通过文本分析、图像识别等技术自动审核生成内容，识别潜在问题。具体步骤如下：收集各类内容数据（如文本、图像、音频、视频），对文本进行词向量化或使用深度学习模型提取特征，对图像和视频内容通过卷积神经网络（CNN）提取特征，音频则通过梅尔频谱图处理。随后，设计审核模型架构，制定审核策略，设置审核阈值和处理方式，确保审核结果准确和合规。最后，设计审核结果的可视化和统计分析功能，实时监控系统状态并反馈审核结果。

2) 伦理评估

伦理评估关注生成内容是否符合道德和社会价值观，防止内容出现歧视、偏见或不当观点，并确保其不对社会造成负面影响。其步骤包括参考法律法规和道德准则制定伦理框架，使用卷积神经网络判别生成内容，最后设计可视化和统计分析功能，监控内容是否符合标准，并计算生成的准确率。

汽标委智能网联汽车分委会

3) 技术性安全评估

技术安全评估检查生成系统的安全性和防护机制，防止恶意攻击或滥用。评估方式包括：

漏洞扫描与安全测试：通过漏洞扫描和常见攻击模型（如对抗性攻击）评估系统防御能力.

数据隐私保护：采取措施确保生成内容不泄露用户个人信息或敏感数据，测试隐私保护机制的有效性。

(4) 数据传输安全评估

本部分关注AI系统本体的通用安全评估及AI模型部署中的通信安全评估。国内在汽车信息安全领域已有即将发布的强制标准，特别是《汽车整车信息安全技术要求》中的第五章和第六章，提出了信息安全管理体系要求及车辆信息安全的一般要求。要求企业建立内部管理流程，包括风险评估与处置、信息安全测试、监测响应和漏洞上报等。同时，标准要求加强对供应商、服务提供商、车辆制造商子组织间信息安全依赖关系的管理，涵盖了后门攻击、安全漏洞、认证与授权等评估内容[55]。

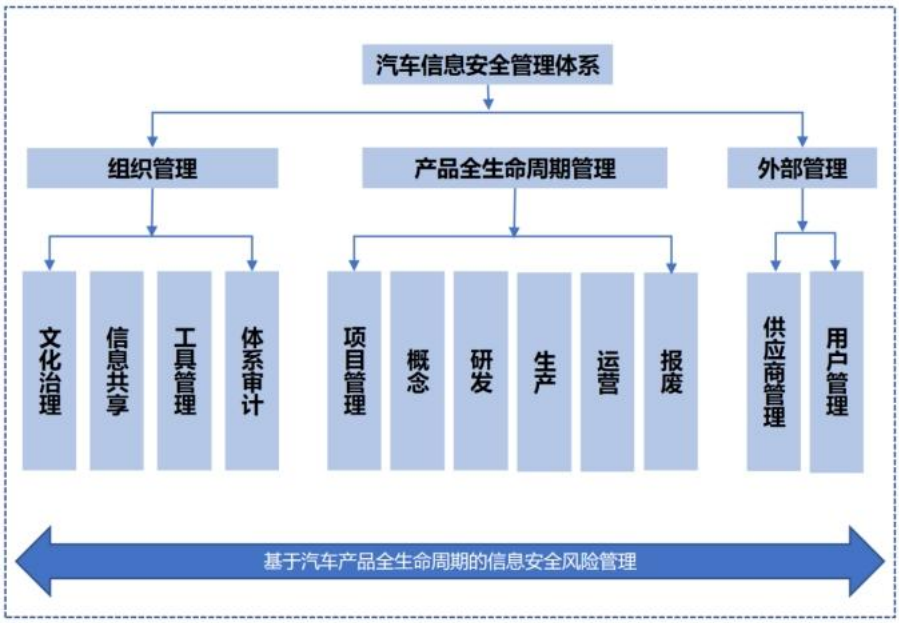


图 10-18 汽车信息安全体系

AI模型通信安全评估是为了评估和确保AI模型在设计、部署和运行过程中的安全性和可信度。这包括防止模型被恶意攻击、保护用户数据隐私、确保模型输出的准确性和可靠性。

1) 认证和授权评估

认证和授权是保护 AI 系统免受未经授权访问的关键机制。评估身份验证和授权机制的安全性和可靠性，确保只有授权用户能够访问 AI 系统和数据。

A. 确定评估范围和目标：明确评估对象（如 AI 系统、数据存储、用户接口等）及目标（如识别认证和授权机制、评估安全性和可靠性等）。

B. 分析身份认证机制：审查密码、双因素认证、生物识别等认证方式，评估其安全性、可靠性及密码强度要求等。

C. 用户身份验证评估：评估注册、登录、注销流程的安全性，模拟非法访问以验证身份验证有效性。

D. 访问控制机制评估：审查 RBAC、ABAC 等访问控制机制，确保只有授权用户能访问系统资源。

E. 授权策略评估：分析用户权限和资源访问权限管理，评估授权策略的安全性和有效性，确保权限被正确分配和实施。

F. 检查会话管理机制：审查会话创建、管理和销毁等过程，评估会话的安全性和可靠性，确保持续验证用户身份。

G. 测试异常情况处理：测试系统对密码错误、登录失败、权限错误等异常情况的处理能力，确保系统及时准确响应并恢复可用性。

通过安全审计和渗透测试，发现潜在漏洞，并确保仅认证用户能访问敏感资源。

2) 识别威胁源

识别威胁源是安全评估的关键步骤，帮助发现潜在威胁并制定防御策略。

A. 能否准确收集信息和情报：通过分析历史安全事件、已知威胁和最新安全情报，识别潜在威胁源。

B. 审查历史安全事件评估：分析过去的攻击方式和漏洞，预测可能的未来

汽标委智能网联汽车分委会

威胁。

C. 分析外部威胁情报：利用威胁情报平台和安全报告，分析外部威胁趋势，评估系统的应对能力。

D. 分析内部威胁：评估系统对内部威胁（如恶意员工或失误操作）的防范能力，重点监控异常事件和实时响应。

3) 加密和数据保护评估

加密和数据保护评估是评估信息安全的重要组成部分，它涉及确保数据在传输和存储过程中得到充分保护的措施。

A. 评估加密算法：确保使用广泛认可的加密算法（如 AES、RSA），评估其安全性和性能。

B. 评估密钥管理：评估密钥的生成、存储、分发和更新，确保密钥的安全管理。

C. 数据传输和存储加密机制评估：评估传输与存储过程中使用的加密技术的可靠性（如数据库加密、文件加密）。

D. 端到端加密评估：确保数据在始发端和终端之间始终受到保护，防止中间人攻击。

E. 审查加密实现和配置：审查加密配置，确保加密功能遵循最佳实践，保护密钥和数据的安全。

10.6 总结与展望

智能网联汽车的迅速发展带来了许多便利，但也引发了关于AI信息安全的担忧。确保智能网联汽车的安全运行，进行AI信息安全测试与评估标准化至关重要。通过对现状、通用要求、测试工具、测试流程和评估示例等方面的研究，得出以下结论：

智能网联汽车AI信息安全测试与评估面临诸多挑战，如敏感数据、对抗攻击、AIGC信息安全及存储传输安全等问题。目前，相关标准尚不完善，行业内缺乏统一的测试和评估方法、流程，使得信息安全难以保障。

汽标委智能网联汽车分委会

为解决这些问题，进行了智能网联汽车AI信息安全测试与评估的标准化研究。在通用要求标准化方面，明确了测试与评估对象及要求，为全产业链企业提供了指导，帮助规范行业测试工作，提升信息安全防护能力。在测试工具标准化方面，确定了工具标准化的目标、现状、软件和硬件要求，并建议优先启动部分标准化工作，以提高工具的可靠性和有效性。在测试流程标准化方面，制定了适用广泛的测试流程，有助于识别安全隐患、保障数据安全、构建防护体系、促进技术创新及提升社会安全意识。在评估示例标准化方面，明确了评估流程、数据管理和评价方法，确保行车安全、保护用户隐私、维护网络安全、推动产业发展并增强公众信任。

此外，汽车标准化创新基地正在推动《汽车信息安全检测实践库》的构建，作为《汽车整车信息安全技术要求》强制性国家标准配套实施的用例库。该库将根据智能网联汽车技术的发展，扩充新的测试方法，满足企业研发与检测认证需求，涵盖车用人工智能的信息安全攻击路径和对抗攻击算法库，为后续的标准化工作提供支撑。

综上所述，推荐进行《智能网联汽车 车用人工智能 信息安全技术要求》标准研究，基于车用人工智能不同业务场景，分析制定各业务场景的信息安全防护技术要求，并形成相适配的测试方法，该标准可作为《汽车整车信息安全技术要求》对于智能网联汽车引入人工智能所带来新风险的技术要求以及检查与试验方法的补充，后续可作为国际提案同步提报至WP29 GRVA CS/OTA 工作组，推动UNECE R155 强制性法规修订。

汽标委智能网联汽车分委会

第十一章 车用 AI 功能安全及标准化研究

11.1 车用 AI 功能安全标准化研究

随着人工智能技术，特别是深度学习和数据驱动模型在自动驾驶系统中的广泛应用，车辆功能安全的内涵和边界正在发生深刻变化[75]。传统功能安全标准（如 ISO 26262[56]）主要应对的是硬件故障或确定性软件缺陷引发的系统失效，而人工智能系统所引入的感知误差[57][71][72]、预测偏差[58]、决策与控制不确定性[59][73][74]等非传统失效模式，已逐渐成为威胁道路车辆安全的新型风险源。在此背景下，车辆 AI 功能安全应运而生，旨在通过系统工程方法确保嵌入 AI 技术的车辆系统在各种预期和非预期场景下依然保持安全可控。

车辆 AI 功能安全是指在智能网联汽车中引入人工智能功能之后，为保证车辆系统在运行过程中即便在模型误判、输入异常、泛化能力不足等情况下，也不会对人类生命、财产或社会环境造成不可接受风险而采取的系统性安全保障机制。这种安全不仅包括传统意义上的“功能性正确”，更强调“行为合理”和“过程可控”。AI 模型在面临复杂动态环境时，其输出并非完全可预测，因此必须从系统层面对 AI 的不确定性进行建模和管理，涵盖其行为边界、决策路径、输入空间等维度的安全约束。

与传统软件模块不同，AI 系统是高度依赖数据驱动的，这使得其失效模式更加隐性，例如由于训练数据分布偏差、感知噪声、对抗样本或未知场景的输入，均可能导致模型行为异常。车辆 AI 功能安全正是面向这些风险，要求通过全生命周期的工程机制对其进行识别、控制和验证。这包括：在需求层明确 AI 功能的角色与安全目标，在设计层加强冗余与容错机制的构建，在验证层引入对抗性测试与不确定性量化，在运行层部署实时安全监控与紧急响应机制，并在维护层建立场景闭环反馈机制，持续优化模型安全性能。

汽标委智能网联汽车分委会

因此，车辆 AI 功能安全不仅仅是技术性问题，更是一个跨学科的系统工程挑战。它需要电子电气架构、AI 建模、安全工程和道路交通法规等多方面的协同，形成从系统架构到工程实践的完整闭环。其最终目标是在保障 AI 系统性能优势的同时，建立一种稳定、可信、可审计的安全保障机制，使人工智能能够真正支持未来智能驾驶系统的安全、稳定、可靠运行。

11.1.1 车用 AI 功能安全标准化研究目的与意义

当前，车用 AI 功能安全面临诸多挑战：人工智能系统的决策机制不透明、对输入环境高度敏感，易受到数据分布变化、感知误差或模型泛化能力不足的影响而导致非预期行为；AI 系统在运行过程中可能出现的非确定性失效模式尚缺乏统一的识别方法与风险建模框架；AI 模型训练、验证、部署全生命周期的安全保障机制尚未建立，不同厂商间在安全指标设定、安全监控接口、安全降级策略等方面缺乏一致性，导致安全验证的可重复性和结果的可比性难以保证。本标准化研究旨在系统梳理车用 AI 功能安全面临的核心问题，明确关键技术环节的标准化需求，探索具备普适性与可操作性的功能安全评估与控制路径。

其意义在于构建统一的安全管理与验证基准，推动行业在模型开发规范、安全目标推导、安全监控机制等方面形成协同机制，提升 AI 系统在复杂交通场景下的安全稳定性；同时为自动驾驶系统的工程落地、产品安全认证、法规合规提供技术基础，支撑智能汽车从技术突破向可靠应用的跨越；最终加快建立面向 AI 的功能安全标准体系，促进智能网联汽车产业的安全可持续发展与社会公众信任的建立。

11.1.2 国内外研究现状

当前，车用 AI 功能安全作为融合传统功能安全、预期功能安全与人工智能安全的新兴交叉领域，正逐步走向标准体系化。国际上，该领域的发展基础可以追溯至通用功能安全领域的核心标准 IEC 61508[60]。IEC 61508 作为功能安全的母本标准，于 1998 年首次发布，是涵盖各类电气/电子/可

汽标委智能网联汽车分委会

编程电子安全相关系统的通用框架。在此基础上，针对不同行业应用需求，衍生出了多个领域专用标准，如铁路行业的 EN 50129[61]、过程工业的 IEC 61511[62]，以及汽车行业广为应用的 ISO 26262[56]。

在汽车领域，ISO 26262 是车用电气/电子系统功能安全的核心标准，2011 年发布第一版，2018 年发布第二版，覆盖从系统设计到软件实现的全流程功能安全开发要求。中国也基于该标准推出了等效国标 GB/T 34590[63]（2022），作为国内统一适配功能安全开发的指导文件。此外，GB 17675[64]对应的是早期ISO 26262起草前我国采纳的IEC 61508等标准的基础性要求，也仍具参考价值。

随着自动驾驶系统逐渐具备感知—决策—控制的闭环能力，安全关注焦点也从“硬件或软件故障引发的失效”扩展到“功能设计本身的局限性所导致的风险”。为应对这类“非故障引发的风险”，预期功能安全（SOTIF）标准应运而生。2015 年，ISO/TC22/SC32/WG8 启动制定 ISO/PAS 21448[65]《道路车辆——预期功能安全》，并于 2019 年发布公开规范（PAS），最终在 2022 年发布为正式国际标准 ISO 21448:2022。该标准在 ISO 26262 的基础上，聚焦自动驾驶场景中由于感知/决策能力受限、环境变化、模型边界等因素引发的风险评估与控制。

针对自动驾驶系统整体安全性的系统性评估，ISO/TC22/SC32/WG13 于 2021 年启动制定 ISO/TS 5083[66]《自动驾驶系统的设计、验证与确认的安全指南》，该标准覆盖自动驾驶安全需求提取、架构设计、验证流程、测试覆盖等多个方面，预计于 2024 年正式发布。

在AI安全方向，ISO/TC22/SC32/WG14 于2021年正式启动了 ISO/PAS 8800[67]《道路车辆——人工智能与安全》 标准的制定工作。该标准专门关注车载系统中使用人工智能技术的电气/电子系统的安全保障问题，尽管不限于自动驾驶系统，但其涵盖的模型失效模式分析、训练数据可信性、模型验证路径、在线监控机制等，均可适用于 ADS（自动驾驶系统）中的 AI

汽标委智能网联汽车分委会

功能安全评估。ISO/PAS 8800 借鉴了 ISO/IEC TR 5469[68]（人工智能系统可靠性）、ISO/TR 4804[69]（ADS 安全指南） 和 ISO 21448 的框架，努力构建一套融合 AI 技术特点的功能安全体系，被业界誉为“集 AI 功能安全之大成”的重要基础标准。

在法规层面，联合国欧洲经济委员会 UN ECE 也于 2021 年正式实施了全球首个针对 L3 级自动驾驶系统的强制性法规 —— UN ECE R157[70]（ALKS 法规），适用于具备自动车道保持能力的车辆系统。该法规 2022 年修订后将适用范围从 M1 类小客车扩展至 M、N 类车辆，允许自动驾驶系统最高工作速度从 60km/h 提升至 130km/h，基本覆盖高速场景的点到点自动驾驶能力，推动了自动驾驶系统安全法规与功能安全标准的协同发展。

在国内，车用 AI 安全相关标准体系建设仍处于推进阶段。除 GB/T 34590 等对 ISO 26262 的等效采标外，国家市场监管总局、工信部、智能网联汽车国家创新中心等单位也已启动对 AI 系统安全、数据可信性、安全监控器、安全指标评估工具链 等相关国家标准或行业标准的研究工作，部分内容正处于征求意见或立项阶段，旨在为车用 AI 功能安全构建适应中国特色道路交通环境的本土标准体系。

11.2 车用 AI 功能安全要求与验证

11.2.1 道路车辆系统安全工程背景下的人工智能

11.2.1.1 目标

人工智能安全生命周期及其活动，以确保人工智能系统的安全相关错误不会导致车辆级别的危险行为的不合理风险。AI 安全生命周期包括：人工智能系统开发和运行期间活动的定义；在系统级安全生命周期的背景下，采用数据驱动的迭代方法来开发、评估和持续保证基于人工智能的功能；安全分析的应用侧重于人工智能系统已达到的性能和潜在的功能不足，以细化人工智能系统的安全要求和设计；和在整个生命周期中系统地收集安

汽标委智能网联汽车分委会

全保证论证的证据。

11.2.1.2 先决条件

AI 系统定义，包括但不限于：人工智能系统功能；和 AI 系统与其余产品的接口，包括但不限于 AI 系统输入的 ASIL 功能（如果适用）。AI 系统语义输入空间的规范；AI 系统语义输入空间分布的规范；人工智能系统句法输入空间的规范；分配给人工智能系统的安全要求。这些顶级安全要求包括但不限于：导致车辆级别不良安全相关行为的各个错误模式（如果适用）；导致车辆级别不良安全相关行为的错误序列模式（如果适用）；人工智能系统检测到错误时的错误反应，包括时间限制（如果适用）。

11.2.1.3 一般要求

1) 应定义描述必要活动的 AI 安全生命周期，用以提供 AI 系统的安全相关错误不能引发整车层面危害行为的不合理风险的证据。

2) 执行中，AI 安全生命周期应基于参考生命周期，并应按照项目特定需求进行裁剪。

3) AI 安全生命周期和相关活动应与 ISO26262 和 ISO21448 定义的系统层面安全生命周期进行整合。

4) 在 AI 安全生命周期的每个阶段，应定义安全工作产品，用于提供证据，证明分配给 AI 系统的安全需求已实现。

5) AI 安全生命周期应包含活动，用于确认在持续保障的运行中，持续实现了分配给 AI 系统的安全需求。

6) 应对 ISO26262-2:2018 进行裁剪，以解决 AI 安全管理问题。裁剪包括但不限于：

- 使用 AI 生命周期代替 ISO26262 生命周期
- 用 AI 安全代替功能安全

汽标委智能网联汽车分委会

- 增加此文档为 ISO26262 系列标准的补充，做为相关标准
- 安全保障论证可作为安全档案使用
- 安全计划也包括此文档的安全活动
- 对 ISO26262-2：2018 的表 1 进行裁剪，关注此文档的安全产品（见表 11-1）

7) 安全保障证据应符合 ISO26262-2.

8) 随机硬件故障，应按照 ISO26262-5-2018 实现随机硬件故障的相关需求。

9) 应按照 ISO26262 系列标准相关部分开发、实现及验证 AI 系统。

10) AI 组件的非 AI 模型部分应按 ISO26262 系列标准相应部分进行开发、实现及验证。

11) 对于是 AI 模型的 AI 组件，还应对 ISO26262 相关部分进行裁剪和应用。

12) 须规范使用的 AI 技能。

13) 须规范使用的 AI 技术。

14) 须分析使用的 AI 技术是否足以单独应用 ISO 26262 来实现 AI 安全。对于这些 AI 技术，须应用 ISO 26262。

15) 对于仅应用 ISO 26262 被认为不足以实现 AI 安全的已利用的 AI 技术。

表 11-1 要求的认可措施，包含要求的独立性等级

认可措施	应用于以下的独立性程度					范围
	QM	ASIL A	ASIL B	ASIL C	ASIL D	
安全计划的认可评审。独立于	-	I1	I1	I2	I3	依照全部安

汽标委智能网联汽车分委会

相关项的开发人员、项目管理和工作成果责任者						全需求中的最高 ASIL 等级执行。
AI 确认和 AI 测试的认可评审。独立于相关项的开发人员、项目管理和工作成果责任者	-	I0	I1	I2	I3	依照全部安全需求中的最高 ASIL 等级执行。
AI 分析的认可评审。独立于相关项的开发人员、项目管理和工作成果责任者	-	I1	I1	I2	I3	依照全部安全需求中的最高 ASIL 等级执行。
安全保障证据的认可评审。独立于安全保障证据责任者	-	I1	I1	I2	I3	依照全部安全需求中的最高 ASIL 等级执行。
AI 安全审核独立于相关项的开发人员、项目管理	-	-	I0	I2	I3	依照全部安全需求中的最高 ASIL 等级执行。
AI 安全审评估独立于相关项的开发人员、项目管理	-	-	I0	I2	I3	依照全部安全需求中的最高 ASIL 等级执行。
——I0：宜执行认可措施；但如果执行，应由与负责创建工作成果的人员不同的人员执行； ——I1：认可措施应由与负责创建工作成果的人员不同的人员执行； ——I2：认可措施应由独立于负责创建工作成果的团队的人员执行，即由不向同一个直接上级报告的人员执行； ——I3：认可措施应由在管理、资源和发布权限方面与负责创建对应工作产品的部门独立的人员执行。						

11.2.1.4 AI 系统的迭代开发范式

AI 安全生命周期的 AI 系统开发阶段涵盖了设计、实现、验证和确认

汽标委智能网联汽车分委会

AI 系统所需的所有活动。AI 系统的迭代开发，由一套性能指标指导，其中包括分配给 AI 系统的与安全需求相关的安全属性。由于需求冲突和 AI 技术的特点，可能无法保障每个单独的性能指标均达到最佳。例如，对于基于 ML 的分类任务，需要在减少误报和漏报结果的基础上对优化进行权衡。因此，迭代开发活动聚焦于使充分满足系统级安全需求的所有性能指标达到最佳状态。这就可能涉及到对 AI 系统的前处理和后处理元素或其他系统级组件的需求的调整。尽管 AI 系统的开发存在固有的迭代性，不同阶段的开发仍聚焦于不同的性能指标集。以下为不同开发阶段的例子：

概念验证阶段：在这个阶段，以最初定义的安全需求集合为背景，设计、实现和测试一个 AI 功能。这一阶段的目标是评估所选的 AI 技术在满足安全需求方面的潜力，同时定义一套减少功能不足的数量或影响的措施（例如：优化模型参数和定义数据收集策略）

AI 系统系列开发阶段：在这个阶段，AI 系统根据所有的安全需求进行迭代开发。对 AI 系统或 AI 模型的错误进行分析，了解其潜在的根源，并确定措施，通过设计和数据选择来减少功能不足，或通过架构措施来减少其影响。此阶段的开发通过使用一个主机平台来进行功能实现和测试，持续到所有与安全有关的性能都达到足够的水平为止。

目标硬件平台部署阶段：在这个阶段，AI 系统被转移到目标软硬件平台上。这其中包括，例如，用于计算结果的数字精度变化，同时考虑定时和内存限制（例如，导致需要修剪计算图），同时需要对潜在的随机硬件故障具有鲁棒性。此开发阶段的重点是在目标执行平台限制的前提下，仍能保证满足安全需求。

部署后的持续开发阶段：在这个阶段，AI 系统根据实际操作过程中的观察学习迭代更新。其中包括在实际环境下对以前未知的触发条件和分布偏移的补偿。在这一阶段的开发中，通过对性能指标集的监测，确保基于功能之前的迭代，改进是单调安全的。例如，以前实现的与安全有关的属

汽标委智能网联汽车分委会

性不会因为新考虑的实施方案的特定属性的优化而降低。此开发阶段还可以包括在操作环境中以“影子”模式运行的功能的开发版本，以便收集合适的数据并评估实际条件下的潜在性能。

11.2.1.5 错误概念、故障模型和因果模型

错误概念通常指人们在认知过程中形成的不符合客观事实或科学原理的观念，可能源于经验偏差、误解或信息误导。例如，认为“重的物体下落更快”就是物理学中的典型错误概念。故障模型主要用于工程、系统维护等领域，是对系统或设备可能出现的故障模式、原因及影响的抽象描述。它通过梳理组件失效的逻辑关系（如电路短路、机械磨损等），帮助预判故障点并制定解决方案。因果模型则聚焦于揭示事物之间的因果关系，通过构建逻辑框架来分析“因”如何导致“果”。它不仅关注事件关联，更强调因果的方向性和必然性，常见于统计学、人工智能等领域。

(1) 因果链

本内容采用故障和错误的概念。如果故障被触发条件激活，就会发生错误。触发条件、故障和错误之间的关系如图 11-1 所示。

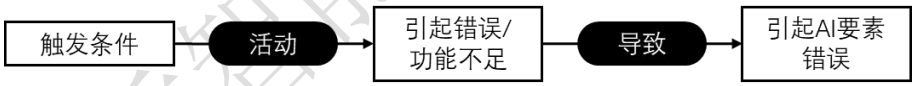


图 11-1 触发条件、故障/功能不足和错误之间的关系

与安全相关的故障，或者更准确地说是规范的不足，可能是用于对象分类的 ML 系统的 AI 训练、AI 验证和 AI 测试数据集中的缺失对象。触发条件将是 ML 系统在实际操作期间遇到该对象。与安全相关的错误是 ML 模型对对象的分类不正确，从而导致错误的分类。

AI 元素的安全相关问题如何被激活并导致安全相关错误，从而对车辆水平产生不良影响的因果链在图 11-2 中展示。

AI 组件的错误可能会通过 AI 系统进行传播，并导致 AI 系统的错误。AI 系统的错误可以通过其余系统进行传播，并可能导致车辆级别的不良安全

相关行为。

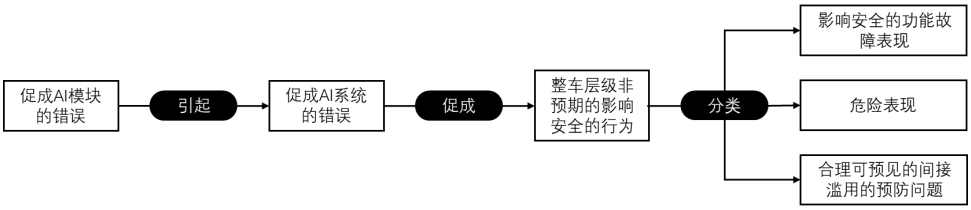


图 11-2 从 AI 模型级别到车辆级别的错误传播

(2) 根本原因分析

可以根据根本原因（即需要激活的故障类别）区分不同的错误类别。错误及其不同根本原因之间的相关性如图 11-3 所示。

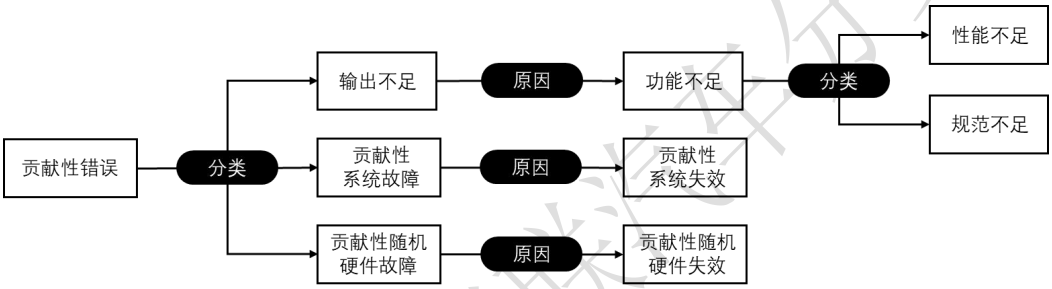


图 11-3 安全相关错误与它们不同根本原因类别的相关性

不同种类的错误的根本原因为：

1) 规范不足：规范不足可能是在人工智能训练、验证和测试数据集中缺少数据集。结果输出不足可能是在缺少数据集的情况下进行错误分类。

2) 性能不足：在特定环境条件下，传感器的范围可能不足。导致的输出不足可能是在轨迹中错误地未检测到障碍物。在 ML 模型的情况下，性能不足可能是由训练和测试数据集相关问题引起的，例如各自输入空间的覆盖不足。这些与数据相关的问题反过来被认为是规范的不足，或者更准确地说是数据规范的不足。

3) 系统性失效的贡献：系统性失效，可能是在软件中除以零或使用不正确的变量名称

4) 随机性硬件失效的贡献：造成接地短路的物理缺陷。在评估安全机制的有效性时，区分不同的错误类别是必要的。同质冗余可以有效检测其

汽标委智能网联汽车分委会

中一个冗余元素中的随机硬件错误，但它不能有效检测系统性错误。不同的故障类别也可以通过针对每个根本原因类别的专门安全措施来进行安全保证论证。

(3) 基于安全影响的错误分类

基于错误对车辆水平产生不良安全行为的潜在影响，错误分类如图 11-4 所示。

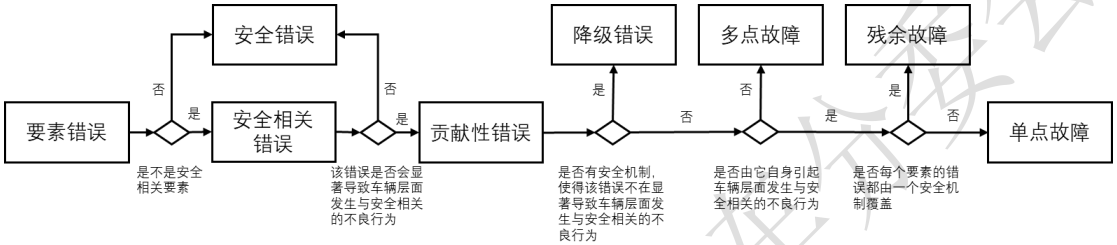


图 11-4 基于可能导致车辆级别的不良安全相关行为的潜力的错误分类方案

该错误分类可以在 AI 系统、AI 组件和嵌入 AI 的系统的安全分析中发挥作用。它可以帮助识别必要的独立性，例如在多点错误的情况下，或特定的关注点，例如由于单点错误。

11.2.2 人工智能系统的安全需求的推导

11.2.2.1 目的

1) 对人工智能系统制定一套完整且一致的安全要求，确保由于残余错误导致的危害行为的残余风险足够低；

2) 基于对开发、验证和确认过程中发现的人工智能系统潜在不足的理解，改进人工智能系统的安全需求；

3) 准确描述人工智能系统在其运行域的局限性(不足)(以升级至更高层级)；

4) 为人工智能系统安全需求的推导定义一个系统性的方法。

汽标委智能网联汽车分委会

11.2.2.2 先决条件

应提供以下信息:

1) 上层安全需求(系统集成商提供)

2) 输入域描述(系统集成商提供)

11.2.2.3 一般条件

1) 应细化高层级输入域定义，创建一个适合数据集生命周期启动的输入域描述，对于高层级输入域中不能被覆盖的用例，应该被识别和记录。

2) 导出的人工智能安全需求应追溯到更高级别的需求，假设和关键场景，或与完善后的输入域定义保持一致，解决功能不足和触发条件的潜在影响因素/根本原因。

3) 应提供推导出的人工智能安全需求充分的理由。

4) 对于不存在由于随机硬件故障、典型的系统性软件故障和其他典型的系统性人工智能开发故障，例如测试数据选择错误，造成的不合理风险。

5) 如人工智能系统用例未完全符合从系统集成商处收到的要求，或需求用例仅满足部分输入域，应作为属性/规范的一部分加以识别和记录。

6) 为支持运营和持续改进，应通过人工智能安全需求，来处理现场作业。

11.2.2.4 从目标到通用要求的工作流程和连接（信息性）

首先，将高层次安全要求（SOTIF、功能安全）转化为初始的 AI 安全要求集合。对于 SOTIF，每个技术 AI 要求通常体现为输入域中错误的发生率低于某个阈值（不会导致 AI 组件输出失效）。为了控制 AI 工程过程中的不确定性，以实现 AI 模型的开发质量，并解决 AI 系统中的安全相关问题以实现 SOTIF，还将产生其他的 AI 安全要求，可以通过参考以往产品的要求或利用可能与本应用相关的 AI 安全属性来生成。

汽标委智能网联汽车分委会

为了控制 Al 工程过程中的不确定性，还将产生其他的 AI 安全要求（定量或定性），可以通过参考以往产品的要求或利用可能与本应用相关的 AI 安全属性来生成。最后，将这些要求分配给不同条款的进一步工程任务。

同时，系统集成商提供的高层输入域（即 ODD）定义可能需要进一步详细说明才能在 AI 安全工程中使用。这导致详细说明输入域的活动。

在 AI 系统工程或运行过程中，可能会发现特定错误（序列）模式的发生率过高。通过了解触发条件或安全相关问题，可以推导出新的 AI 安全要求，以抑制 AI 系统产生的错误，或者（当进一步提高性能不可能时）将性能限制及其相关的触发条件反映给系统集成商。

对于特定错误（序列）模式是否过高的评估或来自现场的观察。一方面，评估了 AI 系统在安全要求方面的残余风险，另一方面，识别了导致违反技术安全要求的 AI 系统中的安全相关问题。

最后，在运行阶段，现场监测可以用于支持发现先前未识别的触发条件和假设违规。

11.2.2.5 关于监督机器学习的需求导出的详细考虑事项

(1) 需求导出和额外的人工智能安全需求

预期功能安全活动在整车级别识别和评估风险、潜在的功能不足及潜在的触发条件，生成的高层级安全需求将被进一步细化为人工智能的安全需求。在 SOTIF 上下文活动中，与性能目标相关的常见高级安全需求之一可以被如下陈述：特定错误模式的发生概率应低于给定阈值。由此导出的人工智能安全规范在详细说明错误模式要求时需考虑人工智能系统的底层计算能力。

示例 1 考虑错误序列模式“连续错误检测到附近行人（假阴性--漏识别）超过 0.5s”。当深度神经网络以 10 帧每秒的速率触发时，上述错误序列模式可以转化为”行人连续出现 5 次以上假阴性”。

对于在特定输入域运行的人工智能功能，其安全规范也可以包含类似于按需失效率（PFD）概念的接受准则，例如“Prob（一种错误模式的发生次数）< α”。当详细设计可用时，尽管如 STPA 之类的方案可以支持识别人工智能组件级别附加安全需求。但某些人工智能组件可能是使用监督机器学习进行设计的，无法进行进一步分解，该条款的剩余部分将聚焦这类情形。

示例 2 人工智能系统详细设计的一个例子可以是使用 5 个冗余的深度神经网络进行感知任务，然后对单个结果进行多数投票。此外，可以使用异常分布检测器来中止多数投票的有效性。有了这样的详细设计，人们可以进一步识别不安全的控制行为（例如，异常分布检测器在雨天置信度过高，从而提供不正确的输出）并创造额外的详细 AI 安全要求。由于输入域的复杂性（例如，自动驾驶）和机器学习的固有性质（即，样本内和样本外存在误差隔阂），不可能直接计算出错的概率。因此，在输入域中测量足够的样本来估计误差的相对频率。输入域内的经验概率和实际概率（类似于 PFD）之间的区别是不确定的，这种区别的不确定性应通过人工智能工程过程的额外要求来控制。图 11-5 说明了潜在的倾向。

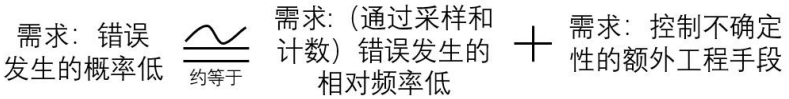


图 11-5 如何理解监督学习的安全需求推导 AI 合成场景能够一些输入条件可控

示例 3 考虑需求“在运行中每小时连续 2 次识别邻近行人为假阴性的次数低于 10 的-6 次”这要求在输入域内的特定错误模式的发生次数足够低。这条对经验概率的要求可被翻译成“在差异性充分的一亿小时驾驶-数据中，深度神经网络组件不能对邻近行人产生两次连续的假阴性识别（漏识别）”。这里的假设是人们可以使用更多的驾驶时间来证明这一要求成立的假设

(2) 创建额外的人工智能安全要求以控制不确定性

对于工程化的机器学习组件，我们可以通过管理工程过程中的不确定

性源来降低风险。在工程化过中每个元素都存在不确定性，我们称之为潜在的影响因素。影响因素可分为观测、标签、模型和操作四个类别。表 11-2 描述了在工程化过程中管理影响因素的典型方法，构建一个抽象的框架，用于管理感知不确定性以确保安全。

表 11-2 在工程化过程中管理影响因素

影响因素类别	不确定性管理方法
观测	(1) 从演绎（自上而下）和归纳（自下而上）的方式中，枚举影响人工智能输出的模式 (2) 为全方位的人工智能性能和安全性（避免伤害）进行数据覆盖策略的定义
标签	减少标注规则和标注工作质量的变化
模型	使用合适的模型选择和事实标准进行调优和评估
操作	检测开发和部署之间的模型性能偏差

应对上述不确定性，表 11-3 描述了一个非详尽的示例通用需求集，适用于特定应用的人工智能系统。

表 11-3 管理影响因素的通用要求示例

影响因素类别	通用要求示例
观测	-推导与推理目标、环境条件和系统配置相关的场景（相关性） -包括可能导致错误的场景（例如，假阴性、假阳性、显著不正确的回归） -考虑系统配置的老化过程（例如，安装位置的变动） -包括在当前产品开发、传统产品开发和运行过程中已经导致错误的具体场景（安全性） -对于每个场景或场景组合指定数据数量（多样性） -解释无法定义目标场景数据数量的合理性

汽标委智能网联汽车分委会

标签	-明确推理目标 -明确标注过程 -明确标注评估策略，即评估时机、评估过程、评估指标和验收标准
模型	-明确模型性能的度量和验收标准 -明确模型选择策略，并根据策略选择合适的模型 -明确超参数搜索策略和条件，例如界限，并记录确定的超参数值 -对 AI 模型进行安全性分析 -介绍检查输出合理性的技术能力 -介绍去噪/量化输入的技术能力 -介绍稳健训练技术的技术能力（例如，针对噪声的训练，量化神经网络） -介绍提供决策解释（可解释性）的技术能力 -对于 DNN，预先全局可解释性（即将决策机制提取为可理解的规则）可能很困难。通过后续局部可解释性（即针对特定输入作出特定决策的原因）可能性更大，可以使用局部线性化或热图[58]等技术。
操作	-介绍识别与已知触发条件匹配或一般条件存在不确定性情况的技术能力 -介绍识别概念/领域漂移的技术能力

(3) 细化人工智能安全工程的输入域描述

输入域定义了工作空间和系统运行的条件，目的是确保其安全设计和安全运行。对于依赖于使用数据驱动方法来建立机器学习模型的基于人工智能的系统，输入域是指数据集要求的定义，这些需求构成了系统整个生命周期中使用的数据的基础。工作空间表征了系统运行的道路或区域。这些条件反映了天气、能见度条件、照明，包括连接性。

输入域的定义是一个迭代过程。对于系统集成商提供的初始定义，它

汽标委智能网联汽车分委会

可能不适用于确保人工智能组件的预期功能。环境因素是基本问题，而系统及其要素根据层次的不同有不同的关注点。输入域描述的细化旨在符合现有系统（如传感器）的能力。其细化也可以考虑新发现的或已知的可能导致性能不足的触发条件。

示例：与其指定人作为识别目标，不如按年龄细化人的类型（老人、儿童、成人），因为设计的人工智能功能应该对老年人给予特别的关注。

由此产生的输入域规范构成了验证和测试目标系统性能的基础。也就是说，输入域限定了人工智能系统的性能预期，减少了与其操作相关的不确定性，从而确保了其安全操作的界限。

(4) 限制 AI 输出不足的发生

在 AI 系统的工程过程中，人们可能会观察到 AI 系统的错误率过高的特定场景(或一组场景)，违反了最初推导的安全要求。您可以再次使用预定义的安全相关属性列表作为检查清单，以检查潜在的根本原因，并随后导出一组具有相关度量的新安全需求。在初始迭代之后的迭代中，安全分析为该活动提供了额外的输入。基于现场或 V&V 活动中不可接受的观察结果，安全分析识别与安全相关问题具有最大关联的安全相关属性。对输出不足和安全相关问题之间的相关性(或者如果适用的情况下，因果性)的洞察可以用作初始迭代之后安全需求细化的基础。新导出的需求可以是定性的，也可以是定量的，对于定量的需求，接受阈值的设计依赖于应用。

注 1:目前，这些与安全有关的特性是概念性的描述，而不是使用形式定义。所带来的的结果是，安全要求的导出不被视为“(逻辑)因果推导”，而是一种“合理猜测”。

注 2:绝大多数机器学习类型的统计模型估计事件的可能性，是从数据中的相关性推断出来的。在这种情况下，我们无法在这些模型的输入和输出(以及系统最终采取的行动)之间建立明确的因果关系。因此，人们应该意识到经典的因果故障分析工具不一定适用于这种情况。

浙江省智能网联汽车分会

如图 11-6 所示，运行过程中的错误率很高，这与性能不足有关。通过进行适当设计的实验(例如，反事实分析或假设测试)，人们可能会发现问题与一个或多个与安全相关的固有 AI 属性(例如，鲁棒性)密切相关，并根据经验在一些可测量的指标中表现出相关性的证据。这项活动可以继续进一步细化表 11-4 所述的影响因素，从而产生合理的要求，以消除性能不足。然而，建议的需求作为对策的有效性仍然应该得到验证。

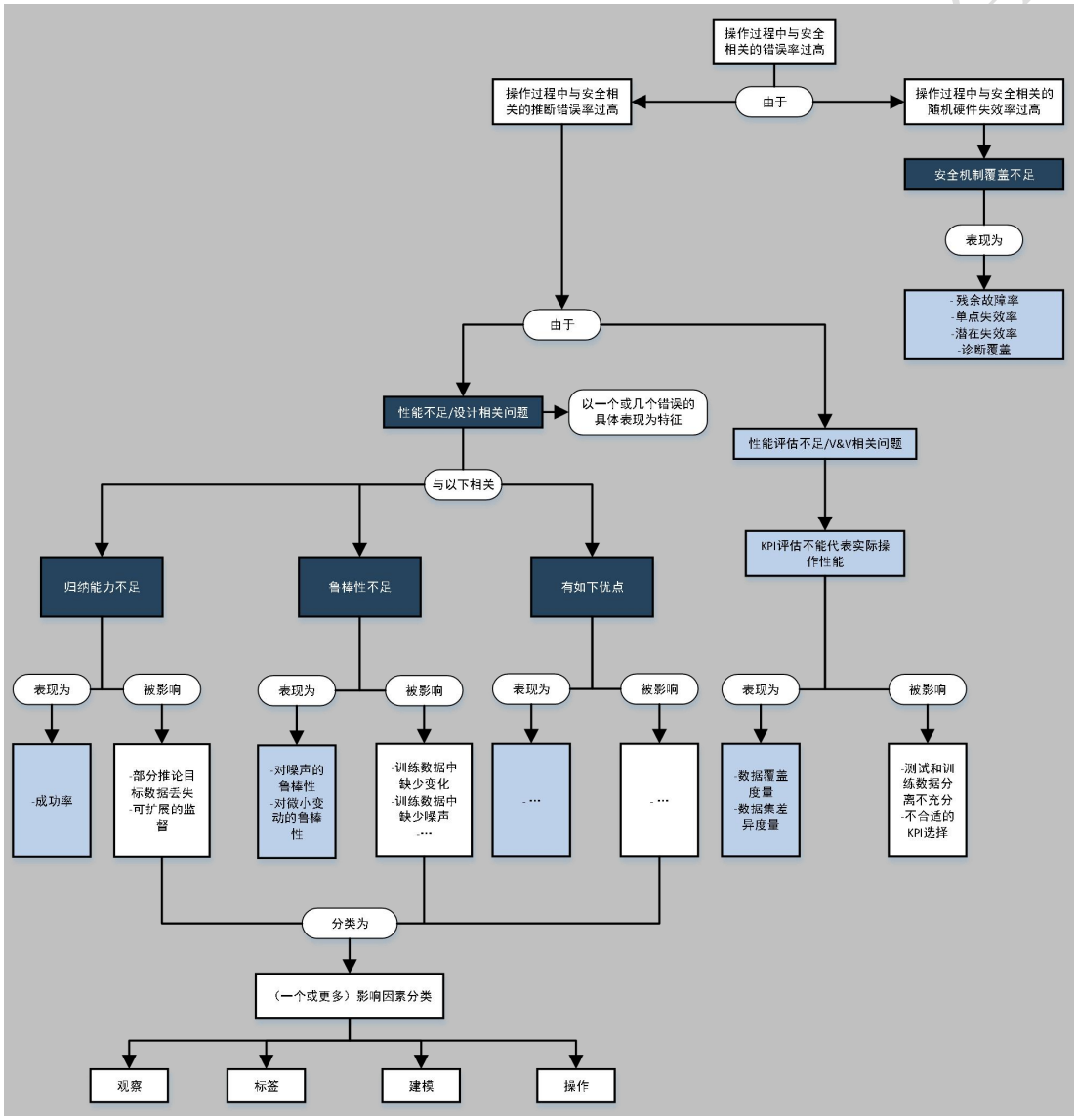


图 11-6 说明如何使用安全相关属性来帮助发现新需求的概念图

表 11-4 影响因素类别

影响因素类别	描述	示例
观察	与表示输入空间的数据相关的影响因素	-培训和测试数据的多样性 -推理目标数据的完整性
标签	与数据标签相关的影响因素	-标签不正确
建模	与机器学习模型本身相关的影响因素	-机器学习模型的选择 -超参数的选择 -训练程序
操作	与操作过程中输入空间内变化相关的影响因素	-对交通标志进行分类的机器学习系统：在其运行阶段引入新的交通标志

表 6–5 举例说明了连接不足、安全相关特性和具体安全要求。

表 6-5 利用安全相关特性得出新的安全要求

从领域观察到的不足	相关的安全相关属性	源于不足的具体安全需求	KPI 的度量	考虑到原始错误模型
在轻微输入变化的情况下，检测到的对象丢失（假阴性）	鲁棒性（模型训练中固有的，决策边界）	对于任何训练数据，添加以 L 无穷大范数 <0.01 为界的高斯噪声，预测的输出类应保持不变，且概率很高（例如 99.5%）	对于训练/测试数据中的每个图像，通过创建具有额外高斯噪声的每个图像的 1000 个变化来评估鲁棒性，并导出错误预测的数量。 如果存在受噪声影响的错误预测数大于 5 的图像，KPI 将返回“违规/错误”。	与假阴性相关的初始更高级别安全要求
当硬件故障发生	鲁棒性（违反随	对于任何训练数	对于训练/测试	与假阴性相关的

汽标委智能网联汽车分委会

时，其中一小部分像素变黑，预测可能总是错误的（假阴性）	机硬件失效）	据，通过任意地将 1%的像素改变为完全黑色，预测的输出类别应该以高概率保持不变（例如，99.5%）	数据中的每个图像，通过将 1%的像素随机设置为黑色来评估鲁棒性，重复 1000 次，并得出错误预测的数量。如果存在受噪声影响的错误预测数量大于 5 的图像，KPI 将返回“违规/错误”。	初始更高级别安全要求
注：L—无穷范数通过取每个输入维度上噪声的最大绝对值来量化噪声。例如，如果 DNN 有三个输入，并且每个输入上的噪声分别为 0.1、-0.2 和 0.15，则 L-无穷大范数=max{|0.1|、|-0.2|、|0.15|}=0.2。使用 L-无穷大范数只是一种选择；存在用于量化噪声的其他规范（例如 L1 或 L2）。				

(5) 指标、测量和阈值设计

定量证据要求定义一组指标和目标阈值。这些指标用于定义、描述和理论分析人工智能系统的特定属性。选择和设计这些指标以及相关的测量方法和目标阈值涉及以下考虑：

1) 为满足安全要求中所引用的人工智能技术的特定属性确定指标和测量方法。

2) 对人工智能系统在这些属性上进行测量和评估。

3) 分析人工智能系统所采用的设计方法对所选指标/测量方法的影响。

4) 定义和创建适用于测量每个属性的合适数据集。这些数据集旨在提供对属性进行统计相关分析，并代表输入领域。

为每个指标选择合适的目标阈值需要进行合理的论证。这个论证可以基于过去的产品经验、行业共识、系统分析、专家判断或实验。决策可能还需要权衡多个因素，考虑到机器学习中的鲁棒性-准确性权衡现象等情况。

例子：以鲁棒性对抗噪声的能力，通过最小扰动改变而导致错误预测为例。在确定验收标准时，一个基本问题是为什么选择使用Ɛ1（例如 0.5），

汽标委智能网联汽车分委会

而不是使用ε2（例如 0.25）。可以通过提供有说服力的论据来合理地论证阈值的选择，例如，为了鲁棒性必须要大于ε1，因为ε1 是从完整输入管道传输的最大可能噪声，或者ε1 基于专家达成一致的实验结果，超过ε1 的噪声对于人类来说也很难进行合理预测，或者ε1 是在类似的标准或产品中使用的数值。

(6) 小结

虽然前面的章节给出了通过安全相关属性辅助推导和细化安全需求的通用原则，但我们还是总结了一些用于需求推导的实用建议。此建议列表虽没有全面穷举，但它们共同阐述了对实现安全至关重要的补充话题。

1) 鉴于深度学习模型的适应性能力无法保证，因此需要明确说明工作空间的定义域和运行条件，通过模型训练的数据来保证模型的性能。

2) 统计学习理论的结果通常基于理想假设， 然而这些理论（如 Vapnik－Chervonenkis 理论）仍可用于推导必需样本数量的下限，从而确保了给定类型深度学习模型（如线性回归）具有较小的泛化误差。 用于模型训练的预估样本量通常只反映了最佳的情况，即训练中使用的数据（样本内）与运行中的数据（样本外）具有相同的分布。

3) 我们还应该考虑人工智能系统的信息安全引发的安全风险。特别是对于使用深度神经网络实现的组件，应研究对抗性攻击及对整个 AI 系统的影响，以了解攻击是否可以被视为是暂时的。 其他特定地区的隐私因素（如欧洲的 GDPR）虽然不会直接对安全产生影响，但可能会间接影响所收集数据的质量并影响最终模型的性能。

4) 不可否认，系统集成商对 AI 组件的错误率进行了调节。 然而在推导为训练监督机器学习模型的具体过程相关需求时，重要的是推导/生成的要求（如标注质量）旨在展示一种正向的趋势，即需求的实现有利于达成目标的错误率。

5) 出于验证的目的，最好是制定一定规范，可以提供接口以理解或部

汽标委智能网联汽车分委会

分理解基于深度学习的组件的响应。其基本原理是使得验证过程，特别是使隔离基于深度学习的黑盒组件的错误变得容易。

6) 在寻求减少特定安全需求的不确定性（如通过提高相关性能）时，我们可能会面临所选 AI 技术所固有的不同性能指标之间的权衡。 如果不能增加总体工作量来实现任何一个相互冲突的需求，就会找到一个（如在鲁棒性和准确率之间）折衷方案。这种限制需要做到系统级别。

7) 在软件工程中，制定约束工程师的规则（指质量规则）来帮助他们生产出引入错误更少的软件，质量规则对机器学习生成的产品在不确定性水平只有一点点影响，至少在具有安全影响的方面是这样的。只有在有充分理由认为质量规则可以提高质量或降低风险时，才应该使用质量规则。

11.2.3 人工智能系统选择和验证

11.2.3.1 目标

本章旨在实现以下目标：a) 选择并证明在人工智能系统中使用了适当的人工智能技术。b) 通过提供证据证明人工智能系统在集成到更高级别的系统时达到了分配给人工智能系统的技术安全要求，从而验证人工智能系统。

11.2.3.2 总体要求

1) 选择最合适的人工智能技术来满足安全要求

2) 应从考虑 AI 系统安全相关的需求和 AI 特性能否被满足的角度选择 AI 技术。

3) 应验证人工智能系统以提供以下证据：符合人工智能系统的安全要求；对不会出现非预期的功能和属性的信心。

4) AI 组件的每个测试用例应包括通过/失败标准，该标准基于分配给 AI 组件的 AI 系统安全要求中提供的阈值和参数（如果适用）的制定。根据人工智能系统的安全要求，为每个人工智能组件制定通过/失败标准可能很

汽标委智能网联汽车分委会

困难。

11.2.3.3 AI 系统的选择

1) 应从考虑 AI 系统安全相关的需求和 AI 特性能否被满足的角度选择 AI 技术。

2) 架构措施和开发措施应降低由于识别到的 AI 组件功能不足导致的违反 AI 系统安全需求的风险。

3) 为了满足 AI 系统或 AI 组件的安全相关特性，应证明适当的架构和开发措施的相关性，并合理说明这些措施的有效性。

4) 应对 AI 系统的架构实体及其输出进行安全分析，以确定是否满足分配给 AI 系统的安全需求。

5) 在合理可行的前提下，开发措施和架构措施应考虑允许运行时检测和减轻与 AI 系统相关的可能导致危害故障行为的分布偏移。

6) AI 组件的预测确定性应在设计时进行估计和校准，以获得高确定性的预测置信度。

7) 当使用同质冗余 AI 组件时，架构措施应防止共因失效。

8) 如果适用，应使用开发措施和架构措施来缓解 AI 组件的输出不足(在运行时)。

9) 架构和开发措施应有助于纠正 AI 组件的危害故障行为。

11.2.3.4 AI 系统的验证和确认

AI 系统的验证适用于 AI 系统安全生命周期的以下阶段。

1) 当定义 AI 系统安全相关要求时，验证确保这些要求是正确的、完整的、以及相互之间是一致的，并且与系统级技术安全概念和安全要求相一致。

2) 在 AI 系统的开发阶段，可以不同形式进行验证，描述如下:在设计

汽标委智能网联汽车分委会

阶段，验证是对样件的评估，例如架构设计、模型或架构方法，从而确保它们符合 AI 系统 对正确性、完整性和一致性的安全要求；在数据生命周期中，每个阶段的数据都要进行验证，以确保足够的正确性、一致性和完整性；在测试阶段，AI 系统的验证是对测试环境中样件和各要素的评估，以确保符合它们的要求。以一种系统性的策划、规范、执行、评估和记录方式进行测试。

(1) AI 部件测试

1）AI 部件测试流程

首先分析分配给 AI 部件的安全要求，以选择最合适的测试方法。安全要求也用于测试生成，其中包括被测 AI 部件的驾驶测试用例，以及测试目的的输入、方法、环境和工具的期望定义。AI 系统的安全要求包括 AI 系统的安全属性和安全性能指标。

接下来是应用于判定测试结果通过/失败（测试谕示（测试指令））的公式化。测试输入从收集的数据中取样，或使用模拟或其他技能生成。评估测试用例，以确保它们有效地覆盖了所考虑的场景空间。

在测试生成和测试用例评估之后，AI 部件将与测试输入一起运行，并生成测试结果。根据通过/失败的判定准则（测试谕示（测试指令））评估测试结果。如果测试失败，则进行安全分析以评估安全方面错误的影响。

当评估由错误引起的风险时，必须考虑要求的目标值和关键性能指标（KPIs）。如果与安全要求相关的相应目标值和 KPIs 不管怎样还是得到了满足，则有可能单个错误的出现不会违背安全要求。

2）测试生成

测试数据生成是人工智能系统验证中的一项重要活动。测试数据集需要包含代表人工智能系统在操作中接收的输入的数据元素。例如，用于感知组件的基于摄像头的 AI 系统接收道路场景作为输入，这些输入依赖于特

汽标委智能网联汽车分委会

征的 ODD，因此测试输入场景需要在整个 ODD 内有足够的覆盖率。人工智能系统的安全需求可能包括一些与功能不足相关的需求，测试这些需求需要基于 ODD 生成边缘场景。系统性地探索输入域，以充分覆盖常规输入和触发功能不足的边缘场景是必不可少的。通常，在输入域上可能会有一个常规和边缘场景的预期分布。

3）测试用例评估

当生成测试用例时，必须对它们进行评估，以确保它们可以充分覆盖人工智能组件的人工智能系统在定义的操作域内的安全需求并有效地覆盖所考虑的场景空间。

(2) 人工智能系统安全确认

和验证不同得是，人工智能系统安全确认是指在人工智能系统集成到更高级别的系统中后，检查是否满足了对人工智能系统的安全要求（来自更高层次系统）的活动。基于 SOTIF 思想，人工智能系统安全确认活动应由定义了确认目标的系统集成商（如整车厂）完成，人工智能系统开发方可能需要支持这一活动。

对于人工智能系统的安全确认重点应该放在对设计运行域内所有相关场景的系统性探索和异常状况的检查上。通过输入空间第一离散化进行的系统随机测试就是此类方法中论证相对完整性的一个例子。

对于深度神经网络，可以在用检测分布外数据的主动学习技术或其他技术的帮助下，用场地测试的方式（例如，通过使用自动驾驶车队进行测试）进行人工智能系统安全确认。

11.2.4 运行期间和持续保证的措施

11.2.4.1 目的

本章节目的是：a) 定义流程要求，以确保 AI 系统部署后的持续安全；b) 定义识别人工智能系统在运行期间的安全风险的措施，以及确保 AI 系统持续安全

的措施；c) 确保系统响应能解决已识别的与 AI 系统相关的安全风险，并确保在发布之前重新批准修改后的 AI 系统。

11.2.4.2 先决条件

应提供以下信息：

1) AI 系统设计及其与车辆系统的接口；

2) AI 系统的使用假设；

3) AI 安全要求，包括安全准则 (例如 KPI)；

4) 架构措施，帮助确保 AI 系统运行中的安全；

5) 数据集要求规范、设计规范和维护计划；

6) AI 系统已知功能不足的验证和确认活动结果(如果有)。

11.2.4.3 一般要求

1) 应定义持续的安全保证流程，包括 AI 运行阶段控制不合理风险的安全活动。

2) 应使用车载和/或非车载安全措施来监控 AI 系统的运行状态，发现与安全相关的异常情况，并防止运行过程中的风险，包括针对现场误用的措施。

3) 应对已识别的安全相关现场事件进行适当评估，并采取应对措施控制风险。

4) 应在实施后对应对措施的有效性进行评估，如果残留风险不可接受，则修改应对措施。

5) 在运行阶段，应进行维护活动，包括现场数据收集、AI 再训练、再验证和再批准，以持续确保和改进 AI 的安全性至合理水平。

汽标委智能网联汽车分委会

11.2.4.4 运行和持续保证的计划

(1) 人工智能系统运行阶段的安全风险

在安全发布阶段，根据开发阶段产生的证据和假设，将剩余风险评估为可接受。 然而，部署后现场风险评估可以检测到与人工智能系统相关的风险升高，这是由于以下原因造成的危害：

1) 开发不确定性，例如：剩余风险估计不正确；以前未知的危险的功能不足；对运行过程中发生的人工智能相关故障的发生估计不正确；

2) 不正确的意外操作特定活动，例如：在维护期间，改装相机或雷达时，未经重新校准或公差比较差。

3) 运行环境的变化，例如：新的交通规则；新的交通设施；新型交通参与者；前期假设的变化。

在运行阶段需要采取特定的措施来应对这些风险并持续确保人工智能的安全。

(2) 运行阶段的安全活动

图 11-7 说明了本章节的活动流程，以确保人工智能在运行阶段的安全。

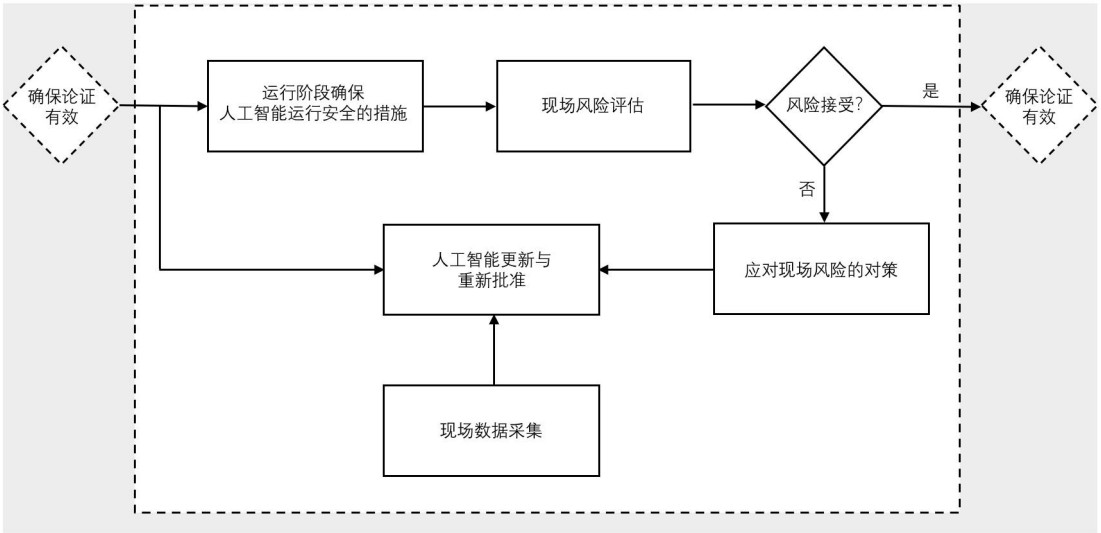


图 11-7 运行阶段的安全保证

运行阶段的安全保障始于在运行阶段确保人工智能安全的措施。 应用技术

措施监控人工智能系统的运行行为，如果检测到异常和危险行为，则采取缓解措施。 引入了一些额外的措施来解决与误用相关的风险，例如用户指南。

评估现场识别的风险。 如果表明已识别的风险水平是可以接受的，则不再采取进一步的活动。 或者，如果发现风险不可接受，则根据评估结果确定对策。

运行阶段的数据采集可用于支持人工智能系统的功能细化和改造。 这包括在危险场景或未知条件下收集的数据。 基于这些数据，人工智能系统可以进行更新和重新批准。

11.2.4.5 AI 系统运行过程中的安全措施

监控、检测和减缓是确保 AI 系统运行过程安全的有效措施，用于评估 AI 系统的错误或不足的行为。这些措施可以依靠内部机制和/或外部机制（例如云监控）。

关于监控和检测，以下事件及相关信息可用于报告和发现 AI 系统的缺陷或错误（如适用）:

1) 与输入空间相关的事件：分布外的数据检测；超出运行环境的检测；概念偏移或特征变化的检测（例如，新对象，不同的行为，新的或改变的规则等）。

2) 与模型行为相关的事件：异常行为检测和异常输出检测；

3) 意外事件/事故分析：AI 系统直接或间接涉及的意外事件或事故（包括公开可得的）可能会造成伤害或触发非常紧急的安全策略（例如紧急操作）。对这些意外事件/事故进行分析，以支持 AI 系统的改进。

除了触发改进活动，运行过程中识别出的 AI 系统错误或不足可能表明开发和安全保证流程、架构措施的无效性，或这些措施使用的假设的不正确性，因此可能需要改进这些措施。

汽标委智能网联汽车分委会

11.3 总结与展望

车用人工智能功能安全正成为保障智能驾驶系统可靠运行的关键课题，尤其在深度学习和数据驱动模型广泛应用背景下，传统功能安全标准已难以应对 AI 系统所带来的感知误差、预测偏差和决策不确定性等非传统失效模式。功能安全不仅需要关注 AI 模型输出的正确性，更需要确保其行为逻辑合理、系统过程可控，需从系统工程角度系统化建模并管理 AI 的不确定性风险。

在安全生命周期管理方面，当前实践已涵盖需求定义、设计冗余与容错机制、对抗测试、不确定性量化、实时运行监控、紧急响应机制，以及结合现场数据进行持续优化等环节。国际上，ISO 26262、ISO 21448、ISO/TS 5083 和 ISO/PAS 8800 等标准为车载 AI 系统提供了分层次的安全保障参考，联合国 ECE 法规也在推动与国际标准的协同发展。与此同时，国内标准体系正在加速构建，以适应中国特色道路环境和智能网联汽车产业需求。

AI 系统开发过程中，强调将安全性纳入迭代开发流程，与传统功能安全体系深度融合，通过错误模式分类与因果分析识别规范不足、系统性失效和性能边界。安全需求推导以系统级需求为基础，结合输入域特征与安全属性进行量化与定性控制；验证方法涵盖设计、测试、运行全过程，重点强化对操作设计域和边缘场景的测试覆盖。运行阶段依赖于风险监测与动态反馈机制，实现系统的持续安全保障和更新验证。

汽标委智能网联汽车分委会

第十二章 车用 AI 伦理安全标准化研究

12.1 车用 AI 伦理安全概述

12.1.1 车用 AI 伦理安全定义

人工智能伦理是新兴概念。2021 年国家新一代人工智能治理专业委员会发布《新一代人工智能伦理规范》，旨在将伦理道德融入人工智能全生命周期，提出六项基本伦理要求，包括增进人类福祉、促进公平公正、保护隐私安全、确保可控可信、强化责任担当以及提升伦理素养，同时针对人工智能在管理、研发、供应、使用等特定活动中的情况，提出十八项具体伦理要求[76]。国家标准 GB/T 41867-2022《信息技术 人工智能 术语》将“人工智能伦理”定义为“开展人工智能技术基础研究和应用实践时遵循的道德规范或准则”。2022 年 3 月，中共中央办公厅、国务院办公厅印发了《关于加强科技伦理治理的意见》，明确指出科技伦理是开展科学研究、技术开发等科技活动需要遵循的价值理念和行为规范，是促进科技事业健康发展的重要保障[77]。根据国家人工智能标准化总体组、全国信标委人工智能分委会 2023 年 3 月发布的《人工智能伦理治理标准化指南》人工智能伦理内涵可从三个层次结构：一是人类行为准则层面，在人工智能技术研发、产品设计及系统应用全过程中需遵循的道德准则及行为规范；二是系统设计机制层面，强调通过技术手段将伦理准则嵌入人工智能本体；三是自主演化能力层面，指人工智能体在深度学习中自主形成的伦理判断机制[78]。

欧盟《可信赖人工智能伦理准则》对人工智能伦理框架进行了系统性建构，其核心内涵包含三个维度：在价值导向上强调人类中心原则，要求维护人类决策自主权并预防技术滥用风险；在运行机制中确立公平可溯原则，主张算法决策系统需消除歧视性偏差并保持可解释性；在技术特征维度提出透明可控原则，强调人工智能运作逻辑应具备可验证性与可干预性。

汽标委智能网联汽车分委会

日本发布的《以人为本的人工智能社会原则》着重指出，人工智能的应用需以全人类和社会公众利益为服务导向。该文件强调，人工智能的发展要以实现尊重人类尊严、包容多元性、推动可持续发展等社会基本发展理念为目标。此文件还清晰界定了人工智能应满足的一系列社会原则，具体包括：以人为本、教育/扫盲原则、隐私保护原则、确保安全原则、公平竞争原则、公平、问责制和透明度原则及创新原则。

车用人工智能伦理安全是指在车用人工智能系统的全生命周期中制定并遵循伦理原则和道德标准，并通过技术手段和管理手段，保障人类权益（生命权、隐私权、公平权等）和确保人工智能产品安全可控、透明可追责，促进公平、公正、和谐、安全，避免偏见、歧视、隐私和信息泄露等问题。

12.1.2 车用 AI 伦理安全原则

车用人工智能应当坚持全国汽车标准化技术委员会于2024年7月发布的《汽车行业人工智能技术标准化发展倡议书》中提出的“坚持以人为本、追求公平公正、遵守法律法规、明确追溯途径、增强透明可释、改善生存现状、保障用户隐私、尊重人类决策”八项基本伦理原则，努力为人类提供安全、高效、便捷、舒适的交通工具与环境，提升车用人工智能系统的可信度与可接受性，推动交通行业创新发展与转型升级。

12.1.2.1 安全性

车用人工智能安全性原则以以人为本为核心，保障人类权益，提升模型与系统的成熟度等，确保多方面安全。在车载数据安全上，要明确车辆数据特别是隐私数据获取的技术要求与必要性验证，鼓励用优质数据训练模型，加强使用监控，采用先进安全措施并不断审查。为保障系统自身安全，在投入使用前对车用人工智能系统进行稳健、可靠、可重复和标准化评估，发布前进行外部安全测试。对于车辆外部环境安全，倡导发展安全可靠的生成式车用人工智能，确保功能稳定、提示及时、处置有效，避免技术滥用。在发展和应用中，坚持以人为本是关键，要全面考虑伦理、法律和社会影响。同时，鉴于技术无国界，国际协同保障车用

汽标委智能网联汽车分委会

人工智能安全性愈发重要。国际社会应达成共识，挖掘潜力并降低风险，让技术符合全人类利益。

12.1.2.2 公正、公平和透明

车用人工智能技术需尊重不同人群，避免因数据局限等产生歧视与偏见，公平公正对待社会成员，且其内部运行机制应可被理解和解释。

无偏见算法：开发与应用车用人工智能系统时，杜绝种族、性别等方面的偏见。严格筛选审查数据集，采用具有地理、社会经济和文化背景多样性的数据集训练，保证系统在各环境下公平表现。

平等对待与公平访问：设计和使用系统时制定统一标准，确保所有用户同条件下待遇相同。让不同群体公平享受技术益处，避免扩大社会不平等，定期评估系统性能与公平性。

算法透明性：公开系统的设计、数据来源、训练过程、决策逻辑等信息，包括算法原理和数据来源，保障数据合法可信。定期开展公众咨询活动，增强用户信任。

算法可解释：系统应能清晰说明决策过程和依据，方便用户理解与监督。开发解释性工具，使用户可直观查询决策依据，增加信任感。

12.1.2.3 可问责

人工智能的可问责性指在其设计、开发、应用等过程中，能对行为和决策解释、评估，事故时明确责任者及内容，并改进升级系统。各类责任主体要有责任感，遵守法规伦理，担起相应责任，促使用户负责使用车用人工智能。

强调研发者责任：研发者要预见问题、规避风险，确保产品安全稳健，记录研发日志以便溯源。保证可控性，赋予用户控制权，提供操作指南，帮助用户及时化解危机。提高可预测性，让用户能预测系统行为，对反常表现给出合理解释。

汽标委智能网联汽车分委会

完善责任与赔偿制度：建立公正透明的责任追究制度，明确责任分配与内容，杜绝推诿。事故调查中各主体积极配合，提供信息，赔偿受害者。要求改进并落实，建立保险制度保障用户。

加强社会监督与伦理评估：修改责任制度时与利益相关者讨论，听取意见。主动接受监管与公众监督，确保符合规范与法规。对系统进行伦理审查评估，保证决策合伦理。

12.1.2.4 隐私保护

本文件所指隐私涵盖用户个人信息及用车中人工智能系统自动生成的数据。车用人工智能尊重保护隐私，依“合法、依规、诚信”原则处理信息，遵守中国的网络安全法、个人信息保护法等相关法律与标准，保障信息数据安全。

数据采集的知情同意与选择，通过多种显著方式告知用户信息处理情况，赋予选择权。

最小化采集使用：仅收集车用人工智能仅收集实现其功能必要最少数据，默认不收集，用隐私保护算法确保安全。

可追溯数据系统：在数据采集、存储、使用、转移等所有处理环节明确责任人，建立授权和追溯机制。

先进隐私保护技术：采用匿名化等技术，对敏感信息额外保护，实施数据保护机制，包括数据加密、访问控制和审计等。

用户数据处置权：用户有权访问、更正、删除数据，控制使用共享，有数据携带权，即将某公司车用人工智能系统收集的个人数据无障碍地直接转移到新的AI 系统。

应急赔偿方案：制定隐私泄露应急预案，及时完善相关管理措施与技术手段，并给予受害者合理的赔偿。

及时更新条例技术：定期审查更新保护措施，根据国家政策和用户反馈意见及时更新隐私保护条例。

汽标委智能网联汽车分委会

12.1.2.5 尊重与自主性

人工智能发展必须坚守"以人为本"原则，在生命安全、身心健康、文化归属三个维度构建保障体系，同时严格防范技术误用与滥用。车用 AI 的基本目标在于：通过替代危险驾驶行为实现出行自由，借助情感交互系统（包括主动关怀、场景化交流等功能）满足多元群体心理诉求，最终达成效率提升与生活品质优化的双重目标[79]。此外，AI 技术应用需构建可持续发展框架：以智能制造、智慧新能源、智慧交通为支点，平衡经济增长、社会进步与生态保护的三角关系。重点应对贫困、资源短缺、环境污染等系统性挑战，在气候变化应对、生物多样性保护、公共安全防护等全球议题中发挥技术赋能作用，形成人—社会—环境协同发展的良性循环。

12.2 车用 AI 伦理安全风险识别、评估与应对策略

12.2.1 车用 AI 伦理安全风险识别

近年来，人工智能技术飞速发展，车用人工智能进展显著。与传统辅助系统相比，它涉及更多感知、决策与控制元素。其功能按智能化程度大致分为车辆健康监测、智能驾驶辅助、自动驾驶三类，自动驾驶等级根据国际自动机工程师学会（SAE）建立的自动驾驶分级标准（SAE J3016）与我国的《汽车自动化分级》（GB/T 40429-2021）进一步划分，可见汽车将走向高度自动化、智能化。

车用人工智能发展目前面临三大核心问题：道路安全风险，如自动驾驶在复杂路况中识别失误可能引发事故；隐私与数据安全，海量训练数据存在泄露和滥用隐患；伦理决策困境，紧急情况下的算法选择易引发道德争议。这些问题贯穿产品研发、应用及服务全流程，需技术规范与法律监管协同解决。

12.2.1.1 主要的风险类型识别

车用人工智能在伦理规范与安全治理中面临多维度的复杂挑战，其风险总体涵盖安全漏洞、算法透明度、责任归属、隐私保护、道德困境和决策歧视六大关键领域[80]。这些挑战既涉及技术系统的脆弱性与不可解释性，

汽标委智能网联汽车分委会

又包含价值判断的伦理争议，需要通过技术约束、法律规制和伦理框架的协同创新来实现系统性风险防控。

表 12-1 主要风险识别表

风险领域	关键风险类型	具体风险描述
安全与操作风险	系统设计与漏洞	软硬件组件设计缺陷或漏洞会使系统不稳定、易受攻击，如智能驾驶系统漏洞可被黑客利用致事故
	网络安全与攻击	通信系统漏洞或受攻击会使系统失效，威胁驾驶与交通
	人为操作与误用	因由人类操作管理，操作错误或滥用会使系统故障，如驾驶员误操作自动驾驶系统可致事故
	供应链安全	由多供应商提供组件服务，某一供应商产品有问题会影响系统安全稳定
透明度与可解释性	算法黑箱	复杂模型（如深度神经网络）因结构复杂导致决策过程不可视，自动驾驶的紧急避让等关键决策缺乏可追溯依据，导致削弱用户信任基础
	数据偏见传导	训练数据偏差会固化模型偏见，特定群体/地域数据缺失可能导致系统性误判，而开发者往往无法清晰披露偏见的形成路径与影响程度
	责任认定困局	黑盒模型存在不可预见的错误风险且难修正，当事故发生时，决策逻辑的不可解释性将导致责任主体难以界定，可能引发法律真空争议
监督与责任风险	事故责任归属	自动驾驶技术发展下，归属是关键，可能涉及制造商、开发商、所有者、驾驶员等

汽标委智能网联汽车分委会

	算法决策责任	车用人工智能依赖机器学习算法，受数据偏差等因素的影响可能决策错误。
	人机交互复杂性	车用人工智能与人机交互复杂，可能导致信息传递不清晰等问题
隐私保护风险	信息收集泄露	车用人工智能收集驾驶员、乘客个人及车辆敏感数据，若被第三方获取，会导致隐私泄露、财产损失与精神困扰
	驾驶行为监控	车用人工智能可能会记录车辆位置等数据，分析可得用户行为习惯等敏感信息，滥用会侵犯隐私
	数据共享风险	第三方数据滥用侵犯隐私权，引发法律问题
道德困境	价值判断抉择	"电车难题"中的生命量化困境，乘客与行人保护优先权争议
	程序化牺牲逻辑	预设的少数牺牲多数决策伦理悖论，个体权利与公共利益的冲突
决策偏见与歧视	数据驱动偏见	历史数据中的种族/性别偏见传导，区域数据不平衡导致决策偏差
	模型结构性歧视	特征表达不足引发的区域/经济地位偏见，算法设计者的主观偏好固化

12.2.1.2 风险成因分析

在车用人工智能的伦理与安全问题中，除一些自主决策的道德困境外，大多数可能导致伦理风险的问题可以归于产品生命周期中各个阶段。

汽标委智能网联汽车分委会

研发阶段的伦理安全问题主要有算法偏见、侵犯隐私、安全漏洞与透明度缺失[81]。开发数据缺乏代表性或存在偏见，以及特征选择不当、标注错误、过拟合等，会导致算法出现偏见；数据收集存储时若存在安全隐患、权限控制不当或共享无保护措施，易侵犯隐私；软硬件设计缺陷、使用未经验证的第三方组件、忽视更新维护，会带来技术安全漏洞；算法 “黑箱” 及团队文档不清晰、沟通不足，导致透明度缺失，增加监管难度并引发信任危机。

生产销售阶段主要涉及的伦理安全风险主要归因于产品缺陷问题以及误导性销售行为。产品制造缺陷主要包括质量问题、供应链问题与工艺技术问题。误导性销售行为主要包括过度宣传与模糊性能信息问题。这些行为都有可能造成人工智能在后续使用过程中出现偏差，甚至引发危害后果进而产生归责问题。例如生产过程中的质量控制措施不足，可能导致产品存在功能缺陷或性能不稳定。这包括生产线的自动化控制系统出错，或是人工检查环节的疏漏。或工艺技术不足以精确处理组件和系统的集成。

产品使用阶段存在用户依赖过度、交互问题与更新维护问题，后续监管阶段面临检测不足、技术迭代兼容及法律法规更新带来的合规问题。车用人工智能使用阶段，伦理安全风险聚焦用户体验与系统表现。消费者因对人工智能功能理解有限易产生过度依赖，如误认自动驾驶系统在全路况和天气下都安全，实则其仅适用于特定环境；用户界面设计不直观会致操作失误、提升事故风险；系统更新维护至关重要，过时软件存安全漏洞，新更新或引入缺陷，均威胁车辆安全。

后续监管阶段，持续监测维护成挑战。需对车用 AI 全生命周期监控以保安全可靠，但监管机构受技术资源限制，难有效监控所有车辆，且法规变化会使原数据处理和隐私保护措施可能不合规。故车企与监管机构需建强合作机制，更新维护系统、合规操作、快速响应问题，以技术更新、用户教育等全面策略应对风险挑战。

汽标委智能网联标准分委会

12.2.1.3 风险类型识别的方法

识别车用人工智能的伦理安全风险是一个关键的步骤，它确保了其在实际应用中能够遵循伦理原则并保护用户的权益。详细的识别过程需要多方机构的合作，以确保全面和细致地评估潜在风险。

识别风险类型需把握关键环节。一是将功利主义等伦理原则转化为数学表达，指导 AI 决策；二是构建风险评估框架，从数据收集到应用评估全流程分析风险并提出对策；三是通过法律政策明确 AI 伦理风险类别与各方规避义务；四是依靠技术社群，发挥其在要素认定和技术问题上的敏锐性；五是结合全球经验与本土实践，明确治理规则，关注动态识别；六是加强用户教育，鼓励用户参与 AI 伦理安全风险识别与治理。风险类型识别方法包括建立系统化的风险评估过程、倡导产品开发者、使用者等多方主体反馈与持续监控等。

系统化的风险评估过程：

风险定义和分类：首先，需要明确定义什么构成伦理安全风险，并据此分类。常见的风险类型包括隐私泄露、决策偏差、算法歧视等。此步骤需要伦理学家和行业专家共同确定哪些问题属于伦理风险，并对它们进行分类。风险映射：利用专业知识映射出人工智能在不同操作环境中可能遇到的具体风险情景。例如，在自动驾驶模式下，系统如何处理紧急避障情景，是否可能牺牲一方利益以保护另一方。映射工作通常需要人工智能开发者、安全专家和行为科学家的合作。潜在风险识别：通过模拟测试和历史数据分析，识别系统可能的风险行为[82]。这一步骤通常涉及大量的数据分析，需要数据科学家和人工智能工程师密切合作，通过机器学习模型和统计工具发现数据或决策中的潜在偏差。

多方主体共治：组织利益相关者会议：定期组织包括用户代表、消费者权益保护组织、法律顾问、伦理学家和技术专家在内的多方利益相关者会议。在会议中，各方可以提出他们对车用人工智能系统可能引发的伦理和安全问题的看法和担忧。收集和分析反馈：通过问卷调查、公开论坛和社交媒体等途径广泛收集用

汽标委智能网联汽车分委会

户和公众的反馈。利用这些信息，评估社会对车用人工智能技术的接受程度，以及人们对可能的伦理风险的敏感性。

持续的监控和审计：实施持续监控系统：在车辆运行过程中实施实时数据监控系统，追踪人工智能决策过程中的关键参数和行为。这可以帮助及时发现问题，如偏差过大、异常行为等。定期安全审计：设立独立的审计团队定期对人工智能系统进行安全和伦理审计，确保系统的运行符合最初的设计伦理标准和法律要求。审计团队应包括外部第三方专家，以保证审计的公正性和客观性。

12.2.2 车用 AI 伦理安全风险评估体系

随着人工智能技术的飞速进步，其在汽车行业的应用变得愈发广泛和深入。为确保车用人工智能系统的安全性、可靠性和伦理合规性，构建一个综合的车用人工智能伦理安全风险评估体系变得至关重要。这一体系旨在全面、系统地评估和管理车用人工智能系统在整个生命周期中的伦理安全风险，包括产品开发、训练、营销、投入使用以及监管等各个环节。

通过这一评估体系，我们期望能够确保从产品设计到监管管理的每一个环节都充分考虑到伦理安全的重要性，并采取相应的预防和管理措施。这不仅有助于企业和政府部门更好地理解和管理车用人工智能的伦理安全风险，降低潜在的安全隐患，还能够提高系统的可信度和用户的信任度，促进车用人工智能行业的健康发展，推动技术创新和社会进步。在保障用户安全和隐私的同时，充分发挥车用人工智能技术的潜力，为社会和经济的可持续发展做出积极贡献。这一体系将是我们迈向更加智能、安全、可靠的未来出行方式的重要一步。

12.2.2.1 评估体系

建立汽车用人工智能伦理安全风险评估体系，需设计全面严谨框架，覆盖产品全生命周期。纵向划分产品开发、训练、营销、使用及监管阶段；横向区分“驾驶任务”与“非驾驶相关任务”。其中“驾驶任务”包含车辆运动控制、目标探测响应等，“非驾驶任务”指驾驶技术解放注意力后衍生的休闲、娱乐等无关事项。

由此可见，“驾驶任务”是车用人工智能伦理应探讨的核心；“非驾驶任务”所涉及的 AI 伦理安全则以通用领域的人工智能伦理为主，其在车用情景中的特殊性，主要体现在与“驾驶任务”相连接时可能触及的伦理安全问题。

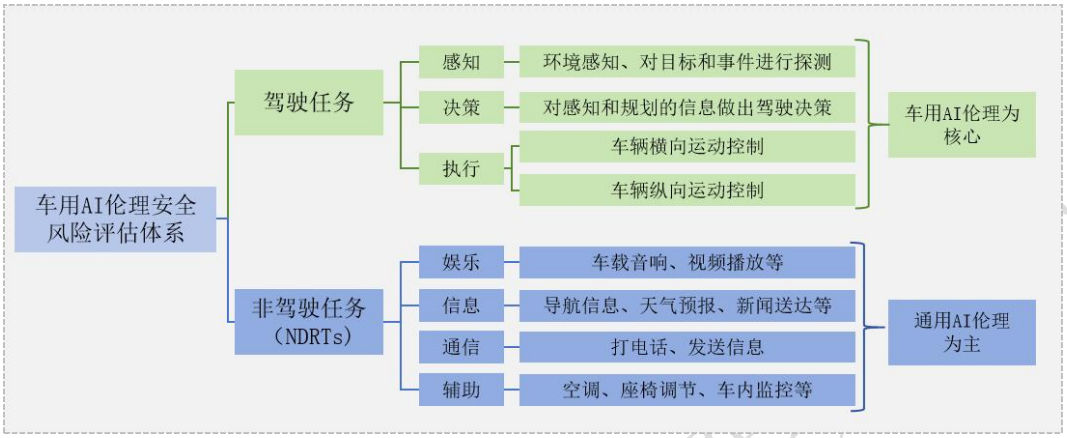


图 12-1 车用 AI 伦理安全风险评估体系框架图

车用人工智能伦理安全风险评估体系需全面覆盖“驾驶任务”全生命周期，从设计到监管各阶段均重视伦理安全并采取对应措施。该体系旨在助力企业与政府理解管理风险，降低隐患，提升系统可信度与用户信任度，最终推动行业健康发展，促进技术创新与社会进步，在保障用户安全隐私的同时，释放车用人工智能技术潜力，为可持续发展贡献力量。

(1) 产品开发与数据训练阶段

产品研发阶段，评估体系涵盖伦理审查委员会、数据管理审核与保护政策评估及动态风险评估模型构建。开发初期组建由多领域专家构成的伦理审查委员会，审查关键决策，规避隐私侵犯与数据偏差，保障数据合规使用。落实数据最小化原则，做好匿名加密处理。搭建专用沙盒，借动态风险评估模型融合定量定性分析，利用故障树分析等识别风险并分级处置[83]。同时，定期扫描测试漏洞，开发可解释模型，提供决策文档，增强透明度。

评估旨在保障数据安全、纠正算法偏见并预防模型过拟合，风险评估系统需涵盖数据集质量与数据偏见评估。训练数据样本特征差异易引发两难：修改数据或致过拟合，影响决策表现；完全依赖则可能滋生算法歧视。平衡数据事实差异

与潜在歧视，需审慎权衡。行业标准倡导构建具备上下文意识的模型，运用加权重采样技术，提升模型透明度与可解释性。同时，应遵循伦理法律要求，依标准处理数据、保护隐私、确保公平，并通过用户教育与反馈机制，优化系统公正性与可靠性。

(2) 营销与投入使用阶段

在车用人工智能的营销和使用阶段，伦理安全评估体系的构建旨在确保所有市场行为的透明度、诚信性，以及产品在实际使用中的安全性和可靠性。这些措施应彻底整合进企业运营和客户交互中，确保所有潜在的伦理和安全问题得到妥善管理。

在产品营销阶段，伦理安全评估体系包括广告宣传审核与双反合规审核。在营销活动中，所有关于车用人工智能的宣传材料和广告都必须经过严格审查，确保其描述准确无误，不夸大其功能。例如，自动驾驶系统的能力应明确标注其操作限制，如“仅限高速公路使用”或“需要在特定天气条件下手动操作”。这种做法可以防止消费者对产品性能产生误解，从而降低由不当期望引起的风险。遵守所有相关的广告和消费者保护法律是车用人工智能营销的基础。

此外，营销团队应接受关于如何负责任地推广高科技产品的培训，这包括理解伦理营销的重要性和如何处理敏感数据的问题。

产品使用阶段聚焦产品监控、用户反馈与技术适配评估。销售后需对车用人工智能持续监控，可搭建实时数据监测系统，捕捉车辆性能异常，一旦发现问题，立即启动事故响应，远程控制车辆、触发紧急停车并通知维修。同时，建立客户服务热线、在线聊天等多渠道用户反馈机制，借用户满意度调查收集需求，推动产品改进与风险管控。此外，鉴于技术与市场变化，制定含软件更新、硬件检查、安全审查的维护计划，保障系统符合最新安全标准，抵御新兴威胁。

(3) 后续监管阶段

对产品监管机构而言，构建有效的车用人工智能安全评估体系意义重大，需融合法律、技术、伦理及消费者保护多维度，确保产品符合社会伦理与法律规范。

汽标委智能网联汽车分委会

监管机构应以政策法规为根基，搭建风险评估系统。联合行业专家制定含数据隐私等原则的伦理准则，构建综合监管框架；要求制造商提交详细伦理风险评估报告，阐述数据处理的伦理挑战与应对策略。对新车用人工智能产品严格开展第三方独立前期审查，上市后通过定期检查与不定期抽查进行动态监管，同时重视反馈、推广教育，并开展国际合作，全方位保障产品合规。伦理风险评估体系需设有效反馈机制，以便消费者、用户等利益相关方报告伦理问题。监管机构应组建快速响应团队，对反馈展开调查并按需纠偏。该机制既能及时发现解决问题，又能提升公众对车用 AI 的信任。为强化行业伦理标准，监管机构需推广伦理教育与意识提升活动。可通过举办研讨会、工作坊、公开课，发布指导手册与最佳实践指南，助力制造商和用户理解践行伦理原则。鉴于车用 AI 技术的全球性，监管机构应参与国际合作对话，以共享经验、协调标准，共同应对全球性伦理挑战，进而更有效地处理跨境数据流动、国际市场准入及全球消费者权益保护等问题。

12.2.2.2 评估方法

就车用人工智能伦理安全风险评估，可以从多个方面与维度共同开展，具体方法如下图所示：

汽标委智能网联汽车分委会

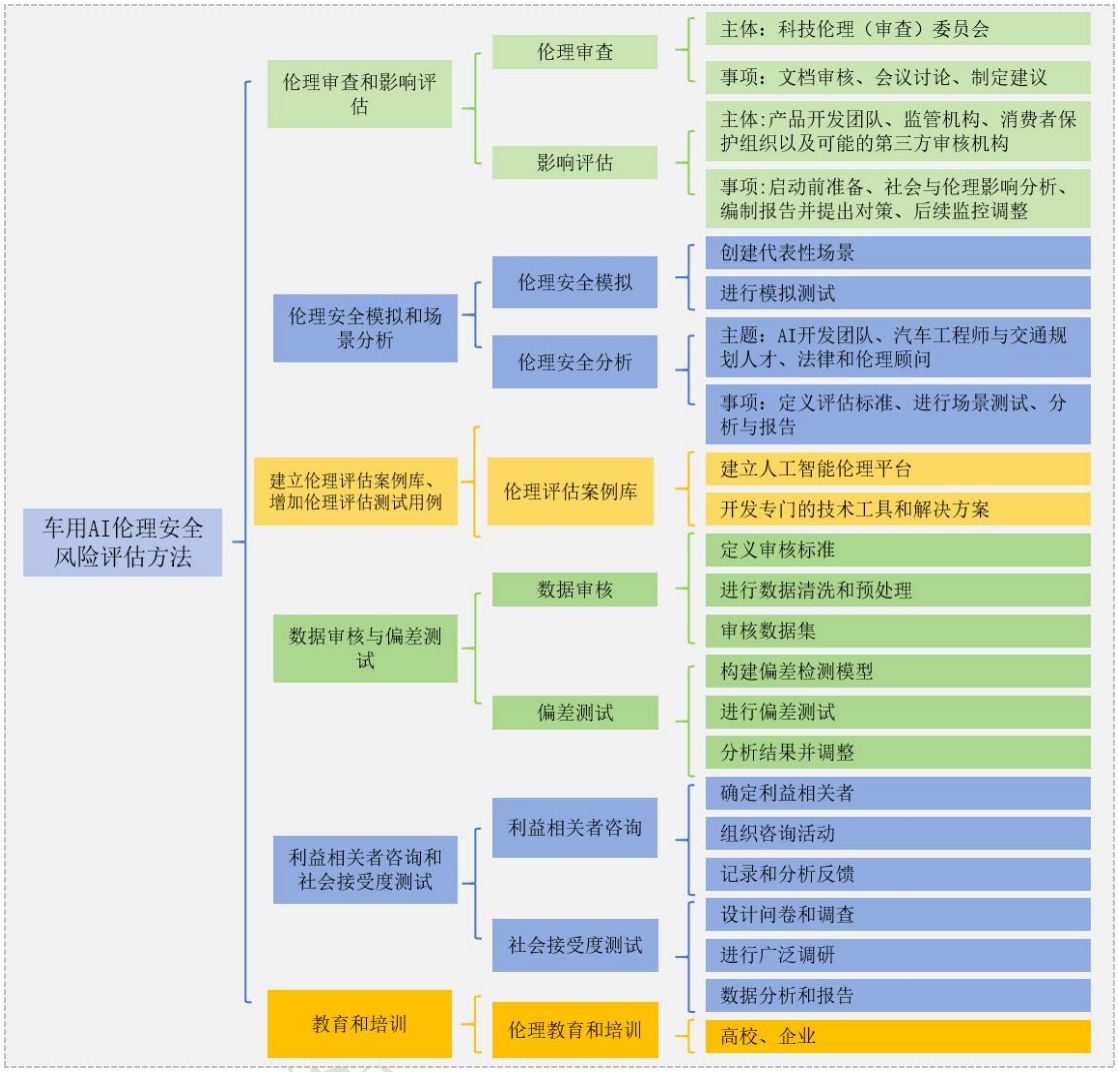


图 12-2 车用 AI 伦理安全风险评估方法框架

(1) 伦理审查和影响评估

伦理审查和影响评估是评估车用人工智能伦理安全风险的重要组成部分，确保这些技术在设计和部署过程中考虑了广泛的伦理、社会和法律问题。

伦理审查由伦理委员会执行，成员涵盖伦理学家、法律专家、技术开发者、社会学家及公众利益代表，委员会可由企业、行业协会、学术机构或政府部门组建。审查流程含文档审核、会议讨论与建议反馈：文档审核聚焦设计文档，评估数据采集处理透明度、隐私保护及算法设计是否存偏见；会议定期探讨系统伦理

问题，如自动驾驶紧急决策对多方影响；最终委员会依审核与讨论结果，提出改进数据处理、提升算法透明度等建议，指导开发团队优化设计。

影响评估是贯穿产品全周期的系统性工作，需产品开发团队、监管机构等多方协作。其进程分为四步：启动前，项目团队在设计初期与监管机构合作，明确车用 AI 系统潜在影响，确保评估框架合规；随后开展社会与伦理分析，运用定性定量方法，联合社会学家、用户代表等剖析产品对特定群体影响及社会伦理问题；接着编制报告，详述风险并提出缓解措施与应对预案；产品发布后，持续监控运行与社会影响，借监管与用户反馈及时处理问题。

(2) 伦理安全模拟和场景分析

在车用人工智能系统的开发和评估过程中，模拟和场景分析是至关重要的，它们帮助设计者、评估者和监管机构理解并预测人工智能系统在现实世界中的行为和决策模式。这一过程通常涉及复杂的技术设置和跨学科合作，以确保评估结果的全面性和准确性。

伦理安全模拟可以通过场景搭建后测试产品在解决对应问题时的反馈，从而评估出产品可能存在的伦理安全风险问题[84]。其评估思路类似建立车用 AI 的伦理测试“沙盒”。创建代表性场景后进行模拟。创建代表性场景：首先需要确定那些能够全面展示车用人工智能潜在伦理风险的场景。这些场景应覆盖广泛的交通环境、突发事件以及交互的复杂性，例如交通高峰时段的城市街道、多车道高速公路、行人密集的过街区域等。场景设计需要与交通规划专家、汽车工程师和人工智能开发人员合作，确保场景的真实性和技术的适应性。进行模拟测试：使用高级仿真软件（如 CarSim 或 IPG CarMaker）进行动态模拟，这些软件能够创建详细的交通环境，并模拟不同的驾驶情况。模拟过程中，人工智能系统的反应被记录和分析，特别是那些涉及伦理决策的场景，如避免事故时的牺牲选择。这一步骤可能需要与软件开发公司、数据分析师以及行为科学家密切合作，确保数据的准确记录和有效解释。

汽标委智能网联汽车分委会

伦理安全分析的生成可结合模拟后对应产出评估报告进行反馈。由人工智能开发团队、汽车工程师与交通规划人才、法律和伦理顾问共同完成。具体包括：定义评估标准：在模拟开始前，明确何种类型的决策或行为会被视为伦理风险。这些标准应基于伦理准则、法律规定以及社会接受度。例如，如何平衡保护乘客安全与保护行人安全的决策，或是在极端情况下对不同类型车辆（如学校巴士与普通小汽车）的避让选择。这需要法律顾问、伦理专家的指导和批准。进行场景测试：在每个场景中运行车用人工智能，观察并记录系统的决策过程和行为结果。这涉及大量的数据采集，如传感器数据、决策逻辑的时间线和最终的执行操作。此外，应该对人工智能的每次决策进行背后逻辑的解析，确保所有行为都能追溯到具体的决策点和原因。分析与报告：场景测试后，需要对收集的数据进行详细分析，识别任何可能的伦理风险或模式。这可能涉及统计分析、行为模式识别和情境重现。最终，这些分析结果将被整理成报告，提供给项目团队、监管机构和相关的伦理审查板，以供进一步的审议和决策。

(3) 建立伦理评估案例库、增加伦理评估测试用例

随着人工智能技术的迅猛发展，其在各个领域的应用日益广泛，为人类生活带来了前所未有的便利。然而，人工智能技术的快速发展也伴随着一系列伦理挑战和问题，如数据隐私、算法偏见、责任归属等，这些问题已经引起了全球范围内的广泛关注。为了应对这些挑战，国外纷纷建立了人工智能伦理平台，以推动AI 技术的健康发展。

“伦理评估案例库”通常指的是收集和整理了各种伦理评估案例的数据库或资源库。这些案例库可以用于教育、研究和实践，帮助人们理解伦理问题、进行伦理分析和决策。它们可能包含不同领域或行业的伦理问题，如医学、商业、工程、人工智能等。在汽车领域，同样也需建立案例库，但同时应注意区分“车用人工智能伦理”与“通用人工智能伦理”。对于以通用人工智能伦理为主的智能座舱部分，可以收集和整理通用的伦理评估案例，并采取通用类测试；对于车用人工智能专门领域的伦理问题，则需建立以智能驾驶为核心场景的案例库，并采用场景库来测试。

汽标委智能网联汽车分委会

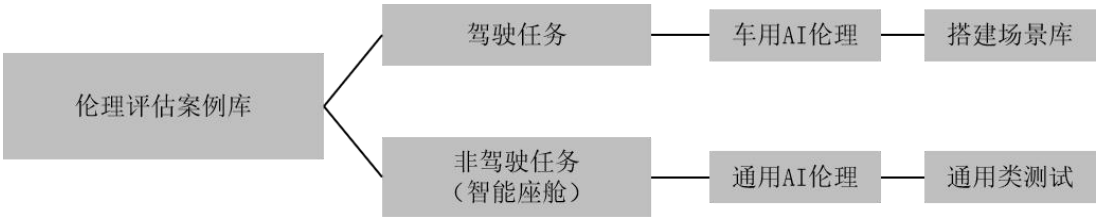


图 12-3 伦理评估案例库框架图

通过对国外人工智能伦理平台的共性进行总结，我们发现这些平台在多个方面展现了共同的特点和趋势。首先，它们均强调制定和遵守一套明确的伦理原则，以确保人工智能技术的研发和应用符合道德和法律要求。这些原则通常涵盖了公平性、安全性、隐私保护、透明度和问责性等方面，为人工智能技术的发展提供了明确的道德指南。

其次，这些平台普遍设立了专门的治理机构或委员会，负责监督人工智能伦理原则的实施，处理与人工智能相关的伦理问题。这些治理机构通过制定规则、提供指导、开展评估等方式，确保人工智能技术的研发和应用符合伦理原则，并及时应对可能出现的伦理风险。

此外，为了支持伦理原则的应用，这些平台还开发了专门的技术工具和解决方案。这些工具可以帮助识别和减少人工智能系统中的偏差，增强模型的鲁棒性和可解释性，从而提高人工智能技术的可靠性和安全性。同时，这些平台也注重员工培训和教育，提供丰富的教育资源，帮助员工提高对人工智能伦理问题的认识和理解。

比如上海开源信息技术协会、AIII 人工智能国际研究院与商汤科技伦理与治理委员会共同打造的人工智能伦理与治理领域的开源案例库。这个案例库面向人工智能行业的专业人士、研究人工智能伦理与治理的政府智库与产业专家以及高校中关注人工智能伦理与治理命题的专业学院，旨在成为该领域的"维基百科"与"案例研究中心"。

除了上海开源信息技术协会、AIII 人工智能国际研究院与商汤科技伦理与治理委员会共同打造的"AI 伦理与治理开源案例库"，国内外还有一些其他的人工

汽标委智能网联汽车分委会

智能伦理与治理相关的案例库或平台，例如人工智能治理公共服务平台：由中国科学院自动化研究所人工智能伦理与治理研究中心联合多家机构研制，提供人工智能风险、伦理与治理的在线检测、评估与服务；旷视科技人工智能治理研究院：旷视科技成立了人工智能治理研究院，发布了全球十大人工智能伦理与治理案例，涵盖人工智能的权利与义务、个人数据与隐私等主题；智善 • 观行 - 人工智能大模型伦理安全观测站：提供了一个案例库，收集和展示人工智能大模型伦理安全相关的案例。

这些平台和案例库的建立，体现了业界对于人工智能伦理与治理问题的关注，旨在通过收集和分析相关案例，促进人工智能技术的健康发展，并为相关研究和实践提供参考和支持。

国外也有类似的人工智能伦理平台。例如，微软公司在人工智能伦理与可信人工智能方面进行了深入的实践，包括提出伦理原则、设立治理机构、开发技术工具和解决方案、制定行动指南以及进行员工培训等。微软的负责任人工智能办公室负责制定公司内部的人工智能伦理规则，并推进立法、规范、标准的制定，确保人工智能技术有助于提升社会福祉；

谷歌也制定了人工智能设计和使用的原则，并为员工提供了科技伦理培训，帮助他们了解如何评估潜在的利害，并遵循道德准则。此外，谷歌还推出了 AI Fairness 360 toolkit，这是一个开源解决方案，用于检测和减少人工智能算法中的偏差；

IBM 发布了《人工智能日常伦理指南》（Everyday Ethics for Artificial Intelligence），用于指导人工智能系统设计者和开发人员系统地考虑 AI 伦理问题，将道德和伦理贯彻在人工智能的全生命周期中；还有 Twitter 等其他科技公司也有自己的人工智能伦理治理机构和实践。

这些平台和实践体现了国外科技公司在推动人工智能技术发展的同时，也在积极面对和解决伴随而来的伦理挑战。通过制定伦理原则、建立治理机构、开发

汽标委智能网联汽车分委会

技术工具和解决方案、提供培训和教育等措施，旨在确保人工智能技术的健康发展，并为社会带来积极的影响。

国外人工智能伦理平台的建立和运作，体现了全球对人工智能伦理问题的关注和应对策略。它们通过制定伦理框架、建立治理机制、技术创新、教育与培训以及促进公众理解等方式，推动人工智能技术朝着有益于社会和人类的方向发展。这些平台不仅为人工智能技术的发展提供了有力的保障，也为全球范围内的伦理对话和合作提供了重要的平台。在未来，我们有理由相信，随着这些平台的不断完善和发展，人工智能将在更广泛的领域得到应用，同时其伦理问题也将得到更加有效地解决。

(4) 数据审核与偏差测试

数据审核和偏差测试是关键步骤，用于确保车用人工智能系统在决策过程中不会因数据偏见引发不公正或不道德的结果。这一过程涉及多个层面的合作，并需要综合使用数据科学、统计分析和机器学习技术来执行。进行数据审核和偏差测试一般由数据科学和机器学习团队、伦理和合规部门、第三方审计、监管机构配合。该类评估方法包括数据处理、模型构建和偏差测试技术。

数据审核是一个全面的过程，旨在确保用于训练和操作车用人工智能的数据集是高质量的、代表性的，并且公正无偏。定义审核标准：首先，需要与数据科学家、伦理专家以及法律顾问合作，定义出一套清晰的数据审核标准。这些标准包括数据的完整性、准确性、时效性以及多样性。例如，确认数据集是否包括了足够的代表各种驾驶环境（城市、郊区、不同天气条件）的样本，是否平衡了各种用户群体（年龄、性别、地区）的代表性。进行数据清洗和预处理：在数据科学团队的协助下，对数据进行清洗和预处理，包括去除异常值、填补缺失值和校正错误数据。此外，对数据进行去标识化处理，确保遵守隐私保护法规。审核数据集：使用统计分析和数据可视化工具，对数据集进行深入分析，检查是否存在偏差或集群现象。例如，分析不同人群的事故记录是否均衡，或者某些特定条件下的数据是否过度集中。

汽标委智能网联汽车分委会

偏差测试专注于识别和纠正由数据集中的偏见可能导致的任何不公正或不道德的决策。构建偏差检测模型：与机器学习工程师合作，开发专门的算法或使用现有工具来识别数据中的偏差。这些模型可以通过各种机器学习技术（如聚类分析、决策树或人工神经网络）进行训练，以预测和识别偏差。进行偏差测试：在保持数据代表性和真实性的前提下，通过引入各种控制变量来测试人工智能模型的输出，检查是否在某些特定人群或情境下表现出偏差。例如，检验模型对不同年龄或性别驾驶者的反应是否存在差异。分析结果并调整：分析偏差测试的结果，如果发现显著偏差，需要调整数据集或算法参数。这可能包括重新平衡数据集、修改特征权重或重新训练模型。所有调整都应记录并验证其效果，确保修改后的模型不再表现出原有的偏差。

(5) 利益相关者咨询和社会接受度测试

利益相关者咨询和社会接受度测试是车用人工智能伦理安全风险评估的重要方式。它们帮助确认这些系统能否得到广泛社会各界的支持和认可，同时也考虑到了不同群体的需求和顾虑。这一过程通常涉及广泛的交流、数据收集和分析活动。但同时应注意商业秘密的保护，一般由政府或市场研究机构主导。

利益相关者咨询旨在从多个视角收集对车用人工智能的意见和反馈，以确保系统设计和实施考虑到了广泛的社会和伦理因素。特别是针对有较高自主性和决策能力的车用人工智能，对于相关技术的广泛调研十分重要。确定利益相关者：首先，需要定义哪些群体是利益相关者。这通常包括车用人工智能的直接用户（如驾驶者）、行业监管者、汽车制造商、科技公司、道路安全专家、消费者保护组织、普通公众以及伦理和法律专家等。确保覆盖所有主要利益相关者是此步骤的关键。组织咨询活动：接下来，组织系列咨询活动，如研讨会、公开会议、焦点小组讨论和一对一访谈等。这些活动应由专业的会议组织者和伦理专家共同策划和实施，确保讨论的广度和深度。在这些活动中，收集参与者对车用人工智能的安全性、道德问题、数据隐私、用户接受度以及预期的社会影响的看法和建议。记录和分析反馈：所有咨询活动的反馈需被详细记录和系统分析。分析结果将帮

汽车零部件智能网联汽车分委会

助开发团队理解社会对车用人工智能技术的期望和顾虑，从而调整设计和实施策略以更好地满足社会需求和伦理标准。

社会接受度测试是通过量化的方式了解公众对车用人工智能技术的接受程度，关注点包括技术的信任度、使用意愿以及对技术带来的改变的态度。设计问卷和调查：设计问卷和调查工具，包含关于车用人工智能系统功能、潜在好处、可能的风险、个人隐私问题等方面的问题。此外，应包括情景模拟，询问受访者在特定情境下的选择和偏好。进行广泛调研：通过在线平台、电话或面对面方式进行广泛的社会调查。确保样本具有代表性，覆盖不同年龄、性别、地区、经济状况和技术熟悉度的人群。这一步骤可能需要与市场研究机构合作，以利用其专业知识和资源进行有效的数据收集。数据分析和报告：分析收集到的数据，评估公众对车用人工智能技术的整体接受度和关注点。将分析结果整理成报告，提供给开发团队、公司管理层和政策制定者，作为决策支持。

(6) 教育和培训

开展人工智能技术和伦理的教育和培训，提高公众对人工智能技术的了解和认识，增强他们的安全意识和自我保护能力。开发者、工程师和决策者应该接受相关的伦理培训，了解伦理原则和价值观的重要性，并学习如何应对伦理挑战。在设计和应用人工智能技术时需要考虑到伦理和社会影响，并遵循道德准则。

斯坦福大学人工智能伦理中心（Stanford Center for AI Ethics）：致力于人工智能的伦理研究、教育和实践，推动人工智能领域的道德发展、麻省理工学院媒体实验室的道德和人工智能倡议（MIT Media Lab's Initiative on AI and Ethics）：探索人工智能技术的道德和伦理问题，为学生提供相关的教育和研究机会；此外，教育部还公布了首批 18 个“人工智能+高等教育”应用场景典型案例，这些案例展示了人工智能技术在高等教育中的多样化应用。案例覆盖了从教学资源集成、交互性实训平台、智能教育大模型到个性化学习功能等多个方面，体现了智能技术在人才培养、教学方法改革和教育治理能力提升中的重要作用；旷视科技人工智能治理研究院发布了全球十大人工智能治理事件，这些事件包括了人工智能的

汽标委智能网联标准分委会

权利与义务、个人数据与隐私、合理的知情同意等主题。该院旨在推动人工智能的健康发展，并与社会各界广泛沟通、与专家学者们展开研究；中国科学技术大学在教学内容中涵盖了人工智能和大数据的基本原理和应用路径，以及人工智能伦理与治理的现状、挑战和解决方案，重点领域包括数据伦理与治理、企业人工智能伦理等。

伦理教育在人工智能发展中扮演着至关重要的角色，它不仅确保了技术的道德合规性，而且促进了从业者的责任意识。通过增强透明度和可解释性，伦理教育使人工智能系统的决策过程更加开放，易于理解，从而帮助预防偏见和歧视的发生，保障了人工智能人权和隐私的尊重。此外，伦理教育还鼓励人工智能技术的发展考虑到环境和社会的长期可持续性，提升了公众对技术的信任度。

其有助于人工智能从业者预见和应对技术发展可能带来的社会、法律和伦理挑战，培养了具有跨学科视角的人才，促进了人工智能技术与人文社科等其他领域的融合。它还促进了国际合作，推动了全球人工智能伦理标准的制定和实施。对于政策制定者而言，伦理教育提供了必要的知识和视角，帮助他们制定合理的人工智能政策和法规，确保了创新的同时考虑社会影响，引导了负责任的人工智能研究和应用。总体而言，伦理教育是推动人工智能技术健康、负责任发展的基石，对于构建人类与人工智能和谐共存的未来起着决定性的作用。

12.3 车用 AI 伦理安全标准化研究

12.3.1 国内现状分析

我国人工智能伦理治理历经规划认可、原则确立和规范细化。值得注意的是，我国已率先开始构建车用人工智能伦理标准，为实现车用 AI 伦理治理标准化做出了诸多努力。与此同时，我国正积极参与全球人工智能伦理治理。2023 年 8 月金砖峰会达成共识，同意建立人工智能研究组，推动制定具有国际共识的治理框架及安全标准。同年 10 月，我国发布《全球人工智能治理倡议》，提出伦理先行的原则，明确构建伦理准则及问责体系。同年 12 月，中国信息通信研究院

汽标委智能网联汽车分委会

发布《人工智能伦理治理研究报告（2023）》，该报告以人工智能伦理治理关切为出发点，通过分析国际治理经验，提出应协调创新发展与伦理监管，完善治理措施并加强国际合作，全面提升各主体伦理风险应对能力。下表为近年我国在人工智能伦理安全治理中采取的行动示例。

表 12-2 人工智能伦理安全治理行动

相关工作	发布时间	发布机构	主要内容
《新一代人工智能发展规划》	2017 年	国务院	提出研究相关伦理问题，建立伦理道德框架
《新一代人工智能伦理规范》	2021 年	国家新一代人工智能治理专业委员会	提出增进人类福祉、促进公平公正、保护隐私安全、确保可控可信、强化责任担当、提升伦理素养等 6 项基本伦理要求
《科技伦理审查办法（试行）》	2023 年	科技部	对科技伦理审查的基本程序、标准、条件等提出了具体要求，将特定自动化决策系统列入复核清单
《人工智能伦理治理标准化指南（2023 版）》	2023 年	全国信标委	深入解析了人工智能伦理准则内涵，为人工智能伦理标准化提出了建设性方案[78]
《汽车行业人工智能技术标准化发展倡议书》	2024 年	全国汽车标准化技术委员会	提出了对汽车行业人工智能未来发展的倡议，包括规范伦理行为，践行基本准则等

汽标委智能网联汽车分委会

12.3.2 国际现状分析

车用人工智能伦理治理已成为国际共识。2021 年，《人工智能伦理问题建议书》通过，该建议书在全球现有人工智能伦理框架之外提供了一份具有公认价值的准则性文书，在阐明治理原则的同时，着力强调了包容、性别平等以及环境和生态保护等问题，要求利益攸关方能够在全球对话的基础上协力承担责任[85]。

同时，各组织也在积极寻求人工智能伦理治理方案，以下为其在人工智能伦理治理方案相关的标准研究工作示例。但值得注意的是，目前仅有 ISO 8800 标准从整车功能角度对道路车辆领域的人工智能安全做出探讨，车用 AI 伦理标准的建设在国际层面仍处于空白状态。

表 12-3 国际组织关于人工智能伦理安全治理研究

相关工作	发布时间	发布组织	主要内容
《人工智能设计的伦理准则》	2017 年	电气电子工程师协会（IEEE）	IEEE P7000，解决系统设计中伦理问题的建模过程； IEEE P7007，伦理驱动的机器人和自动化系统的本体标准； IEEE P7008，机器人、智能与自主系统中伦理驱动的助推标准； IEEE P7010，合乎伦理的人工智能与自主系统的福祉度量标准等
ISO/IEC TR 24028:2020 信息技术 人工智能 人工智能可信度概述	2020 年	国际标准化组织（ISO）	介绍人工智能系统可信赖概念
ISO/IEC TR	2021 年	国际标准化组织	提供了人工智能辅助决策系统偏见的

汽标委智能网联工作分委会

24027:2021 信息技术 人工智能 人工智能系统中的偏见和人工智能辅助决策		（ISO）	评估技术和方法
ISO/IEC TR 24368:2022 信息技术 人工智能 伦理和社会关注概述	2022 年	国际标准化组织（ISO）	通过案例确定了与人工智能有关的若干伦理和社会因素，为人工智能伦理问题的解决提供了建设性意见等。
ISO/IEC 23894:2023 信息技术 人工智能 风险管理指南	2023 年	国际标准化组织（ISO）	提供了人工智能安全风险管理方法等
ISO/IEC TR 5469:2024 功能安全与人工智能系统	2024 年	国际标准化组织（ISO）	提出了实现人工智能技术应用中人身安全保护功能的方法

在美国：2016 年，美国发布《联邦自动驾驶政策：加速道路安全变革》，指出伦理问题是自动驾驶汽车安全评估的十五个因素之一，为车用人工智能伦理安全提出框架性的建议。2018 年，《为未来交通做准备：自动驾驶汽车 3.0》发布，该文件提出了包括安全优先、技术中立、法规现代化在内的六项智能汽车规制原则。2023 年，《关于安全、可靠、可信赖地开发和使用人工智能》行政令发布，明确进行负责任人工智能技术开发，提出安全可靠、保障权利、隐私保护等伦理要求，并对各行政部门如何促进负责任人工智能技术开发和应用作出安排[86]。

汽标委智能网联汽车分委会

在欧洲：德国《自动和网联车辆交通伦理准则》提出了针对自动驾驶汽车的二十条伦理准则，要求以生命利益为先、禁止编程固化、禁止歧视。随后，德国《自动驾驶法案》贯彻了这一理念，将伦理安全作为自动驾驶汽车产品准入的要求之一，强调自动驾驶汽车必须具备伦理安全设计，在两难困境中能够遵循减少和预防事故损害、生命利益优先、禁止歧视三大准则。2019 年，欧盟高级专家组发布《可信人工智能伦理指南》，明确可信人工智能的关键要求，包括保证人类监督、鲁棒性和安全性、隐私和数据治理、透明度、社会和环境福利等。2024 年，欧盟《人工智能法案》通过，该法案强调以人为本、保证透明度等伦理价值要求。

在新加坡：2022 年，新加坡发布《人工智能治理评估框架和工具包》，成为全球首个官方的人工智能检测工具。该工具结合技术测试和流程检查对人工智能技术进行验证，并生成验证报告，涵盖人工智能系统的透明度、安全性及可归责性等人工智能伦理要求。2023 年，新加坡发布国家人工智能战略 2.0，提出人工智能服务公共利益的愿景，再次强调建立一个可信赖和负责任的人工智能生态。

12.3.3 车用 AI 伦理安全面临的挑战

12.3.3.1 车用 AI 伦理安全的必要性

在汽车行业中，人工智能已成为不可或缺的技术，广泛应用于从辅助驾驶系统到动力传动系统管理的各个领域。然而，随着人工智能技术的迅速发展，确保其安全、可靠地运行并与其他车辆兼容面临着重大挑战。

第一，人身安全

车用伦理安全至关重要，因为它直接关系到人类生命的安全。自动驾驶汽车的设计旨在减少交通事故并提升乘客及其他道路使用者的安全性。因此，必须明确自动驾驶汽车在面对危险情况时如何做出符合伦理和道德标准的决策。

第二，隐私安全

汽标委智能网联汽车分委会

人工智能的核心依赖数据与算法，技术升级需海量数据支撑及算法迭代优化。车载信息包含敏感个人隐私，一旦泄露隐患极大，安全防护至关重要。

第三，算法偏见

算法偏见是影响人工智能公正性的主因，在车用人工智能系统中会带来严重安全隐患。其可能源于数据收集、处理或算法设计阶段，使系统对部分群体或情境决策不公，若算法有偏见，会使系统对交通规则的理解和遵守产生偏差，不同情境下对交通参与者区别对待，易引发交通事故或违规，威胁道路安全。同时，系统可能无法准确识别事故风险，如倾向回避特定对象而忽视其他风险，增加事故概率。此外还会降低系统对不同路况和交通情境的适应性，如处理复杂路口或不同驾驶风格时偏向特定决策，忽略其他选项。综上，算法偏见严重影响车用人工智能系统安全，增加事故风险。因此，需采取措施确保算法无偏见，保障车辆和驾乘者安全。因此，必须采取措施确保系统的算法不受偏见影响，从而保障车辆和驾乘者的安全。

第四，责任风险

人工智能主决策能力不断增强，但决策结果仍有失误率。自动驾驶汽车道德困境源于决策自主性转移，美汽车工业学会将其分六个等级，级别越高自主性越强。当人工智能代人类决策执行，如高级别自动驾驶汽车出事故，汽车制造商、软件开发商、车主等相关方谁担责，责任如何认定、划分与承担，给现存法律法规带来新挑战。伦理安全框架可明确责任，公正处理事故责任问题。

第五，法律和监管合规

随着自动驾驶技术的发展，各国政府可能会出台新的法律和规定来管理这些系统。伦理安全措施可以帮助确保车用人工智能系统符合相关法律和监管要求。公众对自动驾驶汽车的接受程度部分取决于这些系统是否被认为是安全和可靠的。通过实施伦理安全措施，可以提高公众对自动驾驶技术的信任和接受度。

汽标委智能网联汽车分委会

总之，车用人工智能伦理安全的必要性在于确保这些系统能够安全、可靠、公平地运行，同时保护用户的隐私和权益，以及遵守相关法律和伦理标准。这不仅对于保护生命财产安全至关重要，也是推动自动驾驶技术健康发展的关键因素。

12.3.3.2 法律法规和标准层面

目前对于通用人工智能，已有一系列相应的伦理治理准则、指南和标准出台，例如国家新一代人工智能治理专业委员会提出的《新一代人工智能伦理规范》、欧盟提出的《可信人工智能伦理指南》、联合国提出的《人工智能伦理问题建议书》、ISO 标准发布的 ISO/IEC TR 24368 等。此外欧盟也通过了《人工智能法案》，进一步在法律法规层面，对 AI 进行监管约束。但针对车用 AI 的伦理安全领域，目前尚未有相应的法律法规或标准颁布，其仍处于较为空白的状态。

在原有法律法规体系中，道路事故中的责任界定已有较为明确的划分，例如由于司机疏忽或违反交通法规导致的事故由司机承担责任；因车辆设计或制造缺陷导致的事故，则需制造商需承担责任等等。而在车用人工智能加入车辆日常使用之后，法律法规如何清晰界定事故中的责任归属将会变得十分复杂，责任认定可能涉及多个方面，包括车辆所有者、制造商、软件开发者等。

车用人工智能在隐私保护上挑战复杂。随着自动驾驶与智能网联汽车发展，其依赖大量传感器和数据感知环境，而这些数据涉及个人隐私。收集的数据涵盖地理位置、驾驶行为等。若信息被滥用，会导致用户被监控、隐私受侵。数据还需与制造商、服务提供商等多实体共享，增加了泄露和滥用风险。如何在保障车辆安全运行时，保护乘客和道路用户隐私，避免个人身份暴露，是面临的伦理挑战[87]。

12.3.3.3 技术层面

开发道德算法旨在让自动驾驶车辆在道德和伦理决策时做出合理选择，但该领域面临着许多复杂的挑战[88]。紧急情况下，车辆可能要在伤害不同个体间抉择，如避行人危及乘客。为道德选择设优先级需权衡不同价值，明确算法规则。然而，目前缺乏统一伦理标准，不同地区规范有别，难以达成一致。设计出道德

汽标委智能网联汽车分委会

算法后，评估其是否满足伦理要求及可实施性，也是技术难题。因人工智能不可解释，系统是“黑箱”，决策过程难理解，公众对车用人工智能信任度不高，并且事故后责任难明确。开发可解释的 AI 模型，让用户了解决策逻辑，增强信任，是挑战之一。车用人工智能还受数据和算法偏见影响，危及公正性与可靠性，导致社会不公和安全风险。训练数据可能有偏差，算法设计也可能忽视少数群体需求。且公平性指标、监管措施及评价机制仍均有所欠缺。

12.3.3.4 社会层面

车用人工智能的普及可能会对传统交通领域的就业产生影响，可能导致大量专业司机（如卡车司机、出租车司机）的工作岗位流失，相关产业（如保险、驾驶培训、维修等）需要适应新的技术环境，经历阵痛期在所难免。自动驾驶技术不仅影响司机岗位，还可能影响物流行业中的其他岗位，如仓储、配送等环节的人工岗位。传统交通管理岗位的需求，如交通警察、道路安全检查员等也会受到较大影响。采取措施来减缓就业影响并促进新的就业机会的出现是车用人工智能伦理在社会层面所需要关注的问题。

对车用人工智能的不合理使用，会对社会价值等方面产生一定的负面影响[89]。用户或不法分子可能非法改装车用人工智能系统，使其绕过安全机制或增加不当功能，造成安全隐患。司机在不适合的环境下（如复杂交通或恶劣天气）仍过度依赖自动驾驶系统，从而导致安全事故，可能会在社会层面造成一定的影响。同时，在未经用户同意在商业上对乘客数据进行滥用等，也是车用人工智能不合理使用所造成的负面影响之一。

公众对自动驾驶技术的道德和伦理接受度是一个重要问题，公众对自动驾驶技术的安全性存疑，担心技术是否足够成熟，能否应对复杂和不可预测的驾驶环境，同时一些驾驶员可能不愿意完全交出驾驶控制权，担心失去对车辆的掌控[90]。如何通过开展广泛的公众教育活动让人们更多地理解和参与到自动驾驶的讨论中，同时推动技术与价格的普及，都需要提供合理的解决方案。

汽标委智能网联汽车分委会

12.4 总结与展望

在当前科技迅猛发展的背景下，车用人工智能的伦理安全面临双重性挑战与机遇。

一方面，自动驾驶技术的高度智能化和自主化使系统承担了前所未有的伦理决策责任。当车辆在无人干预状态下独立处理复杂交通场景时，系统必须精准执行符合伦理原则的决策（如电车难题中的优先级判定），这对安全标准提出了严苛要求。技术发展同时带来解决方案：基于深度学习的感知系统可通过持续优化提升场景识别准确率，强化学习算法使决策模型更适应动态环境；而依托大数据分析和云计算平台，车辆运行数据可实现毫秒级监控与隐患预判，大幅降低事故率并提升交通系统整体效能。

另一方面，技术自主化衍生出新型风险：人类对系统的过度依赖可能放大故障后果，而车联网架构的普及使黑客攻击威胁指数级增长。这要求构建三维保障体系：首先建立覆盖研发-测试-运营全周期的法规标准，其次推动汽车制造商、网络安全机构与伦理委员会跨领域协作，最后通过公众教育提升技术认知。唯有通过技术迭代与制度保障的深度融合，方能在享受自动驾驶便捷性的同时守护人类生命尊严，实现科技与社会发展的共生平衡。

汽标委智能网联汽车分委会

第十三章 车用 AI 安全管理体系标准化研究

13.1 研究背景与范围

13.1.1 研究背景

随着人工智能技术在汽车领域的广泛应用，车用 AI 系统逐渐成为现代汽车的核心组成部分。从高级驾驶辅助系统（ADAS）到自动驾驶功能，AI 技术为汽车的安全性、舒适性和智能化水平带来了显著提升。然而，与此同时，车用 AI 系统也面临着诸多安全挑战，如算法漏洞、数据安全问题、网络攻击风险等，这些问题可能导致严重的安全事故，威胁到驾乘人员和道路行人的生命安全[91]。因此，建立一套完善的车用 AI 安全管理体系标准化框架具有重要的现实意义。目前，全球各国和地区都在积极推动车用 AI 安全管理相关标准的制定和研究。一些国际组织和行业协会已经发布了部分指导性文件，但尚未形成统一的、全面的标准化体系。不同国家和地区的标准在内容和要求上存在一定差异，这给汽车企业的全球化发展带来了一定的困扰。

13.1.2 研究范围

本章节主要聚焦于车用 AI 安全管理体系的标准化问题，探索适用于车用 AI 安全管理体系的构建方法，为车用 AI 安全管理提供可操作的解决思路。首先本章将深入分析国内外相关政策法规和标准，梳理车用人工智能技术在汽车行业的应用及面临的安全挑战。同时详细剖析相关标准在车用人工智能安全管理方面的具体做法和指导。结合 ISO/IEC 42001:2023《信息技术 人工智能 管理体系》和 ISO/PAS 8800:2024《道路车辆 安全和人工智能》标准，构建覆盖企业能力和产品能力的车用人工智能安全管理体系，为行业发展提供有益的参考和借鉴。

汽标委智能网联汽车分委会

13.2 车用 AI 安全管理体系标准化需求研究及挑战

13.2.1 AI 技术在汽车行业的应用

13.2.1.1 驾驶域

在汽车的驾驶域中，AI 技术扮演着至关重要的角色，贯穿数据挖掘、数据标注、数据训练和仿真优化等多个关键过程。数据挖掘作为整个流程的起点，通过收集车辆在各种路况、驾驶场景下产生的海量数据，包括车速、加速度、转向角度、周边环境图像等，为后续分析提供丰富素材，提升驾驶安全和体验的关键信息[92]。

数据标注对挖掘所得的数据进行细致分类和标记，使其能够被计算机有效识别和处理。例如，将图像数据中的行人、车辆、交通标志等进行精准标注，以便模型能够准确学习和识别。数据训练是利用标注好的数据对车用大模型进行反复训练，使其不断优化和提升性能。通过大量的数据训练，模型能够学习到不同驾驶场景下的最佳决策策略，从而帮助提升辅助驾驶的安全性。例如，在遇到紧急情况时，模型能够快速准确地做出制动、避让等决策，避免事故的发生。仿真优化则是在虚拟环境中对训练好的模型进行测试和验证，模拟各种极端和复杂的驾驶场景，发现模型存在的问题并进行优化。通过不断的仿真优化，模型的性能得到进一步提升，推动辅助驾驶向自动驾驶逐步升级，为未来的出行带来更加安全、便捷的体验。

13.2.1.2 座舱域

在座舱域，AI 技术借助自然语言处理和计算机视觉技术，为用户打造了全面的智能交互体验[93]。自然语言处理技术使得车辆能够理解人类的语言指令，实现语音控制车辆各项功能。用户只需说出简单的指令，如“打开空调”“播放音乐”“导航到目的地”等，车辆就能迅速响应并执行相应操作，大大提升了驾驶的便利性和舒适性。

汽标委智能网联汽车分委会

计算机视觉技术则为车辆提供了“眼睛”，能够实时监测驾驶员的状态和行为。例如，通过识别驾驶员的面部表情、眼神、头部姿态等，智能识别驾驶员的疲劳状态。当检测到驾驶员出现疲劳迹象时，系统会及时发出预警，提醒驾驶员休息，有效降低因疲劳驾驶导致的事故风险。此外，计算机视觉技术还可以实现车内物品识别、手势控制等功能，进一步丰富了座舱的交互体验。

13.2.1.3 动力域

在动力域，AI 技术主要致力于提升车辆的能源管理效能。以新能源汽车为例，其核心的三电系统（电池、电机、电控）对车辆的性能和续航里程起着至关重要的作用。AI 技术通过精细化管理三电系统，实现高效节能[94]。

通过实时监测电池的电量、温度、电压等参数，AI 算法能够精准预测电池的剩余电量和续航里程，并根据驾驶习惯、路况等因素优化电池的充放电策略，延长电池的使用寿命。在电机控制方面，AI 技术能够根据车辆的行驶需求实时调整电机的输出功率，提高能量转换效率，降低能耗。同时，在电控系统中，AI 技术可以对车辆的各种电子设备进行智能管理，优化能源分配，进一步提升车辆的整体能源利用效率。

13.2.1.4 车身域与底盘域

AI 技术在车身域的应用通过多模态感知与智能决策算法优化车辆控制逻辑，实现车身功能的智能化升级。在智能环境感知与自适应控制方面，AI 利用摄像头、光线传感器等硬件采集环境数据，经计算机视觉算法实时分析光照、天气及乘员状态，进而自动调节车前/后灯照明模式、使雨刮器根据雨量感知动态调整频率、让天窗基于温湿度数据自动开闭、让空调系统结合驾乘者生物特征实现个性化温控。在多模态交互与场景化控制上，融合语音、手势等多模态交互技术，可通过语音指令控制车窗、电动后备箱等硬件，利用手势识别触发内饰灯氛围切换或车门解锁，智能钥匙结合生物识别自动同步座椅/后视镜预设。在预测性维护与能效优化方面，AI 分析车身部件运行数据，预测电机类部件潜在故障并提前预警，动态管理灯光/空调能耗策略以延长电动车续航。在安全增强型协同控

北京智能网联汽车分会

制上，车门锁与摄像头联动，识别可疑人员自动上锁并触发警报，紧急情况下自动升起车窗并解锁应急通道。AI 技术通过环境感知到数据分析到决策执行的闭环，把传统离散的车身功能转化为协同响应的智能系统，提升了安全性、舒适性与能效，为“人-车-环境”无缝交互提供技术支撑。

在底盘域控制中，AI 技术通过实时分析车速、车身姿态、路面信息等多源传感器数据，结合环境感知与驾驶意图预测，对制动、转向、悬挂等系统进行自适应协同控制。具体应用包括基于路况和车身姿态数据智能调节悬挂，提升通过性与稳定性；预测驾驶场景，提前优化制动力分配和转向助力策略，保障安全；分析历史数据预判部件失效风险并触发维护提示或调整参数。这种动态优化方式显著提升了车辆行驶的稳定性、安全性与舒适性，是智能底盘向“主动预判”演进的核心支撑[95]。

13.2.1.5 其他典型 AI 产品的多元应用

(1) 与 AI 机器人的技术融合

车用 AI 技术的核心技术（感知、决策、控制、交互等）与 AI 机器人有高度重叠。以小鹏汽车为例，其将车用 AI 技术的端到端大模型迁移到 Iron AI 机器人，为机器人赋予了强大的智能能力[96]。在感知方面，机器人能够像汽车一样通过各种传感器实时获取周围环境的信息，包括视觉、听觉、触觉等。在决策方面，端到端大模型能够根据感知到的信息进行快速分析和判断，做出合理的决策。例如，在面对复杂的环境任务时，机器人能够自主规划路径、避开障碍物。

在控制方面，AI 机器人能够根据决策结果精确控制自身的运动，实现灵活的移动和操作。在交互方面，机器人可以与人类进行自然流畅的交流和互动，通过语音、手势等方式理解人类的指令和需求，并做出相应的回应。这种技术融合使得机器人拥有了自主思考甚至推理能力，为机器人的应用场景拓展提供了更多可能。

汽标委智能网联汽车分委会

(2) 车用 AI 芯片的广泛应用

车用 AI 芯片的设计需求（高性能、低功耗、高可靠性）与通用 AI 芯片的研发需求高度契合。以小鹏汽车公司研发的图灵 AI 芯片为例，它能够同时被用于 AI 汽车、AI 机器人及飞行汽车上[96]。在 AI 汽车中，图灵 AI 芯片为车辆的智能驾驶系统提供了强大的计算支持，能够快速处理大量的传感器数据和复杂的算法模型，确保车辆的智能决策和控制的高效运行。

在 AI 机器人中，图灵 AI 芯片使得机器人能够实现高速的运算和智能的决策，提高机器人的反应速度和处理能力。在飞行汽车上，图灵 AI 芯片的高性能和高可靠性特点为飞行汽车的安全飞行提供了保障，能够实时处理飞行过程中的各种数据和信息，确保飞行的稳定和安全。

车用 AI 技术在汽车行业的多元应用，不仅提升了汽车产品的性能和用户体验，还推动了整个汽车行业的 AI 生态快速发展。一方面，汽车企业加大了在 AI 技术研发方面的投入，吸引了大量的科研人才和资金，促进了 AI 技术的不断创新和进步。另一方面，AI 技术的应用也带动了相关产业链的发展，如传感器、芯片、软件等产业。同时，随着车用 AI 技术的不断普及和应用，也催生了新的商业模式和服务形态，如智能出行服务、自动驾驶物流等。这些都为汽车行业的发展带来了新的机遇和挑战，推动汽车行业向智能化、电动化、网联化方向加速迈进。

13.2.2 标准化现状及需求研究

13.2.2.1 ISO/IEC 42001:2023 标准分析

国际标准化组织 ISO 在 2023 年发布了首个专门针对人工智能管理体系的标准 ISO/IEC 42001:2023《信息技术 人工智能 管理体系》。标准采用和其他管理体系标准相同的结构，为开发或使用人工智能系统的企业提供了可认证的管理体系框架，并可应用于不同的行业，通过该框架可开发并部署、持续监控安全的人工智能系统。

汽标委智能网联汽车分委会

ISO/IEC 42001:2023《信息技术 人工智能 管理体系》标准作为一套综合性企业能力管理规范，为人工智能系统安全管理提供了基础指导框架。标准由组织环境、领导作用、策划、支持、运行、绩效评价、改进等章节组成，采用 PDCA 循环（计划-实施-检查-改进），与 ISO 管理体系标准结构一致，如表 13-1 所示。

表 13-1 主要章节介绍[97]

章节	核心要求
4.组织环境	主要介绍了分析组织背景的重要性，以及组织需要做些什么来分析其背景，包括经济、商业趋势、法规要求、伦理期望以及组织绩效等问题
5.领导力	主要介绍了领导力在指导组织通过人工智能管理系统（AIMS）实施过程中的作用以及在不同管理层在促进 AIMS 安装过程中的角色和职责
6.策划	主要介绍了应如何策划建立 AIMS 的过程，并讨论管理者需要采取的进行风险分析的步骤
7.支持	主要介绍了领导层应采取的不同行动，AIMS 的安装准备包括能力框架、沟通策略，以及为什么需要文件化以用于审核和认证目的
8.运行	主要介绍了组织应如何管理、缓解和处理人工智能风险
9.绩效评价	主要介绍了组织如何设计一个绩效评估流程，以便跟踪、监控、测量和评估 AIMS 的整体绩效
10.改进	主要介绍了组织应采取的步骤，以收集关于 AIMS 的反馈。并说明如何利用这些反馈，通过 PDCA（计划-执行-检查-行动）方法来提升 AIMS 的整体绩效

汽标委智能网联汽车分委会

ISO/IEC 42001:2023 标准同时标准定义了 9 大关键控制域并对应提供了 38 项控制措施，具体指导企业如何实施关键控制域，如人工智能风险评估、数据质量管理和伦理审查等，具体如表 13-2 所示。

表 13-2 关键控制域及控制措施[97]

关键控制域	控制措施
A.2 与 AI 相关的方针	AI 政策：组织应制定开发政策文件 AI 系统的指导或使用。 与其他组织的一致性：组织应确定哪些其他政策可能受到 AI 系统目标的影响或适用于这些目标。 AI 政策的审查:应在计划的时间间隔内或根据需要额外审查 AI 政策，以确保其持续的适用性、充分性和有效性。
A.3 内部组织	AI 角色与职责：应根据组织的需求定义并分配 AI 相关的角色和职责。 问题报告：组织应定义并实施一个流程，用于报告在 AI 系统整个生命周期中对其角色的关切。
A.4AI 系统资源	资源文档化：组织应识别并记录在给定 AI 系统生命周期阶段所需的相关资源，以及其他与组织相关的 AI 活动。 数据资源：组织应记录 AI 系统所使用的数据资源信息。 工具资源：组织应记录 AI 系统所使用的工具资源信息。 系统和计算资源：组织应记录 AI 系统所使用的系统及计算资源信息。 人力资源：组织应记录用于 AI 系统开发、部署、运营、变更管理、维护、移交、退役以及验证和集成的人力资源及其能力信息。
A.5 评估 AI 系统影响	AI 系统影响评估流程：组织应建立流程，评估 AI 系统在其整个生命周期中可能对个人、群体或社会造成的潜在影响。 AI 系统影响评估的文档化：组织应记录 AI 系统影响评估的结果，并按

汽标委智能网联汽车分委会

	规定期限保存。 评估 AI 系统对个人或群体的影响：组织应评估并记录 AI 系统在其整个生命周期中对个人或群体的潜在影响。 评估 AI 系统的社会影响：组织应评估并记录其 AI 系统在整个生命周期中的潜在社会影响。
A.6AI 系统生命周期	负责任开发 AI 系统的目标：组织应识别并记录指导 AI 系统负责任开发的目标，并在开发生命周期中考虑这些目标并采取措施实现它们。 负责任 AI 系统设计与开发的流程：组织应定义并记录 AI 系统负责任设计与开发的具体流程。 AI 系统需求与规范：组织应针对新的 AI 系统或对现有系统的重大改进，明确并记录相关需求。 AI 系统设计与开发的文档化：组织应根据组织目标、记录的需求和规范标准，对 AI 系统的设计与开发进行文档化。 AI 系统验证与确认：组织应定义并记录 AI 系统的验证与确认措施，并明确其使用标准。 AI 系统部署：组织应记录部署计划，并确保在部署前满足适当的要求。 AI 系统运行与监控：组织应定义并记录 AI 系统持续运行所需的要素，至少包括系统与性能监控、修复、更新及支持。 AI 系统技术文档：组织应确定不同相关方（如用户、合作伙伴、监管机构）所需的 AI 系统技术文档类别，并以适当形式提供。 AI 系统事件日志记录：组织应明确 AI 系统生命周期中需启用事件日志记录的阶段（至少包含系统使用阶段）。
A.7AI 系统的数	AI 系统开发与增强的数据：组织应定义、记录并实施与 AI 系统开发相

汽标委智能网联汽车分委会

据	关的数据管理流程。 数据获取：组织应确定并记录 AI 系统所用数据的获取与选择细节。 AI 系统数据质量：组织应定义数据质量要求并确保开发及运行 AI 系统的数据符合要求。 数据溯源：组织应定义并记录数据及 AI 系统生命周期中数据来源的追溯流程。 数据预处理：组织应定义并记录数据选择标准及预处理方法。
A.8 相关方信息	用户系统文档与信息：组织应向 AI 系统用户提供必要信息。 外部报告机制：组织应为相关方提供报告 AI 系统负面影响的渠道。 事件通报：组织应制定并记录 AI 系统事件向用户通报的计划。 相关方信息披露：组织应明确并记录向相关方报告 AI 系统信息的义务。
A.9AI 系统的使用	负责任使用 AI 系统的流程：组织应定义并记录 AI 系统负责任使用的流程。 负责任使用 AI 系统的目标：组织应识别并记录指导 AI 系统负责任使用的目标。 AI 系统的预期用途：组织应确保 AI 系统的使用符合其预期用途及配套文档说明。
A.10 第三方和客户关系	责任分配：组织应确保在 AI 系统生命周期中，其与合作伙伴、供应商、客户及第三方之间的责任明确分配。 供应商：组织应建立流程，确保其使用的供应商提供的服务、产品或材料符合组织对 AI 系统负责任开发与使用的原则。 客户：组织应确保其在 AI 系统开发与使用中的负责任举措充分考虑客户的期望与需求。

汽标委智能网联汽车分委会

ISO/IEC 42001:2023 标准适用于不同行业、不同类型和规模的企业，它的突出特点在于强调企业能力的系统性提升和持续改进，并具有强大的扩展性，能够根据企业的特定需求进行灵活定制。在汽车企业构建车用人工智能安全管理体系中，ISO/IEC 42001:2023《信息技术 人工智能 管理体系》标准具有重要的适配性和优势。首先，它能够帮助汽车企业明确各环节参与者的责任归属，以确保人工智能系统的安全运行；其次强调了员工对人工智能治理和开发期间的安全意识，提高企业的整体风险管理能力；最后，作为现有管理体系（如 ISO 27001、ISO 9001）的补充，可帮助企业形成综合性的管理体系，提高管理效率。

但 ISO/IEC 42001:2023 作为通用 AI 管理体系标准，在车用 AI 领域有诸多需针对性细化的关键点。在安全需求层面，存在实时性与可靠性要求以及车规级技术特性方面的不足，如现有标准中风险分级方法未明确与汽车功能安全等级的映射关系，而车用 AI 需满足毫秒级决策响应和功能安全等级 ASIL - D；控制目标未体现车载芯片温域、电磁兼容、振动耐久等硬件标准，需补充车用 AI 硬件全生命周期可靠性管理要求，因此仍需要结合车用 AI 的实际情况进行进一步的细化和补充。

13.2.2.2 ISO/PAS 8800:2024 标准概述

2024 年发布的 ISO/PAS 8800:2024《道路车辆 安全和人工智能》作为道路车辆人工智能安全领域的最新实践指南，针对道路车辆人工智能的风险应对规定了全面、详细的要求和技术指导。其核心内容涉及人工智能安全生命周期、人工智能系统安全要求、人工智能系统安全设计、人工智能系统验证和确认、人工智能系统安全分析、数据相关考虑等方面[98]。标准提出了在车用人工智能安全生命周期每个阶段都应有明确的安全要求和实施指南，确保车用人工智能系统在整个生命周期内都能保持高水平的安全性。例如，在安全需求分析阶段，要求充分考虑功能安全（FUSA）和预期功能安全（SOTIF）等因素，识别潜在风险点等。

对于 ISO/PAS 8800:2024 标准归纳的 AI 系统的与安全相关特性及车用 AI 系统的安全测试相关具体分析请查阅本报告第 8 章。在车用 AI 安全管理体系标准

汽标委智能网联汽车分委会

化中，可基于 ISO/IEC 42001:2023《信息技术 人工智能 管理体系》标准的普适性和可扩展性，结合针对车用人工智能安全的 ISO/PAS 8800:2024《道路车辆 安全和人工智能》标准，指导企业构建覆盖企业能力和产品能力的完善的车用人工智能安全管理体系。

13.2.3 标准化挑战

车用 AI 安全管理体系的标准化面临多重技术与管理协同的深层挑战。在技术层面，现行通用标准（如 ISO/IEC 42001）难以满足汽车行业的特殊需求；同时，AI 特有的算法鲁棒性弱、模型可解释性差、数据偏见等内生风险缺乏量化评估标准，对抗攻击、认知误导等应用风险亦无统一测试框架。在标准协同层面，ISO/IEC 42001:2023、ISO/PAS 8800:2024 与 ISO 26262:2018 及 ISO 21448:2022 的条款交叉重叠，企业需投入大量资源进行差异化整合，易引发管理冗余；更关键的是，AI 全生命周期涉及主机厂、供应商、云服务商等多方主体，但数据归属、算法责任界定与事故追溯机制尚未形成国际共识。

全球化进程进一步加剧挑战，欧盟《AI 法案》侧重人权保护，中国标准强调数据主权，美国推行行业自律，区域合规差异显著增加车企研发成本。此外，车用 AI 安全管理尚未建立国际互认体系，重复测试认证严重延缓产品上市，制约产业全球化布局。因此汽车产业亟需加速构建车用人工智能安全管理体系的标准化框架，以系统性应对智能化转型中的安全挑战。

13.3 车用 AI 安全管理体系搭建思路研究

13.3.1 基于 ISO/IEC 42001:2023 标准的企业能力构建

在企业能力建设方面，构建车用人工智能安全管理体系是一项系统性工程，需要从战略高度明确其核心目标与实施路径。该体系的首要目标是建立一套科学、完整的安全管理框架，通过组织架构优化、流程再造和技术创新等多维度举措，全面提升企业在人工智能技术研发、应用和运营全过程中的安全风险防控能力。具体而言，车用人工智能安全管理体系需要全面覆盖系统全生命周期管理，包括

起始、设计与研发、验证与确认、部署、运行与监控、持续验证、重新评估、退役等环节[99]。在体系设计过程中，需要特别关注五大核心原则：安全性原则确保系统运行稳定可靠；公平性原则避免算法歧视；透明性原则保证决策过程可追溯；可解释性原则要求模型输出结果能被理解；隐私保护原则保障用户数据安全。通过这五大原则的贯彻落实，确保企业在开展智能驾驶、车联网等人工智能相关业务时，能够持续符合网络安全法、数据安全法等国家法律法规，实现业务的安全、合规、稳健发展。

在管理体系的具体构建与实施方面，应当严格遵循 ISO/IEC 42001:2023《信息技术 人工智能 管理体系》国际标准的管理要求，特别是第 4-10 章关于组织环境、领导作用、策划、支持、运行、绩效评价和改进等方面的规定。该管理体系的突出特点是将风险管理作为贯穿始终的核心主线，形成"识别-评估-处置-监控"的闭环管理机制。具体实施路径可分为六个关键步骤：首先，开展全面的组织环境分析，评估政治、经济、社会、技术、环境和法律等外部因素，同时通过 SWOT 分析把握内部优劣势，明确相关方需求，划定管理体系边界，建立风险识别基础框架。其次，强化领导作用，由企业高层直接负责制定《人工智能安全方针》，建立每季度定期评审机制，同时构建矩阵式管理组织，明确研发、测试、运维等各环节的安全责任主体。第三，系统开展风险策划工作，运用 FMEA 等方法识别潜在风险，建立风险登记册，制定分级管控措施，将风险评估结果与战略目标相结合，设定可量化的安全 KPI 指标。第四，夯实资源保障基础，组建专业的安全团队，设计分层级的培训体系，建立跨部门协作机制，搭建风险信息共享平台。第五，严格运行过程控制，建立月度风险评估例会制度，开展系统影响评估（SIA）。最后，构建持续改进机制，通过部署安全态势感知系统实现实时监测，每半年开展内部审核，年度管理评审输出改进计划，形成完整的 PDCA 循环，确保从需求分析到系统退役的全生命周期各环节风险始终处于受控状态。

在控制与实施环节，车用人工智能安全管理体系需要构建全面覆盖通用人工智能安全风险的防控机制，以有效应对 AI 汽车、AI 机器人及 AI 芯片等智能产品的共性安全挑战。根据全国网络安全标准化技术委员会 2024 年发布的《人工

智能安全治理框架》，人工智能系统安全风险可系统性地划分为内生安全风险和应用安全风险两大维度。其中内生安全风险主要包含数据安全风险、模型算法安全风险以及系统安全风险；而应用安全风险则涵盖网络域风险、现实域风险、认知域风险以及伦理域风险，如表 13-3 所示。这些风险在人工智能系统全生命周期的不同阶段呈现出差异化特征：在设计与研发阶段需重点防范算法偏见和系统漏洞；在验证与确认阶段要着重应对数据污染和模型劫持；部署阶段需关注网络攻击和接口安全；运行阶段则需警惕操作风险和决策误判等问题。因此，安全管理体系的控制措施需要针对各阶段的风险特点，建立分层次、差异化的防控策略，通过技术防护、管理控制和应急响应相结合的方式，构建全方位的安全防护网，确保各类风险得到有效管控。

表 13-3 通用人工智能安全风险列表[100]

安全风险		
内生（自身）安全风险	模型算法安全风险	可解释性差的风险
		偏见、歧视风险
		鲁棒性弱风险
		被窃取、篡改的风险
		输出不可靠风险
		对抗攻击风险
	数据安全风险	违规收集使用数据风险
		训练数据含不当内容、被“投毒”风险
		训练数据标注不规范风险
		数据泄露风险
	系统安全风险	缺陷、后门被攻击利用风险
		算力安全风险

汽标委智能网联汽车分委会

		供应链安全风险
应用安全风险	网络域风险	信息内容安全风险
		混淆事实、误导用户、绕过鉴权的风险
		不当使用引发信息泄露风险
		滥用于网络攻击的风险
		模型复用的缺陷传导风险
	现实域风险	诱发传统经济社会安全风险
		用于违法犯罪活动的风险
		两用物项和技术滥用风险
	认知域风险	加剧“信息茧房”效应风险
		用于开展认知战的风险
	伦理域风险	加剧社会歧视偏见、扩大智能鸿沟的风险
		挑战传统社会秩序的风险
		未来脱离控制的风险

13.3.2 基于 ISO/PAS 8800:2024 标准的产品能力构建

在构建车用人工智能安全管理体系时，汽车企业需首先深入理解该标准的核心内容，包括人工智能安全生命周期、人工智能系统安全要求、人工智能系统安全设计、人工智能系统验证和确认、人工智能系统安全分析、数据相关考虑等方面。在具体实施过程中，企业应当结合自身在智能网联汽车领域的研发制造特点，将人工智能安全管理要求与现有的 ISO 26262 功能安全标准和 ISO 21448 预期功能安全标准有机融合，构建统一协调的安全管理体系。这种融合需要充分考虑汽车电子电气架构的特殊性，以及自动驾驶系统等关键应用场景的安全需求，通过差异化的条款适用性评估，确保最终建立的管理体系既符合标准要求，又具备实

际可操作性。特别值得注意的是，在标准条款的转化应用过程中，企业需要组建跨部门的标准实施团队，包括研发、质量、法务等多个职能部门共同参与，以确保标准要求能够准确落实到产品全生命周期的各个环节。

应聚焦于车用人工智能安全生命周期的风险管控，从明确人工智能系统的安全需求、人工智能系统设计与验证确认、评估安全保障措施到部署运营与退役阶段，每个阶段都需融入风险管理的理念和措施。其中在人工智能系统设计与验证确认阶段，标准详细规定了车用人工智能系统的验证与确认方法，为开发和部署安全可靠的车用人工智能系统提供了重要指导。这包括基于 V 模型的人工智能系统设计、人工智能单元设计、人工智能单元集成、人工智能单元验证、人工智能系统确认与验证，以及全周期的人工智能安全分析，如图 13-1 为车用人工智能系统验证与确认环节示意图。同时考虑到数据在车用人工智能系统中的核心地位，标准专门设置了数据相关考虑的章节，对数据采集、预处理、存储、传输等环节提出具体要求。例如，规定数据采集过程中需确保数据质量和代表性；数据预处理阶段需注意避免信息丢失或失真；数据存储需采用加密保护措施；数据传输则需建立安全通道，防止数据泄露或篡改等。

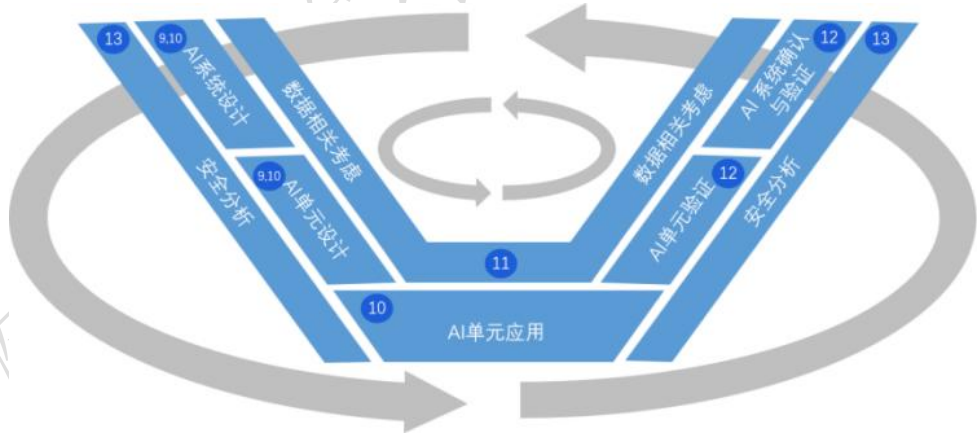


图 13-1 车用人工智能系统验证与确认环节示意图[98]

ISO/PAS 8800:2024《道路车辆 安全和人工智能》国际标准在构建人工智能系统安全分析框架时，融合了传统功能安全和预期功能安全的方法论，并针对人工智能技术的特性进行了系统性扩展。该标准不仅涵盖了传统电子电气系统的安全考量，更重点纳入了人工智能特有的风险维度，包括算法局限性、系统鲁棒性、

泛化性能等关键因素。根据《2024 车用人工智能风险管理白皮书》的研究，车用人工智能安全风险可系统性地划分为四大类：功能安全风险（涉及硬件故障和系统失效）、预期功能安全风险（包括场景覆盖不足和性能局限）、数据安全风险（涵盖数据质量和隐私保护）以及网络安全风险（针对外部攻击和系统入侵）。基于这一分类体系，企业可以结合 ISO/PAS 8800:2024 标准提供的安全需求推导方法论，从多个维度开展全面的风险分析工作，建立精准高效的风险应对机制，确保各类安全风险都能得到专业化的处理。这种结构化的风险管理方法不仅提升了风险识别的全面性，更通过专业分工显著增强了风险应对的效率和效果，为智能网联汽车的可靠运行提供了坚实保障。

表 13-4 车用人工智能安全风险列表[91]

风险分类	风险要素	风险点	影响的阶段	问题归属
内生风险	数据	偏见与歧视	开发阶段	预期功能安全
		隐私泄露	开发阶段 应用阶段	数据安全
		数据质量和完整性不足	开发阶段	数据安全 预期功能安全
		数据格式和协议不兼容	部署阶段	功能安全
		数据残留	销毁阶段	数据安全
	模型	模型偏差	开发阶段	预期功能安全
		鲁棒性弱	开发阶段	预期功能安全
		可解释性差	开发阶段 应用阶段	预期功能安全
		模型残留	销毁阶段	数据安全
		部署环境不匹配	部署阶段	功能安全
		服务不可用	应用阶段	功能安全

汽标委智能网联汽车分委会

		环境配置和凭证泄露	应用阶段	网络安全
		计算环境残留安全隐患	应用阶段 销毁阶段	网络安全
应用风险	数据	信息窃取	开发阶段 应用阶段	数据安全 网络安全
	模型	复用缺陷传导	应用阶段	预期功能安全
		对抗攻击	应用阶段	预期功能安全
		内容隐患	开发阶段 应用阶段	预期功能安全
	环境	网络攻击	开发阶段 应用阶段	网络安全

13.3.3 企业能力与产品能力的协同管理

在汽车企业构建车用人工智能安全管理体系的过程中，ISO/IEC 42001:2023 和 ISO/PAS 8800:2024 两大标准发挥着互补协同的关键作用。ISO/IEC 42001:2023 作为通用性标准，着重于企业整体层面的人工智能管理能力建设，为企业提供包括组织架构设计、制度流程建立、人员能力培养等在内的全方位管理框架；而 ISO/PAS 8800:2024 作为行业技术标准，则聚焦于汽车产品本身的人工智能安全技术要求，特别是在功能安全、预期功能安全等专业技术领域的落地实施。这种"管理+技术"的双轮驱动模式，形成了自上而下的完整管理体系：企业级管理能力为产品级技术实现提供制度保障、资源支持和战略指导，确保技术研发有章可循；同时，产品级技术实践又能反哺企业级管理，通过实际应用场景中的安全数据反馈和经验积累，持续优化企业的管理策略和资源配置。这种双向互动机制不仅实现了管理要求与技术标准的无缝衔接，更通过持续改进循环不断提升企业的人工智能安全管理成熟度，最终构建起覆盖组织管理、产品研发、生产制造、运营服务全价值链的安全管理体系，显著增强企业在智能化转型中的核心竞争力。

汽标委智能网联汽车分委会

在具体实施层面，汽车企业需要建立系统化的协同机制来充分发挥两大标准的合力效应。首先要构建矩阵式的组织架构，成立由质量、研发、信息安全等多部门组成的联合工作组，通过定期召开技术协调会、管理评审会等形式，确保企业级管理要求与产品级技术标准始终保持同步更新。其次要推进流程深度融合，重点对风险管理、变更管理、验证确认等关键流程进行一体化设计，例如将 ISO/IEC 42001 的风险评估方法与 ISO/PAS 8800 的安全分析技术相结合，开发适合汽车行业特点的复合型风险评估工具。同时，企业还需要建立统一的知识管理平台，整合两大标准实施过程中产生的技术文档、案例经验、最佳实践等知识资产，实现跨部门、跨层级的资源共享。通过这些系统化的协同措施，不仅可以避免标准实施中的重复工作和资源浪费，更能形成"1+1>2"的协同效应，全面提升企业车用人工智能安全管理的整体效能，为智能网联汽车的安全可靠运行提供坚实保障。

13.4 总结与展望

本章对车用 AI 安全管理体系标准化进行了全面的探讨，分析了研究背景、现状和范围，阐述了标准化需求和挑战，对 ISO/IEC 42001:2023 和 ISO/PAS 8800:2024 标准进行了分析，提出了车用 AI 安全管理体系的搭建思路。通过本章，为车用 AI 安全管理体系的标准化建设提供了有益的参考。

未来，随着 AI 技术的不断发展和汽车产业的变革，车用 AI 安全管理体系标准化将面临更多的挑战和机遇。需要进一步加强国际合作，制定更加统一、完善的标准化体系，提高车用 AI 安全管理的水平；推动标准的实施和认证，促进汽车企业积极采用标准化的安全管理体系，保障车用 AI 系统的安全可靠运行。

汽标委智能网联汽车分委会

参考文献

[1] 中国汽车技术研究中心. 自动驾驶功能仿真测试标准化需求研究报告［R］. 天津: 中国汽车技术研究中心, 2021.

[2] Abrecht S, Martin A, Brogle M, et al. Deep learning safety concerns in automated driving perception[J]. IEEE Transactions on Intelligent Vehicles, 2024.

[3] 中国汽车工程研究院股份有限公司. 场景驱动自动驾驶商业化发展与展望白皮书［R］. 重庆: 中国汽车工程研究院, 2023-07.

[4] Yang K, Dai Q, He C, et al. Natural-language-driven Simulation Benchmark and Copilot for Efficient Production of Object Interactions in Virtual Road Scenes[J/OL]. arXiv preprint arXiv:2312.04008, 2023. https://arxiv.org/abs/2312.04008

[5] 欧盟人工智能法案（AI Act）［Z］. Brussels: 欧盟官方公报, 2024.

[6] 欧盟议会与理事会. 欧盟整车型式核准框架法规（Regulation (EU) 2018/858）［Z］. 2018.

[7] 欧盟议会与理事会. 通用安全法规（Regulation (EU) 2023/988）［Z］. 2019.

[8] 美国国会. 2020 年国家人工智能倡议法案（NAII）［Z］.

[9] 环球网. 美国发布《联邦自动驾驶汽车政策指南》［EB/OL］. (2016-09-22)［2025-06-30 引用］. https://world.huanqiu.com/article/9CaKrnJXIm9

[10] 日本自动驾驶政策研究机构. 日本 L3 以上自动驾驶发展现状和路线图［EB/OL］. (2024-03-05)［2025-06-30 引用］. https://mp.weixin.qq.com/s?__biz=Mzk0NzE4MTQwMg==&mid=2247484526&idx=1&sn=a8acd519fef0e8084c419d538bcf1e9b

[11] 工业和信息化部, 国家市场监督管理总局. 关于进一步加强智能网联汽车产品准入、召回及软件在线升级管理的通知［Z］. 工信部联通装〔2025〕45 号, 2025-02-28.

[12] 工业和信息化部. 关于开展智能网联汽车“车路云一体化”应用试点工作的通知［Z］. 工信部联通装〔2023〕268 号, 2024-01-15.

汽标委智能网联专分委会

[13] 新京报. 自动驾驶拟受法律规范，《道路交通安全法》修订列入年度立法计划［EB/OL］. (2024-08-27)［2025-06-30 引用］. https://m.bjnews.com.cn/detail/1724729866129482.html

[14] 深圳市人民代表大会常务委员会. 深圳经济特区智能网联汽车管理条例［Z］.2022-06-30.

[15] 北京市第十六届人民代表大会常务委员会. 北京市自动驾驶汽车条例［Z］.2024-12-31.

[16] ISO/PAS 8800:2024. Road Vehicles – Safety and Artificial Intelligence[S].

[17] Salay R, Queiroz R, Czarnecki K. An Analysis of ISO 26262: Using Machine Learning Safely in Automotive Software[C]// ICSE Workshops, 2017.

[18] ISO/IEC 42001:2023. Artificial intelligence — Management system[S].

[19] SAE J3016. Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles[S].

[20] ISO 34501:2022. Road vehicles[S].

[21] ISO 34502:2022. Road vehicles — Test scenarios for automated driving systems — Scenario based safety evaluation framework[S].

[22] UL 4600. Standard for Safety for the Evaluation of Autonomous Products[S].

[23] IEEE 2846-2022. IEEE Standard for Assumptions in Safety-Related Models for Automated Driving Systems[S].

[24] GB/T 40429-2021. 汽车驾驶自动化分级[S].

[25] GB/T 34590. 道路车辆 功能安全[S].

[26] 智能网联汽车标准工作组. 三项智能网联汽车强制性国家标准正式发布［EB/OL］. (2024-09-03)［2025-06-30 引用］. https://www.gov.cn/lianbo/bumen/202409/content_6972669.htm

[27] GB/T 41798-2022 《智能网联汽车自动驾驶功能场地试验方法及要求》

[28] B/T 43119-2023 《自动驾驶封闭测试场地建设技术要求》

[29] GB/T 43759-2024 《智能网联汽车运行安全测试技术要求》

汽标委智能网联汽车分委会

[30] 工业和信息化部，国家市场监督管理总局. 关于进一步加强智能网联汽车产品准入、召回及软件在线升级管理的通知［Z］. 工信部联通装〔2025〕45 号，2025-02-28.

[31] 美国电气电子工程师协会. 《大数据业务安全风险评估标准》 IEEE 2813:2020

[32] 全国信息安全标准化技术委员会.《生成式人工智能服务安全基本要求》

[33] 阿里云.《大模型安全研究报告》

[34] 云安全联盟《AI 安全白皮书》

[35] 浙江大学. AIcert 人工智能安全评测平台

[36] 蒋少良，封正，华莎等：辅助驾驶汽车信息安全检验关键技术研究及测试验证

[37] 汽车信息安全风险评估标准需求研究报告

[38] 邱彬，李广友 智能网联汽车数据安全管理研究[J].汽车工程学报，2022，12（3）：307-313.

[39] 人工智能安全测评白皮书（2021）

[40] 北京市地方标准 《信息安全 人工智能数据安全通用要求》征求意见稿

[41] 人工智能安全框架研究 朱倩倩;唐志敏;王新哲; -《工业信息安全》- 2022-10-18

[42] 基于混合分析的汽车信息安全风险分级方法 赵浩;刘平一;刘天宇; -《汽车实用技术》- 2021-08-25

[43] 2022 人工智能标准化白皮书-20240512235520.docx-原创力文档 - 《互联网文档资源（https://max.book118.）》

[44] 人工智能安全国际标准化进展研究 张琳琳;王腾; -《信息通信技术与政策》- 2021-11-15

[45] 2022 欧盟 AI 网络安全与标准化报告 1.docx-原创力文档 -《互联网文档资源（https://max.book118.）》

[46] 2023 年度欧盟网络安全态势综述 孙频捷; -《中国信息安全》- 2024-01-15

[47] 《人工智能安全标准化白皮书（2023 版）》.pdf-原创力文档 - 《互联网文档资源（https://max.book118.）》

[48] 国家大事 -《中国信息安全》- 2020-09-15

汽标委智能网联汽车分委会

[49] 2024 人工智能 AI 安全白皮书.docx-原创力文档 - 《互联网文档资源（https://max.book118.）》

[50] 具有可延展性和解释性的多响应回归及其相关问题研究吴捷 -《中国科学技术大学博士论文》

[51] 无人系统的视觉感知安全研究 陈小前;姜廷松;姚雯;李超; -《智能安全》

[52] AIGC 时代编辑的身份挑战与角色转型 罗飞宁;刘壮;谢文亮; -《出版广角》- 2023-08-15

[53] 构筑智能网联汽车信息安全标准保障体系 王兆;李宝田;孙航; -《中国信息安全》- 2021-07-15

[54] 2023 年中国人工智能大模型技术白皮书.pdf-原创力文档- 《互联网文档资源（https://max.book118.）》

[55] 国家强标|工信部公布《汽车整车信息安全技术要求》（征求意见稿）

[56] ISO 26262:2018, road vehicles functional safety, ISO, 2018.

[57] X. He, H. Yang, Z. Hu, and C. Lv, “Robust lane change decision making for autonomous vehicles: An observation adversarial reinforcement learning approach,” IEEE Transactions on Intelligent Vehicles, vol. 8, no. 1, pp. 184–193, 2023.

[58] X. Tang, K. Yang, H. Wang, J. Wu, Y. Qin, W. Yu, and D. Cao, “Prediction-uncertainty-aware decision-making for autonomous vehicles,” IEEE Transactions on Intelligent Vehicles, vol. 7, no. 4, pp. 849– 862, 2022.

[59] R. Michelmore, M. Wicker, L. Laurenti, L. Cardelli, Y. Gal, and M. Kwiatkowska, “Uncertainty quantification with statistical guarantees in end-to-end autonomous driving control,” in 2020 IEEE International Conference on Robotics and Automation (ICRA), 2020, pp. 7344–7350.

[60] IEC 61508 电气/电子/可编程电子安全系统的功能安全，IEC,2000

[61] EN 50129 铁路应用－通信、信号和处理系统-信号的安全相关电子处理系统，EN,2018

[62] IEC 61511 Functional safety - Safety instrumented systems for the process industry sector, IEC ,2016

汽标委智能网联汽车分委会

[63] GB/T 34590 道路车辆-功能安全，国家标准,2017

[64] GB 17675 汽车转向系-基本要求，国家标准,2001

[65] ISO/PAS 21448 Road vehicles — Safety of the intended functionality, ISO, 2011

[66] ISO/TS 5083 Road vehicles — Safety for automated driving systems — Design, verif ication and validation ISO,2025

[67] ISO/PAS 8800 Road vehicles — Safety and artificial intelligence, ISO, 2024

[68] ISO/IEC TR 5469 Artificial intelligence — Functional safety and AI systems, ISO IE C, 2024

[69] ISO/TR 4804 Road vehicles — Safety and cybersecurity for automated driving system s — Design, verification and validation, ISO, 2020

[70] UN ECE R157 关于自动车道保持系统(ALKS)车辆认证的统一规定，UN，2024

[71] Yulong Cao, Ningfei Wang, Chaowei Xiao, Dawei Yang, Jin Fang, Ruigang Yang, Qi Alfred Chen, Mingyan Liu, and Bo Li. 2021. Invisible for both camera and lidar: Se curity of multi-sensor fusion based perception in autonomous driving under physical-w orld attacks. In 2021 IEEE Symposium on Security and Privacy (SP). IEEE, 176–19 4.

[72] Yang Sun, Christopher M Poskitt, Jun Sun, Yuqi Chen, and Zijiang Yang. 2022. Law Breaker: An approach for specifying traffic laws and fuzzing autonomous vehicles. In Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering. 1–12

[73] Thomas Laurent, Paolo Arcaini, Fuyuki Ishikawa, and Anthony Ventresque. 2019. A m utation-based approach for assessing weight coverage of a path planner. In 2019 26th Asia-Pacific Software Engineering Conference (APSEC). IEEE, 94–101.

[74] Thomas Laurent, Paolo Arcaini, Fuyuki Ishikawa, and Anthony Ventresque. 2020. Achi eving Weight Coverage for an Autonomous Driving System with Searchbased Test Ge neration. In 2020 25th International Conference on Engineering of Complex Computer Systems (ICECCS). IEEE, 93–102.

汽标委智能网联汽车分委会

[75] Baidu. Apollo. An open autonomous driving platform. https://github.com/ ApolloAuto/apollo.

[76] 国家新一代人工智能治理专业委员会.新一代人工智能伦理规范 [EB/OL]. (2021-09-26) [2025-06-20].

[77] 中共中央办公厅,国务院办公厅.关于加强科技伦理治理的意见 [EB/OL]. (2022-04-10)[2025-06-20].

[78] 国家人工智能标准化总体组,全国信标委人工智能分委会.人工智能伦理治理标准化指南 [EB/OL].(2025-01-19)[2025-06-20].

[79] Xie Yu, Wang Xiaoyi. The Ethical Risks of Artificial Intelligence Emotion and Their Countermeasures[J]. Studies in Ethics, 2024(1): 132-140.

[80] 王东,张振.人工智能伦理风险的镜像、透视及其规避 [J]. 伦理学研究，2021 (1): 109-115.

[81] 陈小平。人工智能：技术条件、风险分析和创新模式升级 [J]. 科学与社会，2021，11 (2): 1-14.

[82] Madry A, Makelov A, Schmidt L, et al. Towards Deep Learning Models Resistant to Adversarial Attacks [J]. arXiv preprint arXiv:1706.06083, 2017.

[83] Chandra A, Yao X. Ensemble learning using multi-objective evolutionary algorithms [J]. Journal of Mathematical Modelling and Algorithms, 2006, 5 (4): 417-445.

[84] Chen H, Yao X. Multiobjective Neural Network Ensembles based on Regularized Negative Correlation Learning [J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22 (12): 1738-1751.

[85] UNESCO. Recommendation on the Ethics of Artificial Intelligence [EB/OL]. (2021-11) [2025-06-20]. https://unesdoc.unesco.org/ark:/48223/pf0000377046.

[86] 中国信通院.人工智能伦理治理研究报告.[EB/OL]. (2023-12)[2025-06-20].

[87] 刘志毅，梁正，郑烨婕。黑镜与秩序：数智化风险社会下的人工智能伦理与治理 [M]. 北京：清华大学出版社，2022.

[88] Gu T L, Li L. Artificial Moral Agents and Their Design Methodology: Retrospect and Prospect [J]. Chinese Journal of Computers, 2021, 44 (4): 632-651.

汽标委智能网联汽车分委会

[89] MCSTAY A. Emotional AI and EdTech: Serving the Public Good?[J]. Learning, Media and Technology, 2019, 45 (3): 299-316.

[90] 高山冰，汪婧。智能传播时代社交机器人的兴起、挑战与反思 [J]. 现代传播 (中国传媒大学学报), 2020 (11): 106-112.

[91] 智能汽车安全技术全国重点实验室.2024 车用人工智能风险管理白皮书[R/OL].[2024-04-28].http://ivstskl.changan.com.cn/?p=4702

[92] Lincan L, Wei S, Wei D. Data-Centric Evolution in Autonomous Driving: A Comprehensive Survey of Big Data System, Data Mining, and Closed-Loop Technologies.[J]. arXiv.2024 Jan

[93] Zhipeng Z. 智能座舱技术现状及发展趋势研究[EB/OL].(2022-08-26)[2025-06-20]

[94] 曹鲁政. 新能源汽车中人工智能技术应用探究 [EB/OL].(2024-12-04)[2025-06-20]

[95] 亿欧智库.2024 中国汽车智能底盘产业技术趋势分析报告.[R].2024-05-31

[96] 小鹏汽车.小鹏 10 年热爱之夜：小鹏 MONA M03 发布，多项核心技术首次亮相[Z/OL].[2025-04-28].https://www.xiaopeng.com/news/company_news/5355.html

[97] ISO/IEC 42001:2023 Information technology - Artificial intelligence - Management system[S/OL].[2025-04-28].https://www.iso.org/standard/81230.html

[98] ISO/PAS 8800:2024 Road vehicles - Safety and artificial intelligence[S/OL].[2025-04-25].https://www.iso.org/standard/83303.html

[99] ISO/IEC 5338:2023 Information technology - Artificial intelligence - AI system life cycle processes[S/OL].[2025-04-28].https://www.iso.org/standard/81118.html

[100] 全国网络安全标准化技术委员会.人工智能安全治理框架[R/OL].(2024-09-09)[2024-04-28].https://www.tc260.org.cn/front/postDetail.html?id=2024090910280

汽标委智能网联汽车分委会

附录

附录 A：国内车用 AI 相关标准明细表

序号	类型	归口组织	标准研究方向	状态
1	国家标准	全国网络安全标准化技术委员会	网络安全技术 人工智能生成合成内容标识方法	发布 GB 45438-2025
2	国家标准	全国信息技术标准化技术委员会	人工智能 管理体系	发布 GB/T 45081-2024
3	国家标准	全国信息技术标准化技术委员会	人工智能 面向机器学习的数据标注规程	发布 GB/T 42755-2023
4	国家标准	全国信息技术标准化技术委员会	网络安全技术 生成式人工智能预训练和优化训练数据安全规范	发布 GB/T 45652-2025
5	国家标准	全国网络安全标准化技术委员会	网络安全技术 生成式人工智能数据标注安全规范	发布 GB/T 45674-2025
6	国家标准	全国网络安全标准化技术委员会	信息安全技术 机器学习算法安全评估规范	发布 GB/T 42888-2023
7	国家标准	全国信息技术标准化技术委员会	人工智能 深度学习算法评估	发布 GB/T 45225-2025

8	国家标准	全国道路交通管理标准化技术委员会	智能网联汽车 运行安全测试技术要求	发布 GB/T 43766-2024
9	国家标准	全国道路交通管理标准化技术委员会	智能网联汽车 运行安全测试项目和方法	发布 GB/T 44850-2024
10	国家标准	全国网络安全标准化技术委员会	网络安全技术 生成式人工智能服务安全基本要求	发布 GB/T 45654-2025
11	国标计划	全国信息技术标准化技术委员会	人工智能 可信赖 第 1 部分：通则	征求意见 20240562-T-469
12	国标计划	全国信息技术标准化技术委员会	人工智能 风险管理能力评估	正在批准 20231740-T-469
13	国标计划	全国智能技术社会应用与评估基础标准化工作组	生成式人工智能技术应用社会影响 服务提供者合规管理指南	起草 20242884-T-469
14	国标计划	全国智能技术社会应用与评估基础标准化工作组	生成式人工智能技术应用社会影响 评估指南	起草 20242891-T-469
15	国标计划	全国汽车标准化技术委员会	智能网联汽车 数据安全管理体系规范	起草 20243203-T-339
16	国标计划	全国网络安全标准化技术委员会	网络安全技术 互联网信息服务深度合成安全规范	起草 20240395-T-469
17	行业标准	中国通信标准化协会	移动智能终端 可信人工智能安全指南	发布 YD/T 4960-2024
18	行业标准	中国通信标准化协会	移动智能终端 人工智能应用的个人信息保护技术要求及评估方法	发布 YD/T 4994-2024

19	行标计划	中国通信标准化协会	人工智能安全治理 生成式人工智能图像检测服务系统能力要求	起草 2024-1352T-YD
20	行标计划	中国通信标准化协会	人工智能 安全治理系统风险管理能力要求	起草 2024-1353T-YD
21	行标计划	中国通信标准化协会	人工智能开发平台通用能力要求 第 2 部分：安全要求	起草 2023-0041T-YD
22	行标计划	中国通信标准化协会	生成式人工智能服务的个人信息保护规范	起草 H-202308286945
23	行标计划	中国通信标准化协会	生成式人工智能终端 安全技术要求	起草 H-202403188928
24	行标计划	中国通信标准化协会	图像视频深度合成鉴伪检测算法引擎安全性评测方法	起草 2022-1490T-YD
25	行标计划	中国通信标准化协会	音频深度合成鉴伪检测算法引擎安全性评测方法	起草 2023-1328T-YD
26	行标计划	中国通信标准化协会	安全大模型能力要求 数据安全领域	起草 H-202409309580

汽标委智

附录 B：国际车用 AI 相关标准明细表

序号	归口组织	标准名称	状态
1	ISO/IEC JTC 1/SC 42	信息技术 信息技术治理 组织使用人工智能的治理影响	发布 ISO/IEC 38507:2022
2	ISO/IEC JTC 1/SC 42	信息技术 大数据参考架构 第 2 部分：用例和导出需求	发布 ISO/IEC TR 20547-2:2018
3	ISO/IEC JTC 1/SC 42	信息技术 大数据参考架构 第 5 部分：标准路线图	发布 ISO/IEC TR 20547-5:2018
4	ISO/IEC JTC 1/SC 42/JWG 2	人工智能 AI 测试 第 2 部分：AI 系统测试概述	批准阶段 ISO/IEC DTS 42119-2
5	ISO/IEC JTC 1/SC 42/JWG 2	人工智能 AI 测试 第 3 部分：AI 系统的验证和确认分析	批准阶段 ISO/IEC DTS 42119-3
6	ISO/IEC JTC 1/SC 42/JWG 4	人工智能 功能安全和人工智能系统 第 1 部分：要求	起草 ISO/IEC AWI TS 22440-1
7	ISO/IEC JTC 1/SC 42/JWG 4	人工智能 功能安全和人工智能系统 第 2 部分：指南	起草 ISO/IEC AWI TS 22440-2
8	ISO/IEC JTC 1/SC 42/JWG 4	人工智能 功能安全和人工智能系统 第 3 部分：应用示例	起草 ISO/IEC AWI TS 22440-3

标委智能网联汽车分委

9	ISO/IEC JTC 1/SC 42/JWG 4	信息技术 人工智能 AI 系统中不确定性量化的指南和要求	起草 ISO/IEC AWI TS 25223
10	ISO/IEC JTC 1/SC 42/JWG 5	人工智能 准确自然语言处理系统的评估方法	起草 ISO/IEC AWI 23282
11	ISO/IEC JTC 1/SC 42/JWG 5	人工智能 与自然语言处理相关的人工智能任务和功能概述	起草 ISO/IEC AWI TR 23281
12	ISO/IEC JTC 1/SC 42/JWG 5	信息技术 人工智能 自然语言处理特定计算方法分类法	起草 ISO/IEC PWI 25526
13	ISO/IEC JTC 1/SC 42/JWG 6	信息技术 人工智能 人工智能系统合格评定方案制定的高层框架和指南	起草 ISO/IEC AWI 42007
14	ISO/IEC JTC 1/SC 42/WG 1	信息技术 人工智能 人工智能概念和术语	发布 ISO/IEC 22989:2022
15	ISO/IEC JTC 1/SC 42/WG 1	信息技术 人工智能 管理系统	发布 ISO/IEC 42001:2023
16	ISO/IEC JTC 1/SC 42/WG 1	人工智能 AI 系统日志记录	草案 ISO/IEC CD 24970
17	ISO/IEC JTC 1/SC 42/WG 1	信息技术 人工智能 人工智能系统方法和能力分类法	起草 ISO/IEC AWI 42102
18	ISO/IEC JTC 1/SC 42/WG 1	信息技术 人工智能 提供人工智能管理体系审核和认证机构的要求	发布 ISO/IEC 42006:2025

汽标委智能网联汽车分委会

19	ISO/IEC JTC 1/SC 42/WG 1	信息技术 人工智能 AI 系统影响评估	发布 ISO/IEC 42005:2025
20	ISO/IEC JTC 1/SC 42/WG 1	信息技术 人工智能 ISO/IEC 42001 的实施指南	起草 ISO/IEC AWI 25651
21	ISO/IEC JTC 1/SC 42/WG 1	信息技术 人工智能 基于 ISO/IEC 42001 的 AI 管理体系审核指南	起草 ISO/IEC PWI 42114
22	ISO/IEC JTC 1/SC 42/WG 2	信息技术 大数据 概述和术语	发布 ISO/IEC 20546:2019
23	ISO/IEC JTC 1/SC 42/WG 2	信息技术 大数据参考架构 第 3 部分：参考架构	发布 ISO/IEC 20547-3:2020
24	ISO/IEC JTC 1/SC 42/WG 2	信息技术 人工智能 大数据分析过程管理框架	发布 ISO/IEC 24668:2022
25	ISO/IEC JTC 1/SC 42/WG 2	人工智能 分析和机器学习（ML）的数据质量 第 1 部分：概述、术语和示例	发布 ISO/IEC 5259-1:2024
26	ISO/IEC JTC 1/SC 42/WG 2	人工智能 分析和机器学习（ML）的数据质量 第 2 部分：数据质量度量	发布 ISO/IEC 5259-2:2024
27	ISO/IEC JTC 1/SC 42/WG 2	人工智能 分析和机器学习的数据质量 第 3 部分：数据质量管理要求和指南	发布 ISO/IEC 5259-3:2024
28	ISO/IEC JTC 1/SC 42/WG 2	人工智能 分析和机器学习的数据质量 第 4 部分：数据质量过程框架	发布 ISO/IEC 5259-4:2024

汽标委智能网联汽车分委会

29	ISO/IEC JTC 1/SC 42/WG 2	信息技术 人工智能 数据生命周期框架	发布 ISO/IEC 8183:2023
30	ISO/IEC JTC 1/SC 42/WG 2	信息技术 人工智能 分析和机器学习的数据概况概述	起草 ISO/IEC AWI TR 25523
31	ISO/IEC JTC 1/SC 42/WG 2	信息技术 人工智能 人工智能系统背景下的合成数据概述	起草 ISO/IEC AWI TR 42103
32	ISO/IEC JTC 1/SC 42/WG 2	人工智能 分析和机器学习（ML）的数据质量 第 6 部分：数据质量可视化框架	草案 ISO/IEC CD TR 5259-6
33	ISO/IEC JTC 1/SC 42/WG 2	人工智能 分析和机器学习（ML）的数据质量 第 5 部分：数据质量治理框架	发布 ISO/IEC 5259-5:2025
34	ISO/IEC JTC 1/SC 42/WG 2	信息技术 人工智能 生成式 AI 应用输出数据质量指南	起草 ISO/IEC AWI 25590
35	ISO/IEC JTC 1/SC 42/WG 2	信息技术 人工智能 分析和机器学习生成数据使用框架	起草 ISO/IEC PWI 42116
36	ISO/IEC JTC 1/SC 42/WG 2	信息技术 人工智能 机器学习训练数据去标识化	起草 ISO/IEC PWI 47559
37	ISO/IEC JTC 1/SC 42/WG 2	信息技术 大数据参考架构 第 1 部分：框架和应用流程	发布 ISO/IEC TR 20547-1:2020
38	ISO/IEC JTC 1/SC 42/WG 3	信息技术 人工智能 风险管理指南	发布 ISO/IEC 23894:2023

汽标委智能网联汽车分委会

39	ISO/IEC JTC 1/SC 42/WG 3	人工智能 神经网络鲁棒性评估 第 2 部分：形式化方法使用指南	发布 ISO/IEC 24029-2:2023
40	ISO/IEC JTC 1/SC 42/WG 3	软件工程 系统与软件质量要求和评估（SQuaRE）— 人工智能系统质量模型	发布 ISO/IEC 25059:2023
41	ISO/IEC JTC 1/SC 42/WG 3	人工智能 神经网络鲁棒性评估 第 3 部分：统计方法使用指南	起草 ISO/IEC AWI 24029-3
42	ISO/IEC JTC 1/SC 42/WG 3	人工智能 AI 增强决策助推	起草 ISO/IEC AWI 25029
43	ISO/IEC JTC 1/SC 42/WG 3	信息技术 人工智能 解决社会问题和伦理考量的指南	草案 ISO/IEC CD TS 22443
44	ISO/IEC JTC 1/SC 42/WG 3	信息技术 人工智能 AI 系统的人类监督指南	草案 ISO/IEC CD 42105
45	ISO/IEC JTC 1/SC 42/WG 3	信息技术 人工智能 AI 系统质量特性差异化基准测试概述	草案 ISO/IEC CD TR 42106
46	ISO/IEC JTC 1/SC 42/WG 3	信息技术 人工智能 AI 系统透明度分类法	批准 ISO/IEC FDIS 12792
47	ISO/IEC JTC 1/SC 42/WG 3	信息技术 人工智能 ML 模型和 AI 系统可解释性与可理解性的目标和方法	发布 ISO/IEC TS 6254
48	ISO/IEC JTC 1/SC 42/WG 3	AI 系统领域工程的术语和概念	起草 ISO/IEC AWI TS 25566

汽标委智能网联汽车分委会

49	ISO/IEC JTC 1/SC 42/WG 3	信息技术 人工智能 生成式 AI 系统风险应对指南	起草 ISO/IEC AWI TS 25568
50	ISO/IEC JTC 1/SC 42/WG 3	信息技术 人工智能 AI 系统可靠性评估	起草 ISO/IEC AWI TS 25570
51	ISO/IEC JTC 1/SC 42/WG 3	人工智能 AI 系统伦理问题记录示例模板	起草 ISO/IEC AWI TS 25571
52	ISO/IEC JTC 1/SC 42/WG 3	人工智能 AI 系统监督	起草 ISO/IEC PWI 18966
53	ISO/IEC JTC 1/SC 42/WG 3	人工智能（AI） 神经网络鲁棒性评估 第 5 部分 方法对其他人工智能算法的适用性	起草 ISO/IEC PWI 24029-5
54	ISO/IEC JTC 1/SC 42/WG 3	人工智能 AI 系统操作设计域（ODD）	起草 ISO/IEC PWI 42108
55	ISO/IEC JTC 1/SC 42/WG 3	信息技术 人工智能 AI 系统可信度事实标签	起草 ISO/IEC PWI 42117
56	ISO/IEC JTC 1/SC 42/WG 3	信息技术 人工智能 AI 系统可靠性	起草 ISO/IEC PWI 42118
57	ISO/IEC JTC 1/SC 42/WG 3	信息技术 人工智能（AI） AI 系统和 AI 辅助决策中的偏差	发布 ISO/IEC TR 24027:2021
58	ISO/IEC JTC 1/SC 42/WG 3	信息技术 人工智能 人工智能可信度概述	发布 ISO/IEC TR 24028:2020

汽标委智能网联汽车分委会

59	ISO/IEC JTC 1/SC 42/WG 3	人工智能（AI） 神经网络鲁棒性评估 第 1 部分：概述	发布 ISO/IEC TR 24029-1:2021
60	ISO/IEC JTC 1/SC 42/WG 3	信息技术 人工智能 伦理和社会关注概述	发布 ISO/IEC TR 24368:2022
61	ISO/IEC JTC 1/SC 42/WG 3	人工智能 功能安全和 AI 系统	发布 ISO/IEC TR 5469:2024
62	ISO/IEC JTC 1/SC 42/WG 3	信息技术 人工智能 分类和回归机器学习任务中不必要偏差的处理	发布 ISO/IEC TS 12791:2024
63	ISO/IEC JTC 1/SC 42/WG 3	信息技术 人工智能 自动化人工智能系统的可控性	发布 ISO/IEC TS 8200:2024
64	ISO/IEC JTC 1/SC 42/WG 4	信息技术 人工智能 AI 系统生命周期过程	发布 ISO/IEC 5338:2023
65	ISO/IEC JTC 1/SC 42/WG 4	信息技术 人工智能 AI 应用指南	发布 ISO/IEC 5339:2024
66	ISO/IEC JTC 1/SC 42/WG 4	信息技术 人工智能 人机协作的用例	起草 ISO/IEC AWI TR 42109
67	ISO/IEC JTC 1/SC 42/WG 4	信息技术 人工智能 有益的 AI 系统	批准 ISO/IEC DTR 21221
68	ISO/IEC JTC 1/SC 42/WG 4	信息技术 人工智能 AI 系统的环境可持续性方面	发布 ISO/IEC TR 20226

汽标委智能网联汽车分委会

69	ISO/IEC JTC 1/SC 42/WG 4	信息技术 人工智能 人机协作框架	起草 ISO/IEC AWI 25589
70	ISO/IEC JTC 1/SC 42/WG 4	信息技术 人工智能 AI 用例和应用的评估指标	起草 ISO/IEC PWI 42113
71	ISO/IEC JTC 1/SC 42/WG 4	信息技术 人工智能（AI） 用例	发布 ISO/IEC TR 24030:2024
72	ISO/IEC JTC 1/SC 42/WG 5	信息技术 人工智能 AI 系统的混合 AI 推理框架	起草 ISO/IEC AWI TS 25258
73	ISO/IEC JTC 1/SC 42/WG 5	信息技术 人工智能 轻量级 AI 系统指南	起草 ISO/IEC AWI TS 42111
74	ISO/IEC JTC 1/SC 42/WG 5	信息技术 人工智能 机器学习模型训练效率优化指南	起草 ISO/IEC AWI TS 42112
75	ISO/IEC JTC 1/SC 42/WG 5	信息技术 人工智能 AI 轻量级建模	起草 ISO/IEC PWI 42107
76	ISO/IEC JTC 1/SC 42/WG 5	信息技术 人工智能 机器学习计算设备概述	发布 ISO/IEC TR 17903:2024
77	ISO/IEC JTC 1/SC 42/WG 5	信息技术 人工智能（AI） AI 系统计算方法概述	发布 ISO/IEC TR 24372:2021
78	ISO/IEC JTC 1/SC 42/WG 5	信息技术 人工智能 机器学习分类性能评估	发布 ISO/IEC TS 4213:2022

汽标委智能网联汽车分委会

79	IEEE C/DC AIM	人工智能模型表示、压缩、分发与管理标准	发布 IEEE 2941-2021
80	IEEE C/AISC Deep Learning	深度学习评估框架与过程推荐实践	发布 IEEE 2841-2022
81	IEEE VT/ITS AV Decision Making	安全相关自动化车辆行为模型假设草案标准	发布 IEEE 2846-2022
82	IEEE VT/ITS AV Decision Making	安全相关模型中其他道路用户合理可预见行为假设应用推荐实践	发布 IEEE 3321-2024
83	IEEE RS/SC Fail-Safe Design	自主与半自主系统故障安全设计标准	发布 IEEE 7009-2024
84	IEEE VT/ITS 2020_WG_AUTOIQ	汽车系统图像质量标准	发布 IEEE 2020-2024
85	IEEE C/AISC XAI	可解释人工智能架构框架指南	发布 IEEE 2894-2024
86	IEEE C/AISC AIFE-WG	机器学习公平性评估方法标准	发布 IEEE 3198-2025
87	IEEE C/AISC AIML-TDF	人工智能与机器学习术语及数据格式标准	草案 IEEE P3123
88	IEEE C/AISC VTCV	计算机视觉应用中机器学习模型漏洞测试推荐实践	草案 IEEE P3157

汽标委智能网联汽车分委会

89	IEEE C/DC DFAD	自动驾驶数据框架标准	草案 IEEE P3184
90	IEEE C/AISC AI Governance	人工智能组织治理推荐实践	草案 IEEE P2863
91	IEEE C/AISC GP-AIM	生成式预训练人工智能模型安全性与可信度要求标准	草案 IEEE P7018
92	IEEE C/AISC AI-RSTR	人工智能风险、安全性、可信度及责任的定义与评估推荐实践	草案 IEEE P3396
93	IEEE CIS/SC XAI WG	可解释人工智能标准	草案 IEEE P2976
94	SAE	地面车辆应用的人工智能数据	发布 J3298_202407
95	SAE	地面车辆应用人工智能用例指南	发布 J3312_202502
96	SAE	驾驶自动化系统命名方法	发布 J3265_202211
97	SAE	道路机动车辆驾驶自动化系统术语分类与定义	发布 J3016_202104
98	SAE	协同驾驶自动化（CDA）功能：遮挡行人防撞感知状态共享	发布 J3251_202308

汽标委智能网联汽车分委会

99	SAE	道路机动车辆协同驾驶自动化术语分类与定义	J3216_202504
100	SAE	协同驾驶自动化（CDA）功能：基于基础设施的规范性协同汇入	发布 J3256_202403
101	SAE	自动驾驶人因工程定义及相关研究主题	发布 J3114_201612
102	SAE	自动驾驶系统测试场地安全规程	发布 J3247_202403
103	SAE	可互操作协同驾驶用例测试框架与测试规程开发流程	发布 J3252_202308
104	SAE	原型自动驾驶系统（ADS）车辆道路测试安全指南	发布 J3018_202012
105	SAE	自动驾驶系统数据记录器规范	发布 J3197_202107
106	SAE	面向协同驾驶与自动驾驶的 V2X 传感器共享技术	发布 J3224_202208
107	SAE	自动驾驶系统（ADS）标志灯技术规范	发布 J3134_201905
108	SAE	Automated Driving System Test Facility Safety Practices	发布 J3247_202403

汽标委智能网联汽车分委会

109	SAE	自动驾驶系统参考架构术语定义	发布 J3131_202203
110	SAE	自动驾驶系统（ADS）安全原则分类与定义	发布 J3206_202107
111	SAE	自动驾驶专属车辆（ADS-DV）残障乘客问题识别指南	发布 J3171_201911
112	SAE	常规工况下自动驾驶系统（ADS）车辆行为与机动操作的本体论及术语集	发布 J3164_202301
113	ETSI	人工智能安全 协作人工智能	已发布 ETSI TR 104 031
114	ETSI	人工智能安全 有效网络防御的关键安全控制：人工智能领域	已发布 ETSI TR 104 030
115	ETSI	许可分布式账本 许可分布式账本的人工智能	已发布 ETSI GR PDL 032
116	ETSI	体验式网络智能 模块化系统的人工智能机制介绍	已发布 ETSI GR ENI 018
117	ETSI	人工智能安全 人工智能/机器学习的隐私方面	已发布 ETSI TR 104 225
118	ETSI	人工智能安全 人工智能处理的可解释性和透明度	已发布 ETSI TS 104 224

汽标委智能网联汽车分委会

119	ETSI	人工智能安全 AI 模型和系统的基线网络安全要求	已发布 ETSI TS 104 223
120	ETSI	人工智能安全 缓解策略报告	已发布 ETSI TR 104 222
121	ETSI	人工智能安全 问题声明	已发布 ETSI TR 104 221
122	ETSI	人工智能安全 人工智能模型和系统的网络安全指南	已发布 ETSI TR 104 128
123	ETSI	人工智能安全 概念验证框架	已发布 ETSI TR 104 067
124	ETSI	人工智能安全 AI 法案与 ETSI 工作计划的映射和差距分析	已发布 ETSI TR 104 065
125	ETSI	人工智能安全 多媒体身份表示的自动化操作	已发布 ETSI TR 104 062
126	ETSI	多媒体身份表示的自动操作	已发布 ETSI TS 104 050
127	ETSI	人工智能安全 数据供应链安全	已发布 ETSI TR 104 048
128	ETSI	人工智能安全 AI 模型的可追溯性	已发布 ETSI TR 104 032
129	ETSI	智能机器间通信(Smart M2M) 人工智能与 oneM2M 架构	已发布 ETSI TR 103 674

汽标委智能网联汽车分委会

130	ETSI	5G 人工智能/机器学习（AI/ML）管理研究	已发布 ETSI TR 128 908
131	ETSI	5G 管理和编排：人工智能/机器学习（AI/ML）管理	已发布 ETSI TS 128 105
132	ETSI	核心网络和互操作性测试（INT） 测试系统中的人工智能（AI）和 AI 模型的测试 人工智能技术在测试中的应用和优势	已发布 ETSI TR 103 748

汽标委智能网联汽车分委会

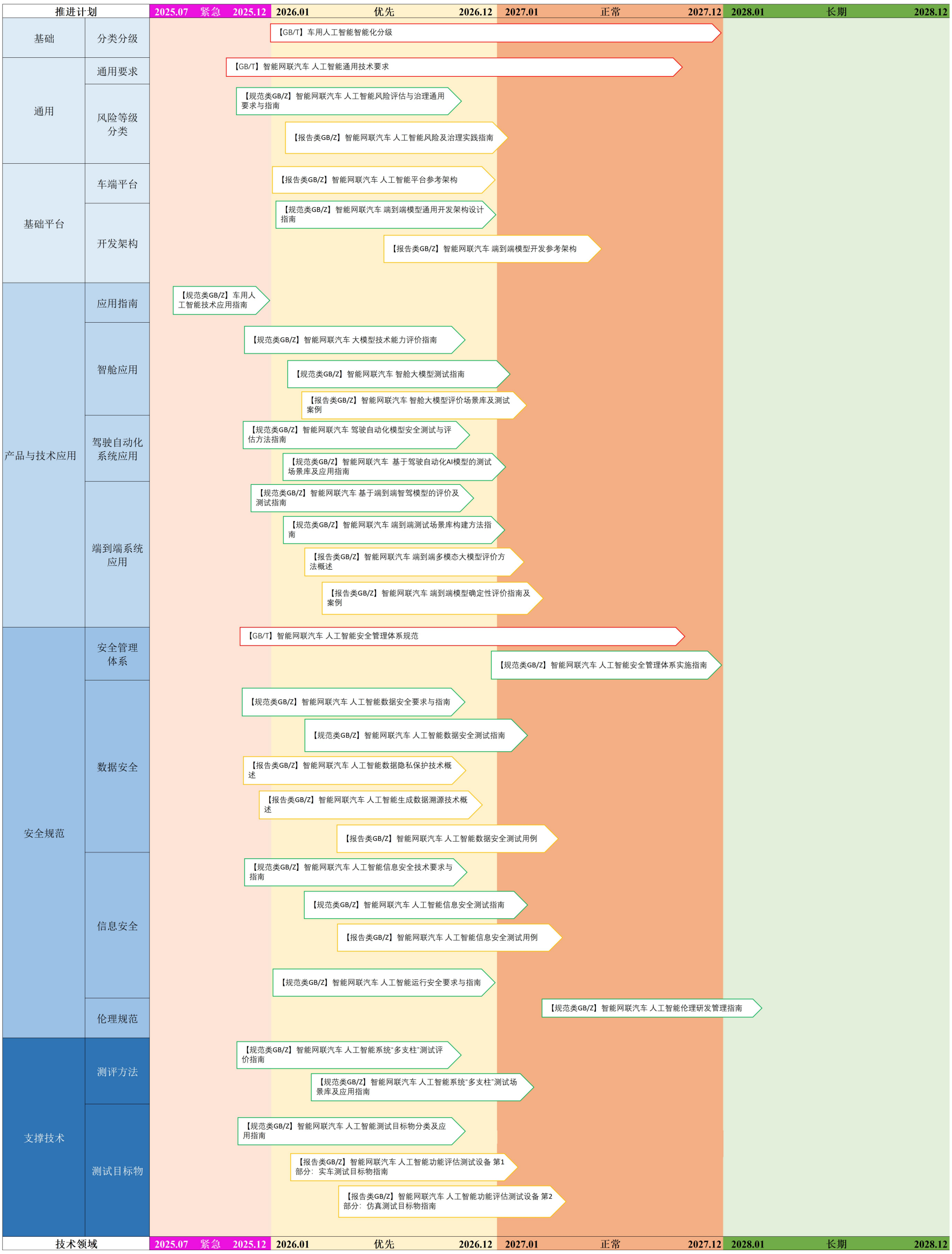
附录 C：车用 AI 领域预计标准研制规划明细表

序号	标准研制方向	目前进度
1	车用人工智能智能化分级	待启动
2	智能网联汽车 人工智能通用技术要求	草案撰写
3	智能网联汽车 人工智能风险评估与治理通用要求与指南	待启动
4	智能网联汽车 人工智能风险及治理实践指南	待启动
5	智能网联汽车 人工智能平台参考架构	待启动
6	智能网联汽车 端到端模型通用开发架构设计指南	待启动
7	智能网联汽车 端到端模型开发参考架构	待启动
8	车用人工智能技术应用指南	立项评审

9	智能网联汽车 大模型技术能力评价指南	待启动
10	智能网联汽车 智舱大模型测试指南	待启动
11	智能网联汽车 智舱大模型评价场景库及测试案例	待启动
12	智能网联汽车 驾驶自动化模型安全测试与评估方法指南	待启动
13	智能网联汽车　基于驾驶自动化 AI 模型的测试场景库及应用指南	待启动
14	智能网联汽车 基于端到端智驾模型的评价及测试指南	待启动
15	智能网联汽车 端到端测试场景库构建方法指南	待启动
16	智能网联汽车 端到端多模态大模型评价方法概述	待启动
17	智能网联汽车 端到端模型确定性评价指南及案例	待启动
18	智能网联汽车 人工智能安全管理体系规范	待启动
19	智能网联汽车 人工智能安全管理体系实施指南	待启动

20	智能网联汽车 人工智能数据安全要求与指南	待启动
21	智能网联汽车 人工智能数据安全测试指南	待启动
22	智能网联汽车 人工智能数据隐私保护技术概述	待启动
23	智能网联汽车 人工智能生成数据溯源技术概述	待启动
24	智能网联汽车 人工智能数据安全测试用例	待启动
25	智能网联汽车 人工智能信息安全技术要求与指南	待启动
26	智能网联汽车 人工智能信息安全测试指南	待启动
27	智能网联汽车 人工智能信息安全测试用例	待启动
28	智能网联汽车 人工智能运行安全要求与指南	待启动
29	智能网联汽车 人工智能伦理研发管理指南	待启动
30	智能网联汽车 人工智能系统“多支柱”测试评价指南	待启动

31	智能网联汽车 人工智能系统“多支柱”测试场景库及应用指南	待启动
32	智能网联汽车 人工智能测试目标物分类及应用指南	待启动
33	智能网联汽车 人工智能功能评估测试设备 第 1 部分：实车测试目标物指南	待启动
34	智能网联汽车 人工智能功能评估测试设备 第 2 部分：仿真测试目标物指南	待启动

汽标委智能网联汽

附录 D：车用 AI 领域标准研制路线图