



中文核心期刊 中国科技核心期刊 中国核心学术期刊
第三届国家期刊奖百种重点期刊 信息通信领域产学研合作特色期刊

ISSN 1009-6868
CN 34-1228/TN

中兴通讯技术

ZTE TECHNOLOGY JOURNAL

<http://tech.zte.com.cn>

第 31 卷 · 总第 184 期 · 2025 年 8 月 · 第 4 期

专题：面向 6G 的高时效智能机器通信技术



创刊 30 周年纪念特刊

ISSN 1009-6868



《中兴通讯技术》第10届编辑委员会

顾问 侯为贵(中兴通讯股份有限公司创始人) 钟义信(北京邮电大学教授)
糜正琨(南京邮电大学教授) 李自学(中兴通讯股份有限公司前董事长)

主任 陆建华(中国科学院院士)

副主任 方 榕(中兴通讯股份有限公司董事长) 李建东(西安电子科技大学教授)

编委

陈建平	上海交通大学教授	唐雄燕	中国联通研究院副院长
陈前斌	重庆邮电大学教授、副校长	陶小峰	北京邮电大学教授
段晓东	中国移动研究院副院长	汪烈军	新疆大学教授、副校长
方 榕	中兴通讯股份有限公司董事长	王 翔	中兴通讯股份有限公司高级副总裁
葛建华	西安电子科技大学教授	王文博	北京邮电大学教授
管海兵	上海交通大学教授、副校长	王文东	北京邮电大学教授
郭 庆	哈尔滨工业大学教授	王喜瑜	中兴通讯股份有限公司执行副总裁
洪 伟	东南大学教授	王耀南	中国工程院院士、湖南大学教授
江 涛	华中科技大学教授	王志勤	中国信息通信研究院副院长
蒋林涛	中国信息通信研究院科技委主任	卫 国	中国科学技术大学教授
金 石	东南大学教授、副校长	邬贺铨	中国工程院院士
李尔平	浙江大学教授	吴春明	浙江大学教授
李红滨	北京大学教授	向际鹰	中兴通讯股份有限公司首席科学家
李厚强	中国科学技术大学教授	肖 甫	南京邮电大学教授、副校长
李建东	西安电子科技大学教授	解冲锋	中国电信新一代信息通信专业首席专家
李乐民	中国工程院院士、电子科技大学教授	徐安士	北京大学教授
李融林	华南理工大学教授	徐子阳	中兴通讯股份有限公司总裁
林晓东	中兴通讯股份有限公司副总裁	续合元	中国信息通信研究院首席专家
刘 健	中兴通讯股份有限公司高级副总裁	薛向阳	复旦大学教授
刘建伟	北京航空航天大学教授	杨义先	北京邮电大学教授
隆克平	北京科技大学教授	易芝玲	中国移动研究院首席科学家
卢光跃	西安邮电大学教授、校长	张 杰	北京邮电大学教授
陆建华	中国科学院院士、清华大学教授	张 平	中国工程院院士、北京邮电大学教授
马建国	中原工学院教授、学术副校长	张 卫	复旦大学教授
毛军发	中国科学院院士、深圳大学校长	张宏科	中国工程院院士、北京交通大学教授
孟洛明	北京邮电大学教授	张钦宇	哈尔滨工业大学(深圳)教授、副校长
尼玛扎西	中国工程院院士、西藏大学教授	张云勇	中国联通云南分公司总经理
石光明	鹏城实验室副主任	赵慧玲	工业和信息化部信息通信科技委常委
史振威	内蒙古大学教授	郑纬民	中国工程院院士、清华大学教授
孙知信	南京邮电大学教授	钟章队	北京交通大学教授
谈振辉	北京交通大学教授	周 亮	南京邮电大学教授、副校长
唐 宏	中国电信 IP 领域首席专家	朱近康	中国科学技术大学教授
唐万斌	电子科技大学教授	祝宁华	中国科学院院士、南开大学教授

砥砺前行三十载，共绘未来新篇章

——祝贺《中兴通讯技术》创刊三十周年



◎ 文 / 方榕（中兴通讯股份有限公司董事长）

值此《中兴通讯技术》杂志创刊三十周年之际，我谨代表中兴通讯股份有限公司，向长期以来关心和支持中兴通讯发展的各界朋友致以诚挚的谢意！向杂志编委、青委及全体工作人员致以衷心的感谢与祝贺！

三十年风雨兼程，三十年春华秋实。《中兴通讯技术》杂志作为公司与行业交流的纽带与桥梁，见证了中兴通讯从初创到壮大的辉煌历程，也记录了通信技术日新月异的变迁与发展。今天，站在新的起点上，我们满怀信心，继续秉承创新精神，深化产学研合作，共同开创更加美好的未来。

回望过往，拼搏创新铸就基石

中兴通讯的诞生和发展，离不开市场的需求和技术的驱动。1985年，中兴通讯在深圳这片改革的热土上破土而出，从一家名不见经传的小厂，逐步成长为全球信息通信行业的领军企业。从2G到5G，每一次跨越都凝聚着中兴人的智慧与汗水。在2G、3G阶段，我们打破海外技术垄断，通过TD-SCDMA、小灵通等技术布局，推动国产移动通信标准形成，助力运营商网络建设；4G时代，

我们提出Pre5G概念，将5G关键技术提前应用于4G网络，使运营商在5G商用前实现网络性能大幅提升；5G时代，我们深度参与标准制定与核心技术攻关，率先完成5G基站与动态智能超表面协同的波束赋形技术验证，提升网络覆盖与能效，在5G-A领域实现三载波聚合外场测试，峰值速率达5.417 Gbit/s，为6G演进提供技术储备。如今，中兴通讯的足迹已覆盖全球160多个国家和地区，我们的产品和服务在全球范围内发挥着连接世界、赋能未来的重要作用。

2018年，中兴通讯曾遭遇严峻挑战。那段艰难岁月让我们更加坚定“向下扎根”的决心。通过五年累计超千亿元的研发投入、数万名工程师的日夜奋战，最终换来了5G基站全球市场份额第二、核心自研芯片性能领先、GoldenDB数据库支撑超10万亿元金融交易的亮眼答卷。这些成就不仅属于中兴通讯，更属于每一位与中国信息通信事业同行的奋斗者。

中兴通讯的成长，离不开对创新理念的坚守，也离不开产学研合作的智慧驱动。我们与清华大学、北京大学等数十所高校开展项目合作，在操作系统、芯片设计、算力架构等领域攻克关键技术；与行业伙伴共同制定5G-A国际标准，推动通感一体、空天地融合等前沿技术落地；更通过《中兴通讯技术》杂志这一平台，汇聚编委、青委等专家学者的智慧，发表高质量成果论文，为行业输送创新思想。这些合作不仅加速了技术产业化进程，更培养出一批兼具学术素养与实践能力的复合型人才。



立足当下，守正出新再创佳绩

今日之中兴，正以“连接+算力”双引擎驱动数智化转型。站在新的历史起点上，面对 5G 建设成熟期与 AI 技术加速融合的挑战，我们通过战略转型，实现了网络、算力、终端和数字能源四大业务领域的协同增长。2024 年，我们政企业务营收同比增长 36.68%，消费者业务增长 16.12%，AI 驱动的智慧业务成为第二增长曲线，占整体营收的 42%。我们发布全栈智算解决方案，包括服务器、存储、数据中心等产品，并与国产开源大模型 DeepSeek 合作推出 AiCube 训推一体机，加速企业 AI 应用落地；数字星云 3.0 平台已服务超 1 000 家合作伙伴，在工业制造、电力、交通等 18 个行业打造智能化标杆项目，助力“人工智能+”价值变现。

2025 年，我们聚焦“稳中求进、守正出新”，加速向“连接+算力”拓展，强化智算生态合作，推动政企领域车路云一体化、大应急等新兴市场布局。我们将继续加大研发投入，重点投向 5G-A、AI 大模型、芯片等领域，以技术创新引领未来发展。在技术路径上，我们将深化 AI 与通信技术的融合，推动基于 AI 驱动的全新端到端网络解决方案（AIR DNA）网络架构的重构，让网络从云原生迈向 AI 原生，实现意图驱动的高阶自治网络；同时，通过“零碳”能源网解决方案，为全球更多 5G 基站注入绿色动能，以科技创新践行碳减排目标。

展望未来，信心满怀共绘新篇

面向未来，中兴通讯将继续秉承“创新引领未来”的理念，深化“全连接+算力”战略转型，加速向“极致 AI 公司”的目标迈进。我们深知，算力普惠是 AI 大规模商用的前提。当前，算力发展受香农定律、冯·诺依曼架构及摩尔定律极限的限制，正面临关键瓶颈。为此，中兴通讯正通过多元路径推动算力普惠：与行业伙伴携手，以算法创新大幅降低训练和推理成本，提升算力使

用效率；依托网络技术创新，实现芯片、服务器、数据中心之间的高速互联，打造超级节点与算力集群，构建“算网一体”的智能底座；推出智算一体机系列产品，直击 AI 商用“最后一公里”难题，让算力资源真正触达千行百业。

我们倡导开放解耦理念，积极共建开源社区，推动建设开放、普惠的算力新生态，让算力如水电般自由流动，赋能 AI 的包容性发展。与此同时，AI 应用正从特定场景向全场景拓展，中兴通讯将围绕“新智能、新体验、新形态”三大方向深耕：以大模型泛化能力破解行业应用碎片化问题，定义“新智能”；聚焦听觉、视觉等感官体验升级，打造沉浸式“新体验”；拥抱 AI 终端、具身智能机器人等新技术形态，探索“新形态”的无限可能。在推动算力普惠与应用普及的进程中，我们更将坚守“科技向善”理念，以技术创新破解隐私保护、数据流通等挑战，确保 AI 安全可靠，加速“连接+算力”战略转型，与广大客户及伙伴全面深化合作，共同构建 AI 商业闭环，迈向“AI 普惠时代”。

在此，我们要特别感谢《中兴通讯技术》杂志的编委团队。各位专家学者以深厚的学术造诣、敏锐的行业洞察和无私的分享精神，为杂志注入思想深度与专业价值。正是专家们的智慧凝聚，让杂志成为产学研协同创新的高地，为行业输送源源不断的创新动能。未来，我们期待与编委、青委们继续携手，以开放合作之姿，在全连接、算力、AI 开放解耦生态的新征程中，共同探索数字文明的新边界。

感谢国家政策的指引，为科技企业营造了充满活力的创新环境！感恩所有合作伙伴和客户的信任，助力中兴通讯努力成为网络连接和智能算力的领导者，让沟通与信任无处不在！

最后，再次感谢所有关心和支持中兴通讯发展的朋友们。让我们以“连接+算力”为翼，以开放生态为基，携手并进，共同开创更加美好的未来！

三十韶华，勇立潮头

——纪念《中兴通讯技术》创刊三十周年



◎ 文 / 黄新明（本刊执行主编）

1995年6月12日，《中兴新通讯》编辑部在合肥成立，零起步创办学术期刊。三十年后的今天，184期中文刊、91期英文刊整齐排列，串联起2G到6G的技术史诗，如同年轮般镌刻着信息通信产业的每一次心跳。三十年，是一段岁月征程，更是一场与时代的对话，而每五年一期的纪念特刊，既是对过往的凝望，亦是对未来的期许。它们是办刊人对岁月的深情告白，也是对初心使命的庄严承诺，并在时光的坐标上，清晰地勾勒出逐浪前行的轨迹。

今天，我们将目光投向过去五年的步履，那些沉淀的成就与变革，恰似星河闪耀，照亮前路。

2020年：筑基固本，守正创新

这一年，中国通信学会首次发布《中国通信学会国内信息通信领域科技期刊分组推荐目录》，《中兴通讯技术》被评为T2级，这份沉甸甸的认可，让我们的脚步愈发坚定。为了提升刊物品质，编辑部以“重组稿、抓编校、促传播、聚资源”为纲，开启了全面提质之路。我们深知，每一篇优质稿件都是杂志的灵魂。通过积极向名家约稿，实现了72%的知名专家撰稿率，让刊物内容更添权威与深度。面对《报纸期刊质量管理规定》的新要求，我

们制定了“3+1”审稿制度，即“三审+外审”层层把关，确保学术严谨性；我们制定了“3+1”校对制度，即“三校+全员审校”，确保每一个差错都无处藏身。这一年，中英文两刊首次实现编排格式统一，善锋软件让参考文献编校自动化，封面设计更以内文插图为灵感，体现科技之美。这一系列创新举措形成全链条质量管控体系，为刊物发展打下坚实基础。

2021年：云端聚智，凝心聚力

疫情阻隔了相聚，却阻不断智慧的交融。这一年，我们组建了新一届（第9届）编委会，107位业内知名专家构建起跨学科专家矩阵。因海外编委参与办刊工作受限，新增国内编委助力支撑起英文刊的学术穹顶。公司内13位技术专家受邀担任产业顾问，为刊物发展注入了新的智慧和力量。在兰州，我们戴着口罩开了一次“特殊”的编委会议；在云端，我们与海外专家隔屏相望，共话科技前沿。为使中英文两刊同频共振，我们统一征集专题，使刊物内容更系统、更深入；为了让读者能够更加便捷地获取刊物信息，我们优化了中英文两刊的独立网站，既拓展了平台，又延伸了刊物影响力。

2022年：以刊为桥，融创未来

把产学研项目合作与办刊工作相结合，是中兴通讯的一招鲜。优秀论文发表在期刊，优秀人才吸引到企业，优秀成果落地到产品，期刊化作了连接学界与工业界的彩虹。这一年，一期产学研成果论文增刊，让更多产学研成果得到了广泛而精准的传播。这一年，我们走进多个编委团队，走进实验室，



宣传中兴通讯的产学研合作资源和学术刊物资源，了解师生的需求和想法，增进了友谊，促进了合作。为托举人才，我们成立了中兴通讯技术杂志社促进产学研合作青年专家委员会，38位来自高校与科研院所的青年才俊，以蓬勃朝气激发产学研合作新动能。他们支撑学术会议，承担产学研项目，帮助我们打开了办刊工作的新局面。

2023年：多元赋能，传播升级

“热点专题”“专家论坛”“企业视界”等栏目特点各异，我们为其量身定制了组稿流程，并反复优化，因为我们深知，没有最好，只有更好。为了更广泛地利用好专家资源，我们在公司内聘请了一批长聘作者，请他们优先供稿，并通过引用论文，提升本刊影响力。另外，随着产学研项目的增加，产学研成果稿件数量大增，刊出论文的质量得到进一步提升。我们还通过产学研项目拓展审稿专家资源，让更多一线专家参与审稿，覆盖更多专业领域。随着三年疫情的结束，我们积极走近科研前线，携刊广泛参展参会。杂志社设计的系列文创产品与刊物交相辉映，成为展台上亮丽的风景线。我们认识到，刊物不仅仅是学术传媒，更是文化载体；我们是“传播者”，更是“生态构建者”。

2024年：荣誉加身，再攀高峰

星光不负赶路人，2024年是收获的一年。中文刊跻身“三核心”期刊：北大中文核心期刊、中国科技核心期刊、中国核心学术期刊，英文刊连续3年入选《世界期刊影响力指数（WJCI）报告》，两刊入选中国通信学会、中国计算机学会、中国电子学会期刊分级目录，两刊入选中国通信学会精品期刊方阵。这些荣誉是对我们多年努力的肯定，更是我们向更高更远目标挑战的信心所在。当刊物在展会上被争相传阅，当产学研经验被期刊主管部门点赞，我们更加笃定：服务产业、回馈社会，是我们刊物未来发展的清晰航标。

回眸三十年历程，我们不仅用6个“五年计划”交出了厚实的答卷，更是在成长的路上收获了一路的风景与感悟。

三十年前，我们以一本中文刊启程；三十年后，中英文两刊并驾齐驱，依托刊物而建立的产学研合作平台硕果累累，为加强校企合作而建立的博士后工作站英才辈出。一路走来，我们见证了中兴通讯从模拟、程控到6G领航的蜕变，陪伴了

无数专家从青涩学者到行业巨擘的成长；我们体会了从手写稿、铅字印刷到云排版、智能审校的技术跃迁。我们是历史的记录者，也是创新的推动者。这是承前启后、使命必达的复兴之旅。

三十年的编委接力，是数百位专家以智慧筑起的灯塔；编辑团队从青丝到白发的坚守，是以青春为墨写就的诗行。而中兴通讯历任领导的信任与投入，让办刊文化深深融入企业血脉——这不仅是资源的支持，更是“科技报国”的坚持。这是青春与情怀的交响。

三十年让我们深知，中国科技期刊的发展任重道远，仍需社会各界戮力同心。企业办刊不应是孤舟独行，而应成为创新链与产业链融合发展的重要引擎。我们期待更多同行者，以期刊为舟，载动产业与学术的星河。这是期刊编辑人心灵的呼唤。

三十年过去了，我们不仅收获了一路的风景与感悟，更有发自内心的感怀与感恩。

感谢中兴通讯历任领导的远见卓识，是你们的战略眼光让“科技公益”的初心始终如炬，是你们的持续支持，让发展蓝图触地生辉。

感谢合办单位与主管部门的指引，是你们的制度规范为航船校准航向，是你们的鼎力支持为征程保驾护航。

致敬历任编委与青委，你们以学术之光点亮前路，是你们的无私奉献和关爱，成就了刊物的过去和现在；是你们的智慧和才华，拓展了学术天地的无垠疆界。

感谢编辑部同仁，你们用匠心守护每一页文稿的纯粹，是你们的辛勤耕耘，让每一期刊物成为精品；是你们的执着坚守，让办刊精神薪火相传。

更要向千万读者致意——是你们的关注和认可，让文字有了温度；是你们的问卷反馈，重塑了内容经纬。你们的每一次翻阅，都是对理想的共鸣。

三十而立，是成熟，更是新生的开始。

站在6G与AI的潮头，我们将继续以刊为舟，载着企业的使命、学者的智慧、编辑的理想，驶向“万物智联”的彼岸。正如诗人所言：“既然选择了远方，便只顾风雨兼程。”

让我们继续携手，在数字旷野，捕捉技术突破的每一个灵感；在产业潮头，记录创新发展的每一步足迹；在时代长卷，书写中国科技期刊的壮丽篇章！

祝愿以三十韶华有序，谱写下一个璀璨三十年！

热烈祝贺 《中兴通讯技术》创刊 30 周年！



侯为贵

中兴通讯股份有限公司创始人

助校企合作建功
促科技创新立业

祝贺《中兴通讯技术》创刊30周年



李自学

中兴通讯股份有限公司前董事长

祝贺《中兴通讯技术》创刊三十周年

星河璀璨映华篇
术道交融汇海川



钟义信

北京邮电大学原副校长

忆往昔，难忘峥嵘岁月
看未来，喜迎壮美风光

贺《中兴通讯技术》
创刊30周年



谈振辉

北京交通大学原校长

贺《中兴通讯技术》创刊30周年

中流砥柱志高远，
兴科重技步履坚；
通研求解根基固，
讯息前沿掌中观；
腾跃卅载书锦绣，
飞歌逐梦谱新篇。


邬贺铨

中国互联网协会
专家咨询委员会主任

贺《中兴通讯技术》30周年
半甲子打造期刊精品
全身心铸就科创平台
三十载坚守通信领航
百年梦再创事业辉煌


郑纬民

清华大学教授

祝贺《中兴通讯技术》杂志创刊 30 周年

《中兴通讯技术》杂志走过波澜壮阔、日新月异的三十载。始终坚守初心，以传播信息通信前沿技术为己任，为行业搭建起一座知识的桥梁；并在促进产学研多维合作方面，如灯塔般照亮了企业、高校与科研机构之间的合作之路；它为年轻一代的科技人才提供了展示才华的舞台，激发了他们的科研热情，为行业的可持续发展奠定了坚实的人才基础。

展望未来，愿《中兴通讯技术》继续秉持初心，紧跟时代步伐，不断拓展视野，为信息通信技术的传播与发展贡献更多力量。



李乐民

电子科技大学教授

祝贺《中兴通讯技术》办刊30周年
聚焦通信与信息前沿技术
探讨市场热点
很有参考价值



毛军发

深圳大学校长

三十而立，新里金刊风采
启航未来，谱写时代新篇
祝贺《中兴通讯技术》创刊30周年


张平

北京邮电大学教授

三十载技术深耕，
而立年再攀高峰！
祝贺《中兴通讯技术》创刊 30 周年


王耀南

湖南大学教授

祝贺《中兴通讯技术》创刊三十周年：
三十载砥砺前行，
推动信息通信跨越发展。
新时代矢志创新，
赋能全球数字化转型！



尼玛扎西

西藏大学教授

祝《中兴通讯技术》杂志
学术影响力不断提升，
为推动我国通信技术发展
作出更大贡献。



高新波

西安电子科技大学校长

三十春秋，在信息通信的广袤天地里辛勤耕耘，
传播前沿科技，搭建起产学研紧密相连的桥梁，助
力科技人才茁壮成长。

看如今，刊物已如一棵枝繁叶茂的大树，扎根
产业实践的土壤，不断汲取创新的养分，在信息通
信领域绽放出璀璨的光彩。

愿未来，继续紧跟时代步伐，勇当行业先锋。
凭借创新的力量推动科技不断向前发展，为数字中
国的建设书写新的壮丽篇章！

恭祝《中兴通讯技术》蓬勃发展，不断创造新
的辉煌！



金石

东南大学副校长

中兴通讯技术深耕信通领域
三十载驰而不息，
产学研协同创新为国广育科技
精英成效斐然。



管海兵

上海交通大学副校长

三十春秋凝思想引擎，
万篇华章织创新罗盘。

谨贺《中兴通讯技术》
创刊三十周年



张钦宇

哈尔滨工业大学（深圳）副校长

欣贺《中兴通讯技术》创刊三十周年！
贵刊创刊以来始终深耕学术前沿，树立产学研用协同创新的标杆价值，助力培育新质生产力，彰显了科技期刊服务国家战略的使命担当。期待与中兴通讯深化产—研—教融合，共育信息产业领军英才！



王 伶

西北工业大学副校长

三十载春华秋实，贵刊以创新为笔、技术为墨，见证并推动了我国通信科技的崛起与跨越，为产学研深度融合搭建了智慧平台。

展望未来，愿《中兴通讯技术》杂志永立潮头，以思想之光点亮创新之路，以学术之力赋能强国之梦！



叶美兰

南京邮电大学校长

三十载务实坚守，
传播信息通信技术前沿创新动态；
新时代拼搏创新，
促进智能时代行业更高质量发展。



陈前斌

重庆邮电大学副校长

三十韶光，学研焕彰；
洞察业象，通讯腾骧。
幸逢盛典，与有荣光；
创新逐梦，再谱华章。



卢光跃

西安邮电大学校长

创新引领三十载
智联未来谱新篇



汪烈军

新疆大学副校长

群贤毕至，汇四海智慧
佳篇云集，领科技前沿
视频《中兴通讯技术》创刊30周年

**周万雷**

澳门城市大学副校长

三十载春秋共振信息脉搏，
半甲子耕耘铸就通信丰碑。

**曾文军**

宁波东方理工大学副校长

值此《中兴通信技术》杂志创刊三十周年之际，我谨以一名科技工作者与读者的身份，向贵刊致以最诚挚的祝贺！三十载春华秋实，贵刊始终秉持“依托专家，服务社会”的办刊理念，以严谨的学术态度、前瞻的技术视野，架起产学研深度融合的桥梁，成为全球通信技术领域不可或缺的思想高地与创新引擎。

三十而立，初心如磐。愿《中兴通信技术》继续以全球视野引领技术变革，以学术深度赋能产业升级，书写更加辉煌的篇章！

中兴通讯成立 40th 周年大事记

艰苦创业

1985—1992 年

- 1985 年 2 月 7 日，深圳市中兴半导体有限公司成立。
- 1986 年 中兴进军通信领域，ZX-60 研发成功，于次年 7 月通过技术鉴定并取得进网许可证，成为中国拥有自主知识产权的第一台程控交换机。
- 1989 年 ZX-500 研发成功，成为中国拥有自主知识产权的第一台数字程控用户交换机。
- 1991 年 12 月，适应中国农话 C5 局端数字改造的 500 门数字局用交换机 ZX500（A）研制成功，并在江苏吴江桃源镇成功开局，开创了国产交换机进入农村市场的先河，在全国掀起了一场农话改革高潮。自此，中兴真正走上了电信设备制造商的道路。

逐步壮大

1993—1996 年

- 1993 年
 - 3 月，ZXJ2000 数字局用交换机研发成功，同年 11 月，通过鉴定并获得邮电部国产数字局用交换机首批入网许可证。12 月，中兴局用数字交换机装机量占到全国农话新增容量 18%（含进口机型），居国内首位。
 - 3 月，航天系统 691 厂、深圳广宇工业（集团）公司与中兴维先通共同投资组建深圳中兴新通讯设备有限公司，首创“国有民营”经营机制。
 - 6 月，创始人侯为贵先生提出了中兴核心价值观——互相尊重，忠于中兴事业；精诚服务，凝聚顾客身上；拼搏创新，集成中兴名牌；科学管理，提高企业效益。
 - 10 月，中兴成立南京研究所，着手研制大容量局用数字程控交换机。次年，成立上海研究所。之后的数年间，相继在北京、西安、成都、重庆、天津、三亚、武汉、长沙及欧美等地设立 15 家研究所，为多元化发展奠定基础。
- 1995 年
 - 3 月，中兴南京研究所开发的 ZXJ10 大容量局用数字程控交换机在江苏射阳镇开通实验局。11 月，ZXJ10 获邮电部颁发的人网许可证，并在专家评审中被认定为国内“目前能与国际一流机型相媲美的最好机型”。
 - 启动国际化战略，史立荣只身参加日内瓦 ITU 世界电信展，中兴首次在国际通信市场亮相。
- 1996 年
 - 在 1996 年第一次市场工作会议上，中兴提出战略上的“三大拓展”：产品结构从单一的交换机设备向多元化产品拓展，目标市场从农话市场向本地网、市话网拓展，由国内市场向国际市场拓展。3 月，成立国际业务部。
 - 11 月，中兴与湖南岳阳签订 ZXJ10 市话局合同，该项目与随后进行的益阳接入网项目、衡阳交换机项目统称“三阳工程”，成为中兴从农话市场向市话市场拓展的标志性工程。
 - 中兴成立 IC 设计部，潜心自主研发核心通信芯片。



高速发展

1997—2003 年

- ▶ **1997 年**
 - 11 月，深圳市中兴通讯股份有限公司成立，并在深圳证券交易所成功上市，成为中国首家大型通信设备上市公司。
 - 中兴通讯参与孟加拉电信公司投标项目，获得 150 万美元的合同，这是中兴通讯第一个正式的国际中标项目。次年 9 月，在孟加拉开通海外第一个交换局。
- ▶ **1998 年**
 - 10 月 17 日，中兴通讯获巴基斯坦全国有线电话网络 9 500 万美元“交钥匙”项目。该项目是当时中国通信企业通过国际竞标在海外获得的最大规模通信工程总承包项目。
 - 12 月，中兴通讯调整组织架构，开始实行准事业部制，设立交换产品事业部、接入产品事业部、移动产品事业部和本部产品事业部 4 个产品事业部，负责产品研发、开发、生产；设立第一营销事业部、第二营销事业部、第三营销事业部 3 个营销事业部，负责销售、维护业务。
- ▶ **1999 年**
 - 1 月，中兴通讯在安徽六安与中国联通签订了第一个 GSM 实验局，并在河北保定开通 1800M 实验局，正式进军无线领域。1999 年 6 月，获得 GSM 入网证。
 - 6 月，中兴通讯成功研发国内首款自主知识产权的全中文双频 GSM 手机 ZTE 189。
- ▶ **2000 年**
 - 中兴通讯综合接入网大规模商用，国内市场占有率超过 40%。
- ▶ **2001 年**
 - 中兴通讯签订尼日尔首个 GSM 海外规模商用合同。随后，在利比亚、埃塞、尼日利亚、印度、斯里兰卡等获得突破，GSM 走向世界。
 - 3 月，中兴通讯发布全球首款机卡分离式 CDMA 手机 ZTE802，突破中国 CDMA 发展的技术瓶颈。
 - 5 月，中兴通讯 CDMA 移动通信系统中标中国联通 110 万线合同，获得 7.5% 的市场份额，正式进入大规模商用阶段。
 - 中兴通讯在巴基斯坦签署首个 CDMA 商用合同，之后产品成功进入亚非欧 60 多个国家的 100 余家运营商网络。
 - 中兴通讯在广西北海和青海西宁开通 PHS（小灵通）实验局。
- ▶ **2002 年**
 - 6 月，中兴通讯 DWDM 产品成功中标中国移动集团省际西部环传输系统（一期）工程。该工程网络覆盖共 18 个省市，总长超 2 万多公里，是当时全球最大的 DWDM 项目，也是当时中国移动唯一由国内厂商承建的 DWDM 系统。
 - 中兴通讯“三大拓展”战略成效初显，营收突破百亿。移动通信领域，PCS、小灵通约占三分之一市场份额；光传输领域跻身国内主流厂商行列，承建了国内外国家干线工程；数据通信领域稳定增长，ADSL 产品市场份额超过 20%。
- ▶ **2003 年**
 - GSM 基站单载频分集接收机获中兴通讯首个中国专利金奖，实现广东省企业在中国专利金奖上零的突破。



昂首全球

2004—2015 年

- 2004 年 · 中兴通讯创新推出全 IP 统一平台，支持 GSM、UMTS、软交换和 IMS 产品，为通信网络融合发展奠定基础。
· 中兴通讯 CDMA 手机在国内市场份额第一，同时在巴西获中国手机出口第一大订单，金额逾 1 亿美元。
· 12 月，中兴通讯在香港联交所主板上市，成为中国首家在香港 H 股上市的 A 股公司，被视为中国证券业的里程碑和 2004 年度中国经济标志性事件。
- 2005 年 · 中兴通讯承建中国电信软交换长途骨干网工程（DC1），标志着全球覆盖节点最多、用户规模最大的固定长途骨干网诞生。
· 中兴通讯于 2005 年打通第一个 TD 电话，并于 2006 年开通业界首例 BBU + RRU 射频拉远站点，彻底解决 TD 天馈工程复杂及成本问题，成为 TD 产业的里程碑事件。
· 12 月 5 日，中兴通讯与法电签署长期研发战略合作协议，成为公司在欧洲市场加强与跨国运营商深度合作的重要里程碑。
- 2006 年 · 中兴通讯 CDMA 等产品在印度 Reliance 和 TATA 等运营商大规模商用，全面突破印度高端市场，使印度成为公司海外收入规模最大的单一国家市场。
· 中兴通讯与埃塞电信签署价值 15 亿美元的合同，建设首张全产品全国网，成为中兴通讯出海的标志性项目。
- 2007 年 · 中兴通讯在中国移动 TD 一期招标项目中取得 51% 的份额，树立 TD 第一品牌形象。参与建设中国第一个 TD-SCDMA（3G）基站于 2004 年入选 100 项“共和国印记”见证物。
· 中兴通讯与高通公司、Aircell 合作 3G ATG，于 2008 年 11 月正式推出业界首个空中移动宽带系统网络平台，建设了全球最早、规模最大的 3G ATG 网络，覆盖美国全境以及加拿大南部。
- 2008 年 · 年初，中兴通讯在业界率先推出 SDR 系列商用基站，成为颠覆式创新的代表，引发了行业跟随。
· 中兴通讯与 CSL 合作在香港承建全球首个基于 SDR 技术的全 IP HSPA+ 网络，是当时全球最快 3G 网络。
- 2009 年 · 年初，中兴通讯推出业界第一款面向下一代 PON 演进的光接入平台 ZXA10 C300。
· 9 月，中兴通讯发布自主研发的明星产品 ZXR10 T8000，是国内自主知识产权程度最高的集群路由器，三大核心芯片均是自主研发。
· 中兴通讯获得德国 E-Plus 无线通信设备订单，无线产品首次突破欧洲高端市场。
- 2010 年 · 中兴通讯率先通过工业和信息化部完整集测试，发布 TD-LTE 商用基站，引领 TD-LTE 产业化进程。
· 中兴通讯为奥地利 H3G 重建 3G 网络，成功实现两网融合，助力客户先后 7 次获得欧洲德语区 / 奥地利最佳质量网络、最佳用户体验网络大奖。
· 中兴通讯 CDMA 全球市场份额达 30%，居全球首位。
- 2011 年 · 中兴通讯独家中标建设 Telenor 旗下马来西亚运营商 Digi 的 2G/3G/4G 统一无线网络，彰显了中兴通讯在复杂大规模网络建设中的卓越实力。
- 2012 年 · 11 月和 2014 年 3 月，中兴通讯先后中标西班牙 Jazztel FTTH（一、二期）项目，项目金额累计超 7.5 亿欧元，是中兴通讯在海外建设的最大规模 FTTH 项目，也是当时欧洲业界最大规模的 FTTH 建设工程，为客户创造了巨大的商业价值。
- 2013 年 · 中兴通讯成功自研 28 nm 工艺 LTE 多模基带芯片。同年，正式推出业界首款小型化、高功率双发四收系列 RRU 产品。
- 2015 年 · 中兴通讯终端在美国整体市场份额跻身前四。
· 中兴通讯发布 Pre5G 预商用基站。
· 中兴通讯营业收入首次突破千亿元。全球无线市场的格局全面提升，4G 发货量增长超过 25%。
· 中兴微电子发展迅猛，业务规模跻身国内行业前三。

凤凰浴火

2016 年至今

- ▶ **2016 年** · 中兴通讯自研的“面向工业智能装备的电信级实时操作系统”与中国探月工程、国产航母等 9 个国家重量级项目同期荣获“中国工业大奖”。同年，该操作系统全球累计发货量突破 2 亿套，成熟商用用于 160 个国家和地区。
 - 在世界移动通信大会（MWC）上，中兴通讯 Pre5G Massive MIMO 荣获 MWC 双料大奖——“最佳移动技术突破奖”及“CTO 选择奖”，并在日本软银实现规模商用。
 - 中兴通讯独家中标意大利 Wind Tre 合并项目，项目金额超过 10 亿欧元。
 - 中兴通讯“第四代移动通信系统（TD-LTE）关键技术与应用”项目荣获 2016 年度国家科技进步奖特等奖。
 - 中兴通讯独家承建孟加拉国家数据中心，2017 年通过 Uptime Tier IV 最高等级设计认证，成为“一带一路”经济合作的标杆。
 - 中兴通讯启动数字化转型，实现与近百家客户在线交易，累计支撑上千亿业绩的达成；上线 iCenter 工具，服务全球 130 多个国家的内外用户。
- ▶ **2017 年** · 2 月，中兴通讯将 5G 作为核心发展战略，品牌标识语确定为“5G 先锋”。同年，发布 5G 全系列预商用产品及解决方案，强化 5G 先锋行业领导者地位。
 - 中兴通讯与中国移动联手在广州大学城开通全国第一个 5G 基站。
- ▶ **2018 年** · 10 月，中兴通讯确定新的愿景和使命，坚持“聚焦运营商，提升客户满意度，坚持技术领先，坚定提高市场占有率，加强人才、合规、内控建设”的战略基本原则，制定并明确了“恢复期、发展期、超越期”的“三阶段”实施路径。
- ▶ **2019 年** · 中兴通讯 SDR 基站平台升级为全制式大容量 BBU（V9200），为业界首个支持容器技术且集成度最高的基站平台。同年，推出 5G 多模基带、多模中频及首颗数模混合芯片，是奠定 5G 格局的重要产品。
 - 中兴通讯 GoldenDB 数据库全项满分通过工业和信息化部标准测评，成为首个满分厂家，在中信银行总行投产，并成为国产分布式数据库金融行业第一品牌。
- ▶ **2020 年** · 开启国内 5G 规模建设，中兴通讯 5G 市场份额超过 30%。
 - 1 月，中兴通讯成功完成 115 亿元 A 股定向增发，为公司 5G 核心产品研发和全球市场布局拓展提供资金保障。
 - 2 月，南京滨江智能制造基地投产，首次提出“用 5G 制造 5G”。至此，公司打造了深圳、西安、长沙、河源、南京五大智能制造基地，其中南京滨江工厂 2024 年荣获国内首个“五星 5G 工厂”称号。
- ▶ **2021 年** · 2021 年起，中兴通讯无线产品在 Telenor 孟加拉、VEON 哈萨克、Orange 喀麦隆、TEF 智利等多国实现大 T 首都核心区域突破。
- ▶ **2022 年** · 中兴通讯服务器突破多家头部互联网企业成主流供应商。
- ▶ **2023 年** · 中兴通讯在 5G 射频芯片等相关领域创纪录地一年荣获 5 项国家科技进步奖（一等奖 3 项、二等奖 2 项）。
 - 中兴通讯全球首发 Real 400G 解决方案并创造网 5 616 km 的超远传输距离纪录。同年，发布全球首个 800G 可插拔和 C+L 一体化光模块，设备容量和集成度业界领先。
- ▶ **2024 年** · 中兴通讯家庭终端全球市场份额四连冠，年发货量突破 1 亿台。

中兴通讯技术杂志社办刊 30th 周年大事记

1995—2000 年

创刊立业的五年

1995 年

6 月,《中兴新通讯》编辑部成立,编委会组建
7 月,《中兴新通讯》正式创办,获省级刊号,季刊出版

1997 年

1 月,《中兴新通讯》改为双月刊

2000 年

7 月,《中兴新通讯》获正式刊号,国内外公开发行,
同时更名为《中兴通讯技术》

2006—2010 年

国际化发展的五年

2006 年

6 月,《ZTE Communications》被中国期刊全文数据库(CJFD)全文收录

2008 年

7 月,《ZTE Communications》被美国《乌利希期刊指南》收录

2009 年

3 月,《ZTE Communications》被美国《剑桥科学文摘(工程技术)》(CSA(T))收录

11 月,《中兴通讯技术》获安徽省优秀期刊奖

5 月,《ZTE Communications》被波兰《哥白尼索引》(Index Copernicus)收录

11 月,《中兴通讯技术》获第 4 届华东地区优秀期刊奖

2010 年

5 月,《ZTE Communications》被中国核心期刊(遴选)数据库收录

创刊

成长

发展

2001—2005 年

快速成长的五年

2000 年

10 月,《中兴通讯技术》被中文科技期刊数据库收录

2001 年

8 月,《中兴通讯技术》获安徽省优秀科技期刊二等奖

2002 年

9 月,《中兴通讯技术》获第 3 届华东地区优秀期刊奖

12 月,《中兴通讯技术》被中国核心期刊(遴选)数据库收录

2003 年

12 月,《中兴通讯技术》荣获首届《CAJ-CD 规范》
执行优秀期刊奖

2005 年

1 月,《中兴通讯技术》荣获第 3 届国家期刊奖百种
重点期刊称号

10 月,《ZTE Communications》获正式刊号,国内外
公开发行

2011—2015 年

确立特色的五年

2011 年

1 月,《ZTE Communications》被中文科技期刊数据库收录

3 月,《ZTE Communications》组建编委会,刊物开始独立组稿、独立运作

12 月,《中兴通讯技术》获工业和信息化部 2010—2011 年度优秀科技期刊奖

2012 年

7 月,杂志社建立编委任期制,并发布编委聘任制度文件

12 月,《中兴通讯技术》获 2012 年华东地区优秀期刊奖

12 月,《ZTE Communications》被英国 INSPEC 数据库收录

2013 年

2 月,中英文两刊引入 DOI 并启用数字优先出版

6 月,《ZTE Communications》被挪威 NSD 数据库收录

2015 年

2 月,明确走办产学研合作特色期刊之路

3 月,《中兴通讯技术》被评为 RCCSE(中国学术期刊评价研究中心)中
国核心学术期刊(A)

6 月,出版创刊 20 周年纪念特刊及画册

8 月,杂志社完成编委会换届,成立中文刊和英文刊第 7 届编委会



1995—2025

创刊 30 周年纪念特刊

2016—2020 年

编校质量提升的五年

2016 年

- 6 月, 杂志社独立网站 tech.zte.com.cn 上线
- 10 月, 《中兴通讯技术》获首届安徽省科技期刊编校质量优秀奖
- 12 月, 《ZTE Communications》启用 ScholarOne 稿件管理平台

2017 年

- 2 月, 《中兴通讯技术》获 2015—2016 年度安徽省优秀期刊奖
- 3 月, 《ZTE Communications》被俄罗斯《文摘杂志》(Abstract Journal) 收录

2018 年

- 8 月, 杂志社完成编委会换届, 成立中文刊和英文刊第 8 届编委会
- 9 月, 《中兴通讯技术》获十佳皖版期刊奖
- 10 月, 《ZTE Communications》被评为 2018 中国国际影响力优秀学术期刊
- 12 月, 《中兴通讯技术》和《ZTE Communications》完成主管、主办和出版单位变更

2019 年

- 2 月, 《中兴通讯技术》和《ZTE Communications》封面和文章首页添加二维码
- 5 月, 《中兴通讯技术》被日本科学技术振兴机构数据库 (JST) 收录
- 6 月, 《中兴通讯技术》入选庆祝中华人民共和国成立 70 周年精品期刊展
- 9 月, 杂志社应邀在第 15 届中国科技期刊发展论坛介绍产学研合作特色办刊经验
- 10 月, 《ZTE Communications》被评为 2019 中国国际影响力优秀学术期刊
- 10 月, 杂志社应邀在 2019 年中国学术期刊未来论坛上介绍产学研合作特色办刊经验
- 11 月, 《ZTE Communications》独立网站 tech-en.zte.com.cn 上线

2020 年

- 3 月, 刊物排版设计开始采用方正学术出版云服务平台
- 3 月, 开展“知识服务”活动, 征集到 5 500 多个目标读者邮箱
- 4 月, 开展“严抓‘三审三校’, 提升编校质量”专项活动
- 8 月, 出版创刊 25 周年纪念特刊

提升

2021—2025 年

传播力提升的五年

2021 年

- 5 月, 西安电子科技大学李建东教授受邀担任中文刊编委会副主任
- 6 月, 杂志社执行主编黄新明受聘为中国科学技术期刊编辑学会第 7 届理事会常务理事
- 7 月, 杂志社聘请 10 位中兴通讯技术专家担任产业顾问
- 9 月, 杂志社完成编委会换届, 成立中英文两刊第 9 届编委会

2022 年

- 2 月, 《中兴通讯技术》和《ZTE Communications》改版
- 2 月, 英文刊入选《科技期刊世界影响力指数报告 (WJCI 2021)》
- 2 月, 中英文两刊入选中国计算机学会《计算领域高质量科技期刊分级目录》T3 级
- 8 月, 中国科协“学清刊壇”项目专家组到杂志社调研
- 9 月, 执行主编黄新明受邀在“第二届长三角科技期刊与科技智库融合发展论坛”分享办刊经验
- 11 月, 中文刊特色栏目“热点专题”入选安徽省期刊“精品栏目”
- 12 月, 英文刊首次入选《科技期刊世界影响力指数报告 (WJCI)》
- 12 月, 《中兴通讯技术》入选中国电子学会《电子技术、通信技术高质量科技期刊分级目录》T3 级
- 12 月, 中英文两刊共计 328 篇论文入选《学术精要数据库》

2023 年

- 1 月, 杂志社组建首届中兴通讯技术杂志社促进产学研合作青年专家委员会, 37 位青年专家成为首批委员
- 2 月, 杂志社“走进高校学术团队”系列活动正式启动
- 3 月, 杂志社执行主编黄新明当选为安徽省科技期刊编辑学会第 7 届理事会副理事长
- 3 月, 首次设计刊物系列文创宣传品, 扩大对外宣传
- 5 月, 杂志社聘请 14 位中兴通讯优秀技术专家组建长聘作者团队
- 11 月, 中英文两刊入选中国通信学会“信息通信精品科技期刊方阵”
- 12 月, 《中兴通讯技术》首次入选《中文核心期刊要目总览》

2024 年

- 4 月, 《中兴通讯技术》入选中国核心学术期刊
- 8 月, 完成编委会换届, 组建中英文刊第 10 届编委会, 编委规模达到 116 人, 含 11 位两院院士
- 9 月, 《中兴通讯技术》连续 16 年入选中国科技核心期刊, 首次入选《中国应用型期刊评价研究报告》
- 10 月, 《ZTE Communications》被国际知名数据库 Scopus 收录
- 12 月, 《中兴通讯技术》和《ZTE Communications》入选中国通信学会发布的《信息通信领域高质量科技期刊分级目录》T2 级
- 12 月, 《中兴通讯技术》和《ZTE Communications》入选《科技期刊世界影响力指数报告 (WJCI 2024)》, 位列 Q3

2025 年

- 2 月, 中英文两刊加入工信科技期刊集群
- 6 月, 中英文两刊入选 2025 BIBF 精品期刊展



中兴通讯技术杂志社第 21—29 次编委会掠影



📍 黄山市

第 21 次编委会

2015 年 8 月 7—8 日，中兴通讯技术杂志社第 21 次编委会在黄山市召开。为纪念创刊 20 周年，中兴通讯创始人侯为贵先生到会并为 6 位资深编委颁发“特别贡献奖”，为杂志社颁发“合作与传播成就奖”。会议增选 15 位编委，完成第 7 届编委会换届。杂志社出版纪念特刊以及创刊 20 周年纪念画册。

📍 长春市

第 22 次编委会

2016 年 8 月 20—21 日，中兴通讯技术杂志社第 22 次编委会在长春市召开。中兴通讯时任董事长兼总裁赵先明到会并致辞。会议宣布，清华大学陆建华院士接替北京邮电大学钟义信教授担任中文刊编委会主任，聘请中兴通讯创始人侯为贵、北京邮电大学教授钟义信、南京邮电大学教授陈锡生 3 人为编委会顾问。



📍 银川市

第 23 次编委会

2017 年 8 月 12—13 日，中兴通讯技术杂志社第 23 次编委会在银川市召开。中兴通讯时任董事长殷一民出席会议。会议进一步明确了杂志社的办刊定位：办一本有观点的通信刊物，创一个有影响的行业智库，建一个有远见的创新平台。

银川

长春

黄山



烟台市

第 24 次编委会

2019 年 8 月 17—18 日，中兴通讯技术杂志社第 24 次编委会在烟台市召开。时任公司董事长李自学到会致欢迎辞。公司总裁徐子阳做了“携手赢未来——5G 时代的中兴通讯发展战略”的演讲。会议介绍了第 8 届编委换届情况，聘请了高文院士担任英文刊编委会主任。杂志社汇报了主管主办单位变更情况及两刊探索产学研特色期刊发展之路的各项举措。

福州市

第 25 次编委会

2020 年 8 月 22—23 日，中兴通讯技术杂志社第 25 次编委会在福州市召开。编委会上，时任公司董事长李自学先生介绍了中兴通讯近两年的发展情况。杂志社执行主编黄新明回顾了办刊 25 年来的发展历程，对年度办刊工作进行了总结和规划。代表们一致认为，杂志社致力于促进产学研合作，刊物特色鲜明，是中国高科技企业办刊的标杆。



兰州市

第 26 次编委会

2021 年 9 月 11—12 日，中兴通讯技术杂志社第 26 次编委会在兰州市召开。聘请西安电子科技大学李建东教授担任中文刊编委会副主任、南京邮电大学糜正琨教授为中文刊顾问。与会代表充分肯定了杂志社办刊成果和产学研合作的价值，一致认为产学研合作是企业办刊的优势和特色，未来需要继续坚持并做到更加开放。



兰州

烟台

福州



编委会掠影



太原市

第 27 次编委会

2022年8月13—14日，中兴通讯技术杂志社第27次编委会在太原市召开。公司执行副总裁王喜瑜出席会议并介绍了中兴通讯最新发展情况。郑纬民、任福继、金石、赵慧玲等16位专家学者针对算力、5G、元宇宙、智能超表面、网络标准等内容做了精彩的学术报告。杂志社提出筹备“促进产学研合作青年专家委员会”。

南宁市

第 28 次编委会

2023年8月19—20日，中兴通讯技术杂志社第28次编委会在南宁市召开。公司时任董事长李自学出席会议。陆建华院士和张平院士分别作特邀学术报告，李建东、张跃平、李向阳、原进宏、陈焰等16位海内外专家学者发表了精彩的学术演讲，本次会议首次采用现场视频直播和图片直播，效果良好。



丽江市

第 29 次编委会

2024年8月25—26日，中兴通讯技术杂志社第29次编委会在丽江市召开。会议总结了办刊成果，分享了两刊获得的荣誉，汇报了中英文两刊专题征集的过程和效果。会议完成了编委换届，10位新编委加入杂志社第10届编委会。黄新明执行主编表示将坚持走差异化发展道路，融合产学研各方智慧，继续探索企业办刊成功之路。



太原

丽江

南宁

目次

中兴通讯技术 (ZHONGXING TONGXUN JISHU)
第 31 卷 总第 184 期 2025 年 8 月 第 4 期

中文核心期刊 中国科技核心期刊 第三届全国期刊奖百种重点期刊 信息通信领域产学研合作特色期刊 中国知网、万方数据、重庆维普等数据库收录期刊 1995 年创刊

热点专题 ▶

面向 6G 的高时效智能机器通信技术

- 01 专题导读..... 张平, 秦晓琦, 马楠
- 03 意图驱动的智简通信..... 戴金晟, 秦晓琦, 秦海龙, 张平
- 13 面向铁路入侵检测的语义通信技术..... 郭疆远, 陈为, 艾渤
- 19 6G 数字孪生信道的三个使能技术:多模态感知、环境知识和大模型..... 于力, 张建华, 蔡逸辰
- 29 端边协同的 6G 内生 AI 网络..... 王志勤, 周吉喆, 韩凯峰
- 34 多智能体协作感知的现状与展望..... 刘根嘉, 陈思衡, 张文军
- 41 基于生成式模型的新型机器语义通信方法研究..... 承楠, 王秀程, 沈学民

名家视点 ▶

- 48 6G:从通信到通感算智融合..... 刘光毅
- 55 面向下一代光网络的光计算技术应用思考..... 李俊杰, 刘宇旸, 霍晓莉

企业视界 ▶

- 64 人工智能技术与应用前沿..... 包义明, 林阳, 屠要峰

综合信息 ▶

- 02 《中兴通讯技术》2026 年专题计划

《中兴通讯技术》2025 年热点专题名称及策划人

1. 6G 立体覆盖技术

西安电子科技大学教授 李建东
西安电子科技大学教授 刘俊宇

3. 6G 网络安全

北京航空航天大学教授 刘建伟
北京航空航天大学教授 王璟璟

5. 网络中的 AI 技术

中国电信新一代信息通信专业首席专家 解冲锋
北京邮电大学教授 孟洛明
清华大学教授 崔勇

2. 智算网络演进及关键技术

中国移动研究院副院长 段晓东
清华大学教授 李丹
电子科技大学教授 虞红芳

4. 面向 6G 的高时效智能机器通信技术

中国工程院院士、北京邮电大学教授 张平
北京邮电大学副教授 秦晓琦
北京邮电大学教授 马楠

6. 新一代光传输技术

上海交通大学教授 陈建平
中国联通研究院副院长 唐雄燕

MAIN CONTENTS

ZTE TECHNOLOGY JOURNAL
Vol. 31 No. 4 Aug. 2025

Special Topic ►	Timely Robotic Communications in 6G Era
	01 Editorial ZHANG Ping, QIN Xiaoqi, MA Nan
	03 Intent-Driven Intelligent and Concise Communications DAI Jincheng, QIN Xiaoqi, QIN Hailong, ZHANG Ping
	13 Semantic Communication Technology for Railway Intrusion Detection GUO Jiangyuan, CHEN Wei, AI Bo
	19 Three Enabling Technologies for 6G Digital Twin Channel: Multimodal Sensing, Environ- ment Knowledge, and Large Model YU Li, ZHANG Jianhua, CAI Yichen
	29 6G AI-Native Network with Edge-Device Collaboration WANG Zhiqin, ZHOU Jizhe, HAN Kaifeng
	34 Multi-Agent Cooperative Perception: Current Status and Prospects LIU Genjia, CHEN Siheng, ZHANG Wenjun
	41 Novel Generative Model-Based Machine-Type Semantic Communications CHENG Nan, WANG Xiucheng, SHEN Xuemin
Expert View ►	48 6G: From Communication to Integration of Communication, Sensing, Computing, and Intelligence LIU Guangyi
	55 Optical Computing: Advancing Next-Generation Optical Network Applications LI Junjie, LIU Yuyang, HUO Xiaoli
Enterprise View ►	64 Frontiers of Artificial Intelligence Technology and Applications BAO Yiming, LIN Yang, TU Yaofeng

期刊基本参数: CN 34-1228/TN*1995*b*16*74*zh*P*¥20.00*6500*10*2025-08

敬告读者	本刊享有所有发表文章的版权, 包括英文版、电子版、网络版和优先数字出版版权, 所支付的稿酬已经包含上述各版本的费用。 未经本刊许可, 不得以任何形式全文转载本刊内容; 如部分引用本刊内容, 须注明该内容出自本刊。
------	---

面向6G的高时效智能机器通信技术 专题导读



专题策划人



张平



秦晓琦



马楠

人工智能在机器终端侧的广泛部署显著提升了机器智能水平，推动6G通信系统中“物联”服务对象从“可联网”的感知设备拓展至“可自主”的智能机器。预计到2030年，以智能车、工业机器人为代表的智能机器通信接入量将超过总接入量的85%。规模庞大的智能机器联网，通过对信息的感知-传输-计算，完成实时协同与自主决策，其性能关键在于对物理世界状态信息的低开销、高时效追踪处理。然而，现有低时延通信体制主要通过叠加物理资源提升传输管道容量，海量智能机器接入使通信系统长期处于重负载状态，沿用传统体制将导致能耗激增，难以实现可持续发展。为此，亟需一次通信范式的革新，以激活智能在物理空间内的高效、有序流动为目标，推动通信网络从“比特管道”向“智能中枢”的根本转变。

在此背景下，“以智促简”的通信理念（ComAI）应运而生。ComAI旨在打破传统通信与智能相分离的设计逻辑，以语义信息论为理论基石，通过定义语义熵、语义互信息等

全新度量，构建“先算后通”甚至“边算边通”的通信体系。该理念将通信系统的设计目标从“语法无误”升级为“语义保真”，强调数据传输对决策控制的有效性，不仅关注传输准确性，更注重关键信息的时效性，实现从语法比特传输到语义意图交互的跃升，以满足智能机器间高效协同的需求。正如语义信息论所揭示：通信系统可沿“熵减”方向优化，通过智能压缩突破香农极限，并借助环境适应实现容量增益。这不仅标志着通信技术的演进，更体现了通信逻辑从“连接万物”向“理解万物”的质变。

为了推动对这一方向的系统化探索，本专题以“面向6G的高时效智能机器通信技术”为主题，围绕“以智促简”的关键理念，汇集了来自通信、人工智能、自动控制等多个领域的研究团队，共同探讨智能机器通信的新范式、新路径和新挑战。专题共收录6篇代表性论文，内容涵盖理论创新、系统设计与典型应用。我们期望这些内容能为读者提供有价值的参考。6G的未来不仅在于“传得更快”，更在于“传得更懂”；不仅在于比特的准确，更在于意图的精准。希望本专题能为这一愿景的实现贡献力量。

策划人简介

张平，中国工程院院士，北京邮电大学教授、网络与交换技术全国重点实验室主任，《通信学报》主编，IEEE Fellow，IMT-2020（5G）专家组成员，IMT-2030（6G）推进组咨询委员会委员，国家自然科学基金委“创新研究群体”带头人；长期致力于移动通信理论研究和技术创新，为中国自主技术成为国际主流做出了基础性的贡献；曾任国家重点基础研究发展计划（“973”计划）首席科学家、国家高技术研究发展计划（“863”计划）主题专家组专家等；获国家科技进步奖特等奖、一等奖，国家技术发明奖二等奖3项，科技进步奖二等奖2项，全国创新争先奖，光华工程科技奖等奖项，入选教育部首批“黄大年式教师团队”。

秦晓琦，北京邮电大学副教授，国家级青年人才，中国工程院“中国工程前沿杰出青年学者”；主要研究方向为智能机器通信

基础理论与关键技术；主持包括国家重点研发计划项目、国家自然科学基金面上项目、北京市自然科学基金项目等多个科研项目；获省部级和全国学会级科学技术一等奖3项，获最佳论文奖3项；已发表论文130余篇，参与制定国际标准2项、行业标准4项。

马楠，北京邮电大学教授、网络与交换技术全国重点实验室成员，先后担任无线通信测试技术北京市重点实验室主任、中国仪器仪表学会信息通信测试仪器仪表专委会副主任委员等；主要从事移动通信理论与测试技术研究，包括复杂场景多模态语义通信理论、无线信道建模与测试技术等；主持4项国家重大攻关项目，包括国家科技重大专项3项、国家重点研发计划课题1项；获国家科技进步奖二等奖、国家教学成果二等奖、全国学会级科学技术奖一等奖，以及中国产学研合作创新奖等奖项。

综合信息

《中兴通讯技术》2026年专题计划

期次	专题名称	策划人
1	6G 关键技术的标准化:Day-1 与未来	易芝玲 中国移动研究院首席科学家
2	广域立体覆盖低空通信技术	金石 东南大学副校长 刘凡 东南大学教授
3	智算网络	赵慧玲 工信部信息通信科技委常委
4	大模型推理中的存算技术	郑纬民 中国工程院院士、清华大学教授 陆游游 清华大学副教授
5	星地太赫兹高速传输技术	洪伟 东南大学教授 唐万斌 电子科技大学教授 郝张成 东南大学教授
6	智能多天线技术	艾渤 北京交通大学教授 章嘉懿 北京交通大学教授

意图驱动的智能通信



Intent-Driven Intelligent and Concise Communications

戴金晟/DAI Jincheng¹, 秦晓琦/QIN Xiaoqi²,
秦海龙/QIN Hailong¹, 张平/ZHANG Ping²

(1. 北京邮电大学泛网无线教育部重点实验室, 中国 北京 100876;

2. 北京邮电大学网络与交换技术国家重点实验室, 中国 北京 100876)

(1. The Key Laboratory of Universal Wireless Communications, Ministry of Education, Beijing University of Posts and Telecommunications, Beijing 100876, China;

2. The State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China)

DOI: 10.12142/ZTETJ.202504002

网络出版地址: <http://kns.cnki.net/kcms/detail/34.1228.TN.20250716.1528.010.html>

网络出版日期: 2025-07-16

收稿日期: 2025-06-05

摘要: 随着通信迈入6G时代, 传统依赖物理资源扩展的通信模式难以满足智能化、泛在化的发展需求。提出一种意图驱动的智能通信系统, 融合认知心理学、信息论与人工智能方法, 以语义 token 为基本单元, 构建面向信息效用的通信范式。该系统集成智能体的感知、认知与反馈能力, 实现异构数据的上下文感知语义建模与压缩传输, 重点突破语义编码、意图解析、鲁棒传输与可信解码等关键技术。所提架构适配人-人感知、人-机控制与多机协同等差异化需求, 支持在带宽受限与信道动态条件下的高效稳健传输。系统梳理了智能通信的研究脉络与核心机制, 为构建高效、泛用、可持续的智能通信体系提供理论支撑与技术参考。

关键词: 智能通信; 意图驱动; 高效压缩; 稳健传输

Abstract: As communication evolves into the 6G era, traditional paradigms relying on physical resource expansion are increasingly struggling to meet the burgeoning demand for intelligent and ubiquitous connectivity. An intent-driven intelligent and concise communications system is proposed, which seamlessly integrates insights from cognitive psychology, information theory, and artificial intelligence. Using semantic tokens as the basic representation units, the system establishes a communication framework centered on information utility. It incorporates intelligent agents with perception, cognition, and feedback capabilities to achieve context-aware semantic modeling and compressed transmission of heterogeneous data. The system's key enabling technologies comprehensively include semantic encoding, intent parsing, robust transmission, and reliable decoding. The proposed architecture supports differentiated communication needs, including high-fidelity human-human interaction, efficient human-machine control, and collaborative multi-agent systems. It is particularly effective and robust under bandwidth-limited and dynamically varying channel conditions. This study reviews the development trajectory and technical foundations of intelligent and concise communication, providing theoretical and practical guidance for building efficient, general-purpose, and sustainable communication systems.

Keywords: intelligent and concise communications; intent-driven; high-efficiency compression; robust transmission

引用格式: 戴金晟, 秦晓琦, 秦海龙, 等. 意图驱动的智能通信 [J]. 中兴通讯技术, 2025, 31(4): 3-12. DOI: 10.12142/ZTETJ.202504002

Citation: DAI J C, QIN X Q, QIN H L, et al. Intent-driven intelligent-concise communications [J]. ZTE technology journal, 2025, 31(4): 3-12. DOI: 10.12142/ZTETJ.202504002

随着通信技术迈入6G时代, 传统以增加频谱、天线等物理资源为核心的发展路径已逼近理论极限。尽管“外延式”扩展在一定程度上提升了系统容量, 但其资源消耗大、部署复杂, 难以支撑未来通信的可持续演进^[1]。同时, 生成式与交互式人工智能技术的迅猛发展, 正在深刻改变通信系统的结构与功能——通信的信源与信宿正由传统的收发模块转变为具备理解与推理能力的通用人工智能体

(AGI Agent)^[2-3]。在此背景下, 继续沿用经典香农通信体制, 将导致人与人、人与机之间的交互效率大幅下降, 无法满足智能泛在时代的通信需求。经典通信理论主要聚焦于“bit”的可靠传输, 但bit本身并不等价于“理解”, 更无法表达“意图”^[4]。在智能体主导的新型通信形态下, 亟需探索如何利用智能体的感知、认知与决策能力, 构建面向信息效用最大化的通信范式。具体而言, 通信系统需突破现有对信号传输的关注, 转向对信息“提取—理解—传输—校正—利用”全过程的柔性适配机制研究, 发掘信息传输的新维度, 并在系统架构与核心机理上实现根本性突破。

基金项目: 国家重点研发计划项目 (2024YFF0509700); 国家自然科学基金项目 (62371063); 北京市自然科学基金项目 (L232047); 北京市科技新星计划

在当前背景下，本文基于认知心理学的启发，结合信息论、人工智能以及系统科学的研究理念，提出构建一种意图驱动的智能编码传输系统。该系统以语义 token 向量作为核心表示单元，能够实现对异构数据的规范化、简约化语义表征，同时集成智能体的“感知—认知—决策—反馈—控制”能力，从而推动通信系统向语义与意图层面升级。针对一人一高保真感知通信、人一机联网控制、多机交互协同等存在差异化需求的场景，本文重点关注以下关键技术方向：通用语义编码与上下文意图解析、意图完备的语义表征适配精简、融合上下文反馈的语义特征压缩编码与联合误差校正、意图对齐的语义解码与可信质量评价机制。

本文着重梳理智能通信技术的发展历程与研究现状，聚焦于以极简压缩、极高韧性、极强泛用为目标的智能编码传输体系，探讨其从“符号通信”向“上下文通信”转型的过程，以赋予通信系统理解语义与执行意图的能力，进而实现物理世界与数字世界的深度融合。

本文阐述了意图驱动的智能通信系统的总体架构与关键技术路线，分析其在有限带宽、有损传输、动态信道环境等现实条件下的适应性。通过对系统的分层设计与核心机制的剖析，揭示了该新型通信范式在提升系统效能、降低通信负担、增强传输鲁棒性等方面的潜力，为通信技术向更高效、更智能、更可持续的方向发展提供理论支撑与技术参考。

1 多学科交叉引领智能通信新时代

1.1 多学科发展的底层逻辑呈现一致性

在全球科技加速融合发展的大背景下，多学科交叉研究正逐渐成为推动科学范式变革的关键力量^[5]。本文聚焦于信息通信范式的转型，即从以离散符号为中心的经典香农通信向以上下文意图为核心的智能通信演进。这一转型不仅契合通信系统自身演进的内在需求，也反映出多个相关学科在方法论与思维方式上呈现出的共通趋势。

如图1所示，系统回顾离散数学、统计物理、人工智能和认知心理学等领域的发展轨迹，可发现其均呈现从个体到群体、从离散到连续、从局部到全局的演化路径。离散数学中，集合论通过抽象个体属性实现系统性描述^[6]；统计物理借助大规模样本统计分析，揭示微观与宏观的关联^[7]；自然语言处理从早期符号匹配发展到上下文建模，显著提升语义理解和泛化能力^[8]；认知心理学的知觉符号系统（PSS）理论，强调动态感知、上下文依赖和情境驱动的认知建模方式，为信息表征和处理提供新认知范式^[9]。

这些演进趋势与当前信息通信瓶颈高度契合。传统通信

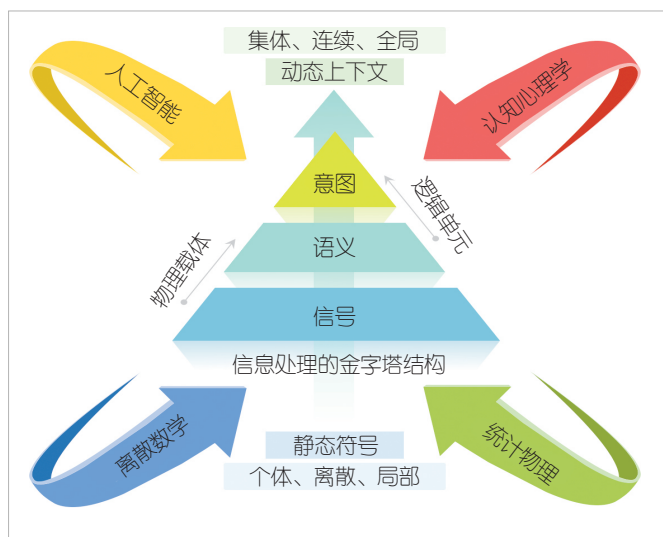


图1 多学科演进路线的共通性启发信息传输范式转变

依赖离散比特传输，难以表达语义依赖性与高层意图。因此，从离散符号向包含上下文信息的动态表示扩展，构建多层次、可解释的信息模型，是通信系统的自然且必要演进方向。

基于此，本文提出以“信号—语义—意图”3层信息结构为基础的智能编码传输框架：信号是物理载体，语义是其抽象组合，意图是决策调控的核心驱动力。如图2所示，类比PSS理论^[9]，该结构使信息传输不再仅限于符号传递，而是围绕意图达成展开，通过语义 token 向量对异构信息进行统一、简约化表达，并利用上下文信息动态调整编码与传输策略，实现端到端的信息效用最大化。

智能通信的提出深刻体现了多学科思想的融合与迁移，其核心理念与多学科在处理复杂系统中的建模方式高度一致，为信息通信范式突破提供了理论依据与实践路径。

1.2 学科交叉驱动下的信息通信范式演进

信息通信技术与人工智能虽起源不同，但近年来呈现出“殊途同归”的融合趋势。从技术演进轨迹看，深度学习领域已率先完成由多样化模型结构向统一自编码器架构^[10]的过渡，并在全局上下文注意力机制^[8]支持下，实现语义层精确建模与意图层深入理解，显著推动语义理解、语言建模及复杂决策生成等任务性能提升。

通信系统虽在物理层与高层协议设计中逐步引入人类自编码器结构，朝着端到端、模块协同优化方向发展，但在信息语义与意图建模方面仍滞后。当前主流的深度联合编码传输框架^[11]，多以迁移深度学习模型为核心策略，虽提升编码效率和鲁棒性，但未改变“以bit为中心”的基本传输对象与机制，难以满足语义对齐与智能交互等新型通信需求^[12]。

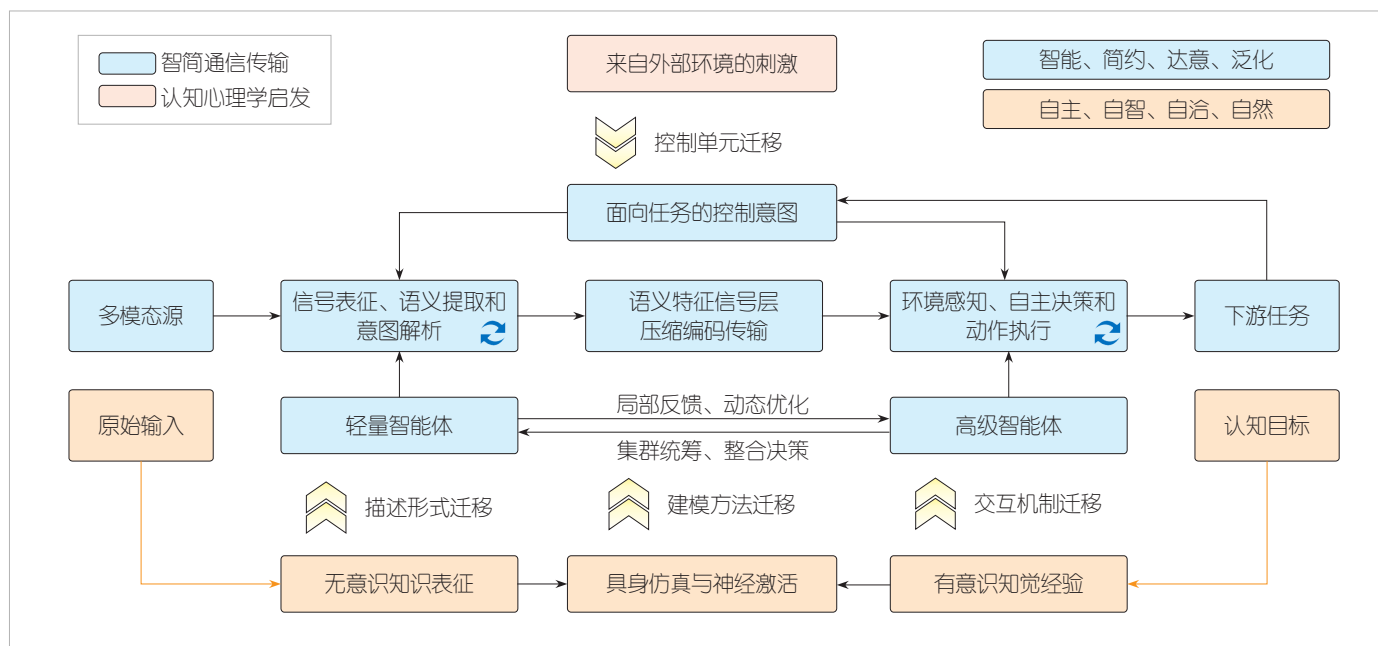


图2 认知心理学启发的智简传输

纵观信息通信与人工智能的研究演进，虽然两者呈现“殊途同归”趋势，但发展速度和深度存在差异。深度学习已从复杂多样的模型结构过渡到基于全局上下文注意力机制的自编码器统一架构，实现了语义级处理和意图级理解。通信系统模型虽然采用类自编码器架构，并朝着模块联合设计方向发展，但在信息的语义和意图解析方面仍存不足^[13]。当前的深度联合编码传输范式是将深度学习模型迁移至信息传输过程，但未从根本上改变信息处理和传输的基本对象^[14]。

本文的研究旨在弥合这一发展代沟，提出将全局上下文作为动态编码对象进行信息处理和传输的新范式，如图3所示。这种方法不仅能深度解析信息的语义和意图，还能通过适度精简模型结构，提高系统的环境感知能力和泛化性，从而适应各类通信场景，实现真正的智能简约通信新范式。这一研究方向有望突破传统通信系统面临的发展瓶颈，为信息通信理论和技术创新发展提供新的思路。

2 智简通信的设计目标与系统架构

为明确意图驱动智简通信的技术定位，需将其与当前主流的语义通信和任务导向通信进行区分。现有语义通信聚焦数据语义的表征与传输，任务导向通信则针对特定任务优化信息编码策略。与之不同，意图驱动智简通信的核心特征在于将编码传输对象从数据语义拓展至包含上下文依赖的复杂业务意图。该技术以语义 token 作为信息表征的基本单元，替代传统 bit 级表示，实现表征粒度的意图自适应调节。此外，系统构建覆盖信息全生命周期的处理机制，不仅实现语

义层面的理解与传输，更强调意图层面的解析与执行，在极简压缩、极高韧性和极强泛用性方面具备显著优势。

2.1 智简通信的设计目标

随着通信系统日益面临带宽紧张、时延敏感与算力受限等多重资源约束，传统以连接为中心的通信范式已难以满足多样化、语义化的智能应用需求^[15]。智简通信系统以“面向意图的智能压缩传输”为核心理念，构建具备“感知—认知—决策—反馈—控制”闭环能力的智能体驱动通信框架，具体如图4所示。在总体设计目标上，智简通信不仅关注数据的传输效率，更强调语义的有效表达与意图的精准对齐。研究重点聚焦于以下关键技术方向：

- 1) 通用语义编码与上下文意图解析：构建跨模态、可泛化的语义表示方法，提取符合通信目标的高价值语义信息，并解析其上下文意图。
- 2) 语义表征的适变精简机制：结合任务需求与资源状态，自适应地调整语义表示粒度，实现信息内容的高效压缩与精简表达。
- 3) 压缩编码与联合误差校正机制：在信号传输阶段，融合语义特征压缩编码与后验反馈校正策略，增强通信系统的韧性与精度。
- 4) 意图对齐与可信质量评价机制：接收端实现语义特征到控制或感知行为的精准映射，建立可量化的主观/客观质量评估体系，确保传输结果与预期意图一致。

基于上述关键模块，智简通信系统从“信息的提取—理

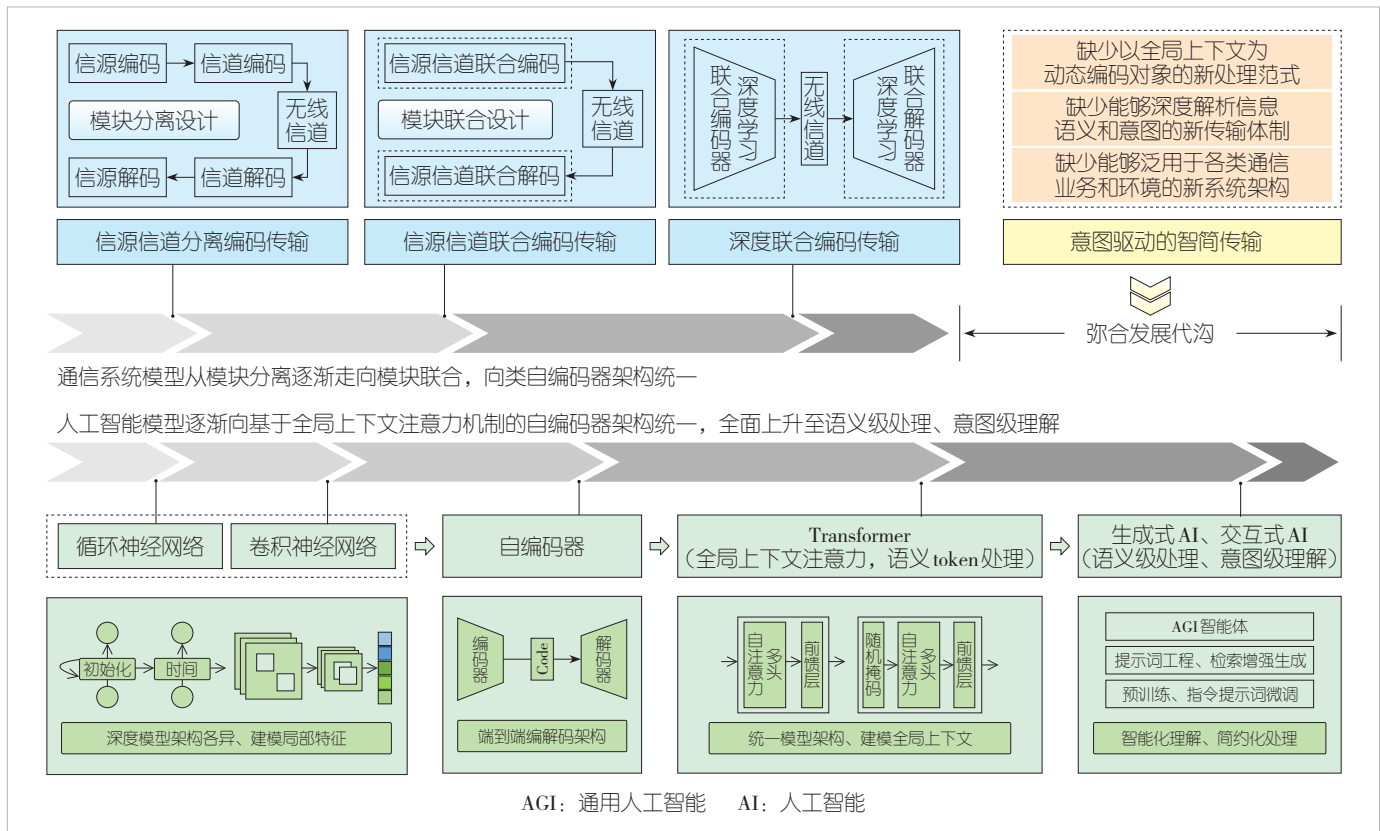


图3 信息通信研究与人工智能研究存在的发展代沟及智简通信的研究定位

解—传输—校正—利用”全过程出发，构建了面向语义与意图层的广义编码传输机制。该机制强调系统的柔性适配能力，可根据不同场景在资源、语义复杂度、控制精度等方面进行动态调整，从而实现以下目标：

- 1) 简压缩：以最小的数据量传达最具意义的信息。
- 2) 高韧性：应对复杂信道和动态环境中的鲁棒传输。
- 3) 强泛用：适配人—人、人—机、机—机等通信需求。

与传统通信系统仅依赖比特级检错纠错不同，智简通信系统创新性地实现了信号空间与语义空间的双重校正机制。信号空间校正继承了经典通信理论中的比特级检错纠错功能，确保了底层信号传输的可靠性。语义空间校正则是系统的创新亮点，通过后验反馈机制，在接收端动态校准语义特征与发送端意图的对齐度，构建意图驱动的端到端质量保障体系。这种双重校正机制不仅保障了传输的准确性，更确保了意图执行的一致性。

综上所述，智简通信是未来通信技术从“连接驱动”向“意图驱动”转型的关键路径。其核心是集成智能体的认知与控制能力，实现更高效、更韧性、更智能的通信系统设计，具有重要的理论研究价值和广泛的实际应用前景。

2.2 智简通信的系统架构

为满足未来通信系统中多样化、智能化的业务需求，本文提出了一种意图驱动的层次化智简编码传输系统架构，如图5所示。该架构在总体设计上呈现出从底层到高层、由比特到语义再到意图的逐级抽象过程，旨在实现高效、韧性、可演进的信息传输机制，且与现有通信体制兼容。

系统架构整体分为3个功能层次：

- 1) 底层（Bit-centric 信号传输层）：对应传统通信系统中以比特为单位的信号调制与物理传输机制，确保底层信号的可靠传递，为上层语义与意图表达提供基础承载。
- 2) 中层（Token-centric 语义传输层）：通信内容以 token 为基本单元进行表达与传输，代表了对原始数据的语义压缩编码结果。通过语义建模与上下文分析，实现对信息“意义”的传输。
- 3) 顶层（Intent-centric 意图调控层）：通过解析业务数据、用户行为及环境状态，提取驱动通信策略调度的高层抽象意图，作为系统智能决策与资源调配的核心依据。意图不仅用于指导语义压缩与 token 筛选，还可动态调整传输策略，实现端到端的信息传输闭环控制。

该层次化架构通过内层支撑外层、外层调控内层的方

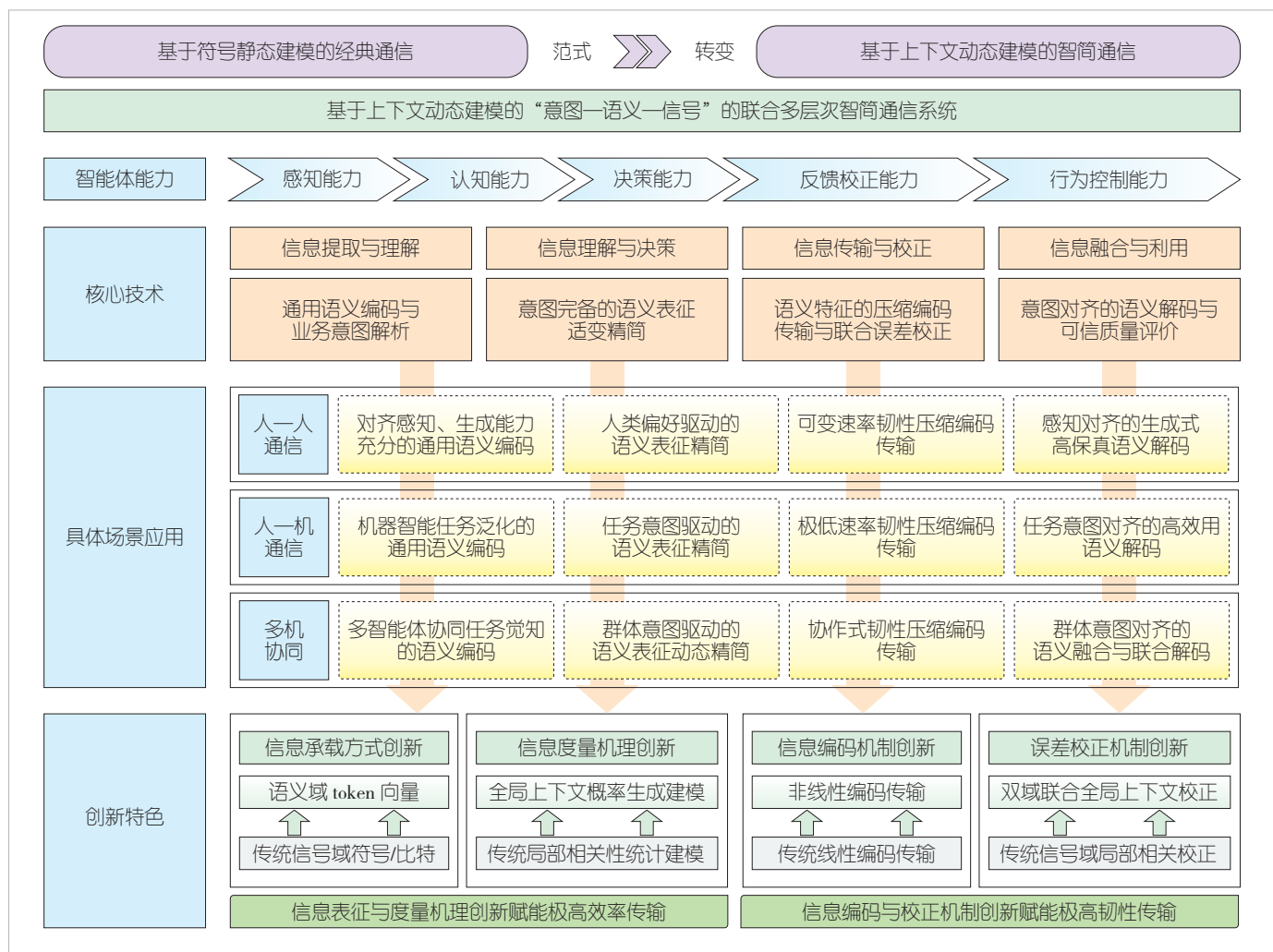


图4 智简通信总体技术路线

式，构建了一个由内而外融合的信息传输体系。同时，它具备良好的可扩展性与演进能力，可在不破坏传统通信协议栈结构的前提下，逐步引入语义与意图驱动的增强功能。

在功能方面，该系统能够适配多种差异化通信场景：

- 1) 面向人—人业务的高保真感知数据传输，强调主观感知对齐与数据压缩保真。
- 2) 面向人—机交互的高效命令与控制指令传输，突出语义精准性与低时延。
- 3) 面向多智能体系统的协同语义共享与意图对齐传输，支撑大规模异构智能体之间的协作控制。

在系统架构设计中，兼容性问题得到了充分考虑。系统采用渐进式演进策略，以现有的比特级信号传输为基础，在其上叠加语义处理与意图解析功能模块。这种分层解耦的方式，确保了语义与意图信息通过底层信号传输层完成物理传递，从而天然兼容现有的通信协议栈。在实现过程中，系统

无须对底层通信基础设施进行颠覆性改造，而是通过在协议栈中插入智能处理模块，实现功能无缝扩展与性能提升。

综上所述，该意图驱动的层次化系统架构为智简通信提供了统一理论框架与技术支持，体现了从“连接为中心”向“语义与意图为中心”的通信范式升级路径。

3 智简通信的技术路线

3.1 面向人—人业务感知的智简通信技术路线

在面向人类用户的高质量数据传输任务中，如何在带宽、时延等多重通信资源受限的条件下有效满足人类感知体验，已成为智能通信系统中的关键挑战之一。近年来，感知对齐的编码传输技术成为研究热点^[16-20]，其核心目标是保障主观感知质量，实现通信资源的高效利用。

在质量建模方面，研究者融合认知心理学与神经科学理论，探索符合人类感知机制的质量评价与建模方法^[21-23]，旨

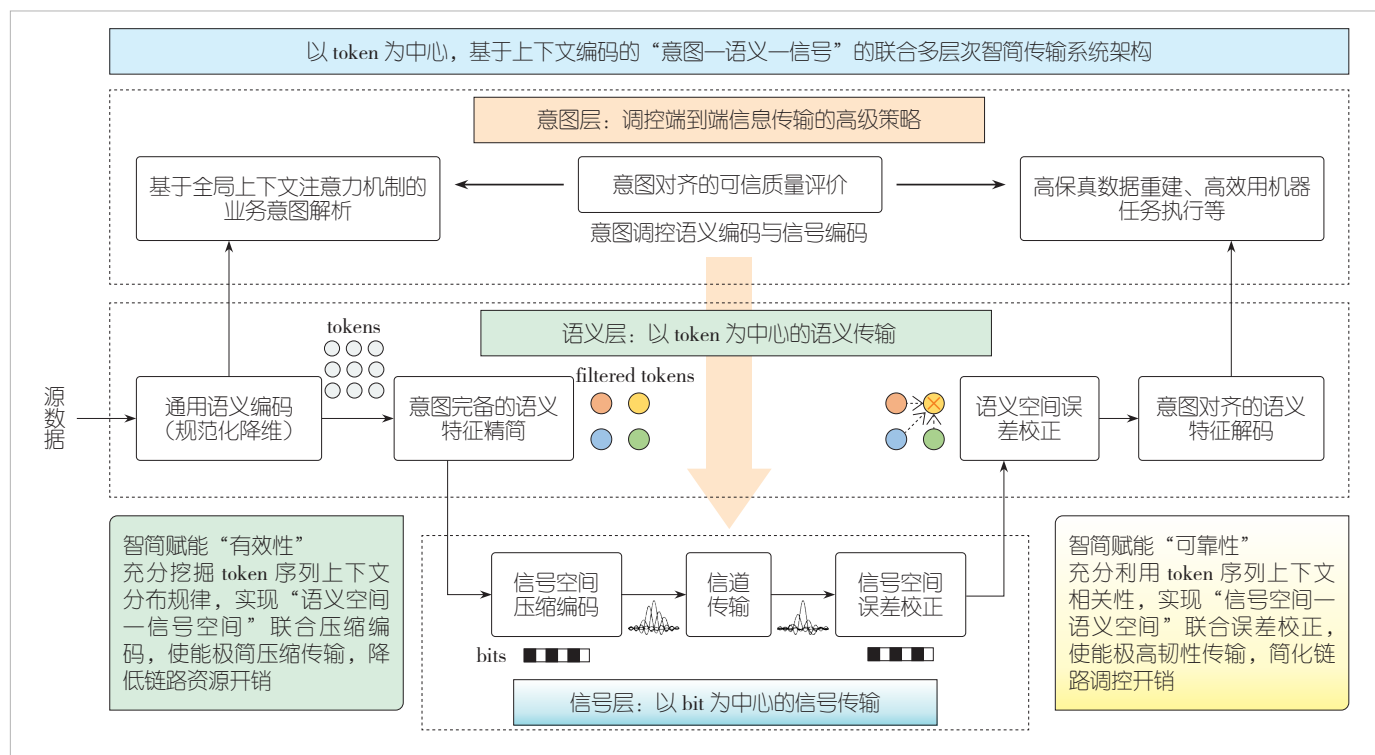


图5 智能通信系统架构

在构建以人类主观感受为核心的传输质量评估体系，通过对用户偏好的智能感知建模，实现从“面向客观失真优化”向“面向主观感知优化”的范式转变。

为应对多模态异构数据在语义表达上的规范性和简约性需求，通用语义编码技术逐渐成为研究焦点^[24-26]。该技术将原始数据映射到对齐人类感知的语义空间，通过生成能力充足的 token 序列完成数据的语义化压缩表达。结合用户感知偏好和意图，对 token 序列进行语义过滤与融合，获取高度简约的语义表示，在保证感知质量的前提下降低数据冗余。

在信号层传输阶段，研究者引入注意力机制与序列建模方法，挖掘 token 序列中潜在的上下文概率结构，支持更高效的变速率压缩编码策略^[16]。结合后验生成采样机制，在接收端实现信号与语义空间的联合误差反馈校正^[20]，提升解码结果对人类语义感知的对齐度，增强端到端传输的有效性与鲁棒性。

如图6所示，感知对齐的智能编码传输机制是通信系统向“以人为中心”智能范式演进的重要方向，未来研究将聚焦于提升感知建模精度、优化语义压缩效率，以及强化解码质量保障。

3.2 面向人一机网联控制的智能通信技术路线

随着人一机协同场景的不断拓展，实现对多类型机器操

作设备的高效、低时延网联控制已成为智能通信系统中的关键课题^[27]。尤其在“人→机”控制链路中，针对强异构性、大规模、低延迟的指令传输需求，构建高效、强韧性的编码传输机制，是实现精准控制与意图对齐的核心挑战。

在系统架构上，近年来的研究开始关注在发送端引入轻量化智能体模块，对实时场景数据进行预处理分析^[28-29]。该模块通过目标导向的感知数据筛选和实时清洗，有效压缩冗余信息，提升下行链路的传输效率。针对人类指令与机器感知信息的融合表达，研究者提出将其嵌入统一的语义表示空间，通过 token 化生成极简语义单元^[26]。

在编码传输阶段，采用全局注意力机制挖掘 token 序列的上下文依赖结构，结合变速率压缩编码策略，实现从语义空间到信号空间的韧性编码与传输，以适应网络动态变化带来的带宽波动，提升关键控制语义的传输鲁棒性。

在接收端，部署的轻量化智能体协同云端高级智能体进行控制意图的解析与执行决策生成。云端智能体基于 token 序列完成意图抽象与高层决策输出，并以 token 形式发送至设备端；设备端智能体则将 token 映射为具体控制命令，实现端到端的意图对齐与执行闭环。

此外，为提升端侧智能体的部署效率与决策质量，相关研究还引入了知识蒸馏与强化学习等手段，以支持智能体的轻量化运行与动态自适应优化^[30]。通过上述机制，系统能够

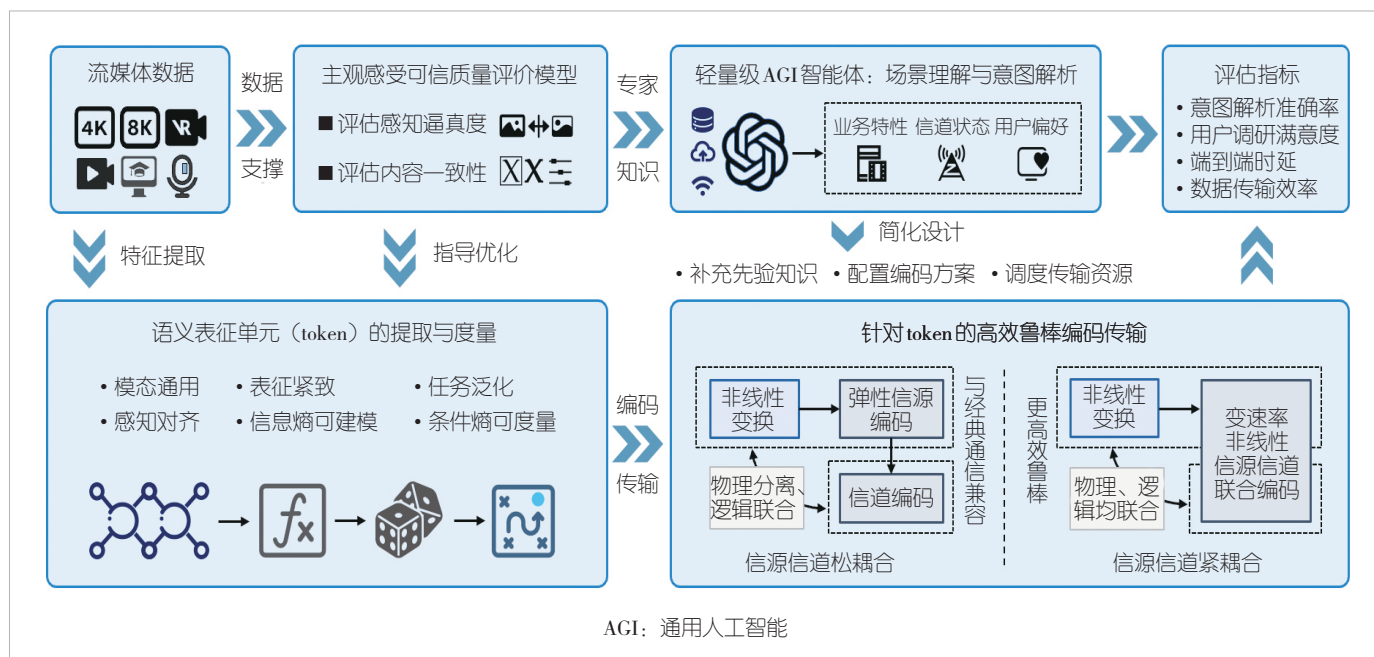


图6 面向人-人业务感知的智能通信技术路线

实现收发端之间的意图快速对齐与闭环反馈,在保障控制时效性的同时,提升人-机交互的稳定性与精度。

如图7所示,面向人-机网联控制的智能编码传输机制通过语义压缩、意图对齐与智能体协同优化,为大规模、低延迟、高鲁棒性的机器通信提供了新范式,展现出极大的研究与应用潜力。

3.3 面向多机交互协同的智能通信技术路线

在智能网联系统中,多智能体的协同感知与联合决策对通信系统提出了更高的语义表达与意图对齐要求。多智能体之间的高效协同不仅依赖于底层通信能力,更需构建面向语义层的智能编码与意图协同机制,以应对多源异构数据、复杂任务需求和时变环境动态带来的挑战。

在语义建模方面,研究者逐步从单智能体语义重要性评估拓展至多智能体场景下的协同语义分析^[31]。基于多模态、多粒度的 token 语义表示,结合注意力机制对异构智能体间的感知数据进行加权融合,实现语义层的时空对齐与意图层的协同映射。这一过程对于实现多智能体之间的高效协同感知与共享认知具有关键意义。

在通信效能评估与优化方面,已有研究构建了基于群体意图驱动的通信评价体系。通过引入图神经网络 (GNN) 与强化学习 (RL) 算法,建立动态连接图结构,捕捉智能体间任务依赖与通信拓扑的演化特征^[32-33]。在此基础上,进一步提出面向任务的动态编码与传输策略,实现通信资源与协

同效率之间的平衡。

在智能体系统的演化机制研究方面,相关工作尝试融合认知心理学中的意图建构理论^[34]与系统动力学建模方法^[35],探索智能体群体的自主演化机制。通过构建稀疏互联的动态语义通信网络,并引入意图理解更新机制与策略自主调度机制,推动智能体间从被动协同走向主动协同,实现多智能体系统中“智慧涌现”式的协作通信范式^[36]。

如图8所示,面向多智能体交互与协同的智能通信机制,融合了语义建模、图结构学习等多学科交叉技术,为实现高效、智能、自适应的多智能体通信系统提供了理论基础与关键技术支撑,是未来智能通信系统的重要发展方向^[35-36]。

3.4 智能通信系统可行性的验证实例

为验证智能编码传输系统的实际可行性和有效性,我们以城市交叉路口的车路协同避障场景为例进行阐述。考虑一个典型的城市十字路口,该路口配备有路侧感知单元,其任务是向接近路口的自动驾驶车辆传输实时感知信息,以辅助安全通行。传统方案需要传输完整的多摄像头视频流(分辨率为 1920×1080 , 帧率为 30 fps, 数据率约 25 Mbit/s)、激光雷达点云数据(约 10 Mbit/s)以及各类传感器原始数据,总数据率超过 40 Mbit/s。这些在 5G 网络峰值条件下尚可承载,但在网络拥塞或边缘覆盖场景中,传输时延达 200 ~ 500 ms,难以满足车辆安全避障的实时性要求 (<100 ms)。

针对上述问题,意图驱动智能通信系统的具体处理过程

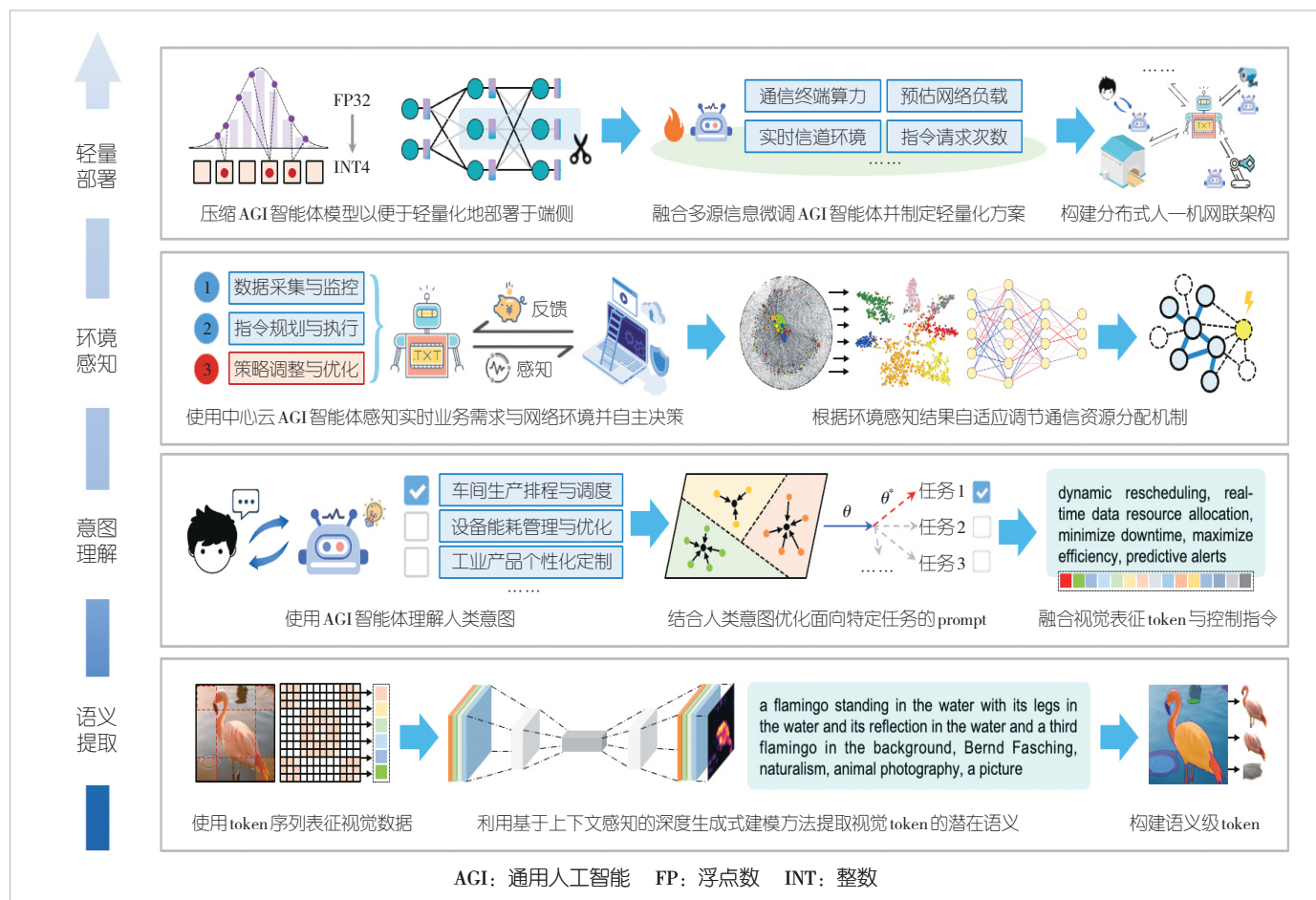


图7 面向人—机业务感知的智能通信技术路线

为：系统首先通过车联网（V2X）获取车辆状态信息，解析其驾驶意图，例如速度为 45 km/h，目标为直行通过路口，距离路口 80 m。基于此意图，系统从海量感知数据中筛选出与该车辆通行路径相关的关键信息：前方车道占用情况、行人穿越动态、信号灯剩余时间等，过滤掉无关车道的车辆信息。筛选后的关键信息被编码为标准化语义 token 序列：

- 交通信号：[LIGHT:GREEN:REMAIN:12S];
- 前方车辆：[VEHICLE:SEDAN:LANE1:DISTANCE:25M:SPEED:35KMH];
- 行人动态：[PEDESTRIAN:ADULT:SIDEWALK:INTENT:WAIT];
- 路面状况：[ROAD:DRY:VISIBILITY:GOOD]。

随后，系统将原始 40 Mbit/s 的数据压缩为仅包含 24 个语义 token 的序列，数据量降低至约 2 kbit/s。根据当前网络状况（可用带宽 5 Mbit/s，往返时延=30 ms）以及信息的紧急程度，系统对 token 序列进行优先级排序。安全关键信息（前方车辆、行人动态）采用冗余编码以确保可靠传输，而

环境辅助信息（如路面状况）则采用普通编码。

车载智能系统接收 token 序列后，结合车辆当前状态生成具体驾驶策略：

- 维持当前速度 45 km/h；
- 监控前方车辆，保持安全车距；
- 准备在信号灯变红前通过路口；
- 注意观察人行道，暂无穿越风险。

相比传统方案，意图驱动的智能通信系统可实现：

- 数据传输量：从 40 Mbit/s 降至 2 kbit/s，压缩比达 99.995%；
- 端到端时延：从 200 ~ 500 ms 降至 35 ~ 50 ms，满足实时性要求；
- 决策准确性：在 100 次测试场景中，系统生成的驾驶决策与基于原始数据的决策一致性达 97%；
- 网络适应性：在带宽受限至 1 Mbit/s 的条件下，系统仍能保持稳定通信。

该实例验证了智能通信系统在确保安全关键信息完整性

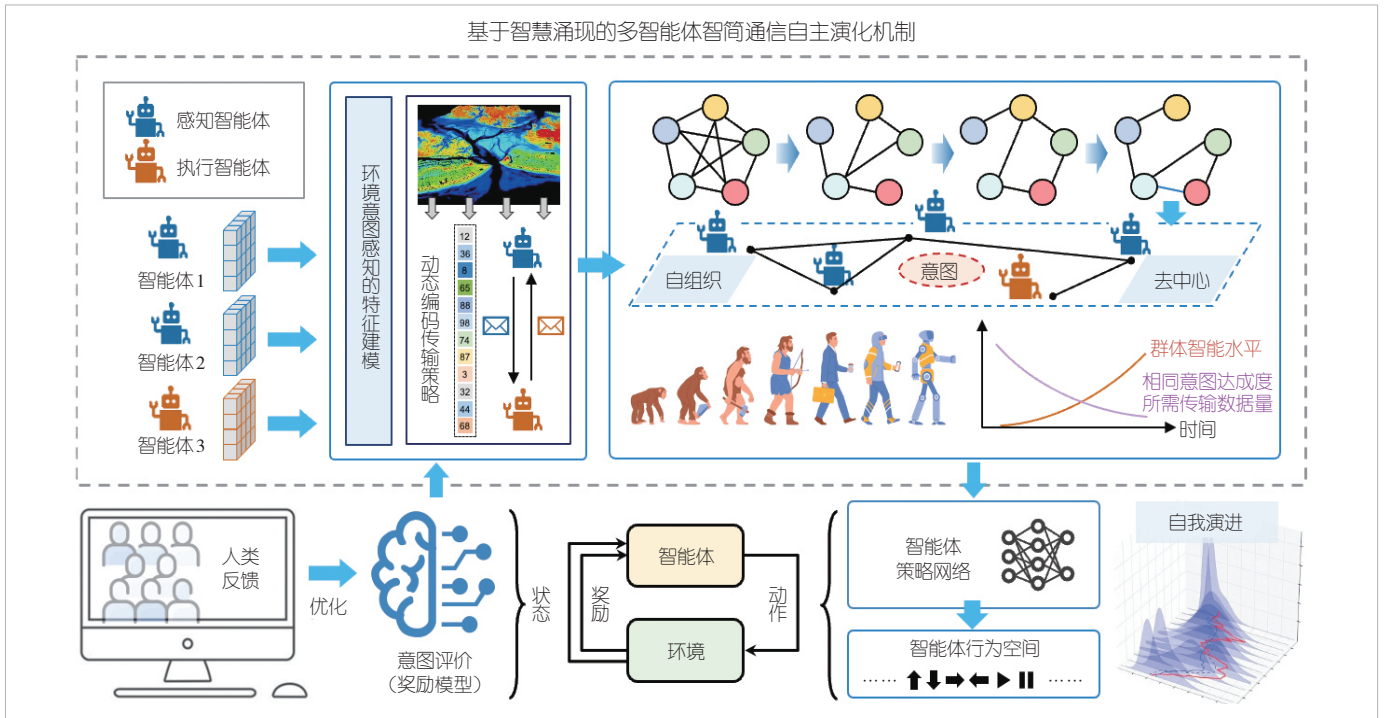


图8 面向多机交互协同的智能通信技术路线

的同时，显著提升传输效率和实时性的能力，充分展现了其在实际应用场景中的可行性和有效性。

4 结束语

本文面向6G时代“万物智联”通信需求，提出一种意图驱动的智能编码传输系统。该系统融合认知心理学、信息论与人工智能方法，以语义token为基本单元，构建信息效用导向的通信新范式。系统集成智能体的感知、认知与反馈能力，支持上下文感知的语义建模与压缩传输，适配人—人感知、人—机控制与多智能体协同等差异化场景，在带宽受限与信道动态环境下表现出高效性、鲁棒性与泛用性，为未来通信系统的智能演进提供理论支撑与技术参考。

参考文献

- [1] ZHANG P, XU W J, GAO H, et al. Toward wisdom-evolutionary and primitive-concise 6G: a new paradigm of semantic communication networks [J]. Engineering, 2022, 8: 60–73. DOI: 10.1016/j.eng.2021.11.003
- [2] CHACCOUR C, SAAD W, DEBBAH M, et al. Less data, more knowledge: building next-generation semantic communication networks [J]. IEEE communications surveys & tutorials, 2024, 27 (1): 37–76. DOI: 10.1109/comst.2024.3412852
- [3] XIA L, SUN Y, LIANG C S, et al. Generative AI for semantic communication: architecture, challenges, and outlook [J]. IEEE wireless communications, 2025, 32(1): 132–140. DOI: 10.1109/mwc.003.2300351
- [4] WEAVER W. Recent contributions to the mathematical theory of

- communication [EB/OL]. [2025–05–06]. https://courses.ischool.berkeley.edu/i218/s15/Weaver_Recent-Contributions.pdf
- [5] 路甬祥. 学科交叉与交叉科学的意义 [J]. 中国科学院院刊, 2005, 20 (1): 58–60. DOI: 10.16418/j.issn.1000–3045.2005.01.014
- [6] GRATTAN-GUINNESS I. From the calculus to set theory 1630–1910: an introductory history [M]. New Jersey, USA: Princeton University Press, 2020
- [7] NISHIMORI H. Statistical physics of spin glasses and information processing: an introduction [M]. Oxford, UK: Clarendon Press, 2001
- [8] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [EB/OL]. (2023–08–02) [2025–05–22]. <https://arxiv.org/abs/1706.03762>
- [9] BARSALOU L W. Perceptual symbol systems [J]. Behavioral and brain sciences, 1999, 22(4): 577–609; discussion 610–660. DOI: 10.1017/s0140525x99002149
- [10] BENGIO Y. Learning deep architectures for AI [J]. Foundations and trends in machine learning, 2009, 2(1): 1–127. DOI: 10.1561/2200000006
- [11] BOURTSOULATZE E, BURTH KURKA D, GUNDUZ D. Deep joint source-channel coding for wireless image transmission [J]. IEEE transactions on cognitive communications and networking, 2019, 5(3): 567–579. DOI: 10.1109/tccn.2019.2919300
- [12] GUNDUZ D, QIN Z J, AGUERRI I E, et al. Beyond transmitting bits: context, semantics, and task-oriented communications [J]. IEEE journal on selected areas in communications, 2022, 41(1): 5–41. DOI: 10.1109/jsac.2022.3223408
- [13] 张平. 语义通信：未来通信系统的智能之道 [J]. 中兴通讯技术, 2024, 30(1): 1–2. DOI: 10.12142/ZTETJ.202401001
- [14] 吕守晔, 戴金晟, 张平. 信源信道联合的新范式：语义通信 [J]. 中兴通讯技术, 2023, 29(2): 2–8. DOI: 10.12142/ZTETJ.202302002
- [15] LUO X W, CHEN H H, GUO Q. Semantic communications: overview, open issues, and future research directions [J]. IEEE wireless communications, 2022, 29(1): 210–219. DOI: 10.1109/mwc.101.2100269
- [16] DAI J C, WANG S X, TAN K L, et al. Nonlinear transform source-

- channel coding for semantic communications [J]. IEEE journal on selected areas in communications, 2022, 40(8): 2300–2316. DOI: 10.1109/jsac.2022.3180802
- [17] WANG S X, DAI J C, LIANG Z J, et al. Wireless deep video semantic transmission [J]. IEEE journal on selected areas in communications, 2022, 41(1): 214–229. DOI: 10.1109/jsac.2022.3221977
- [18] ERDEMIR E, TUNG T Y, DRAGOTTI P L, et al. Generative joint source–channel coding for semantic image transmission [J]. IEEE journal on selected areas in communications, 2023, 41(8): 2645–2657. DOI: 10.1109/jsac.2023.3288243
- [19] TAN K L, DAI J C, LIU Z Y, et al. Rate–distortion–perception controllable joint source–channel coding for high–fidelity generative semantic communications [J]. IEEE transactions on cognitive communications and networking, 2024, 11(2): 672–686. DOI: 10.1109/tccn.2024.3511960
- [20] WANG S X, DAI J C, TAN K L, et al. DiffCom: channel received signal is a natural condition to guide diffusion posterior sampling [J]. IEEE journal on selected areas in communications, 2025, 43(7): 2651–2666. DOI: 10.1109/jsac.2025.3559158
- [21] BLAU Y, MICHAELI T. Rethinking lossy compression: the rate–distortion–perception tradeoff [EB/OL]. (2019–07–30)[2025–05–20]. <https://arxiv.org/abs/1901.07821>
- [22] WASSILIWIZKY E, MENNINGHAUS W. Why and how should cognitive science care about aesthetics? [J]. Trends in cognitive sciences, 2021, 25(6): 437–449. DOI: 10.1016/j.tics.2021.03.008
- [23] MIN X K, GU K, ZHAI G T, et al. Screen content quality assessment: overview, benchmark, and beyond [J]. ACM computing surveys, 2022, 54(9): 1–36. DOI: 10.1145/3470970
- [24] YU Q, WEBER M, DENG X, et al. An image is worth 32 tokens for reconstruction and generation [EB/OL]. (2024–06–11)[2025–05–21]. <https://arxiv.org/abs/2406.07550>
- [25] YAO J, YANG B, WANG X. Reconstruction vs. generation: taming optimization dilemma in latent diffusion models [EB/OL]. [2025–05–21]. <https://cvpr.thecvf.com/virtual/2025/poster/33084>
- [26] QIN H L, DAI J, WANG S, et al. Neural coding is not always semantic: towards the standardized coding workflow in semantic communications [EB/OL]. (2025–05–24) [2025–05–25]. <https://arxiv.org/abs/2505.18637>
- [27] LAN Q, WEN D Z, ZHANG Z Z, et al. What is semantic communication? a view on conveying meaning in the era of machine intelligence [J]. Journal of communications and information networks, 2021, 6(4): 336–371. DOI: 10.23919/jcin.2021.9663101
- [28] YANG W T, LIEW Z Q, LIM W Y B, et al. Semantic communication meets edge intelligence [J]. IEEE wireless communications, 2022, 29(5): 28–35. DOI: 10.1109/mwc.004.2200050
- [29] YANG W T, XIONG Z H, YUAN Y L, et al. Agent–driven generative semantic communication with cross–modality and prediction [J]. IEEE transactions on wireless communications, 2024, 24(3): 2233–2248. DOI: 10.1109/twc.2024.3519325
- [30] TSENG W C, WANG T H J, LIN Y C, et al. Offline multi–agent reinforcement learning with knowledge distillation [EB/OL]. (2022–11–28) [2025–05–25]. <https://dl.acm.org/doi/10.5555/3600270.3600287>
- [31] SYKORA Q, REN M, URTASUN R. Multi–agent routing value iteration network [EB/OL]. (2020–07–09) [2025–05–28]. <https://arxiv.org/abs/2007.05096>
- [32] LIU Y, WANG W X, HU Y J, et al. Multi–agent game abstraction via graph attention neural network [J]. Proceedings of the AAAI conference on artificial intelligence, 2020, 34(5): 7211–7218. DOI: 10.1609/aaai.v34i05.6211
- [33] CHEN S K, DONG J Q, HA P Y J, et al. Graph neural network and reinforcement learning for multi–agent cooperative control of connected autonomous vehicles [J]. Computer–aided civil and infrastructure engineering, 2021, 36(7): 838–857. DOI: 10.1111/mice.12702
- [34] MAO H Y, LIU W L, HAO J Y, et al. Neighborhood cognition consistent multi–agent reinforcement learning [J]. Proceedings of the AAAI conference on artificial intelligence, 2020, 34(5): 7219–7226. DOI: 10.1609/aaai.v34i05.6212
- [35] CHEN F, REN W. On the control of multi–agent systems: a survey [J]. Foundations and trends in systems and control, 2019, 6(4): 339–499. DOI: 10.1561/26000000019
- [36] YANG D, YANG K, WANG Y, et al. How2comm: communication–efficient and collaboration–pragmatic multi–agent perception [J]. Advances in neural information processing systems, 2023, 36: 25151–25164

作者简介



戴金晟，北京邮电大学副教授、博士生导师；主要研究方向为智能网络与智能传输的基础理论与关键技术；主持包括国家重点研发计划、国家自然科学基金、北京市自然科学基金在内的多个科研项目；入选北京市科技新星、中国科协青年人才托举工程、小米青年学者等；发表论文 100 余篇。



秦晓琦，北京邮电大学副教授，国家级青年人才，中国工程院“中国工程前沿杰出青年学者”；主要研究方向为智能机器通信基础理论与关键技术；主持包括国家重点研发计划项目、国家自然科学基金面上项目、北京市自然科学基金项目等多个科研项目；获省部级和全国学会级科学技术一等奖 3 项，获最佳论文奖 3 项；已发表论文 130 余篇，参与制定国际标准 2 项、行业标准 4 项。



秦海龙，北京邮电大学在读研究生；主要研究方向为三维视觉建模、智能编码传输。



张平，中国工程院院士，北京邮电大学教授、网络与交换技术全国重点实验室主任，《通信学报》主编，IEEE Fellow，IMT–2020（5G）专家组成员，IMT–2030（6G）推进组咨询委员会委员，国家自然科学基金委“创新研究群体”带头人；长期致力于移动通信理论研究和科技创新，为中国自主技术成为国际主流做出了基础性的贡献；曾任国家重点基础研究发展计划（“973”计划）首席科学家、国家高技术研究发展计划（“863”计划）主题专家组专家等；获国家科技进步奖特等奖、一等奖，国家技术发明奖二等奖 3 项，科技进步奖二等奖 2 项，全国创新争先奖，光华工程科技奖等奖项，入选教育部首批“黄大年式教师团队”。

面向铁路入侵检测的语义通信技术



Semantic Communication Technology for Railway Intrusion Detection

郭疆远/GUO Jiangyuan, 陈为/CHEN Wei, 艾渤/AI Bo

(北京交通大学, 中国 北京 100044)
(Beijing Jiaotong University, Beijing 100044, China)

DOI: 10.12142/ZTETJ.202504003

网络出版地址: <http://kns.cnki.net/kcms/detail/34.1228.TN.20250715.1637.002.html>

网络出版日期: 2025-07-16

收稿日期: 2025-05-25

摘要: 针对铁路入侵检测中海量视频数据传输效率低、检测精度低的问题, 提出了一种基于视频Transformer的自适应语义通信框架。该框架通过传输与任务高度相关的语义特征并集成信道自适应模块, 实现面向特定任务的高效且鲁棒的语义信息传输。在构建的铁路真实场景视频数据集上进行实验, 将该框架与传统的视频分离编码及视频联合编码方法进行对比。结果表明, 所提出的自适应语义通信框架在不同的高斯白噪声及多种衰落信道条件下均能取得更高的入侵检测精度, 并在低信噪比和复杂信道环境中展现出优越的鲁棒性和性能增益, 为提升铁路智能监控系统的视频分析与传输效能提供了新的技术途径。

关键词: 入侵检测; 语义通信; 深度联合信源信道编码; 信道自适应

Abstract: To address the challenges of low efficiency in transmitting massive video data and suboptimal detection accuracy under complex channel conditions for railway intrusion detection, an adaptive semantic communication framework based on Video Transformer is proposed. This framework achieves efficient and robust task-oriented semantic information transmission by conveying highly task-relevant semantic features and integrating a channel-adaptive module. Experiments conducted on a constructed real-world railway scenario video dataset show that, compared with traditional separate video coding and joint video coding methods, the proposed framework attains higher intrusion detection accuracy under various additive white Gaussian noise and diverse fading channel conditions. Furthermore, it exhibits superior robustness and performance gains in low signal-to-noise ratio and complex channel environments, offering a novel technical approach for enhancing the video analysis and transmission efficacy of intelligent railway monitoring systems.

Keywords: intrusion detection; semantic communication; deep joint source-channel coding; channel adaptation

引用格式: 郭疆远, 陈为, 艾渤. 面向铁路入侵检测的语义通信技术 [J]. 中兴通讯技术, 2025, 31(4): 13-18. DOI: 10.12142/ZTETJ.202504003

Citation: GUO J Y, CHEN W, AI B. Semantic communication technology for railway intrusion detection [J]. ZTE technology journal, 2025, 31(4): 13-18. DOI: 10.12142/ZTETJ.202504003

近年来, 人工智能、物联网和新一代无线通信技术的迅猛发展, 推动了铁路系统的数字化与智能化的深度变革。在“智慧铁路”“数智交通”战略引导下, 铁路行业正加快迈向高度自动化、感知化与智能决策的新阶段^[1]。其中, 视频监控系統作为铁路感知体系的关键组成部分, 广泛部署于站台、咽喉、线路、道口等重点区域, 承担着入侵检测、施工监管、车流监控等重要任务。在这些智能任务中, 海量视频数据的实时、高效传输与分析处理已成为制约系统性能提升的核心瓶颈。一方面, 现有通信系统大多仍以传统

比特级传输为设计理念, 将视频看作像素流处理, 忽视了其承载的“任务语义信息”; 另一方面, 铁路场景中的通信链路环境复杂, 受地形遮挡、高速移动、异构终端功耗等因素的影响, 通信信道常处于低质量状态, 这进一步影响了数据传输的稳定性与下游智能任务的精度。

在铁路智能感知任务中, 入侵检测作为保障行车安全的核心应用之一, 正日益受到重视。该任务旨在及时识别并报警非法进入铁路防护区的人员、动物或障碍物, 避免异物入侵导致的行车事故。入侵检测系统通常依赖于部署在站台边缘、区间线路、道口等重点区域的高清摄像头, 通过对视频流的持续分析, 实现全天候、全场景的智能监控。然而, 铁路入侵检测面临诸多挑战: 一方面, 入侵事件通常具有突发

基金项目: 国家自然科学基金项目 (U2468201, 6221001)

性和稀疏性，要求系统具备更强的视频内容识别与理解能力；另一方面，系统需要在资源受限的边缘设备上实现低时延、高精度的实时检测，并通过通信链路将关键警报信息快速回传至控制中心。因此，实现高效准确的入侵检测，在保障检测精度的同时提升系统效率，是支撑铁路智能安防防控体系的关键技术支柱。

1 铁路入侵检测面临的挑战

传统的铁路入侵检测方法主要涵盖图像差分法和光流法，它们凭借像素域或光流域中的先验知识来检测入侵^[2]，但容易造成次优的检测性能。基于深度学习的入侵检测方法因其巨大潜力备受关注，可基于单帧图像或连续视频帧实现目标检测。基于图像的入侵检测^[3]借助卷积神经网络等神经网络处理特定关键帧来实现入侵识别。而基于视频的入侵检测能够剖析带有额外时间维度的连续视频帧，可提升入侵检测精度，但它依赖复杂的网络来处理视频中的时间关系，这使得在计算和存储资源受限的铁路监控设备上部署与运行该网络面临诸多挑战。

为了应对上述挑战，一种可行的解决方案是将计算负担卸载到云端服务器。在此方案中，监控摄像头捕获的原始视频通过无线信道传输至云端服务器，由其执行后续视频处理和入侵检测。然而，该方式需将完整视频逐帧传输至云端，这不仅带来巨大的带宽压力，而且由于大量图像内容（如空轨道、背景区域）与检测任务无关，存在显著冗余，导致通信效率极其低下。同时，传统视频通信系统普遍采用“视频压缩—信道传输—视频重建—智能识别”的串行架构，其视频压缩算法（如H.264、H.265）主要围绕像素误差进行优化，关注图像客观质量而非任务相关性，导致压缩过程中保留大量无用信息。在低信噪比条件下，基于该架构的视频重建质量大幅下降，进一步降低了后续识别算法（如目标检测、行为分析）的性能，容易引发漏检、误报等问题。此外，视频重建范式对接收端算力要求较高，接收端需顺序完成图像重建与智能推理，造成系统处理链路长、响应延迟高，且任一环节失效均可能影响整体识别效果。由此可见，仅依靠提升压缩效率或增强链路可靠性，难以从根本上解决传统视频通信系统带来的性能瓶颈问题。

作为一种融合通信技术与人工智能的新型范式，语义通信正逐渐受到学术界和产业界的广泛关注^[4-5]。该范式主张在通信过程中

“只传递对任务有意义的信息”，本质上是从“面向信息的压缩”向“面向任务的压缩”的转变。通过在发送端提取任务相关的语义特征，语义通信可以有效减少传输数据量，降低带宽压力，同时保障下游智能任务的精度^[6]。目前，语义通信技术已经在图像分类、目标识别、语音理解等任务中展示出优秀的性能^[7-8]。然而，在铁路这样的强场景约束、高安全要求、高移动特性环境中，语义通信的设计仍面临一系列挑战，例如：语义的有效提取、端侧计算能力的匹配、信道质量波动下的鲁棒性保障等问题，亟需结合场景需求进行定制化设计。

本文将新兴的语义通信技术引入铁路场景并聚焦于铁路入侵检测这一典型智能任务，提出了一种面向铁路入侵检测的语义通信传输框架。该框架结合深度联合信源信道编码技术^[9]，以视频Transformer作为骨干网络，在保障传输效率的同时提升任务识别精度。通过集成一个基于注意力机制的信道自适应模块，该框架可以有效地适应不同信道类型与信道质量，以提升铁路复杂无线环境下的传输鲁棒性。通过构建铁路咽喉区实景视频数据集，本文对所提出框架的检测准确率与传输性能进行了系统评估，验证了该方法在典型铁路通信场景中的有效性。

2 基于语义通信的传输框架设计

2.1 系统模型

本文所提出的传输框架主要包含3个核心组件：

- 1) 发送端语义编码器：部署在前端摄像头设备或轻量化边缘节点中，负责提取原始视频帧中与任务相关的语义特征，并编码为嵌入向量；
- 2) 信道自适应传输机制：通过注意力机制将嵌入向量映射为适配无线信道特点的传输符号，增强抗干扰能力；
- 3) 接收端语义解码器：部署于车站云端服务器或调度中心，负责从接收到的传输符号中还原出语义表示，并完成入侵检测任务。

具体传输流程如图1所示，主要分为3步：

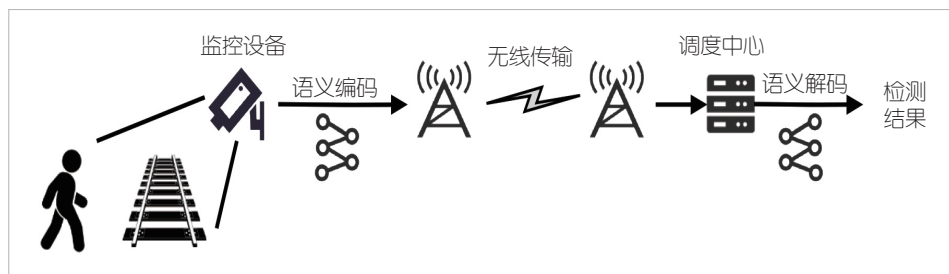


图1 面向铁路入侵检测的语义通信传输框架

1) 首先, 发送端捕获的监控视频可以表示为 $\mathbf{x} \in \mathbb{R}^{3 \times t \times h \times w}$, 其中 t 表示时间步长, 即视频帧的数量, h 表示视频的高度, w 表示视频的宽度。发送端部署的语义编码器包含 N_1 个视频 Transformer 基本处理块。与视觉 Transformer 的处理类似, 为了降低计算量, $\mathbf{x}$ 首先沿着时间和空间维度被划分为 l 个不重叠的尺寸为 $1 \times p \times p$ 的补丁块, 然后被输入至后续神经网络中处理。发送端利用语义编码器, 结合通过信道估计等方法获取的先验信道信息 (信噪比、信道类型等) 实现动态编码, 从原始视频 $\mathbf{x}$ 中提取任务相关的深层视频语义表示 $\mathbf{y} \in \mathbb{R}^{d \times t \times l}$, 其中 d 为视频语义表示的通道数。在发送端的处理过程可以表示为:

$$\mathbf{y} = E(\mathbf{x}, \boldsymbol{\mu}) \quad (1),$$

其中, $E(\cdot)$ 为语义编码函数, $\boldsymbol{\mu}$ 为先验信道信息。随后, 通过将提取到的深层视频语义表示 $\mathbf{y}$ 的每两个实数值组合为一个复数符号, 以得到信道输入信号 $\mathbf{s} \in \mathbb{C}^m$ (m 为信道传输符号数量)。假设一个符号的发送功率为 P , 在进行实际传输之前, 需要对 $\mathbf{s}$ 进行功率归一化操作以满足发送端的平均功率约束, 即:

$$\frac{1}{m} \mathbf{s} \mathbf{s}^* = P \quad (2).$$

2) 发送端将归一化后的信道输入信号 $\mathbf{s}$ 以模拟信号形式直接通过无线噪声信道发送至接收端。

本文考虑两种无线信道下的通信过程, 即加性高斯白噪声信道与衰落信道。加性高斯白噪声信道的传输过程可以表示为:

$$\hat{\mathbf{s}} = \mathbf{s} + \mathbf{n} \quad (3),$$

其中, $\hat{\mathbf{s}}$ 为信道输出信号, $\mathbf{n} \in \mathbb{C}^m$ 为噪声向量, 包含 m 个遵从复高斯分布 $\mathcal{CN}(0, \sigma^2 \mathbf{I})$ 的独立同分布元素, σ^2 表示噪声功率, $\mathbf{I}$ 为单位矩阵。衰落信道的传输过程可以表示为:

$$\hat{\mathbf{s}} = \mathbf{h} \mathbf{s} + \mathbf{n} \quad (4),$$

其中, $\mathbf{h} \in \mathbb{C}^l$ 表示信道增益。

3) 接收端首先接收信道输出信号 $\hat{\mathbf{s}}$, 并将每个复数符号分解为两个实数值以转换为重建的语义表示 $\hat{\mathbf{y}} \in \mathbb{R}^{d \times t \times l}$ 。随后, 接收端通过部署的语义解码器对 $\hat{\mathbf{y}}$ 进行语义解码, 最终通过一个由线性映射层构成的分类头得到识别为入侵/非入侵的概率。与语义编码器类似, 语义解码器包含 N_2 个视频 Transformer 基本处理块, 并结合与发送端共享的先验信道信息实现动态解码。在接收端的处理过程可以表示为:

$$\mathbf{p} = D(\hat{\mathbf{y}}, \boldsymbol{\mu}) \quad (5),$$

其中, $D(\cdot)$ 为语义解码函数, $\boldsymbol{\mu}$ 为先验信道信息, $\mathbf{p} \in \mathbb{R}^2$ 为入侵检测概率向量, 表示对应视频被检测为入侵或非入侵的可能性。具有最高可能性的类别将作为入侵检测结果 c :

$$c = \operatorname{argmax}_i \mathbf{p}_i \quad (6),$$

其中, $\mathbf{p}_i$ 表示 $\mathbf{p}$ 的第 i 个元素。

2.2 基于视频 Transformer 的语义编解码

在铁路智能监控系统中, 入侵检测的核心挑战在于如何从复杂动态的视频流中快速、准确地识别稀疏且突发的入侵目标 (如人员、车辆或障碍物)。传统方法依赖卷积神经网络或光流, 但受限于局部感受野与计算冗余, 难以兼顾时空维度实现联合建模。为此, 本文提出一种基于视频 Transformer 的语义编解码网络, 通过多层次的自注意力机制, 实现对铁路监控视频的高效语义特征提取与任务导向的压缩传输。

给定第 i 层输入特征 $\mathbf{y}_i$, 输出特征 $\mathbf{y}_{i+1}$ 由 $\mathbf{y}_i$ 经一个完整的视频 Transformer 块处理获得, 如图 2 所示。每个视频 Transformer 块包含 3 个核心组件: 空间自注意力模块、时间自注意力模块以及前馈网络。其设计目标是通过多粒度特征

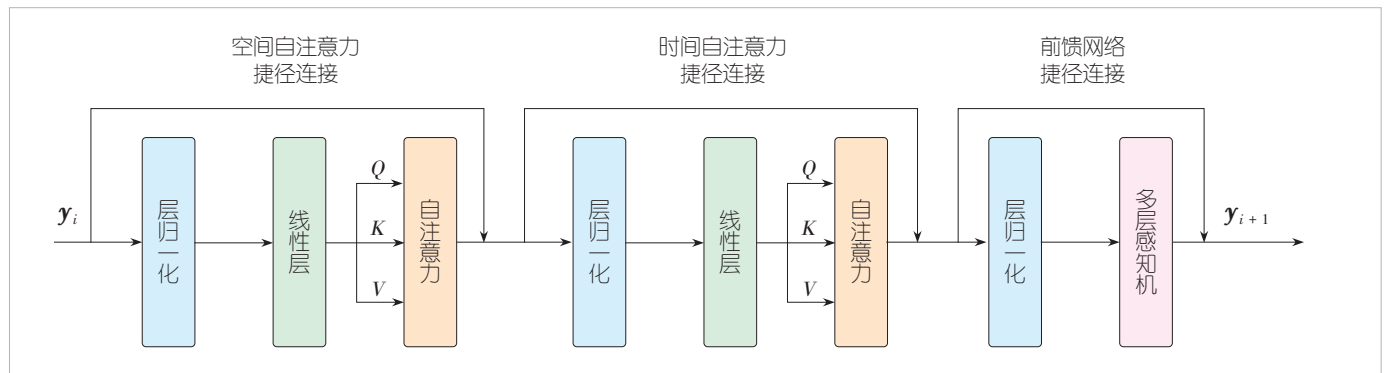


图2 视频Transformer块的网络结构

交互，精准捕捉铁路场景中入侵目标的静态形态与动态行为。其中，空间自注意力模块在空间维度实现注意力操作，聚焦单帧内的局部与全局空间关系，提取入侵目标的形态、位置等静态特征。空间自注意力执行完毕后，引入时间自注意力以建模连续帧间的时序依赖关系，分析入侵目标的运动轨迹与行为模式。最后，利用前馈网络融合时空特征，提升模型的非线性建模能力。考虑到边缘设备存在的算力限制，在本文所提出的传输架构中，发送端部署 $N_1 = 2$ 个视频 Transformer 基本处理块，而接收端部署 $N_2 = 10$ 个视频 Transformer 基本处理块。

2.3 基于注意力机制的信道自适应

铁路环境下的无线信道本质上是复杂且动态变化的，其特性受到多种因素的综合影响，表现为信噪比的时变性以及衰落特性的多样性。这种信道的不确定性和时变性给视频语义信息的可靠传输带来了严峻挑战，严重影响了下游入侵检测任务的精度和稳定性。

为了解决这一核心难题，本文为所提出的语义通信框架集成了一套信道自适应传输机制。该机制的核心目标在于赋予通信系统感知和适应信道变化的能力，通过在发送端和接收端共享并利用先验信道信息，动态优化语义信息的编码和解码，从而在多变的信道条件下最大程度地保障通信的鲁棒性和语义信息的完整性与准确性。此处的先验信道信息不仅包括了量化的信道质量指标，如瞬时或平均信噪比，还涵盖了对信道类型的定性判断，例如当前链路符合的衰落模型。

所提出的信道自适应方法包含一个基于注意力机制的即插即用模块，该模块以在收发双方共享的信道先验信息为输入，实时生成一组缩放向量或仿射变换参数。这些参数用于动态地调整语义编码器与解码器内部的中间层特征，使整个通信系统能够动态地适应当前信道的特定状况。

具体而言，该信道自适应模块所利用的先验信道信息及其处理方式会根据当前无线信道的主要特性进行区分：

1) 针对加性高斯白噪声信道：在该信道条件下，其相对简单且主要受噪声功率影响，信道自适应模块所利用的关键先验信息是瞬时或平均信噪比。信噪比直接反映了信号强度与噪声强度之间的对比关系，是衡量此类信道质量的核心指标。

2) 针对更复杂的衰落信道：对于由多径效应主导，导致信号包络快速且剧烈起伏的衰落信道，仅仅依赖单一的信噪比信息，往往难

以全面而准确地刻画信道的动态特性和对信号的实际影响。信噪比主要反映的是信号平均功率与噪声平均功率的对比，但无法充分体现衰落信道中的信号时变特性。因此，为了更有效地应对这类复杂衰落信道带来的挑战，在继续利用实时信噪比作为重要的信道质量参考的同时，额外引入了“信道类型标签”这一关键的上下文信息。这个标签并非简单的数值，而是对当前信道所属的具体衰落模型（如瑞利、莱斯等）或其所处的典型物理场景特征（如城市峡谷、开阔地、隧道内等）的定性描述或分类。引入信道类型标签，系统能够获得关于信道本质特征的更深层次理解，为后续的自适应调整提供更为精准的依据。

信道自适应模块的结构如图3所示。给定第 $i-1$ 层视频 Transformer 块输出特征 $y_i \in \mathbb{R}^{d_i \times t_i \times l_i}$ ， d_i 表示特征通道数， $t_i \times l_i$ 表示特征分辨率。首先利用多层感知机处理信噪比及信道类型等先验信道信息，生成缩放因子。其中，信道标签 r 为一个整数值，代表对应的信道类型（5个信道类型则对应1—5的信道标签）。信噪比（SNR）为一个浮点数值，表示对应的信道质量。二者首先被合并在一起，组成先验信道信息 μ ，表示为：

$$\mu = \text{concat}(\text{SNR}, r) \in \mathbb{R}^2 \quad (7),$$

其中， $\text{concat}(\cdot)$ 表示合并操作。 μ 经多层感知机处理得到缩放因子 $\alpha \in \mathbb{R}^{d_i}$ 与 $\beta \in \mathbb{R}^{d_i}$ ，表示为：

$$[\alpha, \beta] = \text{MLP}(\mu) \in \mathbb{R}^{2d_i} \quad (8),$$

其中， $\text{MLP}(\cdot)$ 表示由两个线性映射层以及中间的非线性激活构成的多层感知机。

随后， α 、 β 被扩展为 $\mathbb{R}^{d_i} \rightarrow \mathbb{R}^{d_i \times 1 \times 1}$ 并与中间层特征 y_i 进行通道维度的点积与求和，生成调整后的特征 $\tilde{y}_i \in \mathbb{R}^{d_i \times t_i \times l_i}$ 。基于注意力机制的调整过程可以表示为：

$$\tilde{y}_i = \alpha \times y_i + \beta \quad (9),$$

其中， $\times$ 和 $+$ 分别表示通道维度的点积与求和操作。

本文中，每个视频 Transformer 基本处理块后均嵌入一个

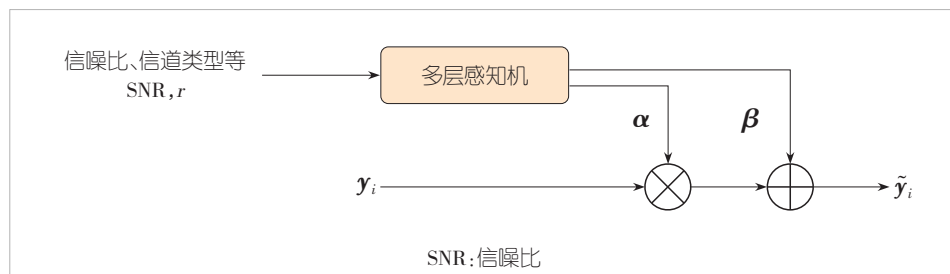


图3 信道自适应模块的结构

信道自适应模块以调整中间特征,从而使得语义编码和语义解码过程能够适配当前信道状态。每一层的处理包含一个视频Transformer块以及一个信道自适应模块。

3 实验结果与分析

3.1 实验设置

本文构建了一个真实的视频监控数据集,以全面评估所提方法性能。数据采集自中国境内65个实际铁路咽喉区(通过云台摄像机在2018年12月—2019年4月采集的监控视频)。这些区域因轨道分支和交叉复杂,安全风险较高。数据集包含9 197个样本,每个样本由8个连续帧构成,并统一调整分辨率至 224×224 像素,用于训练与测试。样本以5:1的比例划分为训练集和测试集,确保入侵与非入侵样本在各数据集中数量均衡。视频Transformer模块中,时空自注意力机制均设置8个注意力头,每个头通道数为64,语义特征的通道数为 $d = 256$ 。

为验证所提出方法的有效性,本文与多种基准方法对比,包括理想化的无差错传输、传统的H.265视频编码结合信道容量传输方案(使用FFmpeg进行H.265编码),以及一种先进的基于深度联合信源信道编码的视频传输方法^[10]。同时,我们还对比了移除信道自适应模块的基础传输框架,以凸显信道自适应模块的性能增益。

本文主要考虑了高斯白噪声信道以及3GPP中定义的3种衰落信道下的视频语义传输。所有模型均在0~20 dB的混合信噪比条件下进行训练和测试,且在性能评估时各方法使用相同带宽资源(传输符号数量一致)。训练中,采用分类任务常见的交叉熵损失函数优化模型参数。

3.2 实验分析

图4展示了不同传输方法在高斯白噪声信道下的入侵检测准确率对比。实验结果表明,本文提出的语义通信框架显著优于依赖视频重建的传统通信方案。具体而言,提出的自适应传输方法在整个0~20 dB的信噪比区间内均表现最佳,即使在0 dB的恶劣条件下,准确率仍高达84.8%,并随信噪比改善迅速逼近约88%的理想传输上限。相比之下,传统的H.265编码结合信道容量的方案(视频分离编码方案)和先进的视频联合编码方法在低信噪比时性能严重恶化(0 dB时准确率分别仅为约66%和69%)。即使在高信噪比下,其准确率上限也与本文方法存在明显差距。这种性能差异源于设计理念的根本不同:传统基于重构的方法侧重于像素级保真,不仅对信道噪声敏感,还传输了大量与入侵检测任务无

关的冗余信息,导致任务精度受损。而本文的语义通信框架通过端到端优化,仅提取和传输与任务高度相关的语义特征,极大地提升了通信效率和噪声鲁棒性。此外,信道自适应机制可以使系统更有效地利用信道信息,以克服信道迅速恶化带来的影响。

图5展示了在信道描述语言(CDL)系列衰落信道及不同信噪比条件下,多种训练策略及是否利用信道类型先验信息对入侵检测性能的影响。首先,混合信道类型训练能有效提升模型的泛化能力。在不输入类型标签的情况下,经混合类型场景训练的模型准确率普遍高于仅在单一特定信道类型(CDL-A、C、D)下训练的模型,这表明多样化训练数据有助于学习更鲁棒的普适特征。其次,当模型能获取并利用当前信道的类型标签时,性能会得到显著提升。输入类型标签

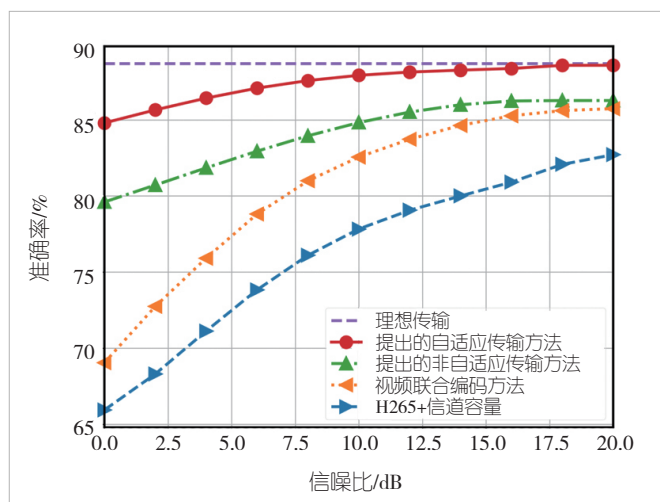


图4 高斯白噪声信道下不同传输方法的性能比较

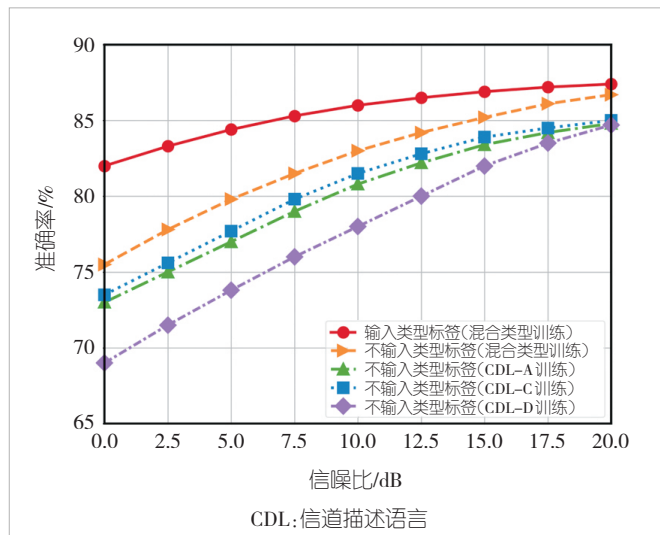


图5 衰落信道下不同训练方法的性能比较

且经混合类型训练的模型,相较于所有不使用类型标签的模型(包括同样混合类型训练但不输入标签的模型),在各信噪比下均取得了更高的检测精度。这说明类型标签作为精确的信道先验,能引导模型进行针对性优化,有效克服特定衰落带来的影响。此外,在特定类型(CDL-A、C、D)下训练的模型间性能差异,反映了不同衰落场景本身的学习难易度与信道特性不同,进一步说明类型标签能够辅助模型训练,从而带来性能增益。因此,赋予模型场景感知能力并结合覆盖多种信道条件的混合训练,是确保在复杂衰落信道中实现高效、鲁棒入侵检测的关键。

4 结束语

本文面向铁路场景下对系统安全至关重要的入侵检测任务,提出了一种基于视频Transformer的自适应语义通信框架。该框架通过直接提取并传输与入侵检测任务高度相关的语义特征,并结合信道自适应机制(考虑信噪比及信道类型先验),有效克服了传统视频传输方法在高带宽需求、信道噪声敏感以及任务信息冗余等方面的局限性。实验结果表明,无论在高斯白噪声信道还是复杂的CDL衰落信道下,本文方法均显著优于传统的视频分离编码方案及先进的视频联合编码方法。在各种信噪比条件下,本文方法均能以更低的传输开销实现更高的入侵检测精度,展现出强大的鲁棒性和传输效率。这项工作为铁路智能化系统的发展提供了新的技术思路,展现了语义通信在提升铁路场景智能任务性能方面的巨大潜力。

参考文献

- [1] 钟章队, 官科, 陈为, 等. 铁路新一代移动通信的挑战与思考 [J]. 中兴通讯技术, 2021, 27(4): 44-50. DOI: 10.12142/ZTETJ.202104009
- [2] CAO Z W, QIN Y, JIA L M, et al. Railway intrusion detection based on machine vision: a survey, challenges, and perspectives [J]. IEEE transactions on intelligent transportation systems, 2024, 25(7): 6427-6448. DOI: 10.1109/tits.2024.3412170
- [3] GONG X, CHEN X, ZHONG Z, et al. Enhanced few-shot learning for intrusion detection in railway video surveillance [J]. IEEE transactions on intelligent transportation systems, 2021, 23(8): 11301-11313
- [4] 吕守晔, 戴金晟, 张平. 信源信道联合的新范式: 语义通信 [J]. 中兴通讯技术, 2022, 27(2): 2-8. DOI: 10.12142/ZTETJ.202302002
- [5] 施雨轩, 吴泳澎, 张文军. 基于信息论的语义通信: 理论与挑战 [J]. 中

- 兴通讯技术, 2023, 29(2): 13-18. DOI: 10.12142/ZTETJ.202302004
- [6] GUO J Y, CHEN W, SUN Y X, et al. VideoQA-SC: adaptive semantic communication for video question answering [J]. IEEE journal on selected areas in communications, 2025, 43(7): 2462-2477. DOI: 10.1109/jsac.2025.3559160
- [7] GUO L, CHEN W, SUN Y X, et al. Digital-SC: digital semantic communication with adaptive network split and learned non-linear quantization [EB/OL]. (2024-12-02) [2025-05-20]. <https://ieeexplore.ieee.org/document/10772628>
- [8] CHENG J M, CHEN W, LI J, et al. Efficient spatio-temporal predictive learning for massive MIMO CSI prediction [J]. ZTE Communications, 2025, 23(1): 3-10. DOI: 10.12142/ZTECOM.202501002
- [9] 许佳龙, 陈为, 艾渤. 基于深度联合信源信道编码的CSI反馈技术 [J]. 中兴通讯技术, 2022, 27(2): 29-33. DOI: 10.12142/ZTETJ.202302007
- [10] WANG S, DAI J, LIANG Z, et al. Wireless deep video semantic transmission [J]. IEEE journal on selected areas in communications, 2022, 41(1): 214-229. DOI: 10.1109/JSAC.2022.3221977

作者简介



郭疆远, 北京交通大学在读博士研究生; 主要研究方向为深度联合信源信道编码、语义通信; 发表论文3篇。



陈为, 北京交通大学教授; 长期从事无线通信、信号处理、人工智能技术研究, 并围绕高维信息感知和处理、海量机器类通信、语义通信、智慧交通开展研究工作; 作为负责人承担国家优秀青年科学基金等多项省部级及以上项目; 获IEEE通信学会杰出讲师、2022年度信息通信领域十大科技进展、首届詹天佑铁道科学技术奖秦驰道专项奖(青年奖)等荣誉。



艾渤, 北京交通大学教授、电子信息工程学院院长, 国家6G技术研发总体专家组专家, 国家自然科学基金委创新群体带头人, 国家杰出青年科学基金、优秀青年科学基金、牛顿高级学者基金、中国科协“求是杰出青年奖”获得者, 中共中央组织部“万人计划”领军人才, 北京市优秀教师, IEEE Fellow, IEEE VTS杰出讲师; 入选斯坦福大学2020-2022年全球前2%顶尖科学家终身影响力榜单, 获省部级科技奖励9项。

6G 数字孪生信道的三个使能技术: 多模态感知、环境知识和大模型



Three Enabling Technologies for 6G Digital Twin Channel: Multimodal Sensing, Environment Knowledge, and Large Model

于力/YU Li, 张建华/ZHANG Jianhua, 蔡逸辰/CAI Yichen

(北京邮电大学网络与交换技术全国重点实验室, 中国 北京 100876)
(State Key Lab of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China)

DOI: 10.12142/ZTETJ.202504004

网络出版地址: <http://kns.cnki.net/kcms/detail/34.1228.tn.20250711.1948.006.html>

网络出版日期: 2025-07-14

收稿日期: 2025-05-16

摘要: 未来6G移动通信系统中, 智能机器类通信(MTC)将在工业自动化、车联网等场景中发挥关键作用, 对通信系统中无线信道建模的精度、实时性与自适应能力提出了前所未有的挑战。数字孪生信道(DTC)作为一种新兴的信道表征范式, 能够在数字世界中构建物理信道的高保真映射, 为6G系统提供动态环境下的感知、预测与决策支持。为推动DTC从理论概念走向实际应用, 进一步提升其多模态数据融合以及多场景泛化能力, 首次将信道大模型(ChannelLM)引入DTC实现框架, 并围绕其核心架构展开关键技术设计与验证。具体而言, DTC架构由多模态环境感知与重构、无线环境知识(WEK), 以及可泛化的ChannelLM三大核心技术支撑, 分别实现对物理环境的高精度建模、环境-信道关系的可解释性构建, 以及基于知识驱动的信道预测与通信策略生成。MTC典型场景的实验结果表明, 所提方案在信道预测精度与场景泛化性方面显著提升, 为DTC技术在6G网络中的应用落地提供了有效的支撑路径。

关键词: 数字孪生信道; 多模态感知; 环境重构; 信道预测; 无线环境知识; 信道大模型

Abstract: In future 6G mobile communication systems, intelligent machine-type communication (MTC) will play a pivotal role in scenarios such as industrial automation and vehicular networks, posing unprecedented challenges to the accuracy, real-time performance, and adaptability of wireless channel modeling. As an emerging paradigm for channel representation, the digital twin channel (DTC) enables high-fidelity mapping of physical channels into the digital domain, providing sensing, prediction, and decision-making support for 6G systems in dynamic environments. To promote the transition of DTC from theoretical concept to practical engineering application, and to enhance its multimodal data fusion and cross-scenario generalization capabilities, this work integrates the channel large model (ChannelLM) into the DTC framework, focusing on the design and implementation of core technologies. Specifically, the DTC architecture is enabled by three key components: multimodal environment sensing and reconstruction, wireless environment knowledge (WEK) construction, and a generalizable ChannelLM. These components respectively support high-precision modeling of physical environments, interpretable construction of environment-channel relationships, knowledge-driven channel prediction, and communication strategy generation. Experimental results across typical MTC scenarios demonstrate that the proposed framework achieves high channel prediction accuracy and strong environmental adaptability, offering a viable pathway for the practical deployment of DTC in future 6G networks.

Keywords: digital twin channel; multimodal sensing; environment reconstruction; channel prediction; wireless environment knowledge; channel large model

引用格式: 于力, 张建华, 蔡逸辰. 6G 数字孪生信道的三个使能技术: 多模态感知、环境知识和大模型 [J]. 中兴通讯技术, 2025, 31(4): 19–28. DOI: 10.12142/ZTETJ.202504004

Citation: YU L, ZHANG J H, CAI Y C. Three enabling technologies for 6G digital twin channel: multimodal sensing, environment knowledge, and large model [J]. ZTE technology journal, 2025, 31(4): 19–28. DOI: 10.12142/ZTETJ.202504004

6G移动通信系统正朝着“泛在无线智能”的愿景^[1-3]加速演进。其中, 智能机器类通信(MTC)被广泛认为是支撑未来智慧城市、自动化工业、智能交通等关键应用场景的

重要通信范式^[4]。MTC强调设备间的高频交互与系统级协同, 对无线网络在空口时延、可靠性、连接密度等关键性能指标提出了更高要求。

然而, 面向6G MTC典型应用场景的无线环境往往复杂多变, 通信过程可控性差。例如, 工业互联网(IIoT)场景

基金项目: 国家自然科学基金项目(62401084, 62525101)

中的移动设备会引发严重的遮挡和时变的通信环境,设备材质电磁特性随频率与温度的改变而发生动态变化,导致信号进一步抖动^[5-7]。车载无线通信(V2X)场景则受到高速多普勒频移和路径遮挡影响,例如建筑物、车辆遮挡使通信环境变化频繁且难以预测,影响通信稳定性。

作为无线信号传播的物理介质,电磁波在无线环境中的传播特性将直接影响通信系统的整体性能。因此,如何在MTC场景中实现对信道的快速、准确、可泛化建模与预测,成为保障系统可靠性的核心课题。近年来,已有部分研究为面向6G的MTC信道表征奠定了初步基础,例如较低信令开销的鲁棒性用户检测和信道估计^[8]、基于边缘云无线接入网络架构的超可靠MTC信道建模^[9]、基于任务权重和用户需求的自适应分配资源等^[10]。但整体来看,当前仍缺乏一套可统一适配多种场景和通信任务的快速、精确信道建模范式。

在此背景下,本文首次提出了多模态环境感知、环境知识与人工智能(AI)大模型技术使能的数字孪生信道(DTC)架构,实现了MTC场景下信道表征的智能演进^[11-25]。通过将物理世界中的信道映射到数字空间,DTC可实现对环境变化与信道衰落状态的实时表征与响应,构建通信系统与环境之间的动态交互机制。这一体系具有精确建模、实时交互、自我更新、任务导向四大关键特性,能够全面复现物理世界状态,精准刻画动态信道演化过程,并在复杂多样异质

的6G MTC场景中,实现基于任务需求的在线预测与系统自适应优化。

所提DTC架构在支撑6G MTC网络方面展现出巨大潜力,但需要在理论框架的基础上进一步明确技术实现流程并完成系统验证。为此,本文进一步引入大模型能力,并围绕环境感知、环境-信道映射、信道预测等核心模块展开系统设计与技术实现。此外,本文在IIoT和V2X典型MTC场景下完成了对DTC的实验验证,有效证明了其在复杂动态环境下的可用性与实用性,为其在6G网络中的工程落地奠定了坚实的技术基础。

1 信道数字孪生架构

为实现面向6G MTC的高可靠低时延信道预测,所提出的DTC架构包含3项主要使能技术:多模态环境感知和重构、无线环境知识(WEK)和信道大模型(ChannelLM),如图1所示。

多模态环境感知与重构技术用于实现物理环境向数字空间的映射。系统通过激光雷达、相机、全球定位系统(GPS)等多种异构传感器获取来自不同模态、互为补充的环境感知数据,并将其融合还原为高精度、完整的三维场景模型。这一技术实现了DTC的精确建模能力,确保系统能够基于真实、细粒度的环境信息开展信道预测与通信策略

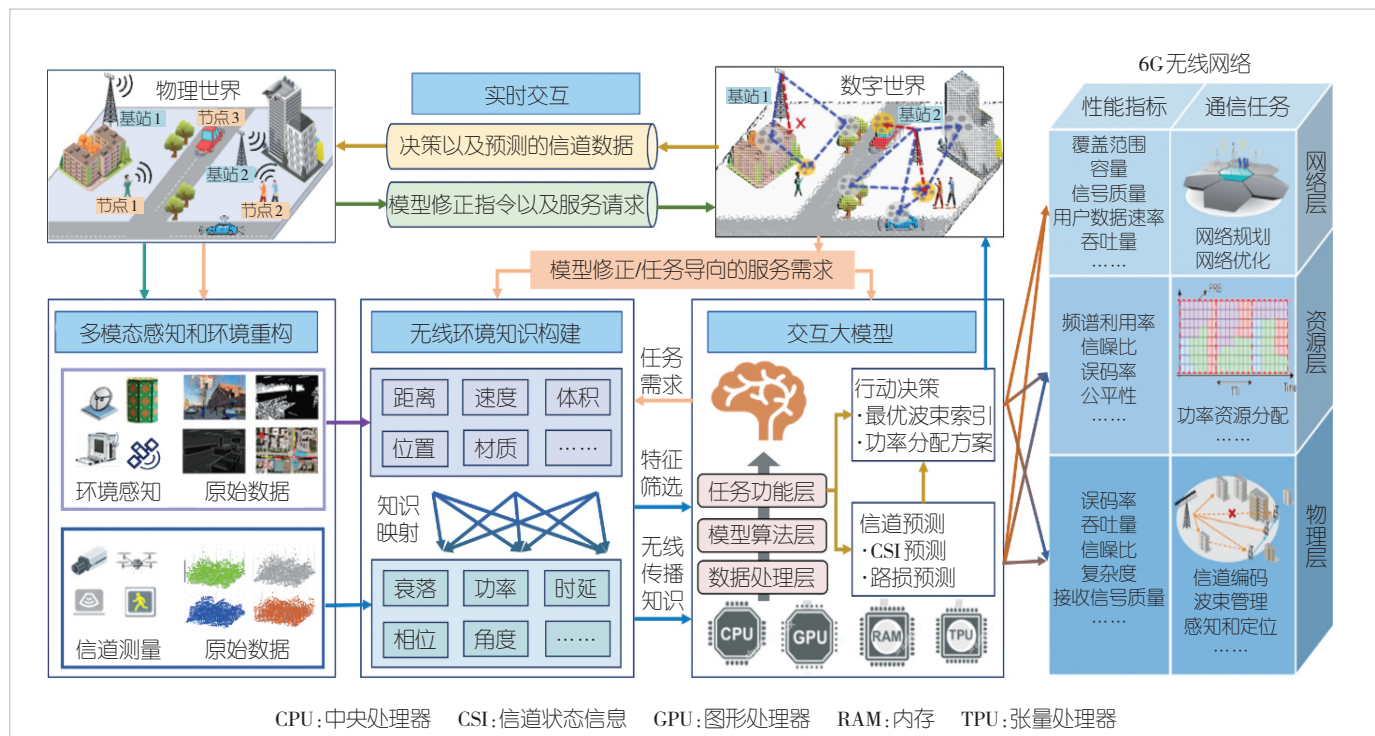


图1 多模态感知、无线环境知识和信道大模型驱动的数字孪生信道实现框架

优化。

WEK建立环境和信道之间本质、固化的映射关系,为后续基于环境信息的信道特性预测提供了基础。这一技术提升了DTC的可解释性、泛化性,并支撑了任务导向性,使其能够根据不同服务质量(QoS)需求进行调整。

然而,由于通信任务、场景、频段等给模型的推理和计算能力带来极大挑战,传统小模型在多模态数据融合、多任务适配等方面难以胜任。为此,本文首次提出ChannelLM以满足多场景泛化、多任务并行、多模态融合等需求。在基础数据与无线传播先验知识的支撑下,模型可实现信道的参数预测、波束选择与功率控制等关键任务。

同时,系统具备动态反馈机制,可根据实时环境状态调整模型推理结果,并对物理世界执行的通信策略进行修正,从而实现DTC的实时交互与自我更新能力。上述架构首次定义了DTC的功能分工。为实现该架构的工程化部署,本文将分别对3项核心使能技术的具体实现方式进行详细描述。

2 信道数字孪生关键技术实现

2.1 多模态感知和环境重构

为了实现对传播环境的准确建模,DTC依赖于多模态感知技术,通过多类传感器协同获取丰富、互补的环境信息,构建完整的三维场景模型。

2.1.1 无线环境的感知

随着感知技术的快速发展,通信系统可以通过单目相机、深度相机、激光雷达(LiDAR)等传感器采集包含丰富环境信息的感知数据。然而,单一模态的传感器在实际部署中往往面临局限性。

例如,单目相机可以捕捉物体的颜色、纹理等丰富的语义信息。然而,其数据缺乏深度信息,这意味着传播环境信息的不完整性。深度相机弥补了单目相机在尺度信息上的不足,能够感知三维空间信息。这对于通信系统在复杂环境中的信道建模和估计非常有帮助。但这类视觉传感器对光照条件较为敏感,夜间或强光照等环境下难以获取准确信息。相比之下,LiDAR作为一种主动传感器,通过发射激光脉冲生成三维点云数据,具有精细的空间分辨率,能够精确反映环境中物体的几何形状、尺寸和位置。此外,LiDAR不受光照条件影响,在极端照明条件下也能提供可靠的环境感知。然而,LiDAR的成本较高,且在远距离时点云数据稀疏,这可能限制其在大规模通信系统中的应用。

为克服上述局限,系统需引入多模态感知方法以获得最

大限度的感知视角,融合不同传感器的优势,减少感知盲区,提高系统的稳定性。在融合策略方面,常见的融合方式包含数据级、特征级和决策级的融合。其中,数据级的融合要求不同模态的数据在时间、空间上严格对齐,对多模态数据的固有错位较为敏感;而决策级融合则忽略了模态间潜在的相关性,难以挖掘中间特征中的深层信息。在此背景下,特征级融合被广泛应用于复杂场景中,它能够在保持信息有效性的基础上,捕捉不同模态之间的互补性,从而实现更优的融合效果。

除了感知数据本身的融合外,传感器的物理部署策略也对实现稳定、有效的环境感知具有重要影响。以视觉传感器为例,相机几何模型用于评估视觉传感器对目标场景的覆盖范围。可通过相机内参和视场角快速计算投影平面上的最大观测区域及其在三维空间中的对应深度范围,从而辅助安装位置优化,以提升覆盖效率并降低冗余。在进行LiDAR设备部署时,需要综合考虑视野遮挡、安装高度、数据处理能力以及环境适应性等多重因素,确保在关键区域实现稳定的三维感知覆盖。在给定系统参数下的最远有效探测距离表示为:

$$R_{\max} = \left(\frac{P_t G_t G_r \lambda^2}{(4\pi)^3 P_{\min} L} \right)^{\frac{1}{2}} \quad (1),$$

其中, P_t 为发射功率, G_t 、 G_r 为发射/接收天线增益, λ 为波长, $P_{\min}$ 为最小可检测功率, L 为路径损耗因子。由于LiDAR垂直扫描角固定,安装高度越高,其在地面的覆盖范围就越大,但点云密度会随之降低。因此,安装高度需根据应用需求在覆盖范围与感知精度之间权衡选择。

合理的系统设计和设备部署是确保环境感知系统稳定运行的基础,也为DTC架构的环境重构模块提供可靠支撑。图2展示了在实际环境中采集的多视角图像与LiDAR点云的数据结果。

2.1.2 无线环境的重构

在获取互补的多模态感知数据后,首先需要对物理传播环境进行建模与重构,以生成可用于信道特性分析的高精度三维场景模型。

已有研究表明,不同环境特征对信道参数具有显著影响。文献[26–28]指出,环境几何结构、散射体密度、天线布局及材料反射吸收特性共同决定信号的路径损耗、延迟扩展和多径分布。表1总结了常见的感知手段,环境特征及对应被影响的信道参数。



图2 北京亦庄自动驾驶区多模态环境感知:多视角图像与LiDAR点云

表1 常见的感知手段、环境信息特征及对应被影响的信道参数

感知手段	环境信息特征	影响的信道参数
深度相机、激光雷达	天线高度	路径损耗、时延扩展、角度扩展、莱斯K因子
深度相机、激光雷达、毫米波雷达、GPS、	Tx和Rx距离	路径损耗、莱斯K因子
深度相机、激光雷达	散射体尺寸	路径损耗、角度扩展
深度相机、激光雷达、毫米波雷达、GPS	散射体位置	路径损耗、角度扩展
摄像头、激光雷达	散射体密度	RMS DS、RMS AS、莱斯K因子
摄像头	散射体材质	路径损耗和时延扩展

AS:角度扩展 DS:时延扩展 GPS:全球定位系统 RMS:均方根 Rx:接收机 Tx:发射机

因此，环境重构不仅要还原三维几何形态和物理属性，更要为提取信道关键信息并将其映射为射线追踪或机器学习模型的输入做好准备。在多源传感器支持下，环境重构过程通常包括数据配准与格式转换、点云融合、深度图还原、图像-几何联合建模等环节。通过对视觉、激光和定位数据的统一处理，能够构建一个与物理世界对应的数字无线环境，进而实现精确的信道预测，并为可泛化大模型提供真实的场景描述，从而在数字孪生系统中支撑后续的智能化决策与优化。

2.2 无线环境知识构建

通过多模态感知与环境重构，系统获得了全面、精准的物理环境模型。然而，这一模型本身仍属于几何层级的信息表示，无法直接用于信道建模与系统优化。为了实现从环境到信道的有效映射，系统需进一步解构环境要素与信道特性之间的内在关系。为此，本文首次提出WEK，以便面向通信任务挖掘结构化、低维度、信道高相关性的环境特征，进而通过利用轻量化的神经网络实现无线环境到信道特性参数的可泛化、高精度推理预测。上述流程建立了环境特征到信道状态之间的可解释映射^[29-30]，从而构成了WEK的核心。

WEK通过理论推导、语义构建等可解释性的过程，挖掘并构建环境和信道之间的本质相关性。从广义角度看，

WEK被认为是从环境信息 $\Phi \in \mathbb{R}^D$ 到信道知识 $k \in \mathbb{C}^J$ 的可解释关系映射 $K(\cdot)$ 的集合，可表示为：

$$\mathcal{K} = \{K(\cdot): \mathbb{R}^D \rightarrow \mathbb{C}^J\} \tag{2},$$

其中， $\Phi \in \mathbb{R}^D$ 表示在物理世界中获得的 D 维环境信息的集合。 $\mathbb{R}$ 包含基站（BS）、用户终端（UT）和散射体的环境信息，例如位置、体积、材料、天线高度等。 D 是与场景相关的维度，随着不同类型的场景而变化。例如，在非视距（NLoS）条件下需要引入遮挡属性，导致其环境信息具有维度比视距（LoS）条件更大的描述维度。 $k \in \mathbb{C}^J$ 表示与特定通信任务相关且具有环境普适性的知识，其中 J 取决于任务的信道需求。 $K(\cdot)$ 表示WEK环境到信道的映射函数，其目标是学习或推导环境信息与信道知识之间的稳定、可解释关系。该映射关系可通过知识图谱、关系图、树形图等多种形式存在。

为解决环境与信道关系构建研究中环境信息冗余、关系不可解释、关系构建复杂度高的问题，文献[31]首次提出了一种电磁特性启发式的WEK构建与应用方法，相关框架如图3所示。该方法通过建模信号在环境中发生的直射、反射、衍射等传播行为，推导环境特征与信道特性之间的数学关系，并量化各类传播机制的贡献权重，形成可用于信道建模的结构化知识表达。

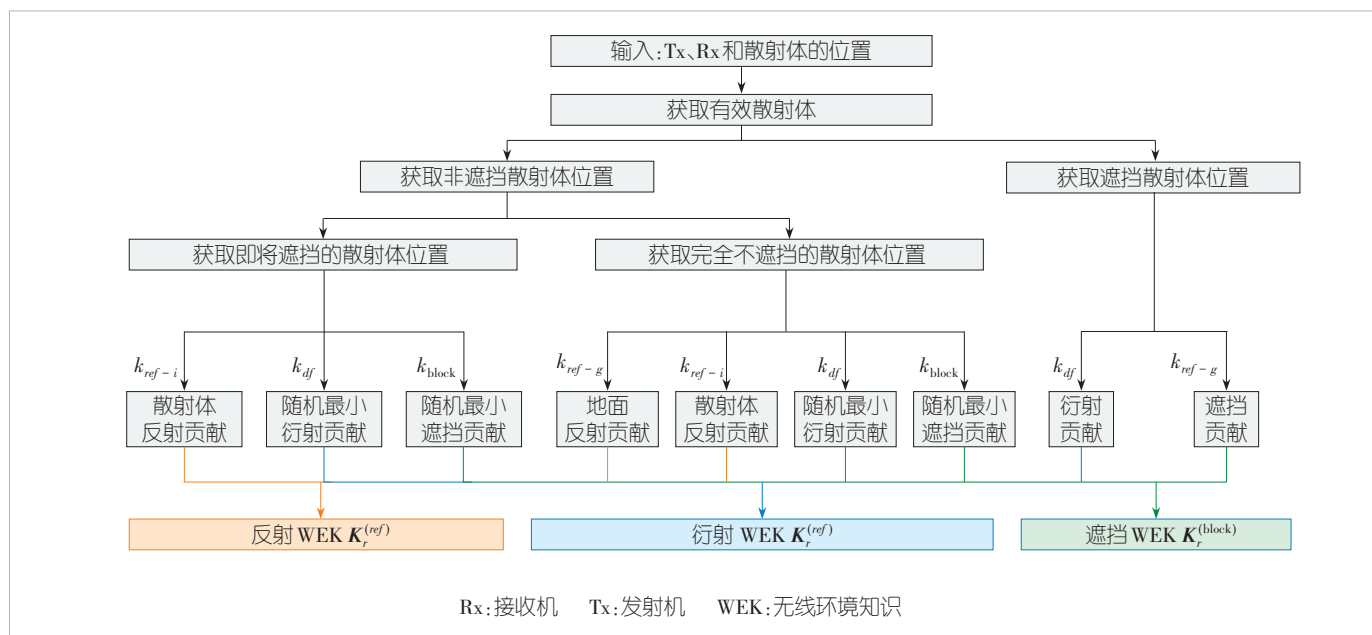


图3 所提WEK构建框架及流程

无线信道由不同传播机制的多径叠加形成，信道模型表示为：

$$h(t) = h_{tas}(t) + h_{raf}(t) + h_{df}(t) \quad (3).$$

$$h_i(t) = \sum_{n=1}^N \beta_{n,i}(t) e^{j(2\pi\nu_{i,n}t + \phi_{i,n}(t))} \quad (4),$$

其中, i 为传播类型, 即直射、反射和衍射。 N_i 是类型 i 产生的信道多径数。 $\beta_{n,i}(t)$ 、 $\phi_{i,n}(t)$ 、 $v_{i,n}$ 分别表示时变路径损耗、相位和多普勒频移。

由于总路径增益的值由多个子路径的叠加决定，这些子路径是根据发射机（Tx）和接收机（Rx）之间的环境结构产生的。因此，构建的WEK由两个部分组成：基于环境信息的传播知识M和知识系数C，具体可表示为：

$$\mathbf{K}(\mathbf{M}, \mathbf{C}): \{P_i, P_r, P_t\} \rightarrow \text{WEK} \quad (5),$$

其中, P_i 、 P_r 、 P_t 分别为第 i 个散射体、第 r 个 Rx 和 Tx 的位置。传播知识根据环境信息解释当前场景中传播模式的贡献程度, 知识系数反映不同场景中传播模式的权重。

电磁波特性的启发式的无线环境知识构建首先需要通过三维环境模型获得Tx、Rx和散射体的位置信息。为判定哪些散射体参与传播过程，我们引入有效散射体检测机制，并基于第一菲涅尔区定义、随机几何理论以及Tx、Rx之间的几何结构，构建散射体椭球模型，设计参与信号传播的散射体判定方法。此外，为了更好地表征NLoS区域的电磁波传播特性，定义遮挡散射体，并设计遮挡散射体的几何判定机

制。这类散射体存在以下几何关系：散射体到Tx和Rx连线的垂直距离小于其对角线长度的一半。

利用几何光学理论,设计真实环境下基于位置信息确定反射点的方法,并根据该反射点,基于几何光学理论推导环境与散射体反射、环境与地面反射之间的数学关系。散射体反射贡献 K_{ref-i} 和地面反射贡献 K_{ref-g} 分别定义为:

$$K_{ref-i} = \frac{k_{ref-i} \|P_t - P_r\|}{\|P_r - RP\| + \|RP - P_t\|} \quad (6),$$

$$K_{ref-g} = \frac{k_{ref-g} \|P_i - P_r\|}{\sqrt{(d_i + d_r)^2 + (h_i + h_r)^2}} \quad (7),$$

其中, k_{ref-i} 和 k_{ref-g} 分别表示散射体和地面反射的知识系数, d_i 和 d_r 分别为 Tx 和 Rx 到地面反射点的水平距离, h_i 和 h_r 分别为 Tx 和 Rx 的天线高度。 K_{ref-i} 和 K_{ref-g} 用于刻画散射体与地面反射对总信道增益的贡献。

基于一致性几何绕射理论, 沿着实际的无线电波传播路径, 散射体开始落入第一菲涅耳区时, 就会发生衍射损耗, 这将导致接收功率和信号质量下降。从散射体的顶点到连接Tx和Rx的距离的垂线称为菲涅耳间隙。以菲涅耳区中心高度为参考, 设遮挡散射体高度为 h_i , 散射体到Tx和Rx的水平距离分别为 d_t 和 d_r , 发射天线高度为 h_t , 接收天线高度为 h_r , 则散射体衍射贡献表示为:

$$K_{df} = k_{df} \left(h_b - h_r - \frac{(h_l - h_r)d_r}{d_l + d_r} \right) \quad (8).$$

基于Tx、Rx和遮挡散射体之间的几何关系,环境与散射体遮挡之间的映射关系表示为:

$$K_{\text{block}} = \frac{\|\mathbf{P}_b \cdot \overrightarrow{\mathbf{P}_t \mathbf{P}_r}\|}{\|\overrightarrow{\mathbf{P}_t \mathbf{P}_r}\|} \cdot \frac{k_{\text{block}}}{\sqrt{l_b^2 + w_b^2}} \quad (9),$$

其中, l 和 w 分别表示散射体的长度和宽度。

上述工作中仅考虑散射体材质单一的室外场景,设置的知识系数 k 皆为常量。在此基础上,文献[32]进一步考虑了IIoT等复杂场景中散射体的不同材质,根据菲涅尔电磁理论,首次提出了反映材质电磁和频率特性的反射和衍射知识系数:

$$\{c_{\text{ref}-i}, c_{\text{df}-i}\} \sim F(\varepsilon) \quad (10),$$

其中, $F(\cdot)$ 表示映射函数, ε 为复介电常数(反映材质的电磁特性)。基于材质相关的知识系数,进一步更新WEK表示方法:

$$K_{\text{ref}-i} = \frac{c_{\text{ref}-i} \|\mathbf{P}_t - \text{RP}\|}{\|\mathbf{P}_t - \text{RP}\| + \|\text{RP} - \mathbf{P}_r\|} \quad (11),$$

$$K_{\text{df}-i} = c_{\text{df}-i} \sqrt{\frac{\|\mathbf{P}_t - \text{DP}\|}{\|\text{DP} - \mathbf{P}_r\| \cdot (\|\mathbf{P}_t - \text{DP}\| + \|\text{DP} - \mathbf{P}_r\|)}} \quad (12),$$

其中, RP和DP分别表示散射体上对应的反射点和衍射点坐标。

所述方法将散射体材质信息引入WEK构建过程,基于菲涅尔电磁理论、几何光学理论和一致性绕射理论,建立了环境与不同电磁传播类型(反射、衍射)以及散射体遮挡之

间的量化关系,实现了对复杂场景的动态自适应与精细化建模。更新后的WEK不仅有效简化了判断和计算过程,还将环境中材质的电磁和频率特性转化为具备物理意义的可解释知识表示,提升了模型的场景泛化能力和信道预测精度。相比于端到端的黑盒学习方式,WEK为DTC提供了一套可物理推导的环境-信道关系图。在此基础上,后续的大模型模块能够在明确的物理机制指引下,完成更具可信性的信道预测与策略决策。

2.3 信道大模型

尽管WEK为通信系统提供了结构化、可解释的知识输入,但要充分发挥其价值,仍需具备强大的模型能力,以便对其输出的环境信息进行深度理解与有效利用。

DTC性能受限于框架中AI模型的参数规模和泛化能力。目前大多数现有模型的参数规模较小,且大多针对特定任务或场景设计,难以满足DTC多任务、多场景的复杂需求。同时,大语言模型在自然语言处理领域展现出强大泛化性和建模能力,为DTC的智能演进提供参考。为实现DTC的多场景泛化、多任务并行、多模态融合,作者团队首次尝试引入大模型到信道研究中,进而提出了信道大模型(Chan-nelLM)。ChannelLM作为DTC的智能核心,能够实时处理多模态输入数据并数字化动态重建物理世界,实现信道特征的深度提取与关键参数的精准预测。依托通信网络的反馈机制,ChannelLM在强大硬件算力支持下通过在线学习不断优化模型输出,实现高效训练与低延迟推理。

图4展示了一个以多模态感知数据和无线环境知识作为

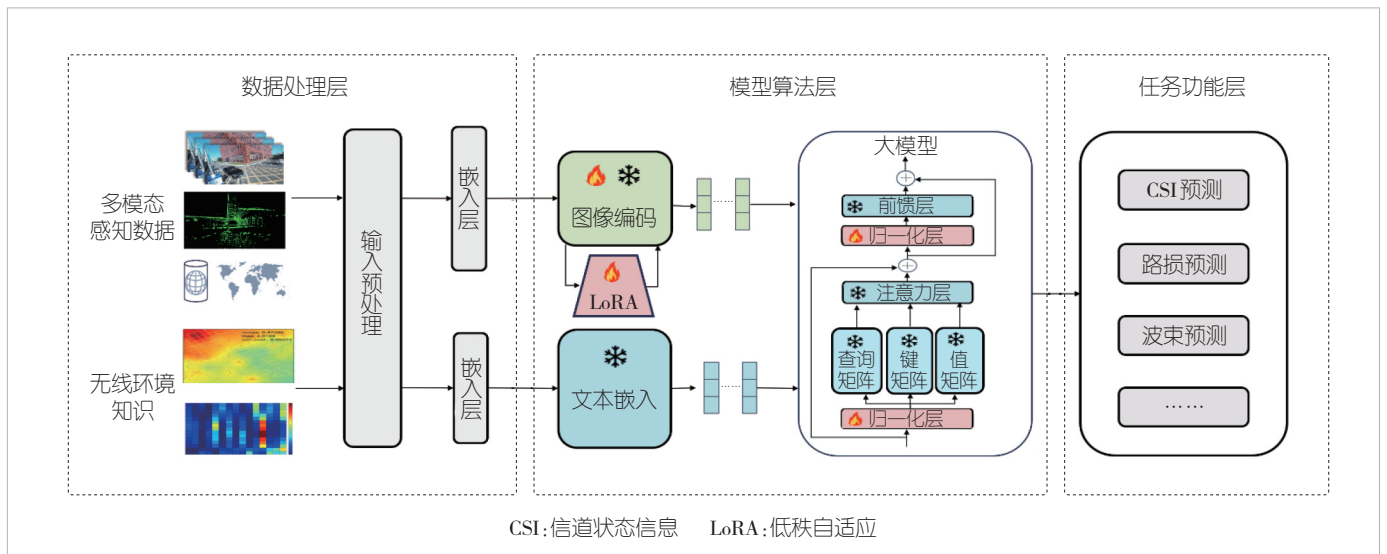


图4 信道大模型原理架构图

输入的 ChannelLM 架构示意图。该 ChannelLM 架构包括数据处理层、模型算法层和任务功能层 3 个部分。

在数据处理层中,由终端传感器获取的多模态感知数据和从 WEK 输出的无线环境知识在经过输入预处理后,由嵌入层分别映射为适用于信道大模型处理的嵌入向量,作为模型算法层的输入。

在模型算法层中,由感知数据和环境知识映射得到的嵌入向量,通过大模型实现多模态融合并提取影响信道传播的关键特征。知识嵌入向量通过分词器将其转化为离散序列,并将离散序列映射为高维连续空间中的嵌入向量,得到文本嵌入。感知嵌入向量通过图像编码模块中的编码器来提取感知数据中的高维语义特征,包括与信道传播相关的遮挡、散射体分布等环境特征,得到能够反映无线传播环境的图像嵌入。在模型的训练过程中,图像编码模块中的注意力权重采用低秩自适应 (LoRA) 技术进行微调,以节省计算资源^[33]。文本嵌入和图像嵌入之间的关系通过大模型模块中的解码器来建立,从而形成统一的传播环境特征表示。在此基础上,大模型利用其强大的表征能力从融合后的环境特征表示中,提取影响信道传播的潜在关键特征,进而实现后续对无线信道关键参数的准确预测。

在任务功能层中,不同的预测任务如 CSI 预测、路损预测、波束预测等,通过特别设计的网络架构和损失函数进行任务定制化训练,并结合每类任务的误差反馈进行权重更新,使模型能够适应多变的无线环境,以提升整体 DTC 系统的泛化能力与鲁棒性。因此,作为信道数字孪生系统中的智能大脑,ChannelLM 所具有的多模态感知融合、实时推理与任务协同的能力,为实现在线、精确的 DTC 系统奠定了基础。

3 仿真与结果

本节进一步通过仿真实验验证所提出关键技术典型 MTC 场景中的实际表现。具体而言,在已构建的多模态环境感知与三维重构基础上,重点对 WEK 构建方法与 ChannelLM 的有效性进行性能展开评估。

3.1 基于多模态环境信息的无线环境知识验证

本节搭建 IIoT 仿真场景,并通过射线跟踪 (RT) 方法构建数据集。基于多模态感知数据,面向路径损耗预测任务,对 2.2 节中的知识系数和 WEK 构建方法进行验证,如图 5 所示。

图 5 (a) 显示了搭建的仿真场景,场景中不同颜色代表不同材质。图 5 (b) 显示了场景中对 Rx 的 WEK 和路径损耗之间的隐性关系。从统计分析的角度看,为了进一步评

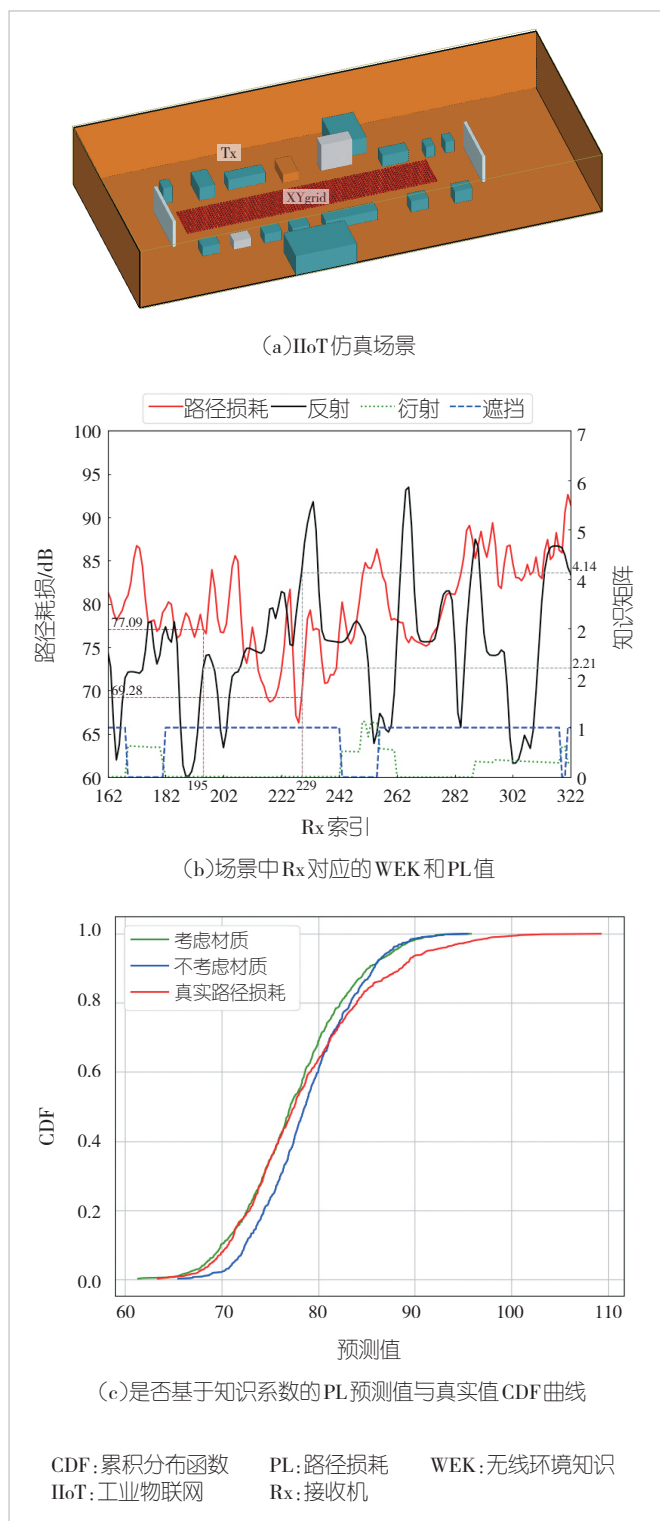


图 5 仿真场景、WEK 构建及 PL 预测结果

估所提出的路径损耗预测框架,我们基于材质系数和默认系数 (设置为 1) 分别构建 WEK。通过卷积神经网络 (CNN) 模型得到的预测值与真实值的累积分布函数 (CDF) 曲线如

图 5 (c) 所示。图 5 验证了 WEK 构建与考虑材质在 IIoT 场景中的有效性。

3.2 基于多模态环境信息的信道大模型仿真验证

在 IIoT 仿真场景中, 基于 WEK 与小模型结合的方法在信道预测任务中取得了良好效果。为了进一步验证其在更广泛场景下的适应性与泛化能力, 本文进一步引入大模型架构, 并在真实采集的 V2X 场景多模态数据集上评估 ChannelLM 框架的性能。

本文分别构建了基于 GPT-2^[34] 的 ChannelGPT^[35-36] 模型和基于 DeepSeek Janus-Pro-1B 的 ChannelDS 模型^[37]。在使用仿真场景数据进行微调训练的基础上, 我们进一步引入少量真实场景数据对上述两种模型进行迁移学习, 从而评估其在小样本条件下的泛化能力。

针对 ChannelGPT 模型在真实环境中的路损预测能力验证, 本文构建了一个多模态真实世界数据集。该数据采集环境位于典型城市 T 字路口, 街道总长度为 375 m, 道路宽度为 14 m, 工作频段设定为 14 GHz。通过彩色 (RGB) 摄像头与 GPS 设备获取环境图像与位置信息, 并结合信道测量装置同步采集路径损耗数据, 共获得 2 600 个样本。在模型训练过程中, 设置训练轮数为 100, 学习率为 0.001。本文采用均方根误差 (RMSE) 作为评价指标, 并绘制 CDF 曲线以展示误差分布情况, 如图 6 (b) 所示。仿真结果表明, 在仅使用 5% 和 20% 真实样本进行训练的条件下, ChannelGPT 在整个测试集上的性能均优于深度神经网络 (DNN) 模型。

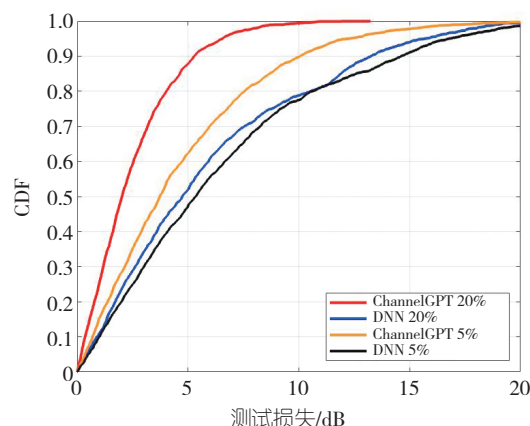
为验证所提出的 ChannelDS 模型在真实场景中的波束预测能力, 本文选用 DeepSense^[38] 数据集中的一个场景进行实验。该场景位于一条宽约 12 m 的双向三车道街道, 车辆类型多样, 通信频段为 60 GHz。车辆端搭载全向天线、GPS 定位模块和 ZED 2 广角摄像头, 最终共采集到 22 500 个样本, 每个样本包含位置信息、车辆视角图像及最佳波束索引。为评估模型在小样本条件下的泛化能力, 我们选取 30% 的样本对 ChannelDS 模型和 3 种基线模型进行微调, 并采用交叉熵损失函数进行优化。各模型在 Top-1 至 Top-5 准确率下的性能对比如图 7 (b) 所示。实验结果表明, ChannelDS 模型的 Top-5 准确率达到 95.1%, 比其他 3 种基线方法高出 14% 以上。

4 未来挑战和展望

随着 DTC 系统架构的建立及关键技术的实现, 其在 6G MTC 中展现出良好应用潜力。为进一步推动 DTC 系统的落地与演进, 未来研究可围绕以下 3 个方面展开:



(a) 路损预测真实场景图



(b) 不同模型路损预测误差的 CDF 对比

CDF: 累积分布函数 GPT: 生成式预训练转换器
DNN: 深度神经网络

图 6 ChannelGPT 模型路损预测仿真结果

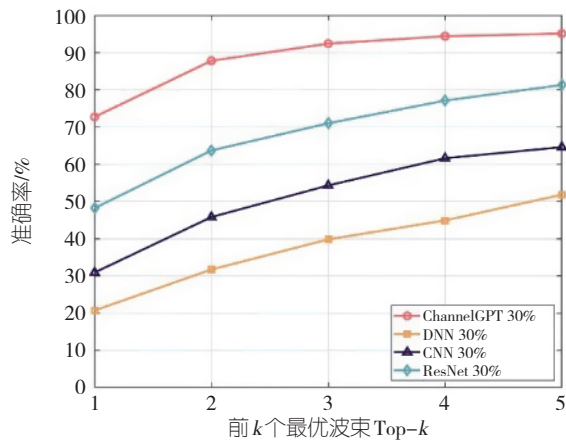
首先, 围绕 WEK 的持续构建与扩展, 进一步通过神经网络、语义提取、数学推导等方法, 提取包含具体环境特征、抽象语义向量在内的多层次无线环境信息。构建具备层次性、泛化性和可更新的无线环境知识表征模型以支持 DTC 的长期演进。

其次, 实现物理世界与孪生世界的实时交互。当前 DTC 仍面临高频率信道更新带来的计算负担、多模态融合的延迟瓶颈及建模精度不足等问题。未来应提升感知-建模-预测流程的集成度与响应速度, 以实现动态环境下的闭环优化。

最后, 探索大模型与信道建模的融合路径。未来可通过结合物理层信号处理技术、引入持续学习和实时反馈等机制, 使 LLM 具备处理更深层次环境信道语义的能力, 支撑更高层级的通信智能决策。



(a) 波束预测真实场景图



(b) 不同模型 Top-1 至 Top-5 波束预测准确率对比

CDF: 累积分布函数 CNN: 深度神经网络
DNN: 深度神经网络 GPT: 生成式预训练转换器

图7 ChannelDS模型波束预测仿真结果

5 结束语

本文围绕 6G MTC 需求, 构建了 DTC 系统的整体框架, 并从多模态环境感知、无线环境知识和可泛化的信道大模型 3 个方面, 系统性设计并实现了支撑 DTC 应用的关键技术。IIoT 以及 V2X 场景的实验结果验证了所提方案在信道建模精度、系统实时性与环境适应性方面的优势, 为 DTC 从概念走向工程落地提供了技术支撑。未来研究将聚焦于知识表达多样化、物理-数字空间实时闭环交互以及大模型与信道建模的融合, 进一步提升 DTC 的智能预测能力和多场景、多任务泛化性。

致谢

感谢北京邮电大学在读硕士研究生赵乙竹、张兆灵、宋垚、王凯, 以及副研究员张宇翔对本文内容提供的帮助, 同时也感谢中国移动研究院对本文相关研究工作的大力支持!

参考文献

- [1] 张建华, 田艺璇, 孙语瞳, 等. 面向 6G 无线环境可预测的感知重构、语义表征及应用 [J]. 中兴通讯技术, 2023, 29(5): 38-44. DOI: 10.12142/ZTETJ.202305007
- [2] ZHANG J H, LIN J X, TANG P, et al. Channel measurement, modeling, and simulation for 6G: a survey and tutorial [EB/OL]. (2023-05-26)[2025-03-21]. <https://arxiv.org/abs/2305.16616>
- [3] 张建华, 王珩, 张宇翔, 等. 6G 信道新特性与建模研究: 挑战、进展与展望 [J]. 中国科学: 信息科学, 2024, 54(5): 1114-1143. DOI: 10.1360/SSI-2023-0259
- [4] LIU G Y, HUANG Y H, LI N, et al. Vision, requirements and network architecture of 6G mobile network beyond 2030[J]. China communications, 2020, 17(9): 92-104. DOI: 10.23919/JCC.2020.09.008
- [5] 张建华, 王玉洁, 唐盼, 等. 工业互联网信道特性与建模研究综述 [J]. 电波科学学报, 2023, 38(1): 3-14. DOI: 10.13443/j.cjors.2022070401
- [6] MO Q M, ZHANG Y X, WANG J L, et al. Multi-frequency channel measurement in smart factories: comparative analysis with 3GPP InF scenarios [C]//2024 IEEE 100th Vehicular Technology Conference (VTC2024-Fall). IEEE, 2024: 1-5. DOI: 10.1109/VTC2024-Fall63153.2024.10757555
- [7] JIANG T, ZHANG J H, TANG P, et al. 3GPP standardized 5G channel model for IIoT scenarios: a survey [J]. IEEE Internet of Things Journal, 2021, 8 (9):7196-7210.DOI:10.1109/JIOT.2020.3048992
- [8] AHN Y, KIM W, SHIM B. Active user detection and channel estimation for massive machine-type communication: deep learning approach [J]. IEEE Internet of Things journal, 2021, 9(14): 11904-11917. DOI: 10.1109/JIOT.2021.3132329
- [9] KHAREL B, LÓPEZ O L A, ALVES H, et al. Ultra-reliable communication for critical machine type communication via CRAN-enabled multi-connectivity diversity schemes [J]. Sensors (basel), 2021, 21(23): 8064. DOI: 10.3390/s21238064
- [10] SHARMILA M, SATYANARAYANA R V S. An intelligent resource allocation strategy for machine type communication environment [J]. International journal of communication systems, 2024, 37(1): e5628. DOI: 10.1002/dac.5628
- [11] WANG H, ZHANG J H, NIE G F, et al. Digital twin channel for 6G: concepts, architectures and potential applications [J]. IEEE communications magazine, 2025, 63(3): 24-30. DOI: 10.1109/MCOM.001.2400213
- [12] ZHANG J H. The interdisciplinary research of big data and wireless channel: a cluster-nuclei based channel model [J]. China communications, 2016, 13(2): 14-26. DOI: 10.1109/CC.2016.7405719
- [13] YU L, ZHANG Y X, ZHANG J H, et al. Implementation framework and validation of cluster-nuclei based channel model using environmental mapping for 6G communication systems [J]. China communications, 2022, 19(4): 1-13. DOI: 10.23919/JCC.2022.04.001
- [14] 张建华, 杨育捷, 张平. 确定智能基站位置的方法、装置及存储介质: ZL201811176132.4 [P]. 2018-10-10
- [15] ZHANG Z, ZHANG Y X, ZHANG J H, et al. Adversarial training-aided time-varying channel prediction for TDD/FDD systems [J]. China communications, 2023, 20(6): 100-115. DOI: 10.23919/JCC.fa.2020-0698.202306
- [16] WANG K, YU L, ZHANG J H, et al. Multi-modal environmental sensing based path loss prediction for V2I communications [EB/OL]. (2024-12-10)[2025-06-20]. <https://arxiv.org/abs/2412.07681>
- [17] LIU G Y, WANG J Z, LI R P, et al. Artificial-intelligence-empowered digital-twin-based network autonomy [J]. Frontiers of information technology & electronic engineering, 2025, 26(2): 157-160. DOI: 10.1631/FITEE.2510000
- [18] LIU G Y, ZHU Y H, KANG M C, et al. Native design for 6G digital

- twin network: use cases, architecture, functions and key technologies [EB/OL]. (2025-03-24) [2025-05-15]. <https://ieeexplore.ieee.org/abstract/document/10937141>
- [19] SHI L Z, ZHANG J H, YU L, et al. Can wireless environment information decrease pilot overhead: a channel prediction example [J]. IEEE wireless communications letters, 2025, 14(3): 861–865. DOI: 10.1109/LWC.2025.3526271
- [20] ZHANG Z, ZHANG J H, ZHANG Y X, et al. AI-based time-, frequency-, and space-domain channel extrapolation for 6G: opportunities and challenges [J]. IEEE vehicular technology magazine, 2023, 18(1): 29–39. DOI: 10.1109/MVT.2023.3234169
- [21] SHEN Z B, YU L, ZHANG Y X, et al. DataAI-6G: a system parameters configurable channel dataset for AI-6G research [C]// Proceedings of IEEE Globecom Workshops (GC Wkshps). IEEE, 2023: 1910–1915. DOI: 10.1109/GCWkshps58843.2023.10464657
- [22] SUN Y T, ZHANG J H, ZHANG Y X, et al. Environment features-based model for path loss prediction [J]. IEEE wireless communications letters, 2022, 11(9): 2010–2014. DOI: 10.1109/LWC.2022.3192516
- [23] SUN Y T, ZHANG J H, YU L, et al. How to define the propagation environment semantics and its application in scatterer-based beam prediction [J]. IEEE wireless communications letters, 2023, 12(4): 649–653. DOI: 10.1109/LWC.2023.3237827
- [24] ZHANG Z, ZHANG J H, ZHANG Y X, et al. Deep reinforcement learning based dynamic beam selection in dual-band communication systems [J]. IEEE transactions on wireless communications, 2024, 23(4): 2591–2606. DOI: 10.1109/TWC.2023.3300830
- [25] WANG J L, ZHANG J H, ZHANG Y X, et al. Radio environment knowledge pool for 6G digital twin channel [J]. IEEE communications magazine, 2025, 63(5): 158–164. DOI: 10.1109/MCOM.003.2400168
- [26] GUAN Y C, ZHANG J H, TIAN L, et al. A comparative study for indoor factory environments at 4.9 and 28 GHz [C]//Proceedings of 14th European Conference on Antennas and Propagation (EuCAP). IEEE, 2020: 1–5. DOI: 10.23919/EuCAP48036.2020.9135294
- [27] ZHANG X, QIU G, ZHANG J H, et al. Analysis of millimeter-wave channel characteristics based on channel measurements in indoor environments at 39 GHz [C]//2019 11th International conference on wireless Communications and Signal Processing (WCSP). IEEE, 2019: 1–6. DOI: 10.1109/WCSP.2019.8928017
- [28] WANG F Y, ZHANG Y X, TANG P, et al. Effects of inaccuracies of indoor environment databases on ray tracing results [C]// Proceedings of IEEE 93rd Vehicular Technology Conference (VTC2021-Spring). IEEE, 2021: 1–7. DOI: 10.1109/vtc2021-spring51267.2021.9448968
- [29] ZHANG J H, CAI Y C, YU L, et al. Wireless environment information sensing, feature, semantic, and knowledge: four steps towards 6G AI-enabled air interface [EB/OL]. (2024-09-28)[2025-03-21]. <https://arxiv.org/abs/2409.19331>
- [30] ZHANG J H, YU L, LIU S Y, et al. Wireless environmental information theory: a new paradigm towards 6G online and proactive environment intelligence communication [EB/OL]. (2024-12-16)[2025-03-21]. <https://arxiv.org/abs/2412.11479>
- [31] WANG J L, ZHANG J H, SUN Y T, et al. Electromagnetic wave property inspired radio environment knowledge construction and artificial intelligence based verification for 6G digital twin channel [J]. Frontiers of information technology & electronic engineering, 2025, 26(2): 260–277. DOI: 10.1631/fitee.2400464
- [32] 张兆灵, 于力, 张宇翔, 等. 无线环境知识表示方法研究——工业互联网场景的应用 [J]. 移动通信, 2025, 49(5): 63–66. DOI: 10.13982/j.cnki.mtc.2025.05063
- [33] HU E J, SHEN Y, WALLIS P, et al. LoRA: low-rank adaptation of large language models [EB/OL]. (2021-06-17) [2025-06-20]. <https://arxiv.org/abs/2106.09685>
- [34] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners [EB/OL]. (2019-02-14) [2025-06-20]. <https://openai.com/blog/better-language-models/>
- [35] YU L, SHI L Z, ZHANG J H, et al. ChannelGPT: a large model to generate digital twin channel for 6G environment intelligence [EB/OL]. (2025-06-06) [2025-06-20]. <https://arxiv.org/abs/2410.13379>
- [36] 张建华, 史廉正, 于力, 等. 信道通用预训练大模型赋能数字孪生信道: 原理与实践 [J]. 无线电工程, 2025, 55(4): 679–686. DOI: 10.1360/SSI-2023-0355
- [37] ZHAO Y Z, YU L, SHI L Z, et al. Multi-modal large models based beam prediction: an example empowered by DeepSeek [EB/OL]. (2025-06-06)[2025-06-20]. <https://arxiv.org/abs/2506.05921>
- [38] ALKHATEEB A, CHARAN G, OSMAN T, et al. DeepSense 6G: a large-scale real-world multi-modal sensing and communication dataset [J]. IEEE communications magazine, 2023, 61(9): 122–128. DOI: 10.1109/MCOM.006.2200730

作者简介



于力, 北京邮电大学网络与交换技术全国重点实验室博士后; 主要研究方向为信道建模和预测、机器学习、环境感知重构、数字孪生、信道大模型等。



张建华, 北京邮电大学教授、博士生导师, 网络与交换技术全国重点实验室副主任, 北邮-移动联合研究院执行院长; 主要研究方向为 6G 移动通信技术、人工智能、数据挖掘、智能信道建模、大规模 MIMO 和太赫兹信道建模、信道仿真仪、OTA 测试等。



蔡逸辰, 北京邮电大学在读博士研究生; 主要研究方向为深度学习、信道预测、信道数字孪生等。

端边协同的6G内生AI网络



6G AI-Native Network with Edge-Device Collaboration

王志勤/WANG Zhiqin, 周吉喆/ZHOU Jizhe,
韩凯峰/HAN Kaifeng

(中国信息通信研究院, 中国 北京 100191)
(China Academy of Information and Communications Technology, Beijing
100191, China)

DOI: 10.12142/ZTETJ.202504005

网络出版地址: <http://kns.cnki.net/kcms/detail/34.1228.TN.20250716.0845.004.html>

网络出版日期: 2025-07-16

收稿日期: 2025-05-28

摘要: 面向6G智能终端AI业务原生、融合感知、智能协同业务需求, 提出了一种端边协同的6G内生智能网络架构。该架构通过分层设计(基础设施层、模型管理层、资源管控层、业务编排层), 实现了端边数据管控、模型动态协同及异构资源融合调度, 具备“通信+计算+数据+模型”一体化服务能力。在内生智能网络架构基础上, 提出了端边智能协同系统评估模型, 并围绕数据管控、模型协同、资源调度3个维度, 提出了高质量数据集构建、数据管控框架、参考模型库、端边模型协同、异构资源融合管控、灵活组网等关键技术, 形成端边智能协同技术体系。

关键词: 通信与人工智能融合; 端边协同; 资源管理

Abstract: To address the requirements of artificial intelligence (AI)-native service, integrated sensing, and intelligent collaboration for 6G AI terminals, an edge-device collaborative 6G AI-native network architecture is proposed. The architecture adopts a hierarchical design (infrastructure layer, model management layer, resource control layer, and service orchestration layer) to achieve core capabilities such as edge-device data management and control, dynamic model collaboration, and integrated heterogeneous resource scheduling, providing integrated "communication + computing + data + model" service capabilities. Based on this architecture, an edge-device AI collaborative system evaluation model is established. Focusing on three key dimensions—data management, model collaboration, and resource scheduling—key technologies are proposed, including high-quality dataset construction, data management framework, reference model library, edge-device model collaboration, heterogeneous resource integration and control, and flexible networking, forming a comprehensive edge-device intelligent collaboration technology system.

Keywords: integration of AI and communication; edge-device collaboration; resource management

引用格式: 王志勤, 周吉喆, 韩凯峰. 端边协同的6G内生AI网络 [J]. 中兴通讯技术, 2025, 31(4): 29–33. DOI: 10.12142/ZTETJ.202504005

Citation: WANG Z Q, ZHOU J Z, HAN K F. 6G AI-native network with edge-device collaboration [J]. ZTE technology journal, 2025, 31(4): 29–33. DOI: 10.12142/ZTETJ.202504005

1 6G端边智能协同研究背景

随着移动通信与人工智能(AI)技术的持续创新发展, 6G技术将与感知、智能、计算、安全等新兴技术深度融合, 这进一步拓宽了网络服务边界, 提升了服务效率, 在5G“万物互联”基础上实现“万物智联”^[1–2]。6G服务对象正逐步向具备融合感知、智能决策、计算能力的新一代智能终端演进, 终端形态也从传统手机扩展到AI手机、智能可穿戴、智能机器人、智能网联车等多元形态^[3], 这将促进6G智能业务多元化、泛在化发展。为满足未来泛在智能业务低时

延、高精度等多维指标与数据高效处理需求, 业界提出将模型部署下沉到边缘侧或终端侧的实施策略: 端侧负责执行本地轻量化智能任务处理, 边侧则专注于处理终端卸载的高复杂度计算任务, 通过端边协同计算显著提升智能业务指标。

当前, 全球标准化组织积极探索面向智能终端的6G端边智能协同关键技术。美国NextG联盟提出要在6G系统设计中考虑终端AI计算能力, 着力解决智能任务在终端与网络算力节点间的协同计算问题, 以应对物联网设备多元数据处理与资源受限的挑战^[4]。欧洲6G-IA协会将协作机器人等智能协作场景纳入6G重点应用场景, 强调6G需要支持AI驱动的空口智能和边缘智能, 研究包括终端在内的边缘设备分布式模型训练和部署技术^[5]。中国IMT-2030(6G)推进组提出6G移动通信网络将提供泛在的AI即服务能力(AIaaS),

基金项目: 北京市自然科学基金资助项目(L253004); 中国科协青年人才托举工程项目(2023QNRC001, 2022QNRC001)

系统开展边缘智能、分布式训练推理等关键技术研究^[6],并针对无线空口物理层领域,深入研究移动通信与AI融合的数据格式与模型设计方案,为业界研发提供模型参考^[7]。2025年,第3代合作伙伴计划(3GPP)Release 20正式启动6G网络架构研究项目,将网络AI框架设计列为核心研究内容,同时探索高效可扩展的数据管控框架^[8]。

学术界积极开展面向数据、模型、资源优化设计的端边智能协同方法研究。文献[9]提出了一种端边协同模型拆分与资源优化方法,降低了系统能耗。文献[10]基于终端移动特性,联合优化端边协同模型拆分、数据压缩和业务接入策略,降低了业务时延。文献[11]研究了云边端计算网络下的带宽、边缘算力与任务分配优化方案,提升网络整体效用。

2 6G端边智能业务特征与需求

端边智能业务以强感知、强计算、强协同为核心,通过数据融合处理、任务协作与实时决策,满足低时延、高精度、高可靠的业务多维指标要求。总体来说,6G端边智能业务发展呈现AI业务原生、融合感知、智能协同三大特征:

1) AI业务原生:大模型、智能体等AI技术与终端深度融合,将催生6G智能终端AI原生业务全新应用范式^[12]。这一趋势对网络服务供给能力提出了更高要求,促使服务需求从传统移动接入向“连接+AI”智能化服务模式转变。因此,网络需精准识别终端AI业务的异构资源需求(如通信、计算),通过一体化决策机制实现智能终端的按需融合服务。

2) 融合感知:随着智能机器人、智能网联车等终端设备自主性与智能化水平的提升,6G智能终端将深度感知周围环境^[13],具备执行智能避障、路径优化等复杂任务的能力。为支持上述业务场景,需要网络提供物理世界数据感知、采集和建模能力,并基于统一数据管控框架为智能终端交互提供高质量数据服务。

3) 智能协同:针对低空经济、智慧工厂等场景,6G智能终端行为模式从“单体智能”向“多智能协作”方式演进,通过多智能终端协同完成无人物流配送、机器人协作搬运等协作任务要求^[14]。为支持上述业务场景,需要网络提供多终端按需组网和动态任务调度能力,支持智能终端任务快速响应和协作融合决策,提升复杂协同任务效率。

由此可见,6G新型智能终端业务对网络资源供给、数据管理、协同机制提出更高要求,亟需优化支持端边协同的6G内生AI网络架构,开展6G端边智能协同关键技术研究。首先,构建内生支持数据模型管控的内生智能网络架构,支撑端边智能协同关键能力,为6G智能终端与边侧算力高效智能协同提供网络承载。其次,面向6G智能终端业务数据、

模型、资源需求,围绕数据管理、模型协同、异构资源融合管控技术维度,提出端边智能协同关键技术。

3 6G内生智能网络架构与系统评价模型设计

6G智能终端具备AI原生、融合感知与智能协同等业务特征,这对端边协同的6G网络架构提出了内生智能的设计要求。为满足融合感知业务需求,6G网络需构建统一的数据处理框架,支持数据采集、传输、存储与共享,从而为终端、网络决策及端边智能协同提供按需数据服务。6G网络需具备端边分布式多级网络全局模型管控能力,通过多节点协同提升智能业务性能指标。为满足多样化AI原生业务需求,6G网络需实现通信-计算-数据-模型的一体化智能服务调度与体验保障,支持动态按需的端边异构资源分配。

3.1 内生智能网络架构

现有移动接入与模型管理体系的分离架构难以满足端边融合感知、智能高效协同及多维服务按需调度的需求。本文提出一种“内生”数据模型管控网络架构,通过统一分布式数据模型管理与异构资源融合调度机制,有效支撑终端复杂智能业务对超低时延、超高精度等性能指标的严苛要求。

如图1所示,从逻辑分层角度,内生智能网络架构包括基础设施层、模型管理层、资源控制层、业务编排层,并通过统一接口对外支撑多元智能应用。

基础设施层由智能终端、边缘计算节点及无线接入网等分布式网络构成,通过泛在网络接入实现动态跨节点协同,为智能业务接入与端边协同提供基础资源支撑。

模型管理层实现分布式端边网络的统一数据与模型管理,核心功能包括:1)构建统一数据管控框架,实现异构设备数据的按需采集、传输、存储、处理;2)实施边侧模型全生命周期管理,包括模型训练、部署、选择、监控、更新等;3)基于终端模型能力感知,建立全局模型部署视图。

资源管控层实现通信、数据、模型、计算多维资源的统一建模与协同管控,通过原子化能力抽象与融合控制机制,支撑端到端异构资源整合及一体化调度。

业务编排层对外接收多样化应用请求,并利用内置AI模型生成“需求解析-资源映射-任务调度”的动态端边协同编排策略,实现业务体验按需保障,保持系统高效运行。

3.2 系统评价模型

当前端边协同推理策略主要聚焦推理时延、推理准确度、能耗等单一指标优化,缺少针对端边多级分布式网络的系统级评估指标,难以实现系统长期效用最优的模型管控与

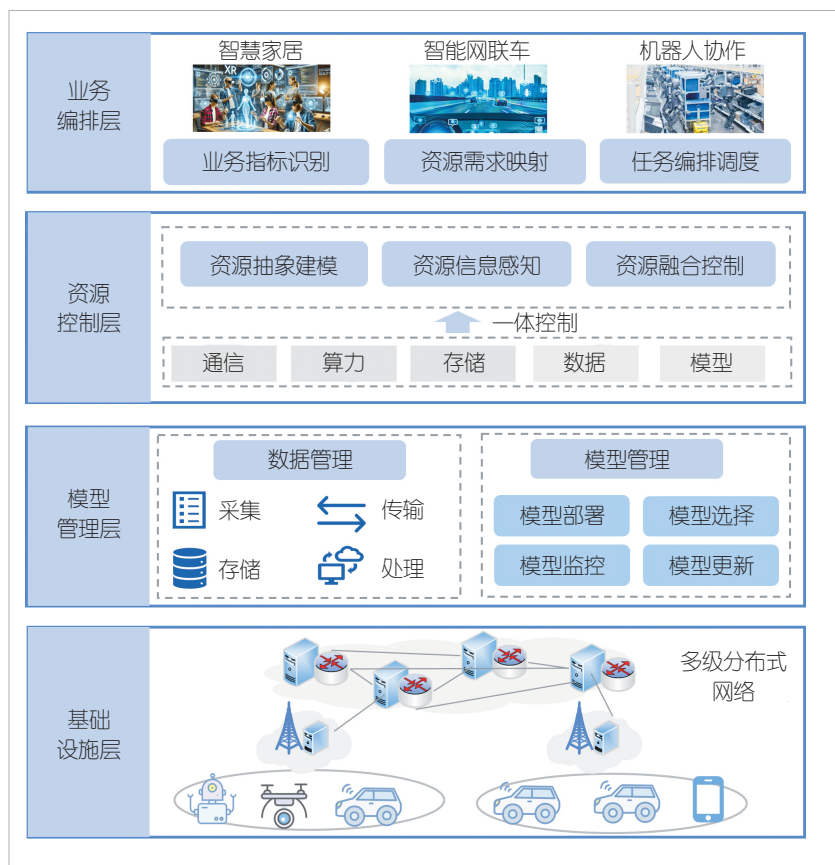


图1 6G内生智能网络架构

多维资源管理机制。针对上述问题，本节面向智能协同与业务多维指标保障需求，设计系统性能评价方法，突破传统单一维度评价的局限性，为系统整体效用优化提供模型基准。

端边协同系统评价模型涵盖基础资源开销、模型协同效率、能效比等多维指标。基础资源开销包括协同推理业务全周期的通信带宽、算力、存储等资源开销。模型协同效率需要考虑推理时延、准确度、模型复杂度等。能效比则聚焦单位任务的能耗，通过构建全过程能耗模型，覆盖通信传输、计算处理、模型更新等环节的能量消耗。为保证不同类型评价指标横向可比较，可利用动态加权、归一化等方法，将不同量纲的指标转换为可比较的数值，建立多维指标的系统综合评价模型。同时，系统评价模型具备灵活性，可根据不同业务场景特征（实时场景可能更关注时延和可靠性，离线场景可能更注重能效比），动态调整权重系数，满足不同场景下系统效用优化目标。

4 端边智能协同的6G数据管控

4.1 数据集构建

大规模、高质量数据是模型研发的基础性资源。一方

面，端边协同模型部署需要适配无线环境，而无线环境具备高时变特性和复杂传播机制，传统方法在信道参数估计精度和实时性方面难以满足端边智能协同业务低时延要求，面向真实无线环境的高质量信道数据采集存在挑战。另一方面，智能网联车、智能无人机、智能机器人等终端决策涉及动态环境中的复杂交互，其不同环境中的适应和泛化能力取决于环境数据多样性，多样化、高质量、真实环境数据采集存在挑战。由此可见，面向真实信道环境和物理世界的数据集构建，是使能6G智能终端交互的关键技术。

面向智能终端的信道数据集构建具体为：首先，基于大规模天线阵列和高性能信号处理平台，建立数据采集硬件基础；其次，开展多样化场景信道数据采集工作，并对原始数据进行降噪、格式统一等预处理；最后，利用智能终端天线方向图，结合实采无线信道参数与宽频信道响应的耦合计算方法，构建高还原度、可扩展的真实场景信道模型。

面向智能终端的环境数据集构建具体为：首先，建立包括感知基站、激光雷达等多模态数据感知平台，整合不同设备感知数据；其次，综合通信信号感知大范围环境轮廓、激光雷达点云测量姿态数据等多源数据优势，基于数据同步与配准算法，提升室内外环境地图构建效率与分辨率。

4.2 数据管控框架

端边智能协同业务需要实现跨设备数据融合与跨域数据共享，但现有数据采集框架主要面向核心网功能，缺乏端边分布式多级网络的跨域数据管控机制。为此，需构建支持数据采集、存储与共享在内的全生命周期管理框架，以保障端边多节点智能协同任务的高效可信交互。

在数据采集方面，建立统一的数据格式标准与规范化采集流程，支撑异构设备的标准化数据采集，并基于数据特征与流量需求自适应选择传输协议，优化采集传输效率。在数据存储方面，依据数据类型与实时性要求设计存储架构与索引机制，支持网络功能模块与跨域节点的高效数据调用。在数据共享方面，定义标准化的跨模块与跨节点数据共享接口，采用服务化调用机制触发共享流程，结合自适应传输协议确保共享效率。

5 端边智能协同的6G模型管理与协同

5.1 参考模型库构建

基于数据采集与建模，亟需面向不同端边智能协同场景构建AI模型。当前，不同用例、不同厂商在技术方案和模型设计方面存在较大差异，制约端边多节点模型智能协同标准化和产业落地。因此，亟需构建6G+AI参考模型库，提供典型场景基准和对比模型，并通过参考模型测试验证，指导工业界实现模型性能与复杂度的最优权衡。首先，构建面向6G典型端边智能协同用例的参考模型，给出面向用例的参考模型设计方案路线以及模型的设计架构和参数，确保模型的可复现性。其次，搭建模型验证评估仿真平台，集成开源数据集与评估工具链，实现全流程测试，同时建立模型评估标准，提供统一评估体系。

5.2 端边模型协同

随着智能网联车、机器人等新一代智能终端发展，智能应用对低时延、低开销、高精度推理等多维指标提出更高要求。然而，终端单体能力受限，无法支撑大模型高精度模型推理；同时，边侧单一静态模型推理受限于动态无线环境，无法自适应动态信道和多样化业务需求。因此，亟需构建端边模型协同机制，综合分布式、差异化端边节点能力，优化模型部署与动态模型协同推理策略。

端边模型部署方面，考虑到边缘节点、终端算力受限且分布式，需要综合考虑边缘节点分布、终端时空分布、业务请求时空分布等特征，设计时空驱动的分布式端边模型部署机制。针对具体业务推理场景，可监控模型协同指标完成情况，在完成指标下降时，基于数据管理框架按需向端边节点下发数据采集任务，并利用增强学习对融合数据和原始部署模型进行微调更新，提升模型对多样化业务适配泛化能力。

端边模型协同推理方面，侧重设计网络辅助的端边模型协同机制，利用分布式多级计算技术使能端边智能模型协同高效推理，支持终端对高算力智能业务需求。具体而言，设计适配动态信道的端侧模型拆分和推理任务分配方法，并结合模型部署状态，优化端边模型选择策略，满足端边推理业务精度、时延、能耗等多维指标要求。

6 端边智能协同的6G异构资源融合管控

6.1 异构资源融合控制调度

6G智能业务对通信、计算、数据、模型

需求多元，且终端分布和业务请求存在动态、时变特性，需要网络侧提供按需资源与服务调度，准确响应终端业务请求。当前，边缘计算业务分流主要根据业务部署和终端位置信息，模型调度主要依赖应用服务商集中管控，难以实现通信、计算、数据、模型等异构资源融合控制调度，制约资源高效使用。因此，亟需构建端边异构资源融合控制与调度机制，设计适配动态业务需求的端边异构资源融合管控机制。首先，对端侧与边侧分布式算力节点资源状态进行实时感知，生成全局资源状态视图。其次，当接收到业务请求时，基于业务需求、终端移动性等，生成最优的数据采集传输、边侧模型选择、通信计算资源分配的一体策略，在保障业务指标的同时提升网络整体资源利用率。

6.2 任务驱动灵活组网

针对机器人协作、智能车联等多智能终端协作任务场景，需要多个智能终端在完成自身智能业务的同时共同完成多终端协作任务。例如，文献[15]面向端边协同车联网场景，优化车车/车路协同下的模型拆分、组网与资源分配策略，以降低AI任务时延。由此可见，面向多智能终端协作场景，需精准识别协作任务在数据、模型等方面的需求，动态组建任务驱动的多终端-多边节点协作网络，并分配共享异构融合资源，实现多智能终端间、端边多节点间高效交互与协作。

7 端边智能协同赋能多智能终端协作

在应急救援、城市安防、智能车联等场景中，需多智能终端协作完成数据融合和统一智能决策。以城市安防为例（具体如图2所示），地面机器人完成地面目标捕捉和识别，无人机完成俯视环境感知和目标识别，智能眼镜完成人类视角画面捕捉和识别，3种终端均完成单一视角信息采集与智能识别业务，但单终端单一视角及受限算力、电池能力，很难支持城市安防快速搜索、全局判断需求。借助6G端边智能协同，可优化调度3种终端的能力，支持边缘节点分担端

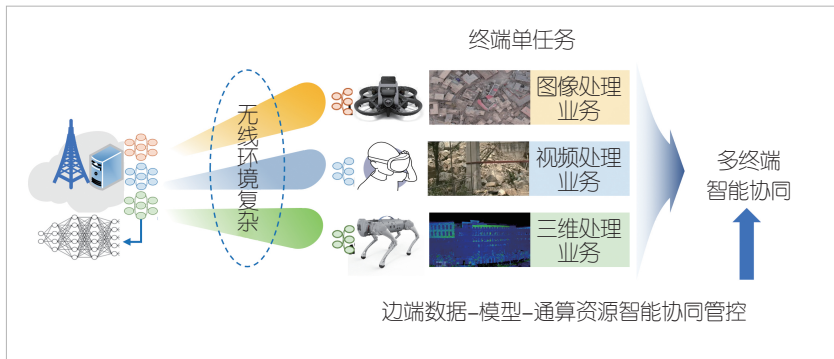


图2 端边智能协同支持城市安防业务场景示例

侧算力需求,实现多业务间协同处理,共同高效完成智能互联与融合决策任务。因此,端边协同的6G内生AI网络是解决复杂智能化业务的重要手段。

8 结束语

随着6G智能终端与智能业务演进发展,端边智能协同已成为支持6G智能终端复杂AI业务需求的有力手段。通过构建统一的端边数据模型管控框架、优化端边智能协同与异构资源融合管控机制,6G网络能够有效支撑AI手机、智能机器人、智能网联车等多形态终端的智能化、协同化业务,应对低时延、高精度等多维指标挑战。未来,随着技术持续演进和产业生态的完善,6G与AI深度融合将进一步释放智能终端的潜力,为垂直行业应用和社会数字化转型提供更强大的技术支撑,开启“万物智联”新篇章。

参考文献

- [1] IMT-2030(6G)推进组. 6G总体愿景与潜在关键技术白皮书[R]. 2021
- [2] IMT-2030(6G)推进组. 6G网络架构展望白皮书[R]. 2023
- [3] 中国信息通信研究院. 新一代智能终端蓝皮书[R]. 2024
- [4] Next G Alliance. 6G technologies for wide-area cloud evolution [R]. 2023
- [5] 6G-IA. White paper: European vision for the 6G network ecosystem [R]. 2024
- [6] IMT-2030(6G)推进组. 6G AI即服务(AlaaS)需求研究[R]. 2023
- [7] IMT-2020(5G)推进组, IMT-2030(6G)推进组. 移动通信与AI融合的数据格式和模型建议书(第一阶段:物理层领域) [R]. 2023
- [8] 3GPP. Study on architecture for 6G system: S2-2506096 [EB/OL]. (2025-05-23) [2025-06-18]. https://www.3gpp.org/ftp/tsg_sa/WG2_Arch/TSGS2_169_Fukuoka_2025-05/Docs
- [9] JIANG W, HAN H C, FENG D Q, et al. Energy-efficient and accuracy-aware DNN inference with IoT device-edge collaboration [J]. IEEE transactions on services computing, 2025, 18(2): 784-797. DOI: 10.1109/TSC.2025.3536311
- [10] LIU Z, TIAN M, DONG M, et al. MoEI: mobility-aware edge inference based on model partition and service migration [J]. IEEE transactions on mobile computing, 2024, 23(10): 9437-9450. DOI: 10.1109/TMC.2024.3366186
- [11] HE Y, FANG J, YU F R, et al. Large language models (LLMs)

inference offloading and resource allocation in cloud-edge computing: an active inference approach [J]. IEEE transactions on mobile computing, 2024, 23(12): 11253-11264 DOI: 10.1109/TMC.2024.3415661

- [12] 中国移动研究院. AI+通信业务白皮书[R]. 2025
- [13] 中国信息通信研究院. 具身智能发展报告[R]. 2024
- [14] 华为终端有限公司. AI终端白皮书[R]. 2025
- [15] LIU Z, DU H, LIN J, et al. DNN partitioning, task offloading, and resource allocation in dynamic vehicular networks: a Lyapunov-Guided diffusion-based reinforcement learning approach [J]. IEEE transactions on mobile computing, 2025, 24(3): 1945-1962. DOI: 10.1109/TMC.2024.3486728

作者简介



王志勤, 教授级高级工程师, 中国信息通信研究院副院长, 中国通信标准化协会副理事长、无线技术委员会主席, 中国通信学会无线移动委员会主任委员; 研究领域为移动通信技术、标准、产业及政策。



周吉喆, 中国信息通信研究院高级工程师、中国通信标准化协会TC5智能体通信网络子组副组长; 主要研究方向为6G网络、网络智能化技术; 主持和参与国家级/省部级项目10余项; 发表论文10余篇。



韩凯峰, 中国信息通信研究院高级工程师; 主要研究方向为面向6G的无线人工智能、通信感知一体化技术; 主持和参与国家级和省部级研究项目10余项; 发表论文60余篇。

多智能体协作感知的现状与展望



Multi-Agent Cooperative Perception: Current Status and Prospects

刘根嘉/LIU Genjia, 陈思衡/CHEN Siheng,
张文军/ZHANG Wenjun

(上海交通大学, 中国 上海 200240)
(Shanghai Jiao Tong University, Shanghai 200240, China)

DOI: 10.12142/ZTETJ.202504006

网络出版地址: <http://kns.cnki.net/kcms/detail/34.1228.TN.20250711.2051.008.html>

网络出版日期: 2025-07-14

收稿日期: 2025-05-08

摘要: 多智能体协作感知作为人工智能与分布式系统交叉的重要方向, 近年来在自动驾驶、无人系统编队、智能监控等领域受到广泛关注。梳理了该领域的研究背景与发展脉络, 重点讨论了通信机制优化、位姿噪声下的鲁棒性、异构协同、安全与隐私保护及仿真平台等研究热点。在此基础上, 探讨了实现高效通信的便捷方案, 并展望了未来发展趋势。可为相关研究提供参考与启示, 推动多智能体协作感知技术的持续进步。

关键词: 协作感知; 多智能体系统; 通信编码; 自动驾驶

Abstract: Multi-agent cooperative perception, as a key interdisciplinary direction between artificial intelligence and distributed systems, has attracted widespread attention in recent years in areas such as autonomous driving, unmanned system formation, and intelligent surveillance. This paper reviews the research background and development of this field, focusing on key topics including communication mechanism optimization, robustness under pose noise, heterogeneous cooperation, security and privacy protection, and simulation platforms. On this basis, this paper further explores a lightweight solution for efficient communication and discusses future research trends. This work aims to provide insights and references for researchers, promoting the continuous advancement of multi-agent cooperative perception technology.

Keywords: cooperative perception; multi-agent system; communication encoding; autonomous driving

引用格式: 刘根嘉, 陈思衡, 张文军. 多智能体协作感知的现状与展望 [J]. 中兴通讯技术, 2025, 31(4): 34-40. DOI: 10.12142/ZTETJ.202504006

Citation: LIU G J, CHEN S H, ZHANG W J. Multi-agent cooperative perception: current status and prospects [J]. ZTE technology journal, 2025, 31(4): 34-40. DOI: 10.12142/ZTETJ.202504006

1 多智能体协作感知背景

感知是智能系统面向物理世界的窗口, 是其执行任务的先决条件。2023年5月更新的《美国国家人工智能研发战略计划》^[1]指出, 增强智能系统的感知能力是一项需要长期投入的基础性研究工作。智能系统需整合各种来源(包括边缘设备)的数据, 以实现态势感知。然而传统单体感知方案无法克服因视线遮挡或远距离导致的感知信息缺失问题。这种物理限制和信息孤岛现象是单一智能体的固有瓶颈, 直接导致自动驾驶等智能系统难以在实际场景中部署落地。

针对上述问题, 借助现代通信技术, 多智能体协作感知技术应运而生。该技术在交通、生产、军事和科学探索等领域展现出显著的应用价值与发展潜力^[2]。传统感知方法通常基于单智能体架构进行环境信息采集。然而, 单一智能体受限于硬件和视野范围, 难以克服物体遮挡、视距约束及设备故障等固有问题。在此背景下, 多智能体协作感知技术通过

信息共享突破单智能体的感知局限, 增强系统容错能力, 成为诸多关键场景的重要解决方案^[3]。多智能体协作感知技术的发展可追溯至20世纪80年代的分布式人工智能研究, 1997年美国国家航空航天局(NASA)的“火星探路者”成功实现了人类首次地外多智能体协作探测, 为后续技术发展奠定了重要基础。进入21世纪后, 随着深度学习技术的突破性进展、5G低延迟通信和协作同步定位与地图构建(SLAM)算法等关键技术的日趋成熟, 多智能体协作感知技术正加速从实验室研究向实际应用跨越式发展。展望未来, 协作感知技术将向在智慧交通、远程医疗与国防等更广阔领域拓展。然而, 多智能体协作感知仍面临诸多挑战, 需提升六大关键能力: 协作通信效率能力、信息融合能力、异构协作能力、协作定位能力、安全隐私防护能力、仿真验证能力。

协作通信效率与信息融合能力是多智能体系统的神经中枢。协作感知面临的关键挑战是如何平衡感知性能与通信带宽。现实中的通信资源有限, 难以支持实时传输完整观测数据。为此, 系统需要筛选并传输紧凑有效的感知信息, 在降

基金项目: 国家自然科学基金项目(62450162)

低通信开销的同时保持良好的感知效果。通常自感知智能体首先根据与协作智能体的距离、自身的感知死角等信息动态连接协作智能体，随后协作智能体会对协作信息进行筛选和压缩，在进行低带宽传输的同时保证较低的感知失真。接收协作信息后，智能体需要对协作信息进行解码，并且通过数据融合模型，将自感知信息与协作感知信息融合，最终达到增强感知性能的目的。

异构协作与协作定位能力构成了系统的骨架支撑。随着协作规模扩大和下游任务复杂度提升，系统亟需具备动态成员管理能力，包括智能体的发现、身份认证与协作性验证等。特别是在多智能体之间存在传感器配置、任务目标和性能差异的异构场景下，如何实现跨异构智能体的有效协作，成为系统可扩展性与稳定性面临的重要挑战^[4-5]。精确的相对位姿估计是协作感知的基础，在全球定位系统（GPS）拒止环境中，需要基于视觉、激光雷达等传感器实现分布式协作定位，同时解决定位误差累积问题^[6]。

安全隐私与仿真验证能力是系统可靠应用的保障。安全方面需建立多层次防护体系：通信链路加密防止窃听，数据完整性校验抵御篡改^[7]，差分隐私技术保护敏感信息，如行人面部等。同时，系统应具备异常检测和自恢复能力，当恶意攻击者通过伪造或篡改传输数据（例如发送虚假障碍物信息）破坏系统可靠性时，系统需识别恶意数据并恢复原始信息。此外，仿真验证平台在降低开发成本和规避风险方面具有重要作用。理想的平台应支持复杂环境建模、智能体行为模拟以及感知算法的测试与评估。目前，英伟达的Isaac Sim、微软的AirSim等平台已初步具备多智能体协作感知的仿真能力，为相关研究与应用提供了有力支撑。

2 多智能体协作感知研究现状

2.1 通信效率

通信效率是协作感知系统实时部署和规模扩展所面临的关键瓶颈，这是因为在协作感知系统中，随着节点数量的增加和任务复杂性的提升，大规模协作不可避免。为保障系统的整体性能和通信资源利用效率，需要设计高效且紧凑的通信策略，以支撑大规模群体间的高效协作。信息完整度最高的通信策略是前期协作^[8]，如图1（a）所示，即在模型推理开始前进行协作，直接将原始观测数据作为协作信息进行传递。这种方法虽然保留了最完整的信息，但带宽开销巨大，难以

直接应用于高频实时的协作通信。为了降低通信量以适应有限的带宽，最直接的方式是进行后期协作，如图1（c）所示，即在模型推理完成后进行协作，仅传输感知结果，此时带宽利用效率较高。但这种方法对协作信息的融合总结能力有限，无法充分发挥协作优势。因此，中期协作开始受到研究者的关注，如图1（b）所示。这类方法传输压缩后的网络中间层特征，在通信开销与感知性能之间取得平衡。例如，V2VNet^[9]和DiscoNet^[10]分别引入熵编码技术与自动编码器，以便更高效地表示特征图；When2com^[11]则通过握手机制排除收益较低的协作对象，将全连接通信图优化为稀疏图结构，从而减少通信冗余。然而，现有方法大多只能采用固定压缩率，无法适应实际应用中的动态带宽情况。为此，一种可行的思路是，通过空间维度的特征筛选降低通信量，根据带宽限制自适应调整需要传输的和任务相关的特征区域。

2.2 位姿扰动下的鲁棒协作

在多智能体协作感知系统中，位姿对齐是实现多个智能体间数据同步和协作任务的重要基础。每个智能体通过其独立的定位系统来估算位姿。这些位姿数据需要在统一的坐标系中进行对齐，才能保证多个智能体的协作效率。然而，在实际应用中，定位系统并非总能提供精准的位姿信息，尤其是在复杂环境中，位姿估计常常受到传感器精度限制或环境干扰的影响，导致定位误差。这些位姿误差会影响智能体间的相对位姿估计，进而削弱协作感知的效果。为了应对这一问题，近年来研究者提出了多种方法来校正位姿扰动。普遍的思路是，根据 N 个视角对 M 个物体级或像素级的锚点

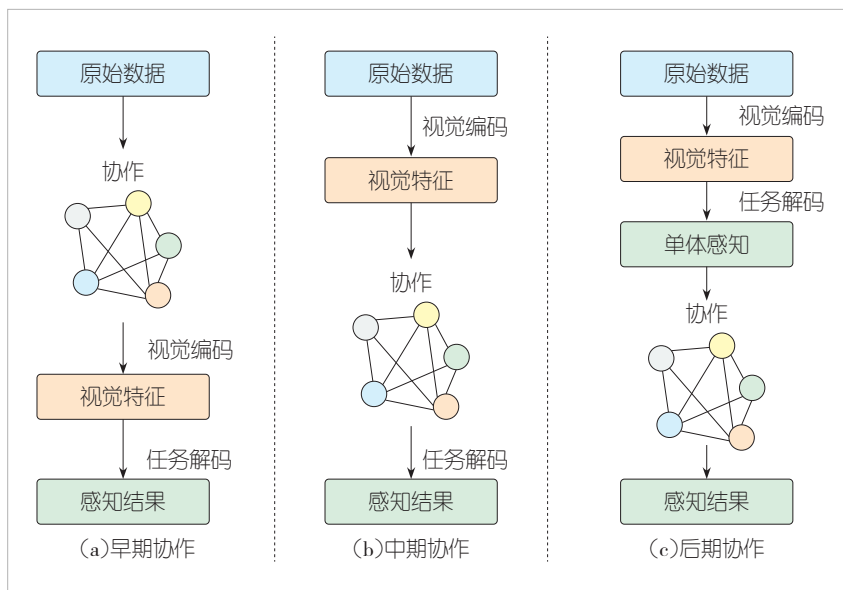


图1 前、中、后期协作示意图

$\left\{ \left\{ a_{ij} \right\}_{j=1}^M \right\}_{i=1}^N$ 的相对观测值, 对齐 N 个不同视角的位姿 $\{ \xi_i \}_{i=1}^N$,

可建模为如下优化问题, 其中 d 是位姿匹配误差度量函数:

$$\min_{\{ \xi_i \}_{i=1}^N} d \left(\left\{ \left\{ a_{ij} \right\}_{j=1}^M, \xi_i \right\}_{i=1}^N \right) \quad (1).$$

例如, MASH^[12] 采用像素级锚点计算相对位姿关系, 通过构建相似度体积, 显式地学习像素级的对应关系, 从而避免使用噪声位姿进行推理。FPV-RCNN^[13] 通过推断关键点的语义标签, 并采用最大共识算法来建立智能体之间的对应关系, 校正位姿误差。CoAlign^[6] 通过智能体-物体图优化, 进一步提高了位姿一致性, 从而减少由定位误差引起的影响。RoCo^[14] 提出了一种创新的无监督框架, 通过迭代的物体匹配和智能体位姿调整, 解决协作感知中由于位姿误差导致的特征对齐问题, 并通过图优化过程最小化关联物体的对齐误差, 从而提升协作物体检测的性能和鲁棒性。尽管上述方法在一定程度上提升了位姿鲁棒性, 但仍存在对初始估计依赖较强、难以应对大尺度场景中高频动态变化的问题。未来研究可探索通过时序一致性约束实现更稳定的跨帧位姿优化。

2.3 异构场景下的协作

异构性是协作感知系统面临的关键挑战之一, 体现在多个层面: 不同传感器模态、不同模型架构和不同参数配置等, 每种异构性都会增加协作信息的融合难度。利用不同尺寸、模态、数据分布的视觉特征进行融合协作, 通常有两种方式: 1) 针对新的异构种类增添新的特征适配器或单独的特征融合模块; 2) 通过数据驱动的方式学习统一的特征空间。HMTiT^[15] 提出了一种异构图注意力网络, 对不同种类的信息融合的学习进行区分, 从而降低异构协作的学习难度。MPDA^[16] 提出多智能体感知领域适配方法, 利用对抗领域对齐的学习框架, 促进异构智能体对之间的特征分布对齐。然而, 随着智能体数量的增加, 上述方法的扩展性受到限制。对此, HEAL^[4] 提出了反向对齐训练策略, 通过固定基础网络的解码器并仅训练编码器来创建异构模型, 实现现有智能体的协作。PolyInter^[17] 通过引入多态特征解释器, 在异构环境中无须训练新解释器, 利用可扩展的学习参数集成新智能体, 显著提高了协作感知的精度并保持了较低参数调整需求。STAMP^[18] 提出了一种可扩展的任务和模型无关的协作感知框架, 通过使用轻量级适配器-还原器对在共享协议域和特定智能体域之间转换鸟瞰图特征, 促进高效的特征共享与融合, 同时降低计算开销。已有方法在特定类型的异构协作中取得了效果, 但在面对传感器类型过多或新设备频繁接入时, 系

统适配可能仍面临困难。后续研究可以考虑构建泛化性更强的特征适配模块, 减少对每种设备单独设计的依赖。

2.4 安全与隐私

协作感知依赖智能体间的信息共享, 虽显著提升了整体感知能力, 但也带来了更多攻击风险与隐私泄露隐患。恶意智能体可能通过伪造、篡改或注入错误感知数据, 干扰系统判断, 甚至引发误识别、漏检等安全事故。因此, 构建协作环境下的安全防护机制已成为研究焦点。目前的主流方法聚焦于异常攻击检测, 通过对多智能体感知结果的一致性分析, 实现对潜在恶意行为的识别与隔离。例如, ROBOSAC^[19] 采用迭代方式并结合匈牙利匹配算法, 在协作过程中实现感知一致性, 以识别并剔除可能的恶意智能体。MADE^[7] 在此基础上引入多测试框架, 通过匹配损失与重建损失的联合判别策略, 实现更稳健的感知一致性验证。尽管上述基于异常检测的策略在识别明显攻击方面表现良好, 但面对更加隐蔽、精准的攻击时, 仍存在较大漏洞。为此, 最新研究开始集中于中间特征级别的一致性检测方向探索, 以提升检测的灵敏度和效率。CP-Guard^[20] 通过引入特征级的恶意智能体检测机制以及双中心对比损失, 在无需验证最终感知结果的前提下有效区分正常与恶意特征, 大幅降低计算开销并增强系统鲁棒性。GCP^[21] 则结合空间一致性与时间行为分析, 利用贝叶斯-霍克测试与运动流重建, 实现对细粒度持续性攻击的高精度检测。目前的方法大多依赖已知攻击模式或规则进行检测, 对于未知类型或动态演化的攻击应对能力仍不足。未来可以考虑引入在线学习或自适应更新机制, 使系统在运行过程中持续优化其安全策略, 提升应对未知威胁的能力。

2.5 环境仿真

在协作感知研究中, 仿真环境已成为连接算法原型与实际道路的重要桥梁。相比高成本、难复现的实车实验, 仿真平台提供了“零风险、可控、可复现”的试验场。研究人员可灵活调节交通密度、天气等变量, 定向构建极端或长尾场景, 并进行多轮回放与消融实验, 系统分析性能瓶颈。

当前仿真方案可大体分为3类技术路线。第1类是端到端数字孪生型平台, 例如以CARLA^[22] 为代表的城市级仿真。作为协作感知场景最常用的仿真平台, 此类平台在视觉与物理真实感上保持高一致性, 支持多车、多传感器和交通行为建模, 可直接用于闭环自动驾驶实验。然而, 其高保真亦意味着对图形处理器 (GPU) 渲染与存储性能的极大消耗, 同时在快速导入真实道路素材、扩展路侧和空中视角方面仍需二次开发。第2类是图像重渲型方法, 以 RoCo-Sim^[23] 等工作

为典型，将高精度三维车辆资产嵌入真实路侧摄像机图像中，实现多视角几何一致渲染。此类技术继承了真实背景的照明与材质信息，生成速度可较端到端平台提高一个量级，尤其适合固定相机与车-路协作的感知-通信研究；但现阶段主要针对静态背景，对大范围动态场景的支持有限。第3类是显式体渲染或 3D Gaussian Splatting^[24] 方案，通过稀疏高斯体或稀疏体素完成连续视图重绘，可较传统基于三角网格的渲染速度提高数十到百倍，并天然支持可微分重定位与外观迁移，已成为快速合成大规模训练数据的新热点。

3 多智能体协作的通信便捷方案

传统基于熵编码和自编码器的压缩技术能在一定程度上减小协作通信量，但在实际感知任务中仍存在明显局限：大量空间区域并不包含关键信息，传输完整的特征图会造成显著带宽浪费；同时，高维特征通道也带来额外的通信开销。本文介绍一种多智能体协作的通信便捷方案，从空间和特征通道这两个维度筛选信息并压缩协作通信量，如图2所示。在空间维度上，采用一种基于置信度补全的信息选择机制：每个智能体根据自身感知结果生成信息评分图，用以评估不同空间区域的重要性，并根据多个协作者的信息评分图筛选关键区域的稀疏但有价值的特征。在特征通道维度上，引入可学习码本进行向量量化，用离散的码本索引代替高维连续特征，从而降低单位地块通信量。

3.1 基于置信度补全的信息选择

为了高效选择对其他智能体有支持作用的紧凑协作消

息，每个智能体应有选择性地分享与任务相关的关键信息，从而使多个智能体所提供的非冗余信息实现互补，满足彼此的信息需求。为实现这一目标，本信息选择策略包含信息评分生成和信息选择优化两个核心步骤。

首先，对于智能体 i ，每个智能体使用信息评分生成器 $\Phi_{\text{generator}}$ 基于自身的特征图 $\mathcal{F}_i$ 生成信息评分图 C_i ：

$$C_i = \Phi_{\text{generator}}(\mathcal{F}_i) \in [0, 1]^{H \times W} \quad (2)$$

图 C_i 反映其不同空间区域上的可用信息量。之后该评分图被广播给所有其他智能体，从而促进智能体之间对彼此可支持信息的全面认知。其中 H 、 W 代表特征图的分辨率， $\Phi_{\text{generator}}$ 由感知解码器实现，通过赋予包含目标区域更高的信息评分，帮助补全漏检内容。

基于所有 N 个智能体共享的评分图 $\{C_i\}_{i=1}^N$ ，每个智能体在本地优化信息选择矩阵，选择最适合支持其他智能体的消息内容。由于每个智能体仅在某些空间区域需要特定信息来实现精准感知，超出部分的信息并不会带来明显收益，因此每个支持者需优先选择那些评分较高且与其他信息来源不冗余的空间区域来提供支持。接收者在信息需求被满足后将停止进一步的消息选择，以避免通信浪费。

具体而言，该问题被建模为一个约束优化问题，其目标是为每对支持者-接收者智能体计算出一个二值选择矩阵，以确定是否在特定空间位置传输信息。记 $M_{i \rightarrow j} \in \{0, 1\}^{H \times W}$ 为在鸟瞰图上定义的二值选择矩阵，其中矩阵中的每个元素表示：智能体 i 是否应在该空间位置向智能体 j 发送信息（1 表示发送，0 表示不发送），求解目标如下所示：

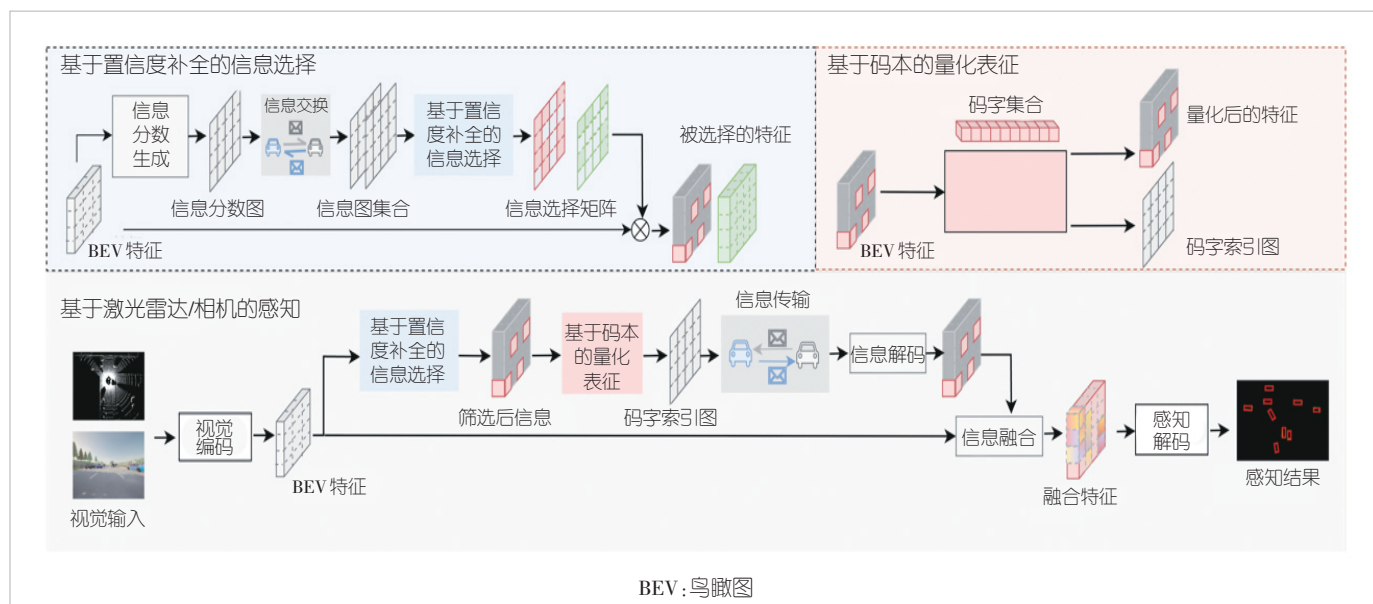


图2 多智能体协作的通信便捷方案

$$\{\mathbf{M}_{i \rightarrow j}^*\}_{i,j} = \operatorname{argmax}_M \sum_{j=1}^N \int_{\min} \left(\mathbf{C}_j + \sum_{i=1, i \neq j}^N \mathbf{M}_{i \rightarrow j} \odot \mathbf{C}_i, u \right) (3),$$

其中, $\sum_{i,j=1, j \neq i}^N \mathbf{M}_{i \rightarrow j} \leq b, \mathbf{M}_{i \rightarrow j} \in \{0,1\}^{H \times W}$, $\odot$ 代表点乘, u 是一个代表信息需求阈值的超参数, b 代表通信量预算, $\int_{\min}$ 计算选择矩阵与阈值之间的较小值。基于选择矩阵 $\{\mathbf{M}_{i \rightarrow j}^*\}_{j=1}^N$, 智能体 i 为 j 发送的信息最终为 $\mathcal{Z}_{i \rightarrow j} = \mathbf{M}_{i \rightarrow j}^* \odot \mathcal{F}_i$ 。

3.2 基于码本的表征量化压缩

为高效表示高维特征图 $\mathcal{Z}_{i \rightarrow j}$, 智能体采用一种基于码本的消息表示方法, 以降低在通道维度上的通信开销。其核心思想是: 利用一个任务驱动的码本, 将高维特征向量近似为码本中最接近的一个码字, 从而只需传输整数形式的码字索引, 无须传输完整的浮点特征向量, 大幅减少通信数据量。

具体而言, 令 $F = \{\mathcal{F}^{(i,s)}\}_{i=1, s=1}^{N,S}$ 表示由所有 N 个智能体在 S 个训练场景中通过观测编码器提取的鸟瞰图 (BEV) 特征集合。令 $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_{n_L}] \in \mathbb{R}^{C \times n_L}$ 表示码本, $\mathbf{D}_{[l]} = \mathbf{d}_l \in \mathbb{R}^C$ 表示第 l 个码字, n_L 代表码字总数, C 代表原始特征维度。码本学习优化目标可表示为:

$$\mathbf{D}^* = \operatorname{argmin}_{\mathbf{D}} \sum_{\mathcal{F} \in F} \sum_{h,w} \min_l \left(\Psi(\mathbf{D}_{[l]}) + (\mathcal{F}_{[h,w]} - \mathbf{D}_{[l]})_2^2 \right) (4),$$

其中, $\Psi(\cdot)$ 表示将原始特征用码字近似表征后的目标检测表现。公式 (4) 中第一项对应下游检测任务的需求, 第二项则反映了原始特征向量与所选码字之间的重构误差。尽管该近似在重构上是有损的, 但对于感知任务而言是无损的, 从而在不影响感知能力的前提下有效降低通信开销。在测试阶段, 基于共享的码本 $\mathbf{D}$, 每个智能体可以将所选的稀疏特征图 $\mathcal{Z}_{i \rightarrow j}$ 替换为一系列码字索引 $\mathcal{I}_{i \rightarrow j}$ 。对于每一个 BEV 空间位置 (h, w) , 其对应的码字索引通过以下方式获得:

$$(\mathcal{I}_{i \rightarrow j})_{[h,w]} = \operatorname{argmin}_l \left((\mathcal{Z}_{i \rightarrow j})_{[h,w]} - \mathbf{D}_{[l]} \right)_2^2 (5)。$$

基于码本的向量量化压缩在降低通信开销方面具有显著优势, 但需要注意多个关键难点。一是, 码本容量的选择需根据具体任务复杂度进行动态调整: 容量过小可能导致信息表达能力不足, 影响感知精度; 容量过大则会增加存储与通信负担, 削弱压缩效果。二是, 码本训练需在特征重建率与任务适配性之间实现权衡。为此, 码本训练同时采用了重建损失和感知损失。这种双重优化机制要求量化后的向量尽可能保留原有信息, 能对关键任务保持高敏感性。

这种基于码本的表示方式具有三大优势: 1) 高效性: 通过传输轻量级的码字索引, 大幅减少通信负载; 2) 适应性: 可通过调整码本配置灵活适应不同的通信资源条件, 较

小的码本提升通信效率, 较大的码本增强感知性能; 3) 可扩展性: 提供统一的共享表示格式, 使得新加入的异构智能体能够通过将其有效的感知特征基加入码本, 轻松融入协作系统。信息接收端将根据索引和共享的码本还原协作信息, 并与本地信息融合用于最终的感知推理。

3.3 实验分析

图 3 展示了置信度筛选和码本压缩在 DairV2X^[25] 真实协作数据集上不同通信带宽下的感知精度。实验设置包括: 检测范围为 204.8 m × 102.4 m, 感知分辨率为 0.4 m × 0.4 m, 最大通信半径为 100 m, 码本容量为 256, 评估指标为在重合度 (IoU) 阈值为 0.3、0.5、0.7 时的平均精度 (AP)。从结果中可以观察到, 该方法展现出以下显著优势: 1) 在所有通信带宽条件下, 置信度筛选+码本压缩方法均优于传统的前期协作与后期协作策略。尤其值得注意的是, 该方法仅使用约 1/256 的通信量, 便可达到前期协作的感知精度。2) 基于置信度筛选的机制使系统具备了自适应通信能力, 即使在 60 倍压缩率下仍保留 95% 的完整通信精度。相比之下, 其他方法往往只能在固定带宽下发挥作用, 缺乏灵活性。3) 基于码本学习的压缩方案显著提升了特征通道维度上的压缩效率, 实现了高达 100 倍的通信压缩比, 充分展现了码本在协作感知中的有效性与可扩展性。

4 未来研究展望

4.1 空天地异构

空天地一体化系统正成为多智能体协作感知的重要发展方向。在灾害应急场景中, 卫星可快速获取全局图像, 无人机提供中尺度巡查, 地面传感器则用于局部高精度监测, 三类智能体形成互补。在智慧交通中, 卫星导航、空中平台监控与地面路侧单元协同, 可实现复杂路况下的实时感知与调度。此外, 该体系在低空物流、森林防火、农业遥感等新型场景中同样具备应用潜力, 已被《数字交通“十四五”规划》与多地“空天信息产业行动计划”纳入重点建设内容。但由于空、天、地智能体之间存在显著异构性, 协作面临多方面技术挑战: 感知数据在分辨率、时间同步性和观测范围上存在差异, 难以直接融合; 通信协议与链路多样, 需应对跨平台互通、链路切换与频谱冲突问题; 计算资源分布不均, 星载/空载平台计算能力受限, 需动态任务分配; 智能体运动特性不同, 轨迹调度复杂。此外, 多域协作中还涉及数据安全、身份认证等系统性难题。

针对这些问题, 未来研究可从以下几个方向推进: 构建

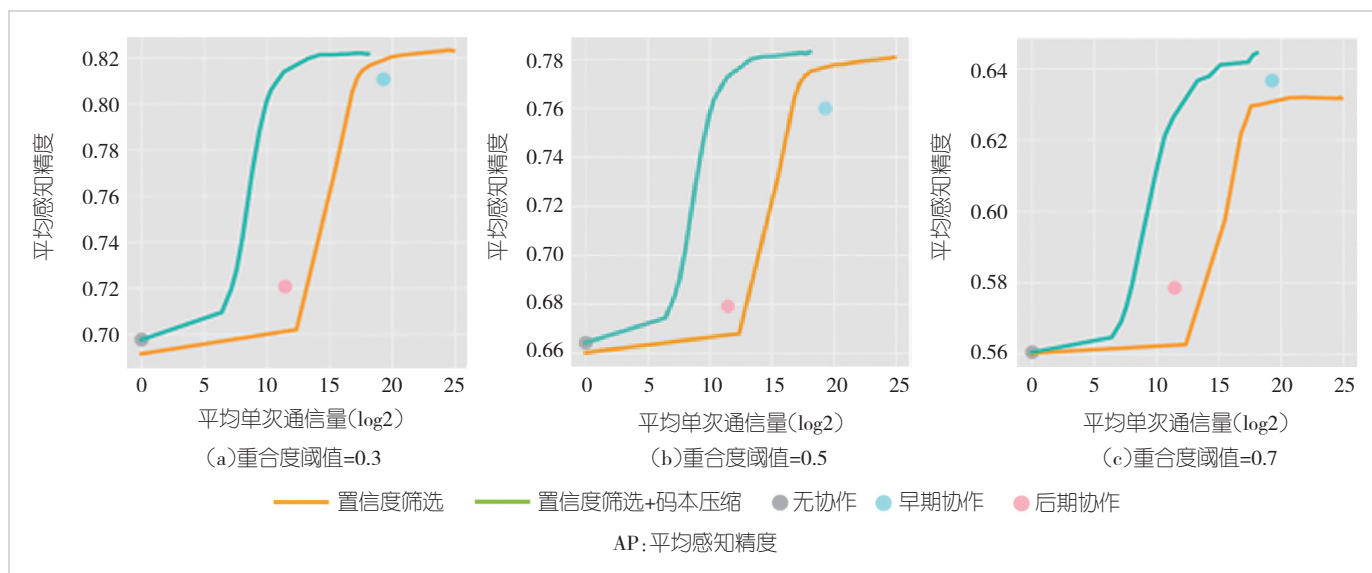


图3 不同通信量下的协作感知精度

统一的多源异构数据融合框架，实现时空对齐与语义整合；设计适用于星-空-地的通信协议栈与弹性链路调度机制；设计感知任务驱动的计算调度算法，实现跨平台的推理负载均衡；开发面向异构运动模型的多智能体协作控制策略，提升任务执行一致性；同时，完善分层安全机制，确保跨域协同过程中的数据隐私与系统安全。

4.2 多具身智能协作

具身智能正从单一智能体系统迈向“多形态、多功能、跨时空”的协作体。在工业、物流、交通等领域，地面车辆、机械臂、空中无人机等多种智能体协同完成复杂任务，展现出显著的效率与灵活性。例如，配送车与空中无人机的协同投递、道路清障机械臂与巡检车的协同运维，以及车与路侧单元组成的混合网络^[26]，均已进入研究视野。这一趋势不仅为传统移动机器人注入新动能，也对感知、任务分配与实时调度提出了全新要求。

与此同时，大规模协作也伴随协作安全、通信拥塞与伦理治理等挑战。集中式规划在大规模系统中难以适用，去中心化的图神经网络与事件驱动通信机制成为研究热点^[27]。目前，缺乏覆盖多种智能体综合任务的统一评测体系，这限制了算法的横向比较与产业落地。未来研究需在安全性、实时性与可评测性之间寻求平衡，并引入更加开放的跨域验证基准，促使学术界与产业界达成共识。

5 结束语

本文梳理了多智能体协作感知的研究脉络与技术发展，围绕通信效率、位姿鲁棒性、异构协作、安全隐私和仿真平

台五大关键问题分析了研究现状，提出了一种面向动态带宽的通信便捷方案，通过空间维度的置信度筛选与特征通道维度的码本量化压缩，实现了通信负载与感知精度之间的高效权衡。实验结果表明，该方法在多种带宽条件和协作场景下均显著优于现有主流方案，具备良好的带宽自适应性和较高感知精度。此外，本文还对空天地一体化协作与多具身系统体的未来发展方向进行了展望，指出协作感知将在异构智能体融合、跨域通信调度与安全协同等方面持续面临挑战。总体而言，本文梳理了协作感知研究中的关键问题，并针对通信效率这一问题提供了一种可行的解决方案。

致谢

感谢上海交通大学在读研究生胡安宁、刘畅行，以及南京理工大学在读研究生晁子晨对本研究工作的支持！

参考文献

- [1] National Artificial Intelligence Agency. Research and development blueprint plan: 2023 update [EB/OL]. (2023-05-23) [2025-04-05]. <https://www.nitrd.gov/pubs/National-Artificial-Intelligence-Research-and-Development-Strategic-Plan-2023-Update.pdf>
- [2] ALOTAIBI E T, ALQEFARI S S, KOUBAA A. LSAR: multi-UAV collaboration for search and rescue missions [J]. IEEE access, 2019, 7: 55817-55832
- [3] CHEN Q, MA X, TANG S H, et al. F-cooper: feature based cooperative perception for autonomous vehicle edge computing system using 3D point clouds [C]//Proceedings of the 4th ACM/IEEE Symposium on Edge Computing. ACM, 2019: 88-100. DOI: 10.1145/3318216.3363300
- [4] LU Y F, HU Y, ZHONG Y Q, et al. An extensible framework for open heterogeneous collaborative perception [EB/OL]. (2024-01-25) [2025-04-05]. <https://arxiv.org/abs/2401.13964>
- [5] MENG F, HUANG Y M, LU Z H, et al. Multi-user mmWave beam

- tracking via multi-agent deep Q-learning [J]. ZTE communications, 2023, 21(2): 53–60. DOI: 10.12142/ZTECOM.202302008.
- [6] LU Y F, LI Q H, LIU B A, et al. Robust collaborative 3D object detection in presence of pose errors [C]//Proceedings of IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2023: 4812–4818. DOI: 10.1109/ICRA48891.2023.10160546
- [7] ZHAO Y H, XIANG Z, YIN S, et al. MADE: malicious agent detection for robust multi-agent collaborative perception [C]//Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2024: 13817–13823. DOI: 10.1109/IROS58592.2024.10801337
- [8] CHEN Q, TANG S H, YANG Q, et al. Cooper: cooperative perception for connected autonomous vehicles based on 3D point clouds [C]//Proceedings of IEEE 39th International Conference on Distributed Computing Systems (ICDCS). IEEE, 2019: 514–524. DOI: 10.1109/icdcs.2019.00058
- [9] WANG T H, MANIVASAGAM S, LIANG M, et al. V2VNet: vehicle-to-vehicle communication for joint perception and prediction [M]//Computer Vision – ECCV 2020. Cham: Springer International Publishing, 2020: 605–621. DOI: 10.1007/978-3-030-58536-5_36
- [10] LI Y, REN S, WU P, et al. Learning distilled collaboration graph for multi-agent perception [EB/OL]. (2021-11-21)[2025-04-05]. <https://arxiv.org/abs/2111.00643>
- [11] LIU Y C, TIAN J J, GLASER N, et al. When2com: multi-agent perception via communication graph grouping [C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020: 4105–4114. DOI: 10.1109/CVPR42600.2020.00416
- [12] GLASER N, LIU Y, TIAN J, et al. Overcoming obstructions via bandwidth-limited multi-agent spatial handshaking [EB/OL]. (2021-07-01)[2025-04-05]. <https://arxiv.org/abs/2107.00771>
- [13] YUAN Y S, CHENG H, SESTER M. Keypoints-based deep feature fusion for cooperative vehicle detection of autonomous driving [J]. IEEE robotics and automation letters, 2022, 7(2): 3054–3061. DOI: 10.1109/LRA.2022.3143299
- [14] HUANG Z, WANG S, WANG Y, et al. RoCo: robust cooperative perception by iterative object matching and pose adjustment [C]//Proceedings of the 32nd ACM International Conference on Multimedia. ACM, 2024: 7833–7842
- [15] XIANG H, XU R S, MA J Q. HM-ViT: hetero-modal vehicle-to-vehicle cooperative perception with vision transformer [C]//Proceedings of the IEEE/CVF international conference on computer vision. IEEE, 2023: 284–295
- [16] XU R S, LI J L, DONG X Y, et al. Bridging the domain gap for multi-agent perception [C]//Proceedings of IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2023: 6035–6042. DOI: 10.1109/ICRA48891.2023.10160871
- [17] XIA Y C, YUAN Q, LUO G Y, et al. One is plenty: a polymorphic feature interpreter for immutable heterogeneous collaborative perception [C]//Proceedings of the Computer Vision and Pattern Recognition Conference. IEEE, 2025: 1592–1601
- [18] GAO X, XU R, LI J, et al. STAMP: scalable task and model-agnostic collaborative perception [EB/OL]. (2025-01-24)[2025-04-05]. <https://arxiv.org/abs/2501.18616>
- [19] LI Y M, FANG Q, BAI J M, et al. Among us: adversarially robust collaborative perception by consensus [C]//Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2023: 186–195. DOI: 10.1109/ICCV51070.2023.00024
- [20] HU S, TAO Y, FANG Z, et al. CP-guard+: a new paradigm for malicious agent detection and defense in collaborative perception [EB/OL]. (2025-02-07) [2025-04-05]. <https://arxiv.org/abs/2502.07807>
- [21] TAO Y, HU S, HU Y, et al. GCP: guarded collaborative perception with spatial-temporal aware malicious agent detection [EB/OL]. (2025-01-05)[2025-04-05]. <https://arxiv.org/abs/2501.02450>
- [22] DOSOVITSKIY A, ROS G, CODEVILLA F, et al. CARLA: an open urban driving simulator [EB/OL]. (2017-11-10)[2025-04-05]. <https://arxiv.org/abs/1711.03938>
- [23] DU Y W, HU A N, CHAO Z C, et al. RoCo-sim: enhancing roadside collaborative perception through foreground simulation [EB/OL]. (2025-03-13)[2025-04-05]. <https://arxiv.org/abs/2503.10410>
- [24] KERBL B, KOPANAS G, LEIMKUHLER T, et al. 3D gaussian splatting for real-time radiance field rendering [EB/OL]. (2023-08-08)[2025-04-05]. <https://arxiv.org/abs/2308.04079>
- [25] YU H B, LUO Y Z, SHU M, et al. DAIR-V2X: a large-scale dataset for vehicle-infrastructure cooperative 3D object detection [C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2022: 21329–21338. DOI: 10.1109/CVPR52688.2022.02067
- [26] LIU G J, HU Y, XU C X, et al. Towards collaborative autonomous driving: simulation platform and end-to-end system [J]. IEEE transactions on pattern analysis and machine intelligence, 2025: 1–18. DOI: 10.1109/TPAMI.2025.3560327
- [27] LIU C X, LIU G J, WANG Z J, et al. CoLMDriver: LLM-based negotiation benefits cooperative autonomous driving [EB/OL]. (2025-03-11)[2025-04-05]. <https://arxiv.org/abs/2503.08683>

作者简介



刘根嘉, 上海交通大学在读博士研究生; 主要研究领域为端到端自动驾驶、多智能体协作等。



陈思衡, 上海交通大学人工智能学院副教授; 主要研究领域为多智能体系统、大模型智能体等; 入选国家级人才计划青年项目, 承担了国家自然科学基金原创探索计划项目及面上项目、科技部人工智能2030重大项目、上海市科委人工智能专项等科研项目; 曾获得IEEE信号处理协会最佳青年作者论文奖、ASME结构检测协会最佳论文Runner-Up奖、2018 GlobalSIP会议最佳论文奖、三菱电机实验室总统奖等; 发表论文100余篇, 谷歌引用10 000余次。



张文军, 上海交通大学讲席教授、国家自然科学基金委数字媒体通信创新群体带头人、杰出青年基金获得者、教育部首批“长江学者”特聘教授、国家百千万工程领军人才、国务院特殊津贴专家、中国电子学会会士、IEEE Fellow; 主要研究领域为生式视频通信与媒体网络融合; 以第一完成人获得国家科技进步奖二等奖4项, 并获中国专利金奖、何梁何利科技进步奖和广电行业CCBN学术贡献奖等荣誉; 发表论文140余篇, 获授权发明专利236项(含35项国际发明专利), 主编工业和信息化部“十四五”规划教材《视频通信》。

基于生成式模型的新型机器语义通信方法研究



Novel Generative Model-Based Machine-Type Semantic Communications

承楠/CHENG Nan¹, 王秀程/WANG Xiucheng¹,
沈学民/SHEN Xuemin²

(1. 西安电子科技大学, 中国 西安 710071;

2. 滑铁卢大学, 加拿大 滑铁卢 N2L3G1)

(1. Xidian University, Xi'an 710071, China;

2. University of Waterloo, Waterloo N2L3G1, Canada)

DOI: 10.12142/ZTETJ.202504007

网络出版地址: <http://kns.cnki.net/kcms/detail/34.1228.tn.20250716.1341.007.html>

网络出版日期: 2025-07-16

收稿日期: 2025-06-10

摘要: 创新性地提出了一种基于隐扩散模型 (LDM) 的语义通信架构, 能够利用预训练生成模型实现无需重新训练的语义特征还原。借助 LDM 固有的去噪能力, 该架构在面对信道噪声扰动和分布外输入时, 依然能够保持稳定的重构性能。此外, 该架构支持灵活集成外部大模型资源, 显著提升了语义通信系统的演进能力。在大规模图像数据集上的实验结果表明, 该方法在低信噪比 (SNR) 条件下仍具有优异的图像恢复与语义保真能力, 尤其在基于学习的感知相似度等语义指标上显著优于现有主流方法。本研究为资源受限设备的机器语义通信提供了全新思路, 具有良好的实用性与推广潜力。

关键词: 语义通信; 隐扩散模型; 机器通信; 机器视觉; 信道鲁棒性

Abstract: A novel semantic communication architecture based on the Latent Diffusion Model (LDM) is proposed. This architecture leverages pre-trained generative models to enable semantic feature reconstruction without the need for task-specific retraining. By exploiting the intrinsic denoising capability of LDM, the proposed system maintains stable reconstruction performance under channel noise perturbations and out-of-distribution inputs. Additionally, the architecture supports the flexible integration of external large-scale models, significantly enhancing the evolutionary capability of semantic communication systems. Experimental results on large-scale image datasets demonstrate that our method achieves superior image recovery and semantic fidelity, even under low signal-to-noise ratio (SNR) conditions. In particular, it significantly outperforms existing mainstream approaches in terms of learning-based perceptual similarity metrics. This study offers a new perspective for deploying machine-type semantic communication on resource-constrained devices and exhibits strong potential for practical applications and large-scale deployment.

Keywords: semantic communication; Latent Diffusion Model; machine-type communication; machine vision; channel robustness

引用格式: 承楠, 王秀程, 沈学民. 基于生成式模型的新型机器语义通信方法研究 [J]. 中兴通讯技术, 2025, 31(4): 41-47. DOI: 10.12142/ZTETJ.202504007

Citation: CHENG N, WANG X C, SHEN X M. Novel generative model-based machine-type semantic communications [J]. ZTE technology journal, 2025, 31(4): 41-47. DOI: 10.12142/ZTETJ.202504007

随着通信系统从 5G 加速向 6G 演进, 物联网中各类监测设备数量激增, 传统以比特无差错传输为目标的通信方式面临严重的带宽瓶颈和高时延问题, 难以满足机器通信对实时性与能效的需求。以智慧城市为例, 海量部署的摄像头与传感器广泛应用于目标检测、人脸识别等高层视觉任务。然而, 由于边缘设备在算力与能耗方面的限制, 终端通常需将大量原始数据上传至云端处理, 这不仅导致通信拥

塞, 还可能造成服务质量明显下降^[1]。为应对这一挑战, 国际电信联盟提出从“传输比特”向“传递语义”的通信范式转型, 旨在支撑超高速率、低时延与高密度连接需求, 从而更好地服务大规模机器通信与人机融合等 6G 关键应用场景^[2]。语义通信通过显著压缩信息冗余、降低通信时延并增强数据安全性, 顺应了 6G “通信-感知-计算一体化”的发展趋势, 为构建高效、智能且可靠的下一代机器通信系统提供了新路径^[3]。

近年来, 得益于神经网络在特征表达与数据压缩方面的

基金项目: 国家重点研发计划项目 (2024YFB907500)

强大能力，基于深度学习的语义通信系统取得了显著进展^[4]。研究表明，神经网络能够通过自动提取数据中的语义特征，显著提升通信效率，甚至实现极低码率下的高保真重构^[5]。然而，当前主流的神经网络语义通信系统仍面临四大核心挑战：噪声泛化能力差、可扩展性不足、系统演化能力弱以及训练成本高。首先，现有系统通常依赖在特定信噪比下训练的神经网络自编码器，当实际信道环境与训练条件存在差异时，系统性能会明显下降^[4]。这一问题的根源在于判别式神经网络难以适应输入数据分布偏移（OOD）导致的特征变化^[6]。具体来说，信道噪声扰动会导致解码端接收到的特征分布与训练数据分布不一致，从而严重削弱系统对未知环境的鲁棒性。更为严峻的是，已有研究表明，在特征空间直接添加噪声将极大削弱神经网络的判别能力，而无线信道噪声正是以这种方式干扰语义特征传输，从理论上限制了现有结构的抗噪能力。其次，神经网络模型的可扩展性存在明显瓶颈。现有系统往往只能在与训练集分布高度一致的数据集上展现优异性能，但在实际应用中，数据分布经常发生漂移，或面临跨场景迁移需求，此时系统性能往往急剧下降^[7]。为提升泛化能力，部分研究尝试训练多个专用模型以适应不同场景。然而，这种方法不仅大幅增加模型存储与系统运维成本，还难以满足实际部署对灵活性和低资源消耗的要求。此外，系统演化能力不足是当前深度学习驱动语义通信面临的关键挑战之一。与计算机视觉等领域可通过大规模预训练模型实现跨任务迁移和持续学习不同，当前语义通信系统仍严重依赖端到端的专用训练流程，难以有效复用现有通用大模型资源，这限制了其在动态通信环境中实现自适应进化和迭代优化的潜力。最后，高昂的训练与部署成本也制约了语义通信的广泛应用。相比于传统通信方案，基于神经网络的端到端系统通常需要大量标注数据进行有监督训练，且需针对不同应用场景反复调整网络结构和优化参数，增加了系统开发、维护与部署的复杂度。

从统计学习角度看，当前语义通信系统面临的诸多挑战，其根源在于主流方法普遍采用判别式神经网络，缺乏对数据先验分布的建模能力。这导致在信道噪声引起的特征分布漂移条件下，系统难以实现鲁棒的语义信息恢复。相比之下，生成式人工智能（特别是扩散模型）因其对复杂数据分布的显式建模能力，能够从任意噪声干扰中恢复高质量语义特征，天然适用于应对无线通信中的信道不确定性和分布外输入等挑战^[8]。受此启发，本文提出基于隐扩散模型（LDM）^[9]的智能语义通信架构。该方法将信道噪声建模与语义特征生成过程有机结合，构建了从噪声“逆演化”到高维语义重构的端到端学习范式。具体而言，所提方法以大规

模预训练扩散模型为基础，将信道噪声视为高斯扰动，并通过扩散-去噪反演过程显式恢复原始语义特征，显著提升了系统对多样化信道条件和复杂数据分布的泛化与适应能力。同时，得益于预训练模型的知识迁移能力，所提系统在低样本甚至零样本条件下仍能实现鲁棒的语义传输，大幅降低了模型训练与系统部署的资源开销。此外，该架构支持灵活集成外部大模型资源，为语义通信系统的持续演化和任务自适应能力提供了坚实基础。

1 系统模型与问题描述

语义通信系统通常包括信源、联合信源信道编码（JSCC）模块、噪声信道、噪声抑制与解码模块以及信宿。其核心目标是通过有效提取、压缩、传输和还原数据中的语义信息，实现高效可靠的端到端传输。具体来说，信源 x 代表待传输的原始数据。该数据首先经过 JSCC 模块，将原始信息压缩为低维语义特征表示 z ，以显著降低数据冗余和带宽占用。随后， z 通过物理信道进行传输，在此过程中受到加性高斯白噪声（AWGN）等信道噪声影响。 y 为接收端实际收到的信号。在信宿侧，系统首先通过噪声抑制模块对 $\hat{z}$ 进行去噪与特征恢复，随后进入解码模块，以最大限度还原原始语义信息，输出信宿重构数据 $\hat{x}$ 。假设信源数据 $x \in R^m$ ，其通过联合编码器 $\mathcal{E}_\theta$ 映射为低维语义特征 $z = \mathcal{E}_\theta(x) \in R^k$ ， $k < m$ 。其通过 AWGN 信道后所接收到的信号 $\hat{y}$ 服从公式（1）：

$$y = z + n, n \sim \mathcal{N}(0, \sigma^2 I) \quad (1),$$

其中， n 为信道噪声， σ^2 为噪声功率。信宿端首先利用噪声抑制模块 $\mathcal{G}_\phi$ 对信号 y 进行处理，获得去噪后的特征 $\hat{z} = \mathcal{G}_\phi(y) \in R^k$ 。之后通过解码器 $\mathcal{D}_\psi$ 还原信源数据，得到重构输出 $\hat{x} = \mathcal{D}_\psi(\hat{z}) \in R^m$ 。语义通信系统的目标是最小化信宿重构数据 $\hat{x}$ 与信源 x 之间的均方误差（MSE）。该优化目标可以表示为公式（2）：

$$\begin{aligned} \min_{\theta, \phi, \psi} \mathbb{E}[\|\hat{x} - x\|_2] &= \min_{\theta, \phi, \psi} \mathbb{E}[\|\mathcal{D}_\psi(\mathcal{G}_\phi(\mathcal{E}_\theta(x)) + n) - x\|_2] \\ \text{s.t. } n &\sim \mathcal{N}(0, \sigma^2 I) \end{aligned} \quad (2),$$

其中， θ 、 ϕ 、 ψ 分别表示编码器、噪声抑制器和解码器的可训练参数。该架构面临两个主要技术挑战：首先，由于 $k < m$ 导致的维度压缩，解码器 $\mathcal{D}_\psi$ 需要求解一个非满秩矩阵构成的方程组，这使得准确恢复信源数据具有本质困难。其次，无线信道的噪声干扰会导致接收端获取的语义特征产生扰动，难以直接还原为高保真原始数据。与显式数据（如图

像或音频)的噪声干扰不同,神经网络对潜空间语义特征向量的噪声扰动表现出更低的鲁棒性^[6]。这是因为潜空间特征通常是高度压缩且分布紧致的抽象表示,高斯扰动更容易导致其分布偏移,从而破坏语义结构的完整性。这种特性使得解码器难以有效捕捉并还原原始信息。因此,在语义通信场景中,如何在潜空间层面对噪声扰动进行精确建模与有效修复,成为决定系统解码质量的关键科学问题。该挑战的核心在于建立能够保持语义结构完整性的鲁棒特征表示方法,以及开发针对压缩特征空间的专用去噪算法。

2 基于去噪扩散模型的智能语义通信系统理论分析

2.1 扩散模型去噪理论基础

扩散模型(DM)最初作为一种高质量数据生成方法被提出,其核心机制是通过多步随机扰动(扩散过程)及其逆向过程(去噪过程),建立数据分布与噪声分布之间的映射^[10]。近年来,得益于对复杂数据分布的强大建模能力和优异的去噪性能,DM在信号处理领域展现出重要应用价值,特别适用于高斯噪声环境下的信号恢复任务。DM的工作原理是通过在数据空间中逐步注入高斯噪声,构建一个从真实数据分布到标准高斯分布的马尔可夫链。设原始信号为 x_0 ,经过 t 步高斯噪声扰动后得到 x_t ,该正向扩散过程可用随机微分方程(SDE)描述^[11]:

$$dx_t = f_t x_t dt + g_t dw_t \quad (3),$$

其中, f_t 和 g_t 是和扩散次数相关的漂移项和扩散项, w_t 是标准Wiener过程。对于高斯噪声扩散,通常设 $f_t x_t = h_t$, $g_t^2 = 1$,则该过程可简化为下述积分方程形式:

$$x_t = x_0 + \int_0^t h_t dt + \int_0^t dw_t \quad (4),$$

其中, $\int_0^1 h_t dt + x_0 = 0$ 。通常可令 $h_t = -x_0$,可以得到公式(5)^[12]:

$$x_t = (1-t)x_0 + \sqrt{t}\epsilon \quad (5),$$

其中, ϵ 是标准高斯变量。要实现去噪,需对上述扩散过程进行反向建模。逆扩散过程同样可以用SDE来描述,其目标是从含噪观测 x_t 逐步还原出干净信号 x_0 。逆过程可表述为:

$$dx_t = [f_t x_t - g_t^2 \nabla_{x_t} \log p_t(x_t) + g_t d\bar{w}_t] \quad (6),$$

其中, $g_t^2 \nabla_{x_t} \log p_t(x_t)$ 表示关于当前状态的概率密度梯度, $g_t d\bar{w}_t$ 为独立于正向过程的Wiener过程。在实际神经网络实

现中,常采用离散时间的迭代形式。对于一步逆扩散,逆过程可简化为:

$$x_{t-\Delta t} = x_t - \frac{\Delta t}{t} \epsilon + \sqrt{\Delta t(t-\Delta t)} \tilde{\epsilon} \quad (7)。$$

在扩散模型的数学框架中, ϵ 代表噪声预测项,而 $\tilde{\epsilon}$ 是为了平衡方差在去噪过程中引入的独立标准高斯噪声。基于通用逼近理论,一个具有足够容量且经过充分训练的神经网络能够逼近任意分布映射关系。这一理论保证使得扩散模型理论上具备去除任意强度高斯噪声的能力。DM的本质即“去噪扩散概率模型”(DDPM)^[10]。大量实验研究表明,该模型能够在不同噪声水平和多样化的数据分布条件下实现高质量去噪。具体而言,当预训练的扩散模型接收到受高斯噪声污染的信号时,只需根据噪声强度选择合适的扩散步数(即参数 t),即可通过逐步去噪过程恢复出高保真度的原始信号。

这一理论特性为扩散模型在通信系统中作为通用噪声抑制模块的应用奠定了理论基础。该模型具有3个显著优势:首先,其对不同强度的噪声均表现出良好的适应性;其次,能够有效建模复杂的数据分布特征;最后,在各种条件下都能保持稳定的去噪性能。这些特性使扩散模型特别适用于处理无线通信信道中的噪声干扰问题,为其作为通信系统噪声抑制模块的可行性提供了理论支撑。

2.2 扩散步骤与噪声信号之间的关联

传统基于神经网络的去噪方法存在明显的局限性,通常需要针对特定信噪比(SNR)分别训练专用模型。当面对OOD噪声水平时,这类方法往往会出现明显的退化。而DM在理论上具有对任意信噪比高斯噪声自适应去噪的能力。该能力的核心机制在于:DM能够根据接收信号中的噪声能量特征,动态确定最优的扩散步数 t 。这一自适应调整过程确保了模型去噪能力与实际信号噪声特性精确匹配,从而保证最优的去噪效果(如图1所示)。

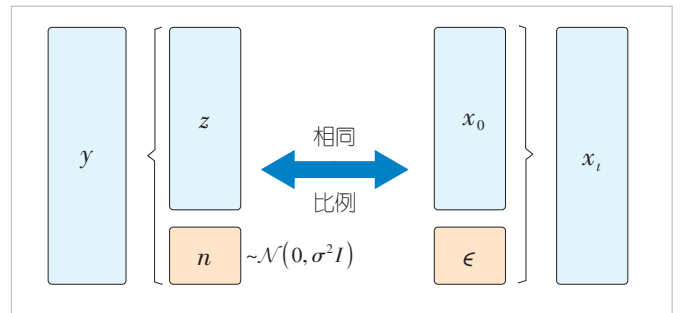


图1 扩散去噪比例示意图

具体而言，接收信号的SNR可通过公式（8）来定义：

$$\text{SNR} = \frac{\mathbb{E}[\|z\|_2^2]}{\mathbb{E}[\|n\|_2^2]} \quad (8)。$$

为了实现自适应去噪，需要确保实际接收信号的SNR与DM输入的 x_t 相匹配。为此，需使得二者满足：

$$\frac{\mathbb{E}[\|z\|_2^2]}{\mathbb{E}[\|n\|_2^2]} = \frac{\mathbb{E}[\|(1-t)x_0\|_2^2]}{\mathbb{E}[\|\sqrt{t}\epsilon\|_2^2]} \quad (9)。$$

假设信号发射功率 $\mathbb{E}[\|x_0\|_2^2] = \gamma$ ，那么对于任意 $\mathbb{E}[\|z\|_2^2]$ 和信道噪声强度 σ^2 有，当 t 满足公式（10）时，公式（9）成立：

$$t = \frac{2 + \frac{\mathbb{E}[\|z\|_2^2]}{\sigma^2\gamma} - \sqrt{\frac{\mathbb{E}[\|z\|_2^2]}{\sigma^2\gamma} \left(\frac{\mathbb{E}[\|z\|_2^2]}{\sigma^2\gamma} + 4 \right)}}{2} \quad (10)。$$

公式（10）可通过如下方法证明。首先对于 $\mathbb{E}[\|(1-t)x_0\|_2^2]$ 有：

$$\mathbb{E}[\|(1-t)x_0\|_2^2] = (1-t)^2 \mathbb{E}[\|x_0\|_2^2] = (1-t)^2 \gamma \quad (11)。$$

相似的有：

$$\mathbb{E}[\|\sqrt{t}\epsilon\|_2^2] = t \quad (12)。$$

因此，原方程可写作：

$$\frac{\mathbb{E}[\|z\|_2^2]}{\sigma^2} = \frac{(1-t)^2 \gamma}{t} \quad (13)。$$

上述关于 t 的一元二次方程。通过求根公式可以得到：

$$t = \frac{2 + \frac{\mathbb{E}[\|z\|_2^2]}{\sigma^2\gamma} - \sqrt{\frac{\mathbb{E}[\|z\|_2^2]}{\sigma^2\gamma} \left(\frac{\mathbb{E}[\|z\|_2^2]}{\sigma^2\gamma} + 4 \right)}}{2} \quad (14)。$$

如图2所示，实际通信系统中接收信号的统计分布往往与DM训练时的输入分布存在偏差。若直接将观测信号输入预训练的DM，可能引发OOD问题，导致去噪性能下降。为解决这一问题，需要通过对接收信号进行线性缩放处理，使其统计特性与DM扩散过程中的变量分布保持一致。这种预处理方法能够确保DM的去噪能力在不同信噪比条件下均能得到有效发挥。

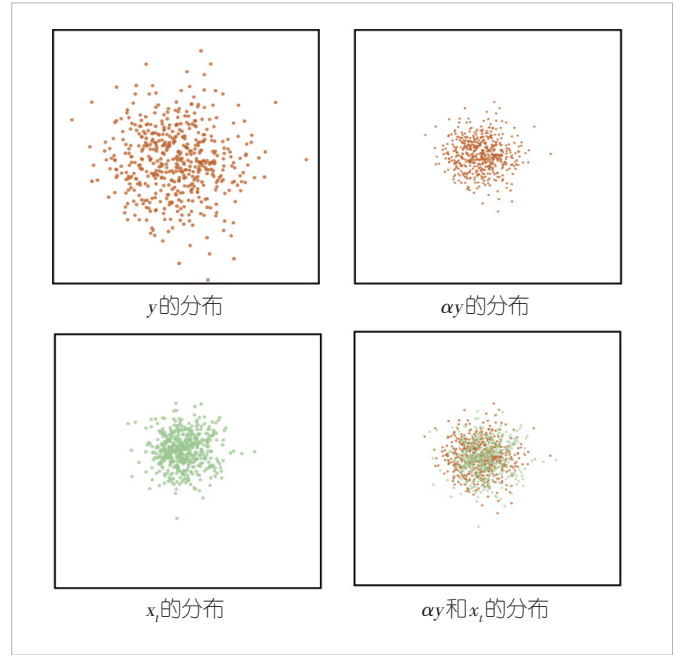


图2 数据分布示意图

通过缩放 y 至 αy ，可以让 αy 的分布与 x_t 一致，这要求 α 满足如下关系：

$$\alpha = \sqrt{\frac{(1-t)^2 \mathbb{E}[\|z\|_2^2] + t}{\mathbb{E}[\|z\|_2^2] + \sigma^2}} \quad (15)。$$

公式（15）的推导过程如下：

首先，期望 $\mathbb{E}[\|\alpha y\|_2^2]$ 可表示为：

$$\mathbb{E}[\|\alpha y\|_2^2] = \mathbb{E}[\|\alpha(z+n)\|_2^2] = \alpha^2 (\mathbb{E}[\|z\|_2^2] + \mathbb{E}[\|n\|_2^2]) \quad (16)。$$

同时， $\mathbb{E}[\|x_t\|_2^2]$ 的计算如下：

$$\mathbb{E}[\|x_t\|_2^2] = \mathbb{E}[\|(1-t)x_0 + \sqrt{t}\epsilon\|_2^2] = (1-t)^2 \mathbb{E}[\|x_0\|_2^2] + t \quad (17)。$$

通过求解 $\mathbb{E}[\|x_t\|_2^2] = \mathbb{E}[\|\alpha y\|_2^2]$ 即可得到 α 的表达。

2.3 基于LDM的语义通信架构

根据上述理论分析，针对语义通信系统在不同信噪比环境下易受分布外噪声干扰、模型泛化能力不足等问题，本文提出了一种以LDM为核心的端到端语义通信架构，如图3所示。该架构充分融合了变分自编码器（VAE）与DM的各自优势，实现了信源语义信息的高效压缩、任意信噪比下的鲁棒恢复及系统低成本可持续演化。在本架构中，信源数据 x

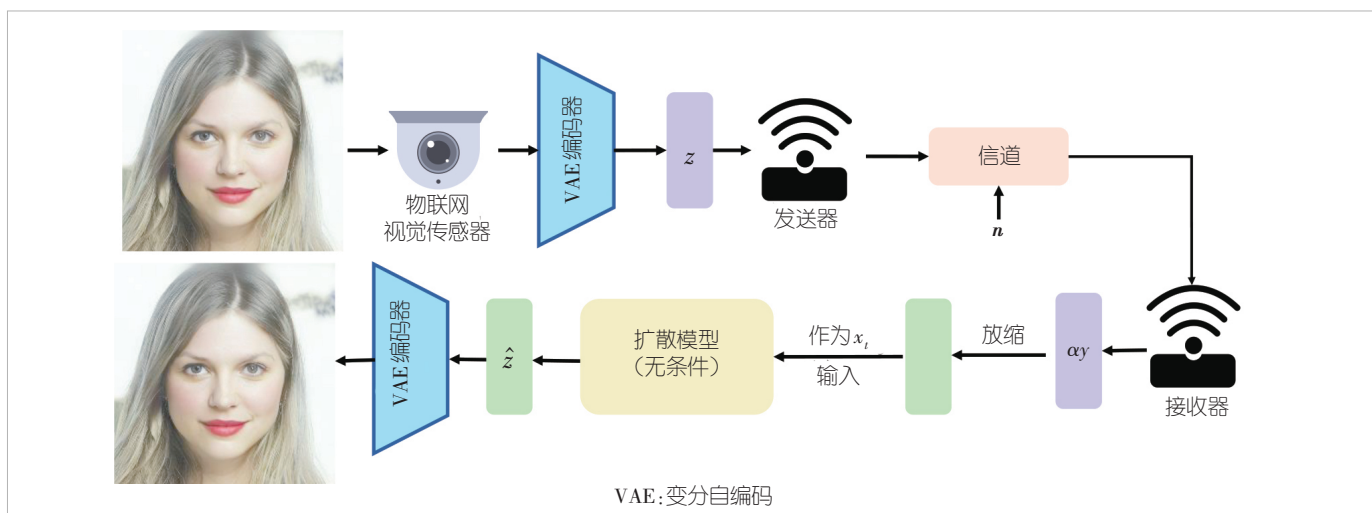


图3 基于隐扩散模型的语义通信系统结构图

首先通过VAE编码器得到隐变量 z ，实现数据语义层的紧凑表示，从而显著降低传输带宽和冗余信息。隐变量 z 经过加性白噪声信道（AWGN）传输后，接收端获得信号 y 。由于信道噪声的影响，隐变量在传输后与原始分布存在差异，直接进行恢复可能导致性能退化，尤其在信噪比剧烈变化时，传统神经网络方法往往难以适应这种分布漂移。

为此，本文引入了LDM作为信号的去噪主干模块。接收端根据当前观测信号的噪声方差 σ^2 ，首先自适应地确定扩散去噪的步数 t 。为了确保输入数据的分布与LDM训练阶段一致，我们对 y 施加了缩放变换 αy 。经过缩放和步数自适应调整后的接收信号与 t 一起输入预训练的LDM。通过扩散逆过程，我们实现了对高斯噪声的逐步抑制，最终获得去噪后的隐变量 $\hat{z}$ 。去噪后的 $\hat{z}$ 经过VAE解码器还原为重构数据 $\hat{x}$ ，从而实现信源到信宿的高保真语义信息恢复。

该架构的核心优势在于其对任意SNR条件下噪声的自适应处理能力，以及对预训练大模型能力的高效继承。首先，本文提出的LDM语义通信架构通过联合自适应的去噪步数与缩放因子机制，能够动态调节模型在不同SNR条件下的去噪策略，从而有效抑制接收语义特征中的残余噪声。该机制有效克服了传统神经网络架构在信道条件变化下容易产生OOD输入的问题，增强了系统在复杂信道环境下的鲁棒性与泛化能力。其次，本文的方法通过引入预训练的生成式大模型，有效利用其对数据先验分布的深层建模能力，实现对带噪语义特征的最大后验恢复，从而提升了带噪语义特征的恢复能力。相比之下，现有基于条件生成模型的方法在处理语义传输任务时，往往因条件编码器难以适配多样化噪声分布，导致通过信道传输后的语义特征形成OOD输入，严重削弱了生成模型的重构性能。而本文所采用的LDM将

噪声视为潜空间中的扰动变量，借助其训练过程中已学得完整数据分布信息，可直接实现跨信道条件的高质量语义还原，无须针对不同信噪比场景重新训练或微调。

3 实验设置与结果分析

3.1 实验设置

为验证所提出的基于LDM的语义通信方法在机器通信任务中的有效性，本文选取与人脸图像识别相关的语义传输场景作为代表应用，并基于大规模高质量人脸图像数据集CelebA-HQ开展仿真实验。该数据集包含多样化的人脸图像，具备丰富的身份语义与结构特征，能有效模拟物联网终端中摄像头采集的人体视觉信息。考虑到所提方法无须重新训练，直接采用了公开发布的预训练LDM，未针对特定信道或任务进行任何微调。

具体而言，原始图像首先通过VAE编码器映射至低维潜空间表示，随后在该潜空间中引入高斯信道噪声，以模拟无线信道的干扰过程。最后，通过LDM的反向扩散过程对受扰语义特征进行去噪恢复，并经VAE解码器还原图像或提取语义特征。为全面评估所提方法的鲁棒性与语义保真性能，本文设计与两种典型深度JSCC方法的对比实验，分别为：

- 1) Deep JSCC^[13]：基于卷积神经网络设计的端到端图像传输结构，具有较强的空间特征提取与压缩能力；
- 2) Deep JSCC-l++^[14]：融合Transformer架构的先进模型，利用注意力机制建模图像中的长程依赖关系，增强复杂信道下的语义保持能力。

上述基线方法均按照其原始论文在CelebA-HQ数据集

上进行复现和训练，并在不同信噪比条件下进行对比测试。为了兼顾像素精度与语义保真，本文采用了两类指标进行评估：1) 图像重建指标，包括峰值信噪比 (PSNR) 和结构相似性 (SSIM)；2) 语义保持指标，即学习感知相似度 (LPIPS)，这是一种基于深度神经网络感知特征的图像相似性度量指标，用于评估重建图像与参考图像在人类视觉感知下的相似程度。

3.2 结果分析

图4展示了在不同信噪比条件下各算法的图像恢复效果。从图4可以看出，所提出的LDM方法即使在完全无训练的条件下，仍能在低信噪比环境中实现较高质量的图像还原。如图5所示，在信噪比低于-7.5 dB的低信噪比场景下，所提出的方法还原的人脸图像在轮廓、结构及语义层面均表现出良好的保持能力，基本能够恢复关键的身份特征。相比之下，两种经典的深度JSCC方法在相同信道条件下的重建结果则出现模糊和失真，表现出明显的退化现象。这一现象可归因于它们依赖判别式训练，在面对OOD信道噪声时缺乏鲁棒性，导致语义信息无法有效还原。

图6展示了不同方法在多个性能指标下的定量比较，这些性能指标包括PSNR、SSIM与LPIPS。在LPIPS指标上，

所提出的方法显著优于所有对比方法，这表明该方法在保持人类感知语义一致性方面具有突出优势。这意味着，即使在高噪声或未知信道条件下，LDM所生成的图像在视觉语义上依然贴近原始图像，更加契合物联网设备中语义感知与识

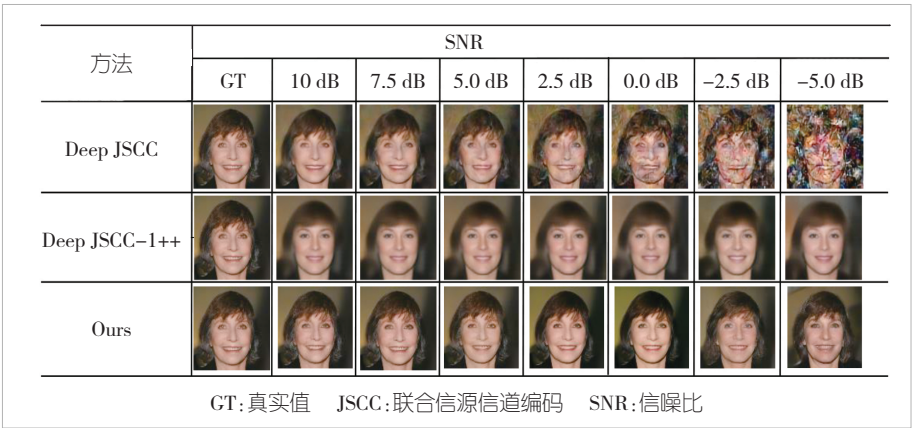


图4 不同信噪比下图像恢复效果对比图

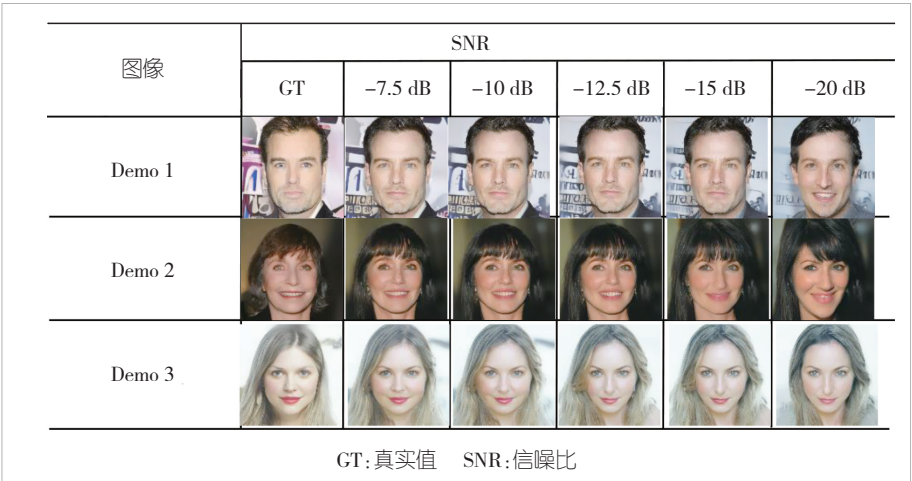


图5 本文所提出方法在低信噪比下图像恢复效果图

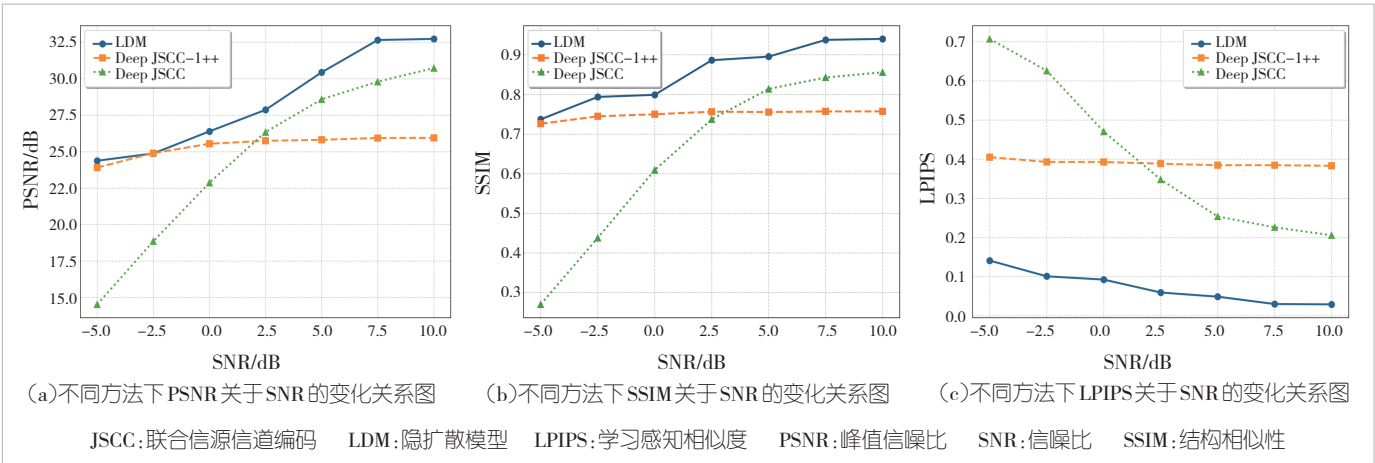


图6 不同信噪比下性能对比图

别任务的实际需求。综上所述,实验结果充分验证了本文方法在语义保真性、噪声鲁棒性及泛化能力方面的综合优势,为资源受限终端实现高效、可靠的语义通信提供了有力支撑。

4 结束语

本文针对6G机器通信场景中的语义通信需求,提出了一种基于LDM的语义通信新架构。该架构利用预训练的LDM对信道噪声进行建模与逆演化,无须重新训练即可鲁棒地还原语义信息,特别适用于资源受限的机器通信场景。在CelebA-HQ数据集上的仿真结果表明,所提方法即使在低信噪比条件下,依然具备出色的图像重建与语义保持能力,并且在LPIPS等感知指标上明显优于现有的典型深度通信模型,验证了其强大的泛化性与实用性。本方法允许直接采用预训练的LDM作为语义通信架构的主干,为后续智能语义通信系统的低成本、可持续演化发展提供了新的思路 and 理论基础。未来,我们将进一步研究模型轻量化与多模态扩展,以推动其在物联网与边缘智能中的实际部署。

参考文献

- [1] 林美玉. 5G网络赋能物联网安全[J]. 中兴通讯技术, 2022, 28(6): 70–74. DOI: 10.12142/ZTETJ.202206012
- [2] DANG S P, AMIN O, SHIHADA B, et al. What should 6G be? [J]. Nature electronics, 2020, 3(1): 20–29. DOI: 10.1038/s41928-019-0355-6
- [3] 廖俊淇, 魏昕, 周亮. 人工智能驱动的跨模态语义通信系统[J]. 中兴通讯技术, 2024, 30(S1): 33–39. DOI: 10.12142/ZTETJ.2024S1005
- [4] ZHANG P, XU W J, LIU Y M, et al. Intellicise wireless networks from semantic communications: a survey, research issues, and challenges [J]. IEEE communications surveys & tutorials, 2024, 27(3): 2051–2084. DOI: 10.1109/comst.2024.3443193
- [5] XU W J, ZHANG Y M, WANG F Y, et al. Semantic communication for the Internet of vehicles: a multiuser cooperative approach [J]. IEEE vehicular technology magazine, 18(1): 100–109. DOI: 10.1109/mvt.2022.3227723
- [6] LECUN Y, BENGIO Y, HINTON G. Deep learning [J]. Nature, 2015, 521(7553): 436–444
- [7] GETU T M, KADDOUM G, BENNIS M. Semantic communication: a survey on research landscape, challenges, and future directions [J]. Proceedings of the IEEE, 2024, 112(11): 1649–1685. DOI: 10.1109/jproc.2024.3520707
- [8] HUANG Y, HUANG J, LIU Y, et al. Diffusion model-based image editing: a survey [J]. IEEE transactions on pattern analysis and machine intelligence, 2025, 47(6): 4409–4437. DOI: 10.1109/tpami.2025.3541625

- [9] ROMBACH R, BLATTMANN A, LORENZ D, et al. High-resolution image synthesis with latent diffusion models [C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2022: 10684–10695. DOI: 10.1109/cvpr52688.2022.01042
- [10] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models [EB/OL]. [2025-05-28]. <https://dl.acm.org/doi/abs/10.5555/3495724.3496298>
- [11] SONG Y, DURKAN C, MURRAY I, et al. Maximum likelihood training of score-based diffusion models [EB/OL]. <https://arxiv.org/abs/2101.09258>
- [12] HUANG Y, QIN Z, LIN X, et al. Decoupled diffusion models: simultaneous image to zero and zero to noise [EB/OL]. [2023-06-23][2025-05-23]. <https://arxiv.org/abs/2306.13720>
- [13] BOURTSOULATZE E, BURTH KURKA D, GUNDUZ D. Deep joint source-channel coding for wireless image transmission [J]. IEEE transactions on cognitive communications and networking, 5(3): 567–579. DOI: 10.1109/tccn.2019.2919300
- [14] BIAN C, SHAO Y, GUNDUZ D. DeepJSCC-1++: robust and bandwidth-adaptive wireless image transmission [C]//2023 IEEE Global Communications Conference. IEEE, 2023: 3148–3154

作者简介



承楠, 西安电子科技大学教授; 主要研究方向为6G按需服务关键技术、空天地一体化网络、智能通信网络; 主持或参与国家重点研发计划项目、国家“863”项目、国家自然科学基金重点项目, 以及加拿大自然科学基金委员会重点基金项目等; 获 IEEE GLOBECOM 2014、IEEE ICC 2015、IEEE WCSP 2015、IEEE ICC 2019 最佳会议论文奖; 发表论文80余篇, 获授权美国专利9项。



王秀程, 西安电子科技大学在读博士研究生; 主要研究方向为智能无线网络技术、生成式人工智能。



沈学民, 加拿大滑铁卢大学教授, 中国工程院外籍院士, 加拿大工程院、工程研究院以及皇家科学院院士, IEEE Fellow, 无线通信网络领域谷歌学术引用次数最高的作者之一; 主要研究方向为空天地一体化网络、车联网、网络安全、人工智能等; 荣获44项国际权威期刊及顶级会议的最佳论文奖; 发表论文600余篇, 撰写专著27本。

6G:从通信到通感算智融合



6G: From Communication to Integration of Communication, Sensing, Computing, and Intelligence

刘光毅/LIU Guangyi^{1,2}

(1. 中关村泛联移动通信技术创新应用研究院, 中国 北京 100080;

2. 中国移动研究院, 中国 北京 100053)

(1. ZGC Institute of Ubiquitous-X Innovation and Applications, Beijing

100080, China;

2. China Mobile Research Institute, Beijing 100053, China)

DOI: 10.12142/ZTETJ.202504008

网络出版地址: <http://kns.cnki.net/kcms/detail/34.1228.TN.20250711.1947.002.html>

网络出版日期: 2025-07-14

收稿日期: 2025-05-15

摘要: 移动通信网络已成为现代社会不可或缺的基础设施。6G作为下一代通信技术, 不仅要在通信性能上有质的飞跃, 更需要实现通信、感知、计算、人工智能(AI)等多个领域的深度融合。首先分析了6G网络的驱动力, 包括业务发展需求、网络建设与运维痛点、AI技术引入的挑战。随后, 详细阐述了6G网络架构的关键设计转变, 包括从多能力烟囱式部署到通感算智多能力一体融合、从单一服务到通感算智多能力差异化服务、从外挂式AI到内生式AI, 以及网络运维从事后高成本处理到事前低成本干预。最后, 介绍了6G网络的关键技术, 包括云原生及服务化无线接入网(RAN)、内生AI以及网络数字孪生。这些技术为6G网络的高效、智能、灵活运行提供了有力支撑。

关键词: 6G; 通感算智融合; 云原生; 服务化RAN; 内生AI; 网络数字孪生

Abstract: Mobile communication networks have emerged as an indispensable infrastructure in contemporary society. As the next-generation communication technology, 6G needs to mark a significant leap in communication performance. More importantly, it needs to achieve the deep integration of multiple fields such as communication, sensing, computing, and artificial intelligence (AI). The article first examines the driving forces behind 6G networks, including the demands of service evolution and the challenges in network construction, operation, and maintenance. Subsequently, it elaborates on the pivotal design transformations in the 6G network architecture, encompassing the transition from a siloed multi-capability model to an integrated fusion of communication, sensing, computing, and AI capabilities, from offering a single service to providing differentiated services leveraging these multi-capabilities, from an added-on AI approach to a native AI, and from post-hoc high-cost handling to proactive low-cost intervention. Finally, the article introduces the key technologies of 6G network architecture, including cloud-native and service-based radio access networks (RAN), native AI, and network digital twins. These technologies collectively offer robust support for the efficient, intelligent, and flexible operation of 6G networks.

Keywords: 6G; integration of communication, sensing, computing, and intelligence; cloud-native; service-based RAN; native AI; network digital twin

引用格式: 刘光毅. 6G: 从通信到通感算智融合 [J]. 中兴通讯技术, 2025, 31(4): 48-54. DOI: 10.12142/ZTETJ.202504008

Citation: LIU G Y. 6G: from communication to integration of communication, sensing, computing, and intelligence [J]. ZTE technology journal, 2025, 31(4): 48-54. DOI: 10.12142/ZTETJ.202504008

在5G持续发展与6G创新布局的关键时期, 移动通信网络正面临网络复杂度持续增加, 行业应用需求趋于定制化、差异化、碎片化, 以及人工智能(AI)技术与网络融合带来的挑战。在此背景下, 本文提出6G^[1]通感算智融合网络^[2], 旨在基于云原生、服务化无线接入网(RAN)^[3]、AI、网络数字孪生等技术, 重塑移动信息网络服务的新范式, 推动业界6G通感算智融合内生研究。

1 6G网络演进的驱动力

技术演进与业务需求双轮驱动移动通信网络的发展。移动通信网络在满足社会日益增长的连接需求的同时, 其自身复杂度也在不断攀升。“数字孪生、智慧泛在”作为6G的社会发展愿景, 正推动移动通信网络向智能化方向演进。通感算智融合已成为6G网络发展的重要趋势^[1], 其驱动力主要源于以下3个方面:

1) 业务发展

面向垂直行业, 现有5G网络解决方案主要依赖专网部署或多种标准化产品的组合来满足差异化需求。然而, 面向

基金项目: 国家重点研发计划项目(2022YFB2902100)

企业（ToB）行业需求愈发呈现出定制化、差异化和碎片化的特征，对带宽、时延、可靠性、安全性、连接密度以及特定服务能力（如定位、计算）的需求呈现高度个性化与定制化特征。但现有面向消费者（ToC）的大一统网络产品形态难以满足这些多样化需求，其架构和功能设计相对固化，缺乏足够的灵活性，难以经济、高效地适配碎片化的行业场景，导致客户定制化需求响应速度慢、部署周期长、整体解决方案成本高、资源利用效率低、管理复杂等问题。因此，网络需要更深层次的“简”，其核心在于实现网络能力与服务的模块化、灵活组装和按需供给，以敏捷、经济、高效的方式响应千行百业的需求，降低行业数字化转型门槛。

2) 网络建设与运维

从现网情况看，5G网络参数和功能繁多，使得运维和管理工作复杂，网络性能难以得到优化与保障。在实际运营中，运营商通常需要投入大量的人力和物力进行繁琐的监控、调整以及故障排查工作。同时，网络中专用设备堆叠部署，导致硬件资源难以实现跨域共享与按需分配。这一问题进一步引发网络建设和维护成本居高不下、运营管理流程繁琐、整体能耗效率难以优化等连锁反应，使得新功能部署速度缓慢，难以快速响应市场和用户的需求。这种复杂性已成为网络持续高效发展的瓶颈，需通过架构层面的根本性设计来破解^[4]。

3) AI的引入

AI已经在网络运维中展现出巨大潜力，但目前数据的获取、治理、隐私保护以及跨厂商设备间的互通性制约了

AI模型训练与部署的效率和效果。面向6G，“智”的引入不应仅局限于网络运维层面的事后优化（运维态），更需要深度融合网络的实时运行过程（运行态），实现感知、决策、执行的闭环自治。更为关键的是，6G将超越传统连接，深度融合感知、计算、AI等多种能力。简单的外挂和叠加会导致效率低下、部署运维成本高昂以及协同效果不佳。因此，网络需要内生的“智”，实现通感算智等核心能力的深度融合与一体化设计，为未来创新业务提供智能、高效、低成本的平台支撑。

2 6G通感算智融合网络

6G网络架构以云化为底座、端到端服务化为基因、内生AI为智能引擎、数字孪生为低成本预验证支持，通过功能的模块化与灵活调用、服务的高效与智能编排管控，以最小的功能集按需适配碎片化、个性化业务需求。该架构旨在打造通感算智一体融合的综合能力体系，满足更加丰富的场景需求。如图1所示，6G网络架构通过四大关键设计转变，构建更加智能、高效、灵活的网络体系，实现通感算智深度融合。

转变1：从烟囱式堆叠到平台化共享

传统网络采用垂直烟囱式架构，专用设备堆叠部署且硬件资源不共享，导致网络成本高、能耗高、资源利用率低、扩展性差。差异化、碎片化的无线算力资源难以匹配6G通感算智网络对低时延、高灵活性的能力需求。为此，需通过异构硬件云化，高效整合各类异构硬件资源，支撑上层高效

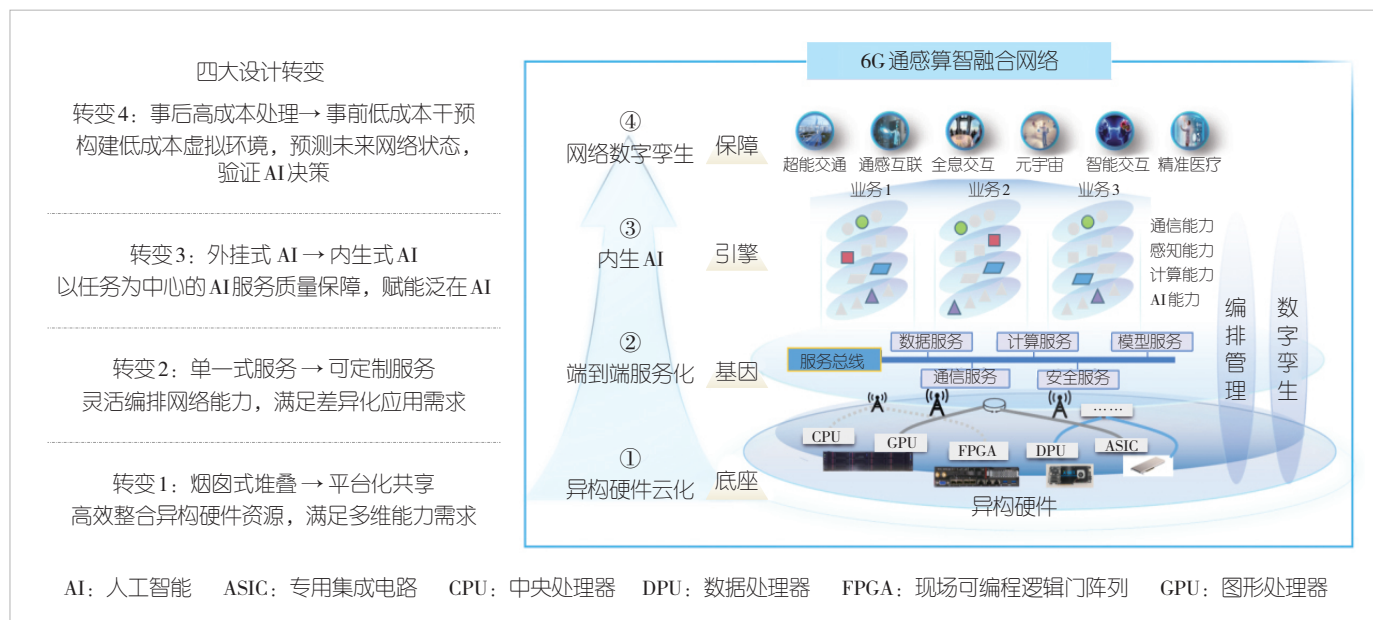


图1 6G通感算智融合网络架构

通感算智算力需求的按需调度，满足不同场景下的高实时、弹性部署需求；同时向下实现算力池化与统一管理，解决无线算力管理难题。

转变2：从单一式服务到可定制服务

5G网络引入服务化架构（SBA）与网络切片技术，实现了一定程度的服务差异化，但其核心理念并未真正延伸至RAN侧。RAN仅通过频率隔离、空口配置满足差异化场景需求，仍存在灵活性不足的问题。6G时代，行业应用的定制化需求更加突出，并将引入通感算智多维网络能力，需要考虑通感算智网络能力的定义与交互，以支撑服务按需提供。端到端服务化可实现网络功能的灵活组合和高效编排，持续满足差异化的服务需求。

转变3：从外挂式AI到内生式AI

“数字孪生、智慧泛在”是6G网络的发展目标。智能网联车、智能终端和智能机器人等智能应用需求呈现更加多样化和复杂化的特点。一方面，需要考虑如何依托6G网络推动AI应用爆发式增长与规模化落地；另一方面，需要探索如何充分利用AI强大的感知、理解、学习能力赋能网络，实现更加高效的通感算智融合设计与运行。未来，6G网络与AI将深度融合，实现内生AI，在6G网络架构内部提供数据采集、数据预处理、模型训练、模型推理、模型评估等AI工作流全生命周期的完整运行环境，同时将AI服务所需的算力、数据、算法、连接与网络功能、协议和流程进行深度融合设计。

转变4：从网络运维的事后高成本处理到事前低成本干预

传统网络运维依赖人工巡检和故障后修复，导致处理成本高且效率低下。例如，5G网络中基站故障的排查与恢复时间可能需要数小时。面对6G通感算智深度融合、AI原生架构以及空天地海一体化组网所带来的更高的网络复杂性与可靠性要求，亟需更高效的运维模式。网络数字孪生技术^[5]可作为低成本预验证的保障，实现预测性主动干预，构建低成本虚拟环境，预测未来网络状态，预验证AI决策。通过预测性维护，系统可在设备性能劣化或潜在故障发生前精准识别风险点，并自动或半自动地执行预防性操作。通过构建动态自治网络，将网络规划、建设、维护和优化工作统一管理，实现从“各环节分开独立”到“数字孪生统一”的转变。

3 6G通感算智融合网络的关键技术

为支持6G网络中的通感算智一体融合，需要重点突破通算资源统一管理、服务化RAN、内生AI和网络数字孪生等关键技术，并形成技术体系。

3.1 通算资源统一管理

计算、AI、感知等能力被视为6G网络核心能力要素，对于满足未来复杂多样的通信需求至关重要。而云原生基站凭借其高度的灵活性、可扩展性与可编程性，为这些关键能力的深度融合提供了坚实基础，使得不同能力能够在基站平台上协同运作，发挥出更大的效能。

面对6G网络中无线算力管理的复杂挑战，需要设计云化异构资源统一管理框架^[2]，实现对异构云化算力资源的全生命周期统一管理。从资源的申请、分配、调度到释放，整个过程都在统一框架下高效进行。这样能够有效解决无线算力资源异构、动态变化带来的管理难题，大幅提升算力资源的利用率和管理效率。

为满足无线网络对低时延的严苛要求，需研发抢占式确定性调度技术，自主编译高实时内核。抢占式确定性调度技术通过对任务进行优先级划分和抢占式调度，确保高优先级任务及时处理；高实时内核可进一步优化系统响应速度和稳定性。两者结合，可实现微秒级响应时延，为工业控制、远程医疗等对实时性要求极高的应用场景提供有力保障。

为满足无线大带宽处理需求，无线云软硬件平衡调优方法被提出。该方法从硬件架构和软件算法两个层面进行优化，满足了400 MHz高带宽基带处理需求，为高清视频传输、虚拟现实等大带宽应用提供支持。

为解决异构硬件底层架构差异带来的复杂性问题，需要突破异构硬件虚拟化技术瓶颈，实现异构资源的池化共享，如图2所示。针对无线多样化平台的软件管理难题，需要突破对象化软件管理方案限制，对服务器、传感器、算力卡等不同类型的平台对象进行抽象建模，实现软件的自动部署、更新和维护，减少人工干预。

3.2 服务化RAN

基于服务化架构，6G通感算智能力将以服务的形式灵活引入RAN，并通过服务化接口（SBI）实现按需调用^[3]。这种设计模式赋予了RAN功能可插拔、网络可定制的特性，使其能够更好地适应未来多样化的业务需求。

为了满足场景的差异化部署需求，“原子化拆分+分子化重构”的场景化网络功能（NF）设计思路是有效的解决方案，如图3所示。通过组合多个原子服务，可以构建出不同的网络功能。这些网络功能可以根据实际部署场景进行灵活配置和组合。

作为RAN迈向服务化的第一步，RAN能力开放将为赋能泛在轻量化智能终端和低时延业务应用提供更多可能。RAN能力开放框架存在两种发展路径：

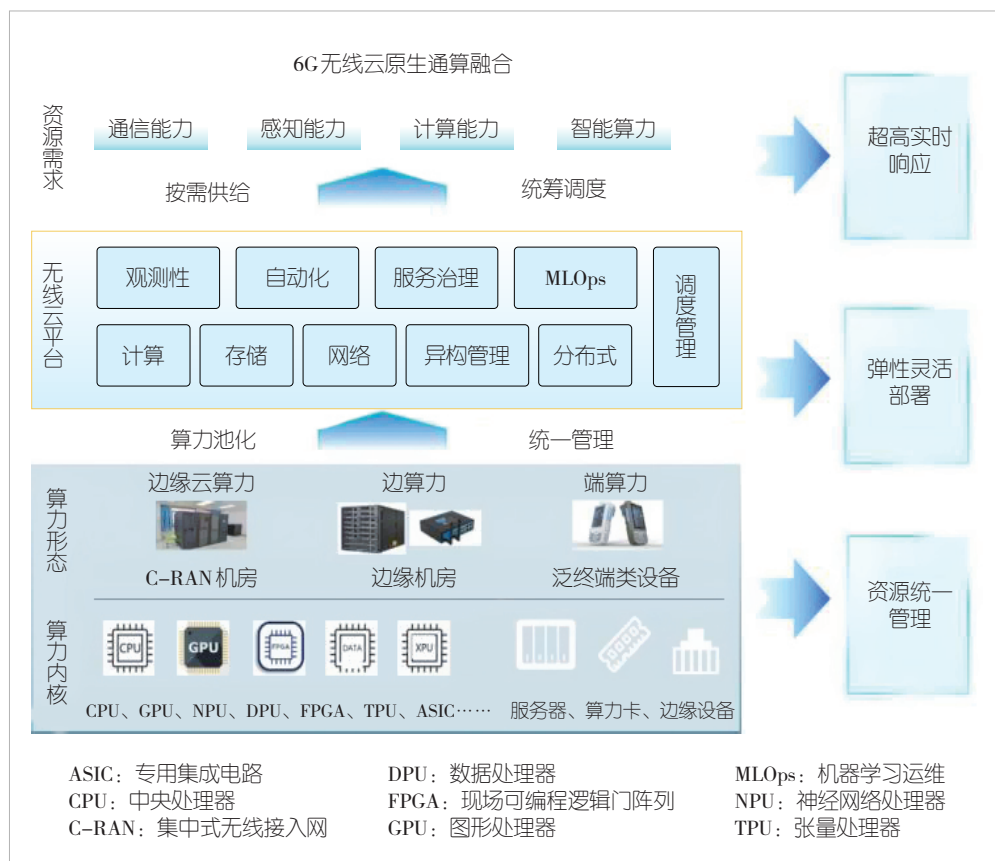


图2 云化异构硬件示意图

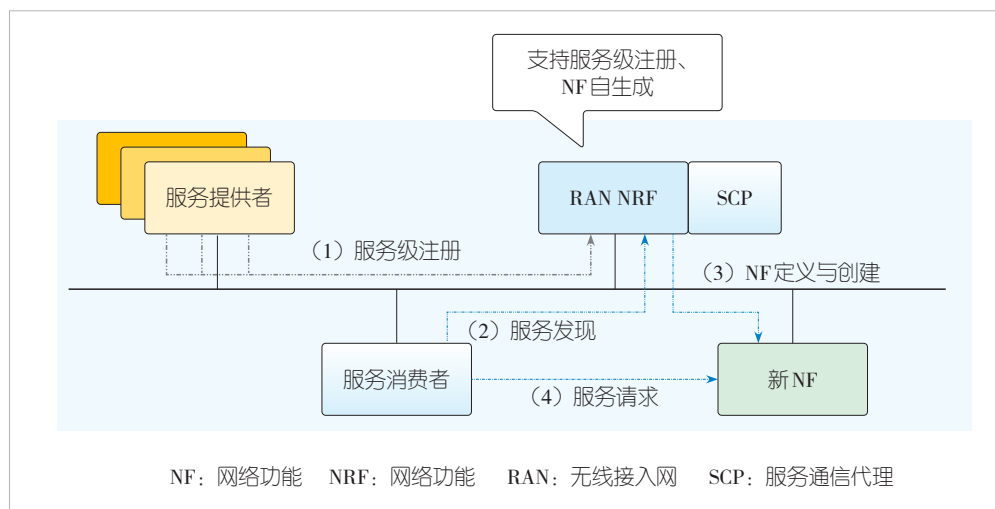


图3 场景化NF自生成

路径1：延续式对外开放，即在不改变现有3GPP能力开放框架的前提下，通过增强网络能力开放功能（NEF）或引入RAN能力开放功能实现RAN能力开放，如图4所示。具体地，该路径又可以通过如下两个方案实现：

• 方案1：RAN能力通过核心网NEF间接对外开放。在这一选项下，RAN与NEF之间可能需要设计服务化接口，

以便于NEF更高效地调用RAN开放能力。二者之间的关系类似于NEF与其他核心网NF之间的调用关系。

• 方案2：在RAN侧引入类似于NEF的RAN能力开放功能。该功能可与应用使能层（如CAPIF）直接交互，实现RAN能力直接对外开放。

路径2：定制化能力开放。如图5所示，该路径与路径1的区别特征在于：

• 特征1：突破NEF唯一开放入口模式，支持NF直接将可开放服务发布到CAPIF。例如，6G RAN功能直接将可开放的服务通过CAPIF-4发布到CAPIF core function，以丰富可开放的网络服务集合。此外，UE和第三方应用也可以将开放服务发布到CAPIF core function，以更好地赋能业务应用。

• 特征2：引入意图驱动的智能编排功能，一方面高效组合开放网络能力以更好地匹配业务需求，另一方面降低客户使用网络能力的门槛。意图驱动的智能编排功能是应用与网络应用程序编程接口（API）之间的桥梁。该功能通过与CAPIF core function交互实现网络API的发现与调用，在实现网络能力的编排组合后，将应用所需能力通过应用使能层API返回给应用。在这种模式

下，应用不能直接发现与调用网络服务，由此降低因网络暴露过多API导致的安全风险。

3.3 内生AI

在6G网络中，通过通算数智融合的编管控技术，完成AI业务需求到底层的多维网络资源的按需映射，实现高效

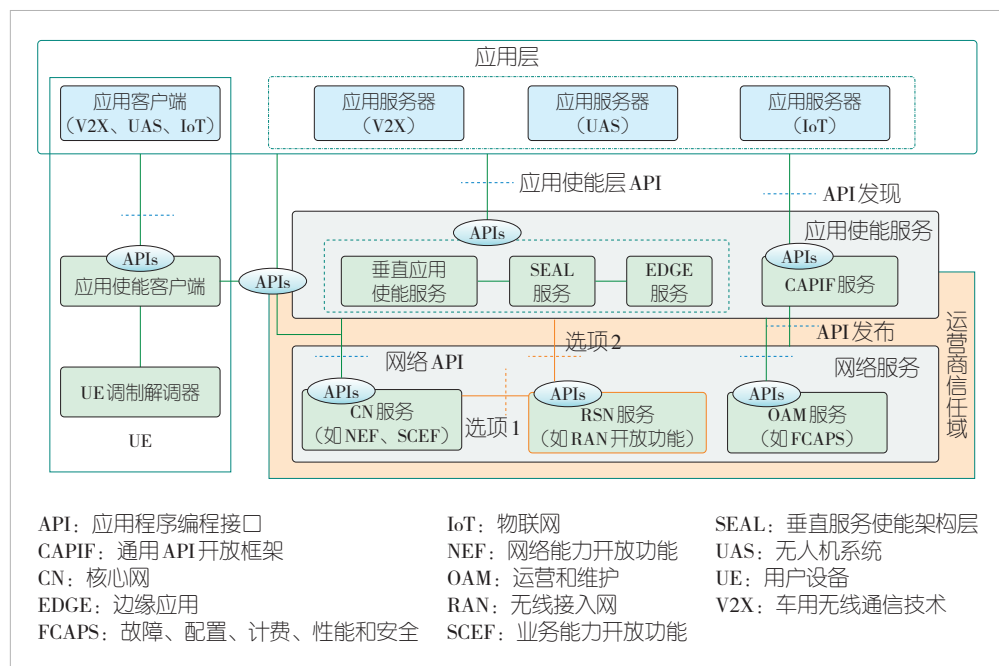


图4 延续式对外开放的RAN能力开放框架

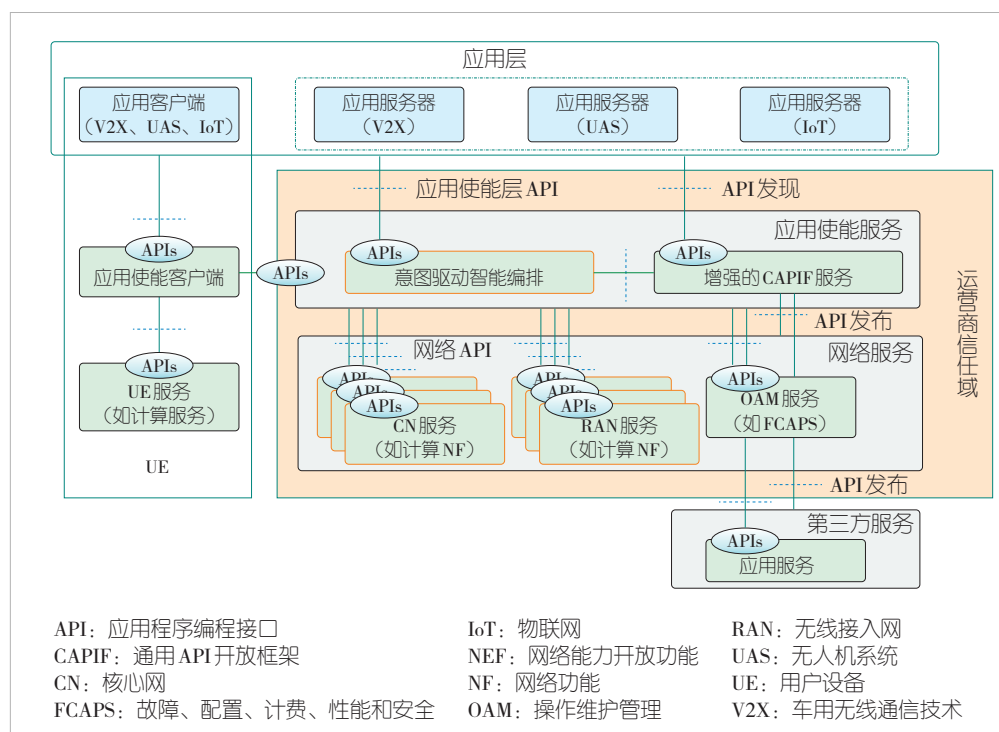


图5 定制化对外开放的RAN能力开放框架

智能决策，满足多样化AI业务需求，如图6所示。该技术通过为不同业务场景提供差异化智能服务，确保服务的高效性和可靠性^[4]。

AI服务编排聚焦于大时间尺度下的批量任务处理，致力于实现AI服务的快速部署。在部署环节，依据AI需求和

执行逻辑，对网络中的AI功能进行合理组合，形成完整的AI服务，并按需部署到网络中。这一过程充分考虑了批量用户的服务质量（用户接受率），确保在满足大量用户需求的同时，保证服务的稳定性和可用性。其中，AI服务功能链编排是核心环节，它能根据AI执行逻辑和网络资源状态，结合服务化RAN设计理念，对网络中的AI功能进行有序编排，形成功能链。这种编排方式能够灵活适应不同业务场景的差异化需求。

AI任务管理主要针对小时间尺度下的单个任务，通过调度保障单个用户的AI服务质量（QoAIS），如时延、准确性等关键指标。在调度方面，充分考虑通信、计算、数据与智能各环节的耦合性，进行AI任务到功能的精准调度。根据用户需求、信道条件和计算资源状态等，动态调整任务调度策略。

AI多维资源控制技术聚焦于实时控制网络中的连接、计算、数据和模型四要素资源，确保每个环节的执行符合通信时延、计算时延等关键指标要求。根据任务的实际需求和网络的实时状态，合理分配和调度资源，提高资源利用率，同时保障AI服务的质量。针对通信、计算、智能与数据独立控制导致的效率低、延迟高等问题，在控制面与用户面设计了相应的多维资源融合机制。在控制面，建立统一的控制，实现对通信、计算、数据和智能资源的集中管理和协同调度，提高资源利用效率；在用户面，通过用户面映射功能，实现通算资源的实时置换。当用户发起通信请求时，用户面依据用户时延要求和当前通信时

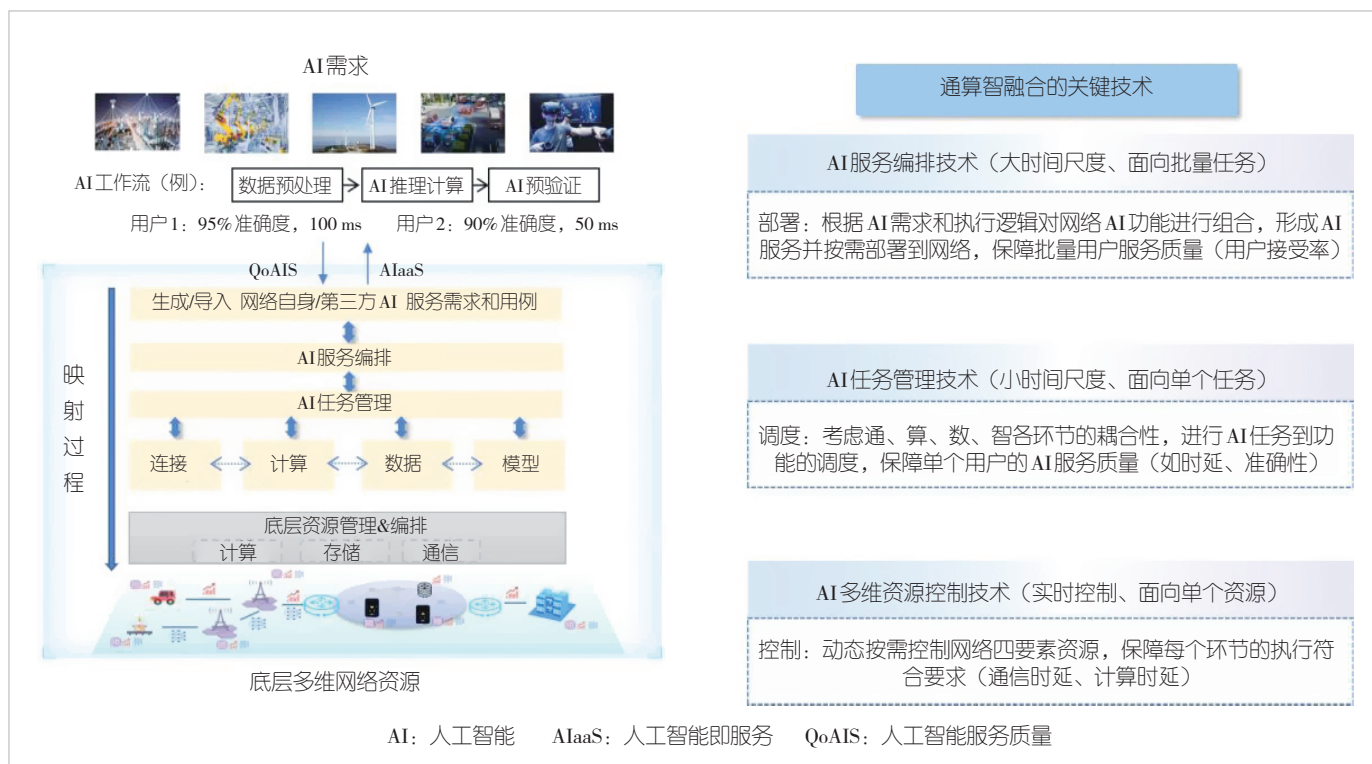


图6 通算数智融合的编管控技术

延, 在计算资源池中动态筛选计算资源, 以满足用户需求。

3.4 网络数字孪生

在6G网络向智能化、高效运维演进的过程中, 网络数字孪生 (NDT) 利用数字化技术创建物理网络的虚拟映射, 能够对物理网络的行为和性能进行精准模拟, 为简化多维能力的6G网络管理、实现AI决策预验证的智能化提供强有力的支持^[5]。

网络数字孪生的功能实现依赖于多项关键技术。高效数据采集和生成技术是NDT的基础。6G网络规模庞大且结构复杂, 要实现物理网络与孪生网络的大规模数据同步, 必须依赖高效的数据采集和生成技术。该技术能够快速、准确地收集物理网络中的各类数据, 包括设备状态、流量信息、用户行为等, 为后续模拟分析供应充实的数据资源。

高精度孪生建模技术是NDT的核心。该技术通过构建精准映射物理网络的高精度孪生模型, 力求在时空维度上全方位地再现物理网络的状态、功能和性能, 为网络提供一个与现实世界高度契合的虚拟镜像。高精度孪生建模技术使用融合注意力机制和神经网络的建模方法, 能够深入挖掘网络中节点和边的关系, 实现高效建模, 极大地提高了网络推理和策略预验证的准确度。在由多数字孪生 (DT) 模型组成的网络全局DT模型中, 不同网络节点和边的DT模型根据各

自物理网元的架构和功能定制而成, 缺乏对网络整体功能的深入理解和优化, 导致网络DT模型效率低下。将扩散模型引入由多个DT模型串联组成的网络模型中, 不仅可以实现高效的多模型协同优化, 达成建模精度与成本之间的平衡, 还能显著提升NDT建模优化的准确度。在资源有限的情况下, 依然能够保证构建出高质量的网络DT模型, 这对于NDT在大规模网络中的实际应用至关重要。

智能编排和管理技术是NDT实现高效运作的关键。通过智能解析方法, 采用多模态智能解析技术对DT任务请求进行深度分析, 能够精准地将复杂的任务请求拆解为3个核心要素: 数据、孪生模型与计算需求, 为后续的资源分配和任务部署提供依据。在编排阶段, 依据智能解析得到的三要素信息, 结合网络的拓扑结构, 运用智能编排算法可对DT模型进行自动化部署。通过智能解析与编排的协同, 可实现NDT的智能闭环管理, 使得网络能够动态适应任务需求的变化, 为NDT的高效运行和广泛应用提供有力支撑。

高性能仿真和预验证技术是NDT的重要应用方向。NDT支持对网络AI决策以及性能指标进行预验证, 通过高性能仿真技术, 在虚拟环境中模拟网络在不同场景下的运行情况, 评估AI策略的有效性和可行性。在6G网络中, AI策略的更新速度较快, 若不及时验证其可行性, 可能会导致网络

性能恶化。NDT能够提前预测未来的网络状态，对AI策略进行全面、深入的预验证，及时发现潜在问题并进行优化调整，确保AI策略在实际应用中能够发挥最佳效果。

4 结束语

6G网络作为未来通信技术的发展方向，其通感算智融合的特性将极大地推动各行业的数字化和智能化转型升级。通过构建以云化为底座、端到端服务化为基因、内生AI为智能引擎的新型网络架构，6G可实现从传统通信网络向智能服务平台的本质跨越。通过云原生、服务化RAN、内生AI以及网络数字孪生等关键技术，6G网络将实现更加智能、高效、灵活的运行，可以有效解决5G网络复杂度高、能耗大、对业务适应能力弱等行业痛点。未来，随着技术的不断进步和应用场景的不断拓展，6G网络将在更多领域发挥重要作用，推动社会向更加数字化、智能化的方向发展。

参考文献

- [1] IMT-2030(6G)推进组. 6G无线系统设计原则和典型特征 [R]. 2023
- [2] I C-L, HUANG J, CRESPI N. 6G access and edge computing – ICDT deep convergence [M]// BERTIN E, CRESPI N, MAGEDANZ T. Shaping Future 6G Networks: Needs, Impacts, and Technologies. Hoboken: Wiley-IEEE Press, 2022: 187–220. DOI: 10.1002/9781119765554.ch12

- [3] LI N, LIU G, ZHANG H, et al. Micro-service-based radio access network [J]. China communications, 2022, 19(3): 1–15
- [4] LIU G, LI N, DENG J, et al. The SOLIDS 6G mobile network architecture: driving forces, features, and functional topology [J]. Engineering, 2022, 8: 42–59
- [5] DENG J, ZHENG Q, LIU G, et al. A digital twin approach for self-optimization of mobile networks [C]//2021 IEEE Wireless Communications and Networking Conference Workshops (WCNCW). IEEE, 2021. DOI:10.1109/WCNCW49093.2021.9420037

作者简介



刘光毅，中国移动通信集团有限公司集团级首席专家、6G总监，中关村泛联院首席科学家，教授级高工，中国通信学会会士，享受国务院政府特殊津贴专家，科技部中青年科技创新领军人才，6G推进组无线组副组长；曾主持和参与国家级研发项目10余项，2016年荣获国家科技进步奖特等奖，2015年获中国通信学会科技进步奖特等奖，2017年所在团队入选央企楷模，2020年入选全国劳模，先后荣获中国通信标准化协会科学技术奖一等奖和二等奖共5项、国家知识产权局颁发的中国专利奖金奖1项和优秀奖3项；发表学术论文200余篇，申请发明专利500余项，参与国际标准制定20余项，出版专著《6G重塑世界》《6G：从通信到多能力融合的革命》等。

面向下一代光网络的光计算技术应用思考



Optical Computing: Advancing Next-Generation Optical Network Applications

李俊杰/LI Junjie, 刘宇阳/LIU Yuyang, 霍晓莉/HUO Xiaoli

(中国电信股份有限公司研究院, 中国 北京 102209)

(Research Institute of China Telecom Corporation Limited, Beijing 102209, China)

DOI: 10.12142/ZTETJ.202504009

网络出版地址: <http://kns.cnki.net/kcms/detail/34.1228.tn.20250716.0844.002.html>

网络出版日期: 2025-07-16

收稿日期: 2025-05-26

摘要: 光计算作为一种基于光子进行信息处理的新型计算范式, 凭借高并行性、低能耗和大带宽等优势, 正成为下一代光网络演进的关键支撑技术。从光计算的架构形式与系统层级双重路径出发, 系统解析其技术发展脉络, 并归纳关键支撑要素与集成模式。在此基础上, 聚焦下一代光网络中的典型应用场景, 重点探讨光计算在光信号处理、网络智能优化、通感一体及智算体系等领域的融合路径, 揭示其在构建新型信息基础设施中的潜在价值与发展方向。

关键词: 下一代光网络; 光计算; 新型信息基础设施

Abstract: Optical computing, as a novel computing paradigm utilizing photons for information processing, is emerging as a key enabling technology for next-generation optical networks due to inherent advantages including high parallelism, low power consumption, and large bandwidth. The technological evolution of optical computing can be analyzed through dual perspectives of architectural forms and system hierarchy, with critical supporting elements and integration paradigms systematically summarized. Typical application scenarios in next-generation optical networks demonstrate optical computing's integration pathways in optical signal processing, network intelligence optimization, integrated sensing-communication, and intelligent computing systems. These developments reveal significant potential value and clarify future directions for constructing advanced information infrastructure.

Keywords: next-generation optical network; optical computing; new information infrastructure

引用格式: 李俊杰, 刘宇阳, 霍晓莉. 面向下一代光网络的光计算技术应用思考 [J]. 中兴通讯技术, 2025, 31(4): 55-63. DOI: 10.12142/ZTETJ.202504009

Citation: LI J J, LIU Y Y, HUO X L. Optical computing: advancing next-generation optical network applications [J]. ZTE technology journal, 2025, 31(4): 55-63. DOI: 10.12142/ZTETJ.202504009

随着6G、云计算、人工智能等业务蓬勃发展, 下一代光网络承载着前所未有的流量增长与多样化业务需求, 正朝着超高速率、超大容量、超大规模及高度智能化方向演进^[1]。为支撑未来光网络的新兴业务及其运维, 应对数据处理和计算需求的激增, 传统电计算遭遇算力与能效的双重瓶颈。通信网络对实时性、智能化的要求不断提高, 需在更低时延下处理海量数据。这些因素促使业界重新审视信号处理方式, 推动在光网络中引入新型计算技术以突破现有瓶颈。

光计算(又称光子计算)凭借光信号承载与处理信息, 被视为突破电子计算瓶颈的有力候选方案^[2]。与传统电子计算相比, 其具备多维并行、超低时延及低功耗等显著优势。相关研究显示, 光子集成电路现已能够实现超高速人工神经

网络, 有效加速机器学习与人工智能应用, 在医疗诊断、电信网络及高性能计算等诸多领域展现出巨大潜力^[3]。

近年来, 光计算技术受到学术界和产业界的广泛关注, 相关研究与应用竞争日趋激烈。2014年前后, 多个团队提出可扩展硅光神经网络架构; 2017年, 基于硅光平台的光神经网络计算芯片初步演示成功^[4]。这些系统引入片上电子电路进行校准和控制, 解决了光学器件高敏感性、低稳定性问题, 实现多级光计算单元级联和基本非线性功能。同时, 高速调制器、低损耗波导、相变存储材料、光放大器等新型光器件不断涌现, 为构建复杂光计算系统提供元件支撑^[5-6]。在产业界, 大量初创公司从高校实验室孵化, 聚焦光计算芯片商业化落地^[7]。谷歌、Meta、微软、百度等通过投资初创

公司布局光计算，英伟达等大型企业也投入研发。面向人工智能（AI）加速的光计算芯片与板卡已研制成功并试用，有望短期内用于云端AI推理加速。例如，美国Lightmatter公司研发的硅光AI芯片采用多级马赫-曾德尔干涉仪（MZI）阵列架构，结合电子芯片构建光电混合计算系统。测试数据显示，在双向编码器表示Transformer（BERT）模型自然语言推理任务中，该系统在同等功耗条件下，计算速度是英伟达A100图形处理器（GPU）的5倍以上^[8-9]。中国初创企业曦智科技开发的光电混合架构光子计算处理器在组合优化问题求解中展现出显著优势，其计算延迟可低至3 ns，较传统GPU提升两个数量级^[9]。当前，光计算技术已实现从实验室研发到产业化应用的重要跨越。

本文系统分析光计算技术的3个技术路径——模拟光计算、数字光计算与光电混合计算，重点解析其架构特征、关键技术及应用场景。基于此，进一步聚焦下一代光网络的核心应用领域，深入探讨光计算技术的应用潜力与发展前景。

1 光计算架构、技术与场景分析

光计算技术可从架构类型、关键技术和系统集成3个维度进行分析。在架构层面，分为模拟光计算、数字光计算和光电混合计算3种类型；技术层面聚焦器件、芯片与模块的开发；系统层面则着重整体架构集成与平台化应用实现。

1.1 架构分类

1.1.1 模拟光计算

模拟光计算是利用光模拟信号的连续幅值或相位直接执

行数学运算的计算方式，其典型代表包括光学神经网络、光学矩阵计算和光学伊辛机等系统。与数字计算侧重离散逻辑不同，模拟光计算借助光学干涉、衍射等物理过程实现矩阵运算加速，具备高度并行性和极高吞吐量，在以线性矩阵运算为主导的深度学习等任务中，展现出显著的性能优势^[10]。

模拟光计算可细分为相干光计算和非相干光计算两种形式。相干光计算借助激光相干性的幅值和相位信息，通过干涉实现精准加权求和，如基于MZI阵列的光矩阵乘法器；非相干光计算则以光强度（功率）作为计算量，常通过光电探测将光强转为电信号读取计算结果。相干方案精度高，但对相位稳定性要求严格；非相干方案结构简单，更易集成。近年来的研究聚焦于在片上集成光路中实现大规模光学矩阵运算单元。例如，平面集成式光子神经网络（PNN）架构广泛采用MZI、微环谐振器（MRR）、相变材料（PCM）等器件构建阵列，实现矩阵向量乘法等线性计算^[5]。清华大学团队提出的“太极”分布式光芯片结合光学干涉与衍射两种机制，集成了上百万个等效的“神经元”，在保持可重构通用计算能力的同时实现了大规模并行计算，系统能效可达到160 TOPS/W（每瓦特万亿次运算）^[11]，如图1（a）所示。利用微环阵列可对不同波长的光信号施加独立的幅度权重，以实现多波长并行的矩阵计算。2023年，上海交通大学团队基于可调谐微环阵列研制了高阶光子张量处理芯片（PTFP），基于波长、空间等维度实现了更复杂的张量运算，加速深度学习神经网络计算^[12]，如图1（b）所示。自由空间光学也是模拟光计算的重要路径之一，例如利用空间光调制器（SLM）和透镜构建光学傅里叶变换和卷积计算器，或构造空间全光神经网络。自由空间互连式PNN可支持规模庞大

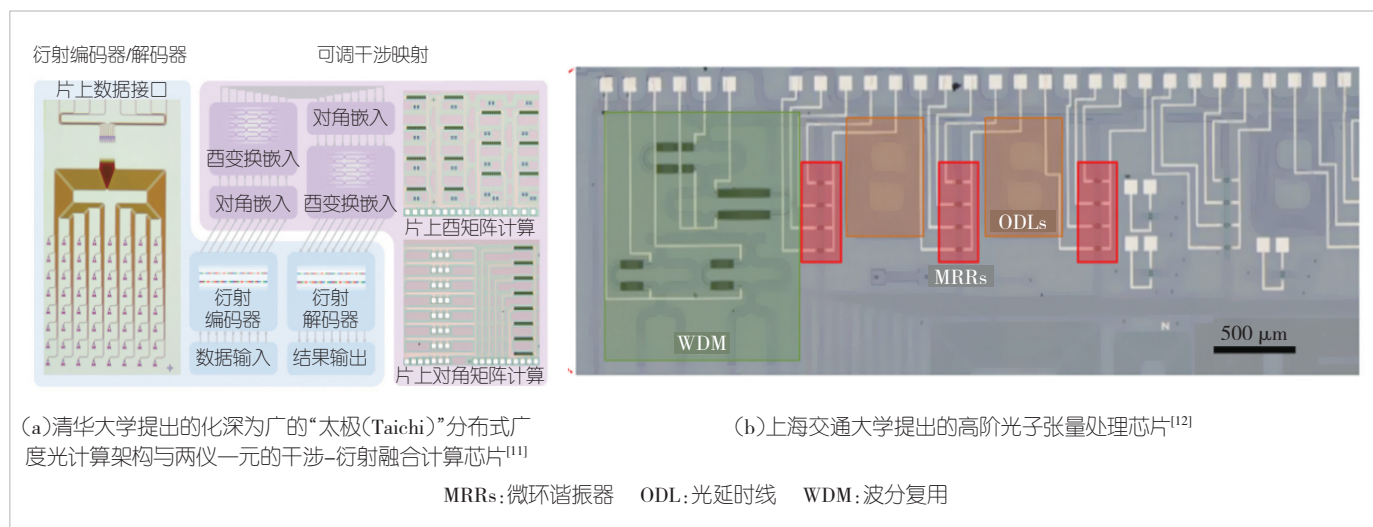


图1 典型技术方案与实现

的矩阵运算，但体积较大且难以集成，适合作为实验验证和特定高端应用（如超高速信号处理）的方案^[13]。

此外，光学伊辛机作为模拟光计算的重要方向，针对组合优化离散问题展示出强大潜力。伊辛问题求解可映射为光学系统的能量极小化过程，通过模拟自旋相互作用的光学器件网络快速收敛到最优或次优解^[7]。例如，利用光纤环路延迟和非线性增益介质构建的相干伊辛机（CIM），可在硬件上直接模拟伊辛模型的演化以高效寻优。在2015年，斯坦福大学与东京大学联合团队提出了测量反馈型相干伊辛机方案，通过高速光电探测和反馈电路模拟伊辛模型演化过程，并于2016年成功利用该系统求解了典型非确定性多项式难题（NP-hard）——最大割问题。华中科技大学在2025年提出的单芯片集成光电耦合伊辛计算机取得了重要突破。该系统采用差分驱动微环调制器（MRM）实现高效非线性响应（4 mW/spin），通过Sagnac结构光矩阵实现实对称耦合，可在150 ns内完成四自旋问题的地面态求解，芯片面积仅为0.01 mm²^[14]。上述研究表明，此类光学计算系统在求解网络路由优化、资源调度等组合优化难题时，相比传统电子计算能实现指数级的加速和更优的能耗表现。因此，在需实时求解复杂网络优化的场景（如网络动态路由计算）中，光学伊辛机有望成为传统算法的重要补充。

模拟光计算在精度和容错性方面仍面临挑战。模拟信号易受噪声和器件不稳定性影响，因此如何在大规模光计算系统中保持计算精度成为关键问题，而这通常需要借助混合集成的电子控制电路对光学单元进行校准补偿^[15]。此外，模拟光计算实现非线性激活功能往往需要引入光电转换或光学非线性器件，如利用光电二极管+放大电路实现整流线性单元函数（ReLU）激活，或采用饱和吸收体、相变材料等实现纯光学的阈值响应。目前，基于锗硅探测器和微环调谐的光控激活器等片上可编程的光学非线性单元已实现突破，能够实现低至毫瓦级阈值的ReLU和Sigmoid函数，有效缓解了多层光网络中信号级联衰减的问题。随着器件工艺的进步和光电融合设计的优化，模拟光计算的精度和稳定性有望逐步提升。

1.1.2 数字光计算

数字光计算是一种采用光子来表示二进制信息并实现逻辑运算与数字信号处理的技术方案。该技术旨在构建与电子计算机逻辑门阵列类似的全光逻辑电路体系，以实现通用数字计算功能。

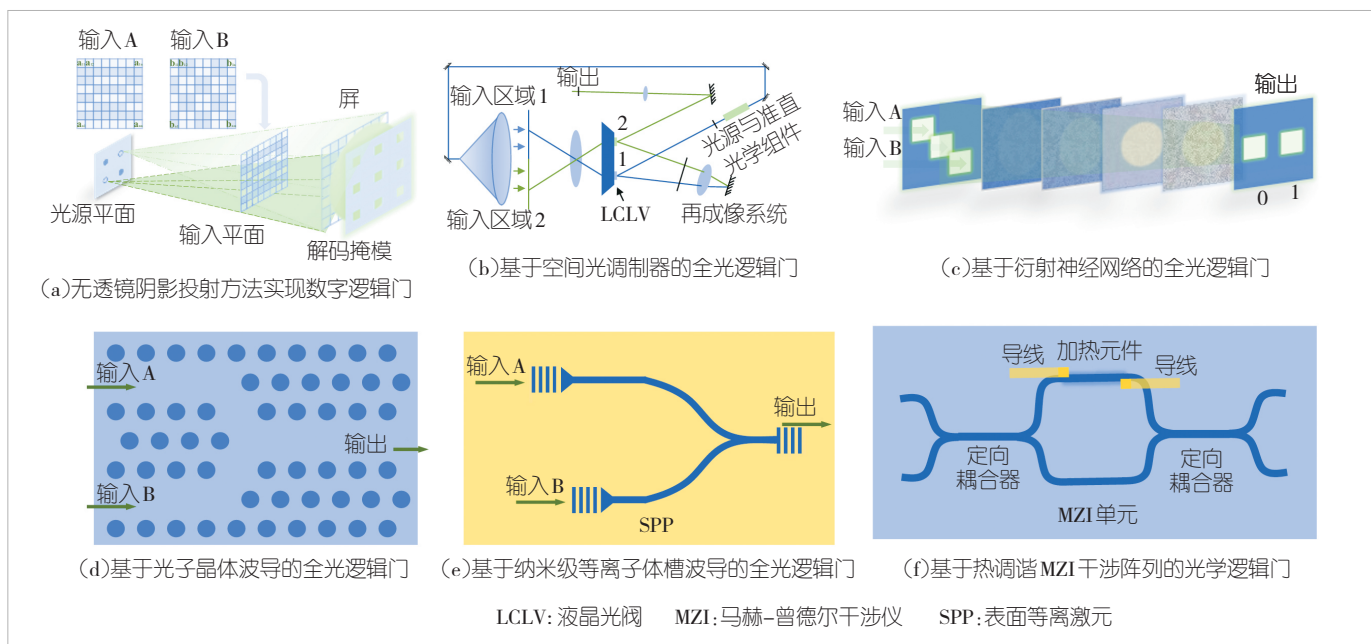
数字光计算展现出显著优势：光的开关响应速度极快且无需电容充放电过程，构建稳定的全光逻辑门电路可能实现

超越互补金属氧化物半导体（CMOS）技术的高速并行数字处理能力。然而实际发展面临诸多挑战，核心瓶颈在于全光非线性效应和反馈控制的有效实现^[6]。早期研究主要利用光学非线性效应（如四波混频、增益饱和等）实现全光开关和触发器，但面临材料非线性阈值高、集成困难等问题。为提高与硅光工艺的兼容性并降低损耗，中山大学等团队提出基于多模干涉（MMI）的集成光学逻辑门方案，通过精确设计波导结构使干涉输出匹配预期逻辑结果，实现了硅基集成多种基本逻辑门^[16]。近年还出现了基于线性光学的新型全光逻辑实现途径：利用衍射型深度神经网络（D²NN）训练超表面光学元件，实现对输入光信号的任意可编程线性变换，从而完成与逻辑门等价的映射。东京大学团队利用定制超表面构建了具有完备逻辑功能的衍射光学层，能在图像平面上对256个像素并行执行逻辑运算，并通过更换衍射层在硬件上动态重构AND（与）、OR（或）、XOR（异或）等16种逻辑功能^[17]。类似地，还有研究利用空间光调制器实现可重构的全光逻辑门，并基于光子晶体波导、等离子纳米波导、热调谐MZI阵列等不同机理演示了多种光学逻辑门。图2总结了部分全光逻辑门的范式，包括透镜阴影投射法、SLM方法、衍射神经网络、光子晶体、等离子槽波导、热光MZI阵列等路径。这些成果表明，全光数字计算在器件层面是可行的。尤其在光通信领域，数字光计算可用于光学信号处理和网络路由等应用。例如，以光逻辑门直接解析分组头信息、执行错误检测等，从而减少光电转换开销，实现真正的全光网络。

当前数字光计算与通用可编程计算机仍存在显著差距，主要瓶颈在于缺乏高效的光存储和时序控制机制^[4]。电子计算中依赖触发器和存储器保持逻辑状态，而纯光系统中光子难以静止存储，现有光学缓存和随机存取存储器尚未成熟。尽管光学缓存、光子晶体存储等方案被提出，但其读写速率和集成度仍远逊于电子存储器。此外，全光电路要实现复杂逻辑往往需要多级门阵列和反馈回路，面临光路损耗、信号放大和噪声累积等严峻挑战。目前全光逻辑演示多限于固定功能或小规模系统，可编程能力较弱。因此，数字光计算短期内更适合作为专用光信号处理加速器（如高速全光开关矩阵、光学加密/编码器件等），尚难取代通用中央处理器（CPU）与GPU。

1.1.3 光电混合计算

鉴于纯光模拟与纯光数字方案均存在局限性，当前研究趋势正转向光电混合计算架构。该架构将光计算单元与电子器件紧密结合，充分发挥二者优势，构建切实可行的混合系

图2 全光逻辑门的范式^[6]

统。在光电混合计算中，核心思路是让光子承担大规模并行的线性计算，以充分发挥其高速、低功耗的优势；而电子器件处理存储、控制以及非线性计算等任务，为系统提供成熟可靠的补充。

光电混合计算已成为当前光计算原型的主流实现方案。例如，硅光神经网络芯片通常在片上集成光调制矩阵阵列和光电探测器，同时配套CMOS电路实现对各光调制单元的调控、漂移补偿以及对探测信号的后续非线性处理。Light-matter公司的光计算加速器正是基于此思路的典型产物，其核心是一个可等效执行矩阵乘法的多级MZI光子芯片，外围结合高性能数字电子芯片，从而实现“光计算+电读取/控制”的协同工作。从体系结构上看，光电混合计算本质上是一种异构计算范式，即在同一系统中融合不同物理机制的计算引擎，各司其职完成复杂任务。为提升这类架构的性能，学术界正探索更紧密的光电融合设计，包括共封装光学（CPO）系统（将光引擎与电芯片高密度互连封装）以及“存算一体”光电单元（将光学计算单元与非易失光存储集成）等。

未来可扩展的光子计算机可能采取的混合路线是：在电子器件擅长的领域仍保留电子方案，将最耗时耗能的矩阵运算等部分交由光子高效完成。《Nature Photonics》中的综述指出，实现大规模神经形态光计算需要高度紧密的光电子融合，包括混合光学-电子存储架构，以根据应用灵活选择易失/非易失、数字/模拟的存储方案^[4]。换言之，光计算的未来并非取代电子计算，而是与之融合共生，构建光电协同的新型计算形态。通过发挥光的优势并弥补其短板，光电混合计算为光计算技术走出实验室、服务实际应用提供了现实途径。

表1针对上述3种架构进行了对比，不同光计算架构在运算机制、应用目标与当前产业落地路径上呈现明显差异。模拟光计算侧重于高通量AI推理，数字光计算探索全光逻辑通用处理，混合计算则在工程可行性与系统兼容性上具备突出优势。

1.2 关键技术分析

为了支撑上述光计算的各类架构，需在器件、芯片与模块三大领域进行技术创新。

表1 典型光计算架构类型对比分析

架构类型	运算机制	优势	挑战	应用场景
模拟光计算	干涉、衍射、衰减模拟线性运算	超高并行性、大吞吐量、极高能效	精度受噪声影响、缺乏非线性机制	智能推理、图像识别、伊辛优化
数字光计算	全光逻辑门执行二进制运算	高速、可构建通用逻辑	缺乏光存储、非线性控制难	光通信逻辑、加密处理
光电混合计算	光做矩阵乘法、电处理控制与非线性	可重构、易部署、工程可实现	光电协同复杂、封装和热控困难	智能加速、边缘计算

1.2.1 器件

器件层面涵盖了光计算所需的各类光子器件和关键材料，是构建上层计算功能的基础。

高速光调制器和相位调制器通过输入电子信号调控光信号的幅度或相位，实现计算系数或数据的编码。常用的方案包括基于电光效应的硅基 MZI 调制器、铌酸锂调制器，以及依赖热光效应的微环调制器等。这些调制器需具备高带宽、低插损和低功耗，以满足大规模阵列集成需求。

激光光源为光计算提供载波信号，其中多波长平行光源（例如光频梳或多路激光阵列）可有效提升频谱并行度。近期发展的片上集成可调谐激光阵列技术，为单芯片集成稳定的多波长光源^[15, 18]提供了可能。

光学互连和波导实现光信号在芯片内外的传输与分配，其中低损耗、高密度的硅波导以及三维光纤耦合技术保障了光计算模块间高效通信。光探测器（如锗硅光电二极管）完成光信号至电信号的转换，用于数据读出或反馈控制。其关键性能指标包括高响应度、低噪声和高速响应，部分光计算架构还要求探测器具备模拟域非线性函数的直接执行能力。

光计算系统不仅依赖核心器件，还需光学非线性元件（如饱和吸收器、非线性光纤、光放大器等）以实现阈值逻辑或信号放大，以及可重构光学元件（如相变存储材料、可调谐滤波器等）用于动态调整计算参数。此外，高精度的光学封装、散热与控制组件也是其重要组成部分。

总之，光计算系统的性能上限由器件成熟度直接决定。尽管光子器件在速度和并行度上具有天然优势，但在尺寸、工艺成熟度和集成复杂度等方面仍需改进。推动器件向更小型化、低功耗和易于大规模制造方向发展，是光计算的长期任务。

1.2.2 芯片

芯片级聚焦于将多种光子器件集成，构建可执行特定计算功能的光计算芯片或光子电路^[19-20]。

典型的芯片级实现通常基于硅光子或 III-V 材料平台，集成矩阵运算单元（如 MZI 阵列或微环阵列）、光源耦合、光探测阵列等，构成完整的计算加速器。例如，麻省理工学院等机构研制的光学矩阵乘法芯片包含了 $N \times N$ 的干涉仪网络，可对输入向量与片上编程的矩阵进行并行乘积运算。芯片级设计需综合考虑光路布局 and 电子电路布局，以减少光传输损耗和串扰，并实现有效的热控和校准。硅光子技术因可兼容 CMOS 工艺，被认为是构建高速、低延时光计算芯片的理想平台，在 AI 高算力场景中具有重要的潜在应用价值。

近年来，全球已研发出多款原型光计算芯片，包括用于

简单图像分类、算力达数十 GMAC/s（每秒执行的十亿次乘积累加运算）的硅基神经网络芯片，以及光子卷积加速器芯片、光学张量处理芯片等。这些芯片在手写数字识别数据集（MNIST）等任务上取得了接近 90% 的推理准确率，算力达到传统同规模电子网络的数倍到数 10 倍。尽管目前光计算芯片多为专用加速器，功能相对单一，但其在实际任务中的可行性已得到初步验证^[15, 21]。

随着工艺改进和设计优化，未来的光计算芯片有望实现更高的计算密度。例如，有研究引入多维复用（波长复用+模式复用+偏振复用），使一款 8×8 MZI 网格的矩阵计算芯片在执行矩阵-矩阵乘法加速时，等效算力达到 40 GMAC/s，是相同规模传统电子网络的 2 倍。芯片级的进展表明，在几平方毫米的光片上实现光万亿次级别计算已成为现实，为上层模块和系统提供了功能完备的基本单元。

1.2.3 模块

模块级是指将一个或多个光计算芯片与辅助电路、封装光学等集成为可供系统直接使用的计算模块或板卡，类似于 CPU 主板、GPU 加速卡，但其核心计算单元为光计算芯片。

模块级设计关注电源、控制接口、散热以及与现有数字系统的对接。光计算芯片通常需要精密光纤耦合、激光光源供给、温度稳定控制等，因此将其制成标准化模块可极大方便系统集成。目前已有公司推出面向数据中心服务器的光计算加速卡，采用外围组件互连高速通道（PCIe）等接口直接插入主机，作为 AI 推理加速器使用。这些模块一般包含光计算芯片（或多个芯片）、驱动电路板、光源（可能通过光纤引入或板载激光器）、光电转换及控制应用特定集成电路（ASIC），以及必要的固件软件。通过标准化模块，光计算可以像 GPU 那样被现有计算机系统调用，例如协助处理深度学习推理、实时信号处理等任务。

模块级的重点在于兼容性和易用性，即如何将光计算能力透明地提供给用户应用，无需用户了解底层光学细节。这涉及开发相应的软件栈、编译工具，将神经网络模型映射到光计算单元上执行^[22]。此外，模块级还要解决多芯片协同问题。一个模块可能集成多个光芯片或多个并行光通道，以获得更高吞吐量；如何高效调度这些资源并保证模块间通信低延迟，是当前面临的重要工程挑战。

尽管如此，当前已有少量光计算模块进入测试阶段，表明光计算正迈出实验室，向产品化又进了一步。

表 2 展示了光计算关键技术的对比，从器件到模块的推进顺序反映了其产业化落地的进程。其中，器件与芯片是性

能的基础，而模块构成了部署单元。整体上看，光计算正从低层硬件创新逐步走向高层系统集成与服务平台构建，为通信网络的多层级算力融合提供了坚实基础。

1.3 系统及平台集成

基于上述技术与架构，光计算需考虑如何融入实际网络并创造价值。其融合方式通常可分为计算系统与光计算的融合应用（系统层面），以及上层平台与光计算的融合应用（平台层面）。

1.3.1 系统

系统级着眼于光计算在完整计算系统中的引入和部署形态，包括光计算与现有计算/网络基础设施的融合方式。例如，在数据中心的服务器和加速集群中部署光计算板卡，形成“CPU+GPU+光计算加速器”异构计算服务器；或者在超算系统中，引入光计算节点构成部分功能单元，实现更高性能和能效比的系统设计^[23]。

从系统架构角度出发，光计算单元的网络互联协同至关重要。在高速互联计算集群里，可借助光互连网络，让多个光计算加速器直连，凭借光的低延迟、高带宽特性，构建分布式光计算系统。在边缘计算场景中，光计算芯片可嵌入网络设备（如路由器、基站）内，作为本地 AI 处理单元，打造出“计算+通信”融合的智能网络节点。当系统级部署光计算时，接口与编排是关键。需制定明晰的接口协议，使上层软件能便捷调用光计算资源，并将其纳入系统资源管理与调度体系。可采用容器或虚拟化技术，将光计算加速呈现为可编排服务；或研发新操作系统支持，把光计算纳入异构计算调度体系。

目前，光计算硬件尚处于原型阶段，尚未实现系统级大规模部署的成熟应用。但学术界已开始探讨相关模型，例如利用光计算加速数据中心网络中的繁重任务（如 AI 推理服务），建立“光计算云”；或在超级计算系统中，将光子加速模块用于特定迭代算法以提升整体性能。可以预见，随着单节点光计算性能的提高和可靠性的增强，系统级采用光计算

将有望从小规模试验走向大规模推广。在许多国家，互联网数据中心和超算中心也在关注采用光子加速技术来降低能耗，提升性能。

1.3.2 平台

平台级是指在更宏观的信息基础设施和应用平台中融合光计算能力。例如，未来企业构建的“算力网络”或“智能云网”平台，可以将光计算模块作为算力池的一部分，通过网络进行统一调度，为用户提供更高效的计算服务。

在这一层级，光计算不再被视为单一设备，而是作为整个信息平台的组成要素参与资源编排。例如，在面向智能城市的云平台中，后台拥有中央处理器（CPU）/GPU 集群以及若干光计算加速阵列。通过调度策略，将某些高耗能 AI 任务（如视频分析、机器视觉）分配到光计算阵列执行，以降低延迟和功耗；而其他任务则交由通用 CPU/GPU 处理，以保证灵活性。

在未来边缘云平台中，分布在各地的边缘节点可能搭载光计算加速器，用于实时处理本地数据，然后通过高速光网络回传或协同决策。这种架构需要新的平台软件支持，包括异构算力的统一编程接口、分布式任务拆分与调度算法，以及能根据任务特性智能选择最优计算资源（光还是电）的中间件。宏观来看，平台级视角意味着要从战略层面规划光计算在网络架构中的位置。例如，在哪些节点部署光计算单元、如何与现有的光传输系统对接、怎样提供面向业务的服务接口等。这一层级往往与国家或行业的发展战略相关，如“东数西算”工程强调将计算任务在全国范围内高效调配，“算网融合”要求计算资源和网络资源的一体化编排。在这些顶层规划中，光计算有望作为绿色高效算力的重要组成部分被纳入，从而支持国家在人工智能、边缘计算、碳中和等方面的宏大战略需求。

综上所述，从架构、技术与应用层级 3 个维度展开的分析来看，光计算技术各层级的现状和挑战存在差异。在关键技术层面，器件、芯片、模块整体处于技术攻关和工程化起

表2 光计算关键技术对比分析

技术	功能作用	代表器件/方案	关键技术指标	主要挑战
器件	构成基本计算与调控单元	光调制器、微环谐振器、光探测器等	带宽、响应时间、插损、尺寸	微纳加工精度、材料稳定性、低功耗一致性
芯片	集成基本计算功能的最小系统单元	MZI 矩阵、微环阵列、张量处理核	算力密度、误差率、封装密度	热漂移补偿、信号完整性、封装耦合
模块	提供标准接口与辅助功能的计算板卡	光计算主板、PCIe 卡、边缘插卡	模块功耗、接口兼容性	光电接口标准不统一、系统级散热与耦合

MZI:马赫-曾德尔干涉仪 PCIe:外围组件互连高速通道

步阶段。器件和芯片层科研创新比较活跃，但成熟商用尚需时日；模块层则需要实际工程来体现价值。在应用场景方面，系统层已有原型验证，展现特定领域应用前景；平台层属前瞻性探索，短期内落地难度大，但从长远看潜力巨大。

2 面向下一代光网络的光计算应用分析

光计算的兴起既是机遇也是挑战。下一代光网络将不再仅仅是传统的传输管道，而是面向 AI 时代、实现网络与智能深度融合的关键基础设施。光计算作为利用光子进行超高速、低功耗信息处理的技术，有望成为推动光网络智能化演进的重要途径，使网络从以传输为中心转变为集传输、计算、感知于一体的智能网络。

从光计算的具体应用场景来看，其未来潜在应用方向主要分为以下几方面：光电信号处理、网络智能优化、通感一体以及智算体系建设。

2.1 光电信号处理

光计算可用于直接处理高速光信号，从而提升光网络中信号传输和处理的效率与性能。在骨干网和城域网中，当前的信号处理主要依赖高速数字信号处理（DSP）芯片和电交换设备。然而随着速率攀升，电子 DSP 的功耗和复杂度急剧上升，已成为限制系统性能提升的瓶颈。因此，利用光计算器件直接对传输中的光信号进行均衡、滤波、纠错等处理，可减少光电转换和数字 DSP 负担。例如，全光神经网络可用于非线性噪声抑制和光性能监测等。这类应用可望通过在光域完成复杂运算来降低节点处理延迟和能耗^[24]。

目前，在光节点中采用波长变换、光再生等光学处理手段来卸载电子信号处理，已有效提高了网络性能。这表明，光计算用于信号处理可进一步改善传输质量和效率：比如全光的信道均衡与纠错，可以在不增加电子开销的情况下提升信号传输可靠性。因此，在高速率、大容量传输场景下引入光计算型信号处理单元，有潜力缓解电子瓶颈，使光网络支持 Tbit/s 级的信号以更低延迟、更高可靠性传输，从而满足未来业务对带宽和实时性的要求。

2.2 网络智能优化

通信网络的规划和资源调度涉及大量复杂的组合优化问题，例如路由和波长分配（RWA）、业务调度、流量工程等。这些问题在大规模网络中非常复杂，传统电子计算往往难以实时求解。例如，中国电信拥有全球最大的基于可重构光分插复用器（ROADM）的全光骨干网，该网络包含近 700 个节点，系统带宽超 1 200 Tbit/s，总长度达 40×10^5 km。

在如此庞大的干线网络中，实现极速、精确、高效的网络规划与资源调度是一项非常艰巨的工程。

光计算擅长并行处理大规模计算任务，在解决这类网络优化问题上具有独特优势。研究表明，对于不要求精确全局最优解的复杂计算任务（如神经网络推理或组合优化），光计算凭借大规模并行和低时延特性，有望超过经典电子计算的效率。面向实际网络，可考虑引入“光计算资源编排器”，利用光子处理器快速求解网络优化模型^[19]。

网络的路由规划、资源分配是复杂的组合优化问题，传统算法在网络规模增大时计算耗时长且难以实时优化。光学伊辛机等模拟计算设备可用于求解网络最优路由、流量工程等 NP-hard 问题，在较短时间内逼近最优解。例如，将网络拓扑映射为伊辛模型，由光计算加速求解，可用于动态线路调度或多约束路由计算，加速自优化网络的决策过程。又如，基于光子的等值模拟退火或伊辛模型计算器，可在极短时间内得到接近最优的网络资源分配方案，从而实时优化光纤链路的路由选择、频谱分配或多接入边缘计算任务的调度。这样的光计算网络优化引擎将使网络控制决策速度大幅提升，提高网络对动态业务变化的响应能力和资源利用效率。

随着网络规模和算力网络协同需求的增长，光计算有望成为未来智能网络的大脑之一，支撑更自优化、更高效的网络运营。

2.3 通感一体

未来网络（如 6G）强调通信与环境感知的融合，即利用同一套网络基础设施同时提供通信服务和环境/态势感知。“通感一体”光网络目前已成为研究热点，其协同架构需要能够处理海量感知数据并与通信过程实时联动。

光计算在其中具有广阔的应用前景：一方面，光纤本身可作为分布式传感器，例如探测光缆振动以进行安防或地质监测；另一方面，光计算单元可以高速处理这些传感信号并提取有用信息，而不必完全依赖电层计算，从而实现通信信号传输与感知数据处理的同步进行。例如，在光纤通信系统中引入光通感协同优化器，利用光计算实时分析沿光纤传回的后向散射信号以探测环境变化，同时优化调整通信波形或频谱资源，以在不牺牲通信质量的前提下提高感知精度。又如，在自由空间光通信或可见光通信中，光计算可以用于快速处理光学雷达（LiDAR）或成像信息，与通信信号共同作用，实现边通信边探测目标的功能。

通过这些手段，光网络将不仅传输信息，还可感知网络所处的环境状态，为安全监测、交通管理等提供实时数据支

撑。尽管通感一体化目前仍处于研究阶段，但基于光计算有望突破实时数据处理瓶颈，助力打造面向智慧城市、智慧交通的光通信与光感知融合平台。

2.4 智算体系建设

光计算在智能算力支持方面具有广阔的应用前景，可作为AI和高性能计算任务的加速器，提升算力供给能力。光子的并行计算特性特别适合矩阵运算等AI核心计算。例如，在大型神经网络推理中，矩阵乘法是主要瓶颈之一，而光计算可以通过光学干涉实现大规模矩阵并行乘积，大幅降低延迟和能耗。

在智算数据中心互联（DCI）场景下，光计算的作用主要体现在两方面：一是支持数据中心内部的算力扩展，二是优化跨数据中心的传输效率。首先，光计算芯片有望成为未来数据中心算力池的新成员。在超大规模数据中心中，AI训练和推理任务占据了相当比重且是电力消耗大户。若在服务器内或机架级别部署光计算加速卡，可大幅提高AI任务性能/功耗比，降低数据中心电源使用效率（PUE）。目前，业界的光计算加速器已经在理论与试验上展现出显著优势。例如，中国科学院上海光学精密机械研究所与北京大学共同研制的超高并行光计算芯片，在50 GHz光学主频下，理论峰值算力可达2 560 TOPS^[25]。对于数据中心（如智算业务）而言，引入这类光计算加速资源有双重收益：既能提升云服务的AI处理能力，又可削减能耗，降低运营成本。因此，可在自有互联网数据中心（IDC）中试点试验性的光计算服务器，用于特定AI应用（如通信网络优化仿真、视频内容分析等），评估其性能和节能效果。其次，在数据中心互连网络优化方面，光计算也有用武之地。数据中心集群间的流量调度、负载均衡也是复杂的优化问题，需要实时决策。在DC-Mesh组网的链路中，如果能利用光学优化器实时计算最优流量分配，或者利用光子加速器动态压缩/加密传输数据，将提升DCI网络的效率 and 安全性。例如，可以构建全光数据压缩模块，对冗余数据进行光学变换和编码压缩，然后在另一端以光学方式解压，提高链路有效利用率。

另一个典型的场景是边缘计算。随着移动边缘计算（MEC）的部署，人们开始在本地汇聚点（如5G基站机房、汇聚局所）部署计算资源，以提供低时延服务。如果能够使用光计算加速器承担部分边缘AI推理或大数据处理任务，那么边缘计算节点的性能密度便能得到显著提升，同时功耗也会相应降低。例如，在视频监控与安全领域，城域网边缘节点需要实时分析海量摄像头视频，采用传统方案将视频回传至云端进行分析，这会面临带宽和时延方面的巨大压力。

但如果在边缘局所放置光计算视觉加速模块，就可以在本地对视频流进行高速并行的目标识别和异常检测，然后仅传送结果或元数据，这样可极大缓解骨干压力并实现毫秒级响应。在工业互联网、车路协同等业务中，靠近现场部署光计算单元处理本地感知数据，能够减少云往返，提高系统自治能力。

将光计算融入网络，能够高效构建算力与网络融合的新型基础设施。在信号传输优化、网络智能控制、通信感知融合和算力供给提升等诸多方面，光计算为实际网络拓展出了广阔的应用前景。此时，网络不再仅仅是传输通道，更成为提供算力服务的平台，成功打造出兼具“高速率”与“智能化”的下一代光网络，在AI和算网融合时代占据优势。

3 结束语

面向未来，光网络的内涵有望从单纯的连接拓展为计算与连接紧密协同、光网络与光计算深度融合的智能化光网络。光计算技术以其独特优势，有望成为突破现有计算瓶颈的新路径，对于支撑下一代光网络的演进具有重要意义。然而，目前光计算技术仍处于探索阶段，其未来发展充满不确定性和挑战。从技术层面看，光计算器件需要在稳定性、集成度和成本上取得突破，新的材料与工艺或许是实现高性能光计算芯片的关键；光计算系统架构需要解决光电融合与可编程性等问题，形成完善的软硬件生态。从应用层面看，光计算要真正发挥作用，还需寻找与电子计算互补且经济可行的应用场景。例如，针对光网络中的特定高价值场景（如超低时延控制、超大规模网络优化），光计算才能体现出足够的优势。

展望未来，光计算的发展路径将呈现多样化趋势。在技术演进层面，光电混合计算或将成为主流方向，即在充分利用光计算的优势的同时，通过电子电路补足控制和通用计算功能。这种折中方案在可预见的未来更具可行性，已有多项研究和产品朝该方向稳步推进。在应用拓展上，光计算有望在通信网络的细分领域逐步实现落地，如全光交换节点、实时光信号处理模块、光子AI加速卡等，完成从实验室走向商用的跨越。同时，光计算与其他新兴领域的融合也值得关注，例如与量子计算、硅光子技术和神经形态计算的交叉，将可能催生出全新的计算架构。在国家和产业层面，对光计算的投入正不断增加，科研机构和企业正协同攻关，力求在这一战略性前沿领域取得关键突破。

总体而言，面向下一代光网络的光计算技术应用仍处于孕育阶段，但其潜在价值巨大。一方面，它为突破网络中电子计算的瓶颈提供了新思路；另一方面，它为网络功能演进

(如智能化、通感算一体化)提供了新的技术手段。未来几年内,我们有望看到光计算在局部场景的试点应用和关键技术的逐步成熟。未来,随着技术的演进和应用的拓展,光计算或将成为下一代网络体系中不可或缺的一环,有望从边缘试点走向规模部署,在光电信号处理、网络智能优化、通感一体以及智算等各类场景中发挥价值。

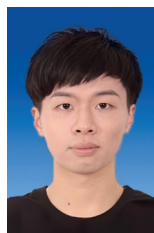
参考文献

- [1] Omdia. Road to 2030: AI and the future of network services - traffic outlook and Implications [R]. 2024
- [2] 中国信息通信研究院. 光计算技术与产业发展研究报告(2023年) [R]. 2023
- [3] 中国信息通信研究院. 信息光子技术发展与应用研究报告(2024年) [R]. 2024
- [4] SHASTRI B J, TAIT A N, FERREIRA DE LIMA T, et al. Photonics for artificial intelligence and neuromorphic computing [J]. Nature photonics, 2021, 15(2): 102 - 114. DOI: 10.1038/s41566-020-00754-y
- [5] 毕岩峰, 吴星宇, 张璐矾, 等. 模拟光计算的发展与应用 [J]. 科学通报, 2024, 69(34): 5028 - 5041. DOI: 10.1360/tb-2024-0692
- [6] 李敬城, 张文凯, 董文婵, 等. 光学逻辑计算进展与挑战 [EB/OL]. (2025-03-24) [2025-04-23]. <https://www.opticsjournal.net/Articles/g241964/EarlyPosting>
- [7] 梁林俊, 鲁楠, 王扬. 光计算技术发展与应用研究 [J]. 信息通信技术与政策, 2023, 49(6): 40 - 46. DOI: 10.12267/j. issn. 2096-5931.2023.06.006
- [8] Lightmatter. Products [EB/OL]. (2025-05-12). <https://lightmatter.co/products/envise/>
- [9] STONE L. Lightmatter announces envise, a photonic accelerator for AI [EB/OL]. (2021-03-15) [2025-05-22]. <https://aibusiness.com/verticals/lightmatter-announces-envise-a-photonic-accelerator-for-ai>
- [10] HUA S, DIVITA E, YU S, et al. An integrated large-scale photonic accelerator with ultralow latency [J]. Nature, 2025, 640 (8058): 361 - 367. DOI: 10.1038/s41586-025-08786-6
- [11] 马国庆, 周常河, 朱镕威, 等. 光计算的发展趋势: 模拟或数字? [J]. 中国激光, 2023, 50(5): 1-13. DOI: 10.3788/CJL221209
- [12] XU Z, ZHOU T, MA M, et al. Large-scale photonic chiplet Taichi empowers 160-TOPS/W artificial general intelligence [J]. Science, 2024, 384(6692): 202 - 209. DOI: 10.1126/science.adl1203
- [13] XU S, WANG J, YI S, et al. High-order tensor flow processing using integrated photonic circuits [J]. Nature communications, 2022, 13(1): 7970. DOI: 10.1038/s41467-022-35723-2
- [14] WETZSTEIN G, OZCAN A, GIGAN S, et al. Inference in artificial intelligence with deep optics and photonics [J]. Nature, 2020, 588(7836): 39 - 47. DOI: 10.1038/s41586-020-2973-6
- [15] WU B, ZHANG W, ZHANG S, et al. A monolithically integrated optical ising machine [J]. Nature communications, 2025, 16(1): 4296. DOI: 10.1038/s41467-025-59537-0
- [16] 黄莹. 基于片上神经网络的硅基集成光子计算芯片 [D]. 杭州: 浙江大学, 2023
- [17] LI Z, CHEN Z, LI B. Optical pulse controlled all-optical logic gates in SiGe/Si multimode interference [J]. Optics express, 2005, 13(3): 1033 - 1038. DOI: 10.1364/opex.13.001033
- [18] MASHIKO R, NARUSE M, HORISAKI R. Diffraction casting [EB/OL]. [2025-04-20]. <https://researching.cn/ArticlePdf/m00090/2024/6/5/056005.pdf>
- [19] 韩溪林, 于潇, 易淑兰, 等. 片上“通感算一体化”光子网络助力打造高智能化警用无人机集群 [J]. 警察技术, 2025(1): 76-79. DOI: 10.3969/j.issn.1009-9875.2025.01.020
- [20] 成骏伟, 江雪怡, 周海龙, 等. 光电智能计算研究进展与挑战 [J]. 中国激光, 2022, 49(12): 1219001. DOI: 10.3788/CJL202249.1219001
- [21] 王俊, 杨晓飞. 光子芯片研究进展及展望 [J]. 世界科学, 2020(12): 29 - 31
- [22] 陈蓓, 张肇阳, 戴庭舸, 等. 光学神经网络及其应用 [J]. 激光与光电子学进展, 2023, 60(6): 0600001. DOI: 10.3788/LOP222304
- [23] 符庭钊, 孙润, 黄禹尧, 等. 片上集成光学神经网络综述(特邀) [J]. 中国激光, 2024(1): 403-418
- [24] 淡一航. 面向光学计算的可集成光子器件及系统研究 [D]. 北京: 北京邮电大学
- [25] 陈宏伟, 于振明, 张天, 等. 光子神经网络发展与挑战 [J]. 中国激光, 2020, 47(5): 80 - 91. DOI: 10.3788/CJL202047.0500004
- [26] YU X, WEI Z, SHA F, et al. Parallel optical computing capable of 100-wavelength multiplexing [EB/OL]. (2025-05-17) [2025-05-22]. <https://elight.springeropen.com/articles/10.1186/s43593-025-00088-8>

作者简介



李俊杰, 中国电信股份有限公司研究院副院长, 教授级高级工程师, 中国电信股份有限公司光传输专业首席专家; 长期从事光传输与光网络等技术研究与应用; 发表论文50余篇。



刘宇韬, 中国电信股份有限公司研究院工程师; 主要从事光通信与光网络等前沿技术研究与应用; 发表论文50余篇。



霍晓莉, 中国电信股份有限公司研究院传输网络研究中心总监, 教授级高级工程师, 中国电信股份有限公司二级首席专家; 长期从事光传送网络技术研究与应用; 发表论文40余篇。

人工智能技术与应用前沿



Frontiers of Artificial Intelligence Technology and Applications

包义明/BAO Yiming, 林阳/LIN Yang, 屠要峰/TU Yaofeng

(中兴通讯股份有限公司, 中国 深圳 518057)
(ZTE Corporation, Shenzhen 518057, China)

DOI: 10.12142/ZTETJ.202504010

网络出版地址: <http://kns.cnki.net/kcms/detail/34.1228.TN.20250711.1947.004.html>

网络出版日期: 2025-07-14

收稿日期: 2025-05-10

摘要: 人工智能(AI)正以前所未有的深度和广度重塑人类社会的生产与生活方式。在AI技术层面,大模型的架构创新、研发范式变革、多模态大模型算法以及大模型训推基础设施不断突破。同时,AI工程化应用显著加速,端侧AI和具身智能等新型产品与应用取得了重要进展。DeepSeek通过在模型算法和工程优化上的系统级创新,为资源受限环境下探索通用人工智能开辟了全新路径。未来,人工智能的核心竞争力将取决于:AI算法及模型架构的持续创新,AI基础设施对算力、存储、网络等资源的高效综合调度能力,AI应用的快速迭代与场景化落地能力。以大模型技术为主线,系统分析梳理了大模型架构与研发范式、AI基础设施、大模型应用技术等领域的最新进展与未来发展方向,为人工智能研究与实践提供参考。

关键词: 大模型技术; 大模型训推基础设施; 强化学习; 思维链; 智能体

Abstract: Artificial intelligence (AI) is reshaping the production modes and lifestyles of human society with unprecedented depth and breadth. At the AI technology level, there are continuous breakthroughs in the architectural innovation of large models, changes in R&D paradigms, multimodal large model algorithms, and large model training and inference infrastructure. At the same time, the engineering application of AI has been significantly accelerated, and new products and applications such as end-side AI and embodied intelligence have made important progress. DeepSeek has opened up a new path for exploring general artificial intelligence in resource-constrained environments through system-level innovations in model algorithms and engineering optimization. In the future, the core competitiveness of artificial intelligence will depend on: the continuous innovation of AI algorithms and model architectures, the efficient and comprehensive scheduling capabilities of AI infrastructure for computing power, storage, network and other resources, and the rapid iteration and scenario-based landing capabilities of AI applications. This article takes large model technology as the main line, systematically analyzes and sorts out the latest progress and future development directions in the fields of large model architecture and R&D paradigms, AI infrastructure, and large model application technology, and provides a reference for artificial intelligence research and practice.

Keywords: large model technology; large model training and inference infrastructure; reinforcement learning; chain of thought; AI agent

引用格式: 包义明, 林阳, 屠要峰. 人工智能技术与应用前沿 [J]. 中兴通讯技术, 2025, 31(4): 67-74. DOI: 10.12142/ZTETJ.202504010

Citation: BAO Y M, LIN Y, TU Y F. Frontiers of artificial intelligence technology and applications [J]. ZTE technology journal, 2025, 31(4): 64-74. DOI: 10.12142/ZTETJ.202504010

近年来,主流大模型在技术路径上呈现出显著分化的趋势。OpenAI的GPT-4系列凭借Transformer架构的深度优化在自然语言处理任务中确立了标杆地位。Google的Gemini通过多模态融合技术实现了文本、图像和语音的统一建模。Meta的Llama3则以开源策略推动了社区生态的繁荣发展。然而,这些模型普遍依赖海量参数规模和巨额算力投入来追求性能提升,逐渐形成了“大力出奇迹”的行业共识。

2024年底,DeepSeek通过在大模型算法和工程优化方面进行系统级创新,打破了AI在扩展定律下的预期天花板,为大模型的高效训练与推理提供了有效参考,并为在受限资源下探索通用人工智能开辟了新的道路。这标志着技术竞争

焦点已转向架构创新、算法效率与场景适配的深度融合。

1 大模型技术

如表1所示,以Transformer架构为核心的人工智能技术前沿创新主要集中在3个方向:1)大模型架构创新,主要对大模型中的关键组件进行优化;2)大模型研发范式创新,以数据工程、大模型后训练、强化学习等手段,推动大模型逼近通用人工智能(AGI);3)多模态大模型,大模型的能力从自然语言建模扩展到图片、视频、音频等多种模态的建模、理解和生成,向“世界模型”迈进。通过这些创新,Transformer及其变体模型不仅在性能上实现持续突破,还在

表1 近年来大模型技术前沿创新

技术创新类型	细分类别	方法/论文
模型架构创新	Attention层优化	MHA ^[1] 、GQA、Flash Attention ^[2] 、Selective Attention ^[3] 、MLA ^[4] 、NSA ^[5]
	MLP层优化	Gshard、DeepSeek-MoE ^[6] 、HMoE ^[7]
	其他算子优化	DyT ^[8] 、Scalable Softmax、RoPE ^[9] 、YaRN ^[10] 、Self-Extend、MRoPE ^[11]
	非Transformer架构创新	Mamba、Nemotron-H ^[12] 、Hunyuan-T1 ^[13] 、RWKV
研发范式创新	强化学习与思维链	InstructGPT、PPO、CoT ^[14] 、PRM ^[15] 、OpenAI o1 ^[16] 、GRPO ^[17] 、DeepSeek-R1 ^[18] 、DAPO ^[19] 、VAPO ^[20]
多模态大模型	多模态架构创新	Flamingo ^[21] 、DeepSeek-VL2 ^[22] 、Chameleon ^[23] 、Emu3 ^[24] 、Stable Diffusion ^[25]
	多模态技术创新	ViT ^[26] 、Q-Former ^[27] 、DiT ^[28]
CoT: 思维链		MoE: 混合专家
DAPO: 解耦剪切与动态采样策略优化		MRoPE: 多模态旋转位置编码
DiT: 扩散转换器		NSA: 原生稀疏注意力
DyT: 动态双曲正切函数		PPO: 近端策略优化
GQA: 分组查询注意力		PRM: 过程监督奖励模型
GRPO: 群体相对策略优化		RoPE: 旋转位置编码
HMoE: 异构混合专家		RWKV: 响应加权键值
MHA: 多头注意力		VAPO: 基于价值的增强近端策略优化
MLA: 多头潜在注意力		ViT: 视觉转换器
MLP: 多层感知机		YaRN: 另一种RoPE扩展方法

效率、成本和适用性方面实现了显著优化，为人工智能的进一步发展奠定了坚实基础。

1.1 大模型架构创新

1) 注意力机制创新

Transformer 架构之所以成为当前人工智能革命的核心驱动力，其根本原因在于注意力机制对序列建模的范式革新。自注意力层通过动态计算 Token 间相关性实现了全局依赖捕获，而多头注意力（MHA）^[1] 则进一步提升了模型对不同维度和距离的交互模式学习能力。然而，这种强大的建模能力也带来了较高的计算复杂度 $O(N^2)$ 和显存占用问题。为了降低计算资源的消耗，同时保持模型的性能，分组查询注意力（GQA）通过对查询（Query）进行分组，使组内查询矩阵共享键值（Key, Value）矩阵，从而有效减少计算量。Flash Attention^[2] 则提出分块计算与重计算策略，将内存访问量由二次方降至线性级。选择性注意力^[3] 通过动态筛选和压缩来减少计算量。DeepSeek 提出的多头潜在注意力（MLA）^[4] 则颠覆性地将键值矩阵进行低秩压缩，在无损的前提下将内存开销降低超 50%。

2) MLP层创新

在 Transformer 架构中，由 MLP 构造的前馈网络（FFN）通常占据整个模型 50% 以上的参数量。因此，针对 MLP 层的优化和创新成为提升模型效率的关键方向，其中稀疏化技术尤为重要。例如，混合专家（MoE）门控机制通过选择性地激活部分 FFN 层的参数，在降低推理复杂度的同时，实现对输入 Token 的“分而治之”。2020 年，Google 提出的 Gshard 首次将 MoE 机制引入 Transformer 架构。然而 Gshard 中的 MoE 存在专家负载不均衡和专家能力不足的问题。为了解决这些问题，DeepSeek^[6] 提出了细分小专家的方法，并在图形处理器（GPU）设备粒度上设置 Token 路由上限，实现了细粒度的负载均衡。DeepSeek 和 HMoE^[7] 等模型还引入异构差异化专家，通过结合细分领域与通用领域知识，使 MoE 模型能够更好地满足不同任务需求。这种异构专家的设计显著提升了大模型的泛化能力和领域适应性。

3) 其余算子创新优化

除了 Attention 层与 MLP 层，Transformer 架构中其他关键算子的优化也取得了显著进展。在归一化层，何凯明等提出的动态反切归一化（DyT）^[8] 使用动态缩放的 Tanh 函数替代传统的 LayerNorm 和 RMSNorm，大幅提升了计算效率。这一创新标志着算子优化从“统计学驱动”向“几何学驱动”的转变。在位置编码方面，苏剑林等提出的旋转位置编码（RoPE）^[9] 成为相对位置编码的重要里程碑，显著提升了大模型的外推能力，支持了长序列建模的实现。Qwen2.5-VL^[11] 等将 RoPE 算子从纯文本建模推广到多模态领域，使模型在长视频理解等任务中表现卓越。

4) 非Transformer架构创新

近年来，具有线性复杂度的大模型架构不断涌现。例如，Mamba 架构基于选择性状态空间模型（Selective SSM），能够根据输入内容动态调整模型参数，对敏感内容进行重点建模，同时选择性地遗忘无效信息。这一特性使 Mamba 在处理长序列任务时表现出色。此外，Mamba-Transformer 混合架构也逐步成熟。英伟达的 Nemotron-H^[12] 和腾讯的 Hunyuan-T1^[13] 充分结合了 Mamba 处理长序列的能力与 Transformer 的局部建模能力，实现了速度和精度的高度平衡。而接收加权键值（RWKV）架构在 Transformer 中融合 RNN 的能力，通过动态状态演化大幅度降低了大模型的边缘部署成本。

1.2 强化学习与思维链

大模型的训练旨在提升人工智能对世界的理解能力。在预训练（PT）- 监督微调（SFT）- 人类反馈强化学习

(RLHF)的大模型研发范式下,基于Next Token Prediction的PT和SFT本质上是一种“行为克隆”,其所学知识来源于人类语言,因此模型的能力受限于标注或无标注的训练数据。然而,近年来互联网中可用的高质量训练数据几乎已被消耗殆尽,大模型的能力也因此遇到了瓶颈。RLHF结合近端策略优化(PPO)算法在生成式预训练转换器(GPT)系列模型中的成功应用,为大规模人类反馈学习注入了新知识,同时显著提升了大模型的安全性。然而,RLHF的训练成本很高,因为人类对齐所需要的大量奖励信号需要人工标注。此外,基于这些奖励的大模型后训练也耗资巨大。

为突破大模型训练效率瓶颈,使其通过自我学习获得超越人类的知识能力,另一种更为通用的强化学习范式——自我博弈(Self-Play)迅速发展。这种范式催生了OpenAI o1和DeepSeek R1等具有自我纠偏和链式思考能力的大模型,推动了新一波人工智能革命浪潮。在强化学习的理论与实践,智能体(Agent)通过执行动作(Action)从环境中获取奖惩信号(Reward),进而进行自我调整并更新状态(State)。在RLHF任务中,通常采用PPO算法训练大模型,这一过程需要构建4个模型,即Actor模型、Reference模型、Reward模型和Critic模型。其中,Actor和Critic模型需要实时训练与更新参数,Reference和Reward模型仅进行推理以提供参考信息。为了让大模型获得自我纠偏能力,以Google和OpenAI为代表的研究团队在过去两年研究出了一套较为成熟的强化学

习范式。这一范式以PPO算法为基础,结合思维链(CoT)^[14]与过程监督奖励模型(PRM)^[15]等技术,使得大模型在训练和推理时具备了显式思考和自我纠偏的能力。

如图1(a)所示,CoT通过将复杂问题拆解为多步骤的中间推理链,引导模型显式展示“思考过程”,从而提升解决复杂任务的准确性。例如,在数学问题中,模型需先列出公式,再计算中间结果,最后推导出答案。PRM则对思维链的每一步推理结果进行评价和验证。这种密集的奖励反馈信号,显著提升了大模型在长思维链任务中的稳定性、安全性与准确性。尽管上述方法取得了良好的性能效果,但过程监督的密集计算,以及蒙特卡洛树搜索等思维链涉及技术的搜索广度,大幅增加了强化学习的成本。为了降低强化学习算法PPO的复杂度和资源消耗,DeepSeek提出了群体相对策略优化(GRPO)^[17],通过对组内多候选输出进行归一化奖励计算当前的相对优势,从而去除了Critic模型,其训练速度相较PPO提升30%,显存占用减少50%。此外,DeepSeek R1还设计了一种基于规则的奖励策略。结合答案准确性奖励和思考过程的格式约束奖励,在经过GRPO训练后,DeepSeek R1模型具备自我反思能力,出现了“Aha Moment”训练流程,如图1(b)所示。近期,字节跳动也提出了解耦剪切与动态采样策略优化(DAPO)^[19]与基于价值的增强近端策略优化(VAPO)^[20]算法,为其大模型Seek-Thinking V1.5提供了更优的强化学习方案。其中,

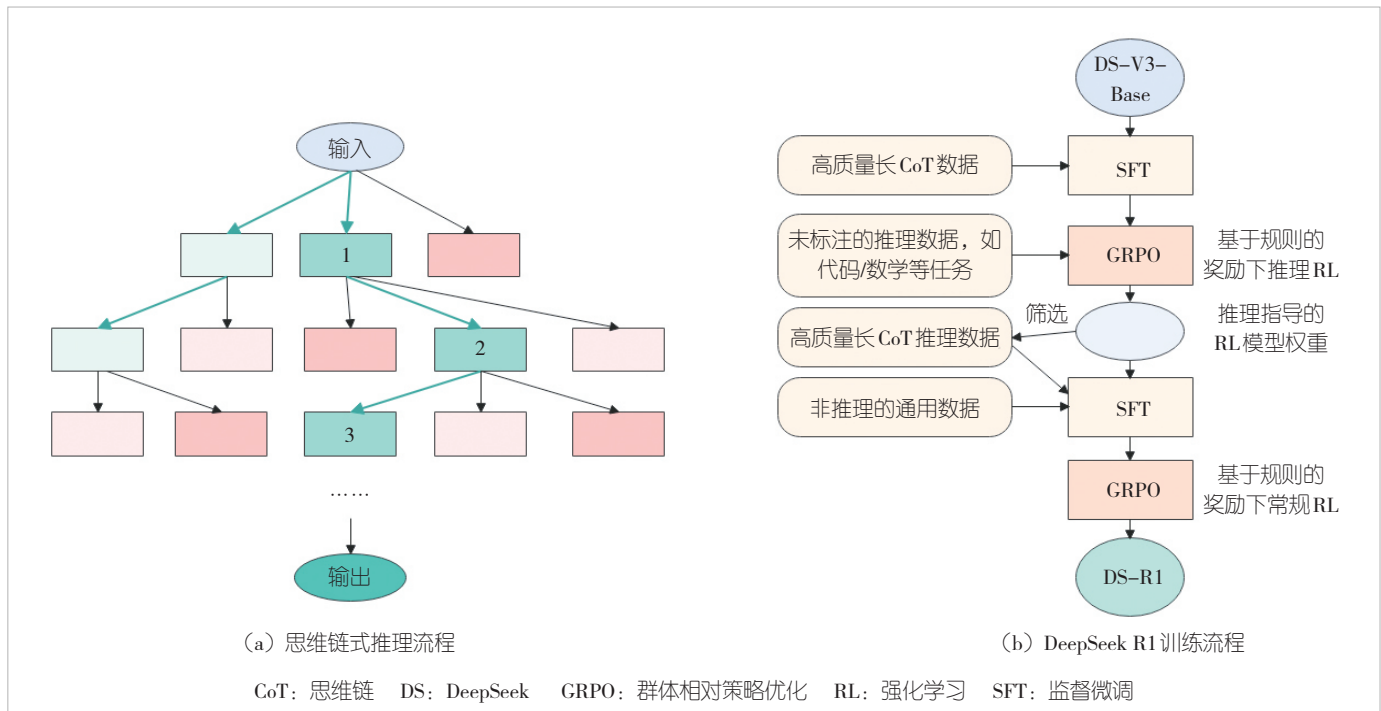


图1 CoT式推理与DeepSeek R1模型训练流程

VAPO 是基于价值方法的 SOTA 方案，而 DAPO 则是无价值方法的 SOTA 方案。

1.3 大型多模态模型

早期的多模态模型主要聚焦于特定任务，通用性较差。基于 Transformer 的视觉语言预训练模型通过对比学习图像与文本特征，需要针对下游任务进行微调。多模态大语言模型（MLLM）将非文本输入和输出融入大语言模型（LLM），赋予 LLM 多模态信息处理能力。MLLM 的典型架构通常包含 5 个组件，即模态编码器、输入映射器、LLM、输出映射器和模态生成器。Flamingo^[21]是首个在视觉语言领域将 LLM 与视觉模型结合的 MLLM。DeepSeek-VL2^[22] 利用 MoE 架构提升推理效率，由视觉编码器、视觉语言适配器和 MoE 语言模型组成。

尽管 MLLM 通过复用语言模型和其他模态模型降低了训练成本，但其跨模态能力较为受限。最近两年，统一多模态模型得到了快速发展。这类模型通过单一 Transformer 架构，统一处理多种模态数据，无需额外的对齐过程，通过预测下一个离散表示实现了跨模态的推理与生成。Chameleon^[23]基于 Transformer 模型统一处理图像、文本、代码的离散 Token。Emu3^[24]的文本、图像和视频使用同一个离散 Token 空间。

MLLM 在多模态学习中依赖模态编码、模态对齐和模态生成等多项核心技术。模态编码将不同模态的数据转换为统一特征表示，包括词嵌入模型、预训练语言模型、预训练音频模型、基于 ViT^[26]或卷积神经网络（CNN）的视觉模型等。模态对齐将不同模态的特征进行对齐，包括由线性层与非线性层交替组成的 MLP，通过查询向量从视觉模型中提取视觉特征的 Q-Former^[27]等。模态生成是指特定模态的生成输出。扩散模型是视觉生成的主流方法。Stable Diffusion^[25]是基于扩散模型的图像生成技术。Diffusion Transformer^[28]是在 Stable Diffusion 的基础上，将主干 U-Net 替换成 Transformer

的图像生成模型。

MLLM 通常采用分阶段训练策略，模态对齐负责弥合不同模态之间的差距；指令微调使模型理解和遵循多模态场景下的指令；偏好学习将模型的输出与人类的偏好对齐。多模态 Prompt 工程包括多模态 CoT 和多模态检索增强生成（RAG）^[29]。其中，多模态 CoT 通过在输入中添加显式的思维链提示，引导模型逐步推理；多模态 RAG 将检索到的多种模态信息融入上下文，帮助模型全面理解任务背景并生成更优质的输出。

随着技术的快速迭代，MLLM 展现出强大的跨模态处理能力，能够高效完成复杂的多模态任务。这些模型在图像生成、视频理解、音频分析等领域推动了下游任务性能的边界。

2 大模型训推基础设施

大模型训推基础设施通过“算力解耦”与“资源池化”破解了大模型成本困局。近期，DeepSeek 创新众多 AI Infra 技术，大幅降低了上千亿参数模型的训练成本。未来 10 年，人工智能企业的核心竞争力将取决于 AI Infra 对算力、存储、网络的综合调度能力。

2.1 大模型高效率训练

随着参数规模的迅速飙升，大模型训练过程面临的性能瓶颈愈发显著，主要体现在内存墙、通信墙和存储墙 3 个方面。其中，内存墙尤为突出，主要受限于 GPU 显存容量。为突破上述性能瓶颈，表 2 总结了相关优化技术与解决思路。

1) 大模型训练显存优化

大模型训练优化技术的发展与硬件（如显卡）的迭代相辅相成，其中分布式并行计算是解决 GPU 显存溢出问题的核心方法。Pytorch 框架中的数据并行将一个批次中的样本分发到不同 GPU 上同时进行计算。Megatron-LM 提出张量并行技术，将模型中间张量切分到多张 GPU。流水线并行技术

表2 大模型训练效率优化技术

训练效率优化	优化技术或思路	方法/论文
显存优化	分布式并行训练	数据并行、张量并行、流水线并行 ^[30-31] 、序列并行 ^[32-33] 、专家并行 ^[34]
	ZeRO 技术	DeepSpeed ZeRO、CPU Offload
	其他显存优化技术	自动并行搜索 ^[35] 、混合精度训练 ^[36] 、梯度累积
通信优化	节点间 GPU 通信优化	InfiniBand、RoCE
	节点内 GPU 通信优化	PCIe、NVLink
	计算-通信重叠	Centauro ^[37] 、DualPipe ^[38]
存储优化	高效文件 I/O	DeepSeek-3FS ^[39]
	异步检查点 I/O	DLRover ^[40]

CPU: 中央处理器 GPU: 图形处理器 I/O: 输入/输出 PCIe: 高速串行总线接口 ZeRO: 零冗余

如 Gpipe^[30]、ZeroBubble^[31]和 DeepSeek DualPipe^[38]等,将模型的不同层级按流水线方式分布到多 GPU 中执行。DeepSpeed 推出的零冗余优化 (ZeRO),将模型参数、优化器状态及张量梯度切分到多个 GPU 中存储。近年来,随着大模型训练的序列长度不断增长,序列并行成为大模型增强长下文建模能力的重要技术,以 DeepSpeed Ulysses^[32]和 Ring Attention^[33]两种序列并行 (SP) 技术的应用为基础,全球大模型得以支持 128k 甚至百万 Token 长度的序列建模。随着稀疏专家 (MoE) 架构模型的发展,专家并行技术被用于提升 MoE 模型的训练效率。为避免对上述多种并行方式的参数进行低效的人工调优,中兴通讯与北京大学联合研发了 Galvatron^[35]等自动并行优化技术,能够通过动态规划算法自动搜索最优并行参数。此外,面向显存优化的其他手段还包括 FP8 低精度训练^[36]、梯度累积和参数量化等。

2) 大模型训练通信优化

大模型分布式训练的效率通过 GPU 的“吞吐量 (Throughput)”来衡量,而 GPU 之间的通信效率对训练效率起着决定性作用。在训练过程中,数据、模型参数以及中间结果需在不同 GPU 之间同步传递,通信延迟直接影响计算与通信的协同效率,从而可能导致资源闲置。为此,通信优化主要从两个层面展开:在硬件层面,多种高带宽组网协议推动着 GPU 集群通信效率的不断提升,例如以 InfiniBand 和基于融合以太网的远程直接内存访问 (RoCE) 为代表的节点间 GPU 通信协议,和以 PCIe 和 NVLink 为代表的节点内 GPU 通信协议;在软件框架层面,计算-通信重叠技术通过细粒度的算子拆分与编排来最大化通信效率,减少因数据传输导致的计算停滞时间。

3) 大模型训练存储优化

在大模型训练中,存储系统扮演着关键角色:模型训练需要频繁地从存储中加载数据,同时需要实时将模型权重

(检查点) 保存到存储中,以避免因训练出错导致的算力浪费和研发进度延误。因此,高效的文件输入/输出 (I/O) 系统对于大模型训练至关重要。DeepSeek 设计的 3FS^[39]系统,采用内存、固态硬盘 (SSD)、分布式集群三级缓存结构,大幅提升了数据读取吞吐量,为大模型训练提供了高效的数据加载能力。为提升检查点读写速度,异步检查点技术如 DLRouter^[40]等通过先将模型权重保存至内存,再在后续训练过程中异步写入磁盘,显著减少训练与存储操作的冲突,提升了存储和训练效率。

混合精度训练和量化是提升大模型训练效率的关键手段。混合精度训练通过结合单精度 (FP32) 和半精度 (FP16) 格式,在训练过程中动态调整数值精度,既减少了显存占用,又加快了计算速度。量化技术则进一步将模型参数和激活值从浮点数压缩为低位宽整数,有效降低模型大小和计算复杂度。

综上所述,通过显存、通信与存储优化技术的协同发展,大模型训练效率得以显著提升,同时推动了硬件与软件技术的进一步演化。这些优化手段为应对大模型训练中的性能瓶颈提供了系统性解决方案。

2.2 大模型高效推理

随着 Agent、RAG 等 AI 应用的快速发展,大模型的高效部署与推理已成为推动 AI 应用落地的关键需求。本节对比大模型高效推理技术,从计算效率、资源利用率等方面分析优化技术及其效果,详细总结见表 3。

系统级优化通过传统网络和存储技术的改进优化推理流程。预计算与编码 (PD) 分离架构通过高性能网络将 Prefill 阶段和 Decode 阶段分开部署到异构集群,从而实现负载分离,显著提升推理吞吐性能。键值缓存 (KV) Cache 扩展利用存储设备实现对 KV Cache 的扩展,有效缓解内存压力,

表 3 大模型推理优化技术

优化类型	技术	原理	效果	应用
系统优化	PD 分离	推理的 Prefill/Decode 阶段分开部署	异构硬件部署,实现更高资源利用率	Kimi、DeepSeek
	多级 KV 缓存	显存、RAM、SSD、存储构成多级 KV Cache 存储	扩展 KV 多级缓存空间,支持更高吞吐率	Kimi、DeepSeek
模型结构	投机采样	用较小模型推理草稿,用大模型验证结果	有效降低大模型解码的次数	Medusa、GPT-4
	MTP	预训练 n 个独立 Decode 网络并行预测 n 个 Token	缩短训练周期,提高推理效率	DeepSeek
	MLA	预训练低秩矩阵,实现 KV 的压缩存储	压缩历史 KV 的存储空间,提高端到端推理性能	DeepSeek
注意力稀疏化	MoBA	对输入序列进行分块,最关键的块参与技术	以较少的存算资源实现接近原生注意力的精度	Kimi
	NSA	融合 3 种策略,选择最关键的 KV 参与注意力计算	极低的存算开销实现优于原生注意力的精度	DeepSeek
长思维链	隐藏空间推理	通过高维向量数学运算完成推理,将结果解码为可读形式	全新思维链范式	Coconut
KV: 键值		MoBA: 混合块注意力	NSA: 原生稀疏注意力	RAM: 随机存取存储器
MLA: 多头潜在注意力		MTP: 多 Token 预测	PD: 预计算与编码	SSD: 固态硬盘

同时进一步优化推理性能。

在模型结构方面,投机采样是一种并行推理技术,通过提前预测多个后续Token来提升推理效率。Meta和DeepSeek提出的多Token预测(MTP)^[41]技术利用多个独立的Decoder输出网络并行预测多个后续Token,从而提升推理速度。DeepSeek的MLA^[4]通过改进Transformer架构,有效压缩KV张量的存储空间。

注意力稀疏化技术旨在降低注意力机制的计算复杂度,通过捕捉推理过程中的稀疏性,仅对关键Token执行注意力计算,就能减少存算开销。Kimi提出的混合块注意力(MoBA)^[42]算法,通过将注意力划分为若干块级区域,显著减少计算复杂度。DeepSeek提出的原生稀疏注意力(NSA)^[5]策略,动态选择关键Token进行注意力计算,进一步优化推理性能。

长思维链推理优化关注在保证推理质量的同时减少资源开销。传统方法通常通过直接压缩推理步骤来降低计算成本,但这可能会影响推理效果。隐藏空间推理则通过潜在空间中的数学运算替代自然推理过程^[43],避免冗长的推理路径,成为长思维链优化的新范式。

综上所述,大模型高效推理需要从系统架构优化、模型轻量化、稀疏化技术以及推理范式创新等多方面入手,综合平衡推理性能、资源利用率与推理的可解释性,才能全面满足AI应用对高效推理的需求。

3 大模型应用技术

3.1 数据与大模型

数据是大模型能力的根基与驱动力^[44]。在预训练阶段,海量高质量语料使模型得以掌握通用知识、语言模式和复杂语义结构;在推理阶段,结合检索增强生成(RAG)技术,动态获取外部知识为大模型提供实时、可信的信息补充,有效提升生成结果的准确性与时效性。

1) 语料工程

大模型的预训练语料主要分为两类:一是由人类创作的原生数据,二是由模型生成的合成数据。随着模型参数规模的指数级扩展,原生高质量数据的增长速度远无法满足大模型训练的需求,合成数据正逐步成为下一代语料构建的关键。

一种主流的合成数据方式是结合知识蒸馏,从强大的教师模型中蒸馏知识用于训练轻量模型,实现性能迁移。在实际工程中,通常将合成数据生成与知识蒸馏联合应用:先利用大模型生成高质量训练样本,突破原生数据在规模和质量上的瓶颈;再通过蒸馏技术提升小模型性能,实现知识的高

效迁移。例如,DeepSeek R1^[18]通过动态温度参数调节,使3B模型性能逼近70B模型。OpenAI基于o1^[16]生成合成语料训练Orion,并通过蒸馏实现对边缘设备的适配,形成从语料生成到端侧部署的闭环优化。

2) RAG

RAG系统通常包含4个阶段:预检索(Pre-Retrieval)、检索(Retrieval)、后处理(Post-Retrieval)和生成(Generation)。Pre-Retrieval阶段将检索源的数据组织为结构化知识库,支持文本、图像、表格等多模态数据。该阶段涉及数据清洗、去噪、切分(Chunking)与知识结构化处理。Retrieval阶段可采用多种检索方式。关键字检索通过倒排索引定位相关文档,利用BM25等算法计算文档得分。稠密向量检索通过查询向量和文档向量之间的距离衡量两者的语义相似性。学习型稀疏向量检索通过模型生成高维学习型稀疏向量,结合了关键词检索和稠密向量检索的优势。当前主流的检索方式为混合检索模式,即融合多种检索机制以提升准确率。Post-Retrieval对检索结果进行重排序、信息压缩(如摘要融合)等处理,以提升上下文信息密度,减少无效片段对生成质量的影响。在Generation阶段,大模型在融合的上下文信息基础上进行内容生成,提供更具事实性与语义丰富度的输出结果。

在实际应用中,RAG已广泛应用于企业知识问答、文档总结、垂直搜索等场景。企业数据通常呈现为文本、图像、表格、图表等多种模态,面对图文混排、多模态融合等复杂任务,多模态RAG成为突破关键。然而,多模态RAG仍面临多方面的挑战,包括异构数据的预处理与标准化、不同模态信息的协同存储与高效整合、跨模态语义融合与对齐、多模态混合检索结果的排序优化与语义一致性控制等。

3) 向量数据库

向量是大模型处理文本、图像等多模态数据的关键,而向量数据库则是专为高效存储与检索高维向量数据(如文本、图像、音视频的嵌入表示)而设计,是AI时代的关键基础设施。向量数据库的核心价值在于:1)支撑多模态数据处理。向量数据库是实现跨模态语义对齐的关键组件,可通过语义匹配与知识融合高效处理多模态数据。2)RAG技术协同。向量数据库通过外置动态知识来降低模型幻觉,提升响应的实时性,同时避免模型参数膨胀与重复训练;3)私域知识管理与库内训推。向量数据库管理私有化知识,可以基于数据库技术框架实现领域大模型的训练和推理,在降低部署复杂度的同时确保数据安全与隐私。

目前主流向量数据库包括Weaviate、Milvus、Faiss、Chroma、Qdrant以及EBASE AI,其关键能力对比详见表4。

表4 主流向量数据库关键特性对比

特性	内置文本嵌入	索引类型	文本搜索	混合检索	元数据过滤	多租户	最大向量维度
Weaviate	√	HNSW	√	√	√	√	65 535
Milvus	×	FLAT、HNSW 等	×	√	√	×	32 768
Faiss	×	FLAT、HNSW 等	×	×	×	×	无限制
Chroma	×	HNSW	×	√	√	×	无限制
Qdrant	×	HNSW	×	√	√	√	无限制
EBASE AI	√	IVFELAT、HNSW 等	√	√	√	√	32 768

AI:人工智能 FLAT:线性扫描索引 HNSW:分层可导航小世界 IVFELAT:倒排文件索引和线性扫描索引

EBASE AI是中兴通讯面向大模型时代自研的AI数据库系统，融合多模数据存储与一站式RAG能力。目前，EBASE AI已广泛应用于中兴通讯大模型系统，在产品问答、代码生成、文档分析等场景中实现快速落地，展示出卓越的能力。

3.2 端侧AI

近年来，在终端部署大模型服务于个人用户，已成为实现“个人智能”的主要路径，即端侧AI。中兴通讯端侧AI的统一技术架构如图2所示，通过在基础软硬件层、技术服务层、业务与应用层进行贯通及创新，推动中兴通讯在端侧

AI领域的技术突破与产品落地，助力行业引领。

1) 基础软硬件层

智能终端作为端侧AI的核心载体，其硬件架构与算力水平正迎来系统性升级。以高通骁龙8 Gen3、联发科天玑9400为代表的神经网络处理器（NPU）采用异构计算架构，大幅提升了大模型推理速度，同时将功耗降低35%。在硬件升级的同时，智能终端的软件生态也呈现深度协同的趋势。原生操作系统集成内核级AI能力，并通过意图框架解析用户指令，触发跨应用服务链。交互技术的创新尤为显著，基于苹果Ferret-UI和谷歌ScreenAI等面向端侧UI的视觉语言模型，支持屏幕元素的语义理解与自动化操作，使终端具备“以意图为中心”的类“端侧自动驾驶”能力，显著提升了智能交互体验。

2) 技术服务层

在个人智能的愿景下，端侧AI服务需要更加贴合用户需求，兼具隐私性、情感化和个性化等多重特性。通过在端侧构建用户画像，并将用户画像数据存储到向量数据库，再结合RAG技术向通用大模型提供支持，是端侧AI优化的创新路径。

端侧向量数据库通过存储-计算-索引的协同设计，能够在端侧AI资源受限场景中实现高效检索与个性化服务。EBASE AI数据库端侧版本，通过对物联网（IoT）设备、智能手机和嵌入式系统等多样化硬件平台的无缝适配，已在资源受限的边缘与终端环境中展现出卓越能力。云边端协同通过异构资源的动态编排和调度，实现了计算负载与隐私安全的全局优化。边缘节点负责高实时性任务，云端专注于模型训练、更新与复杂任务推理，端侧聚焦个性化微调和轻量化的实时执行。

3) 业务与应用层

端侧AI正加速向多形态终端渗透，形成了“核心智能终端+泛边缘设备”的立体化布局。在应用模式上，多模态融合成为端侧AI技术演进的主要路径。在业务场景上，个性化服务能力则成为技术竞争高地。例如，中兴通讯星云助

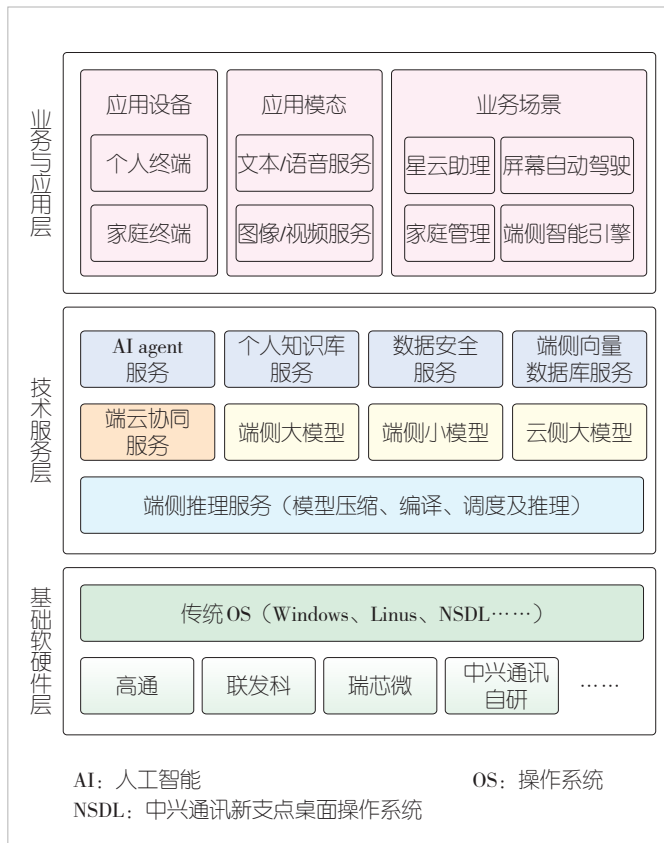


图2 中兴通讯端侧AI统一技术架构图

理中兴通讯智能终端提供端侧自动驾驶等创新能力。这些基于端侧智能引擎的能力将进一步推动端侧AI的产业变革，加速多场景业务需求的落地，助力构建全面智能化时代。

3.3 智能体

大模型技术的飞跃推动了智能体（AI Agent）的快速发展^[45-47]，逐步形成“模型突破→Agent应用涌现→场景扩展”的演进路径。各大厂商纷纷加码智能体技术，推出了一系列具有颠覆性的应用实践。ChatGPT催生了如Copilot这类AI增强助手，能够辅助人类完成写作、编程等任务，但尚不具备自主任务规划能力。而GPT-4强大的推理与规划能力则开启了“自主智能体”时代，衍生出AutoGPT等开源项目，以及全球首个全自动AI程序员Devin。随着OpenAI的o1和DeepSeek R1等具备复杂推理能力的大模型问世，Agent应用逐步具备了“思考+执行+验证”的完整任务闭环能力，为真正意义上的智能代理系统奠定了技术基础。

在应用场景复杂性日益提升的背景下，单一Agent已难以胜任多步推理与并发任务执行。因此，多智能体系统（MAS）应运而生。MAS通常由一个主控Agent协调多个专职Agent协作完成复杂任务。中国发布的Manus代表了智能体系统的最新发展，采用分工协作机制，实现了规划、执行、验证的闭环，突破了单体智能体难以覆盖的应用瓶颈。由于多Agent与Tools工具之间存在大量异步通信与上下文交互，统一的通信协议变得尤为关键。模块上下文协议（MCP）^[48]已成为当前Agent系统中的主流通信标准，显著降低了系统开发与部署复杂度。近期，Google基于MCP推出扩展协议智能体到智能体（A2A）^[49]，支持跨平台、跨厂商的智能体间安全协作，成为智能体社区的研究热点。

随着Agent系统日益智能化，其应用边界不断拓展，具身智能成为新一轮研究焦点。通过将智能体“具身化”，即赋予其感知现实世界并执行任务的能力，Agent技术正从虚拟场景走向真实环境，拓展出更广阔的应用空间。

3.4 具身智能

具身智能是指具备环境感知能力、目标规划能力以及实际执行能力的智能实体，通常表现为具有人类感知与行为能力的机器人系统^[50]。随着宇树科技在四足机器人方面的突破，以及自动驾驶技术的商业化落地，具身智能逐步从实验室走向产业。具身智能被视为迈向AGI的关键路径之一，被英伟达的黄仁勋称为“AI的下一个浪潮”。

大模型的发展为具身智能带来多项核心能力突破，包括自然语言指令理解、视觉感知融合、多步动作规划与反馈式

执行能力。传统虚拟Agent已开始“具身化”，通过集成硬件传感器和控制接口，将模型能力延伸至物理世界。当前主流的具身智能大模型有两类：一是以Google为代表的端到端具身大模型，二是以OpenAI为代表的分层决策模型。前者采用一个视觉-语言-动作（VLA）大模型完成感知、规划、执行全流程，如OpenVLA、TR-X模型基于Transformer架构统一处理视频、指令、动作序列^[51]；后者通过分层模型处理不同任务，上层“大脑”基于预训练大模型负责感知、决策，中间“小脑”基于RL小模型实现动作规划，底层硬件“本体”负责精准执行^[52]。智源研究院发布的大小脑协同系统RoboOS创新性地实现了“云端智能决策+终端实时执行”的协同框架，标志着具身智能正从单体智能向群体智能演进。

在产业应用方面，自动驾驶是目前具身智能落地程度最高的领域。百度“萝卜快跑”采用大模型分层控制架构，其安全性被证实优于人类驾驶员；而特斯拉的完全自动驾驶（FSD）系统则依赖端到端模型，直接将传感器输入映射为控制信号，实现驾驶自动化。

具身智能目前仍面临诸多挑战：高质量训练数据严重匮乏、仿真替代方案的保真度与通用性不足、开发工具与评估体系不成熟、真实部署复杂性高等。具身智能的发展亟需构建开放式生态体系，完善训练平台、通信协议和评估框架，以推动从“实验智能”走向“实用智能”的转变。

3.5 中兴通讯大模型及智算研发实践

中兴通讯在大模型技术创新方面持续深耕，勇于探索关键前沿方向，取得了显著成果。据SuperCLUE最新发布的《中文大模型基准测评2025年5月报告》显示，中兴通讯星云大模型Nebula Coder-V6在推理专项榜单中表现卓越，总分并列第一，同时在综合总榜中斩获银牌，展现出强大的模型能力与工程落地实力。围绕多模态大模型的产业化应用，中兴通讯已在工业制造、智慧教育、医疗健康等多个垂直场景中开展落地探索，未来将继续在强化学习、思维链推理、多模态融合和智算加速等核心技术方向加大投入，推动模型能力持续演进。

在算力基础设施方面，中兴通讯构建了完善的智算方案和AI平台，提供面向大模型研发的全栈AI Infra能力（如图3所示）。平台实现了数据、算力与算法的统一调度与集中管理，覆盖从数据采集、模型预训练、精调、强化学习到推理部署的完整开发流水线，显著提升了大模型研发的效率与稳定性。

在AI数据库与检索增强生成技术方面，中兴通讯基于

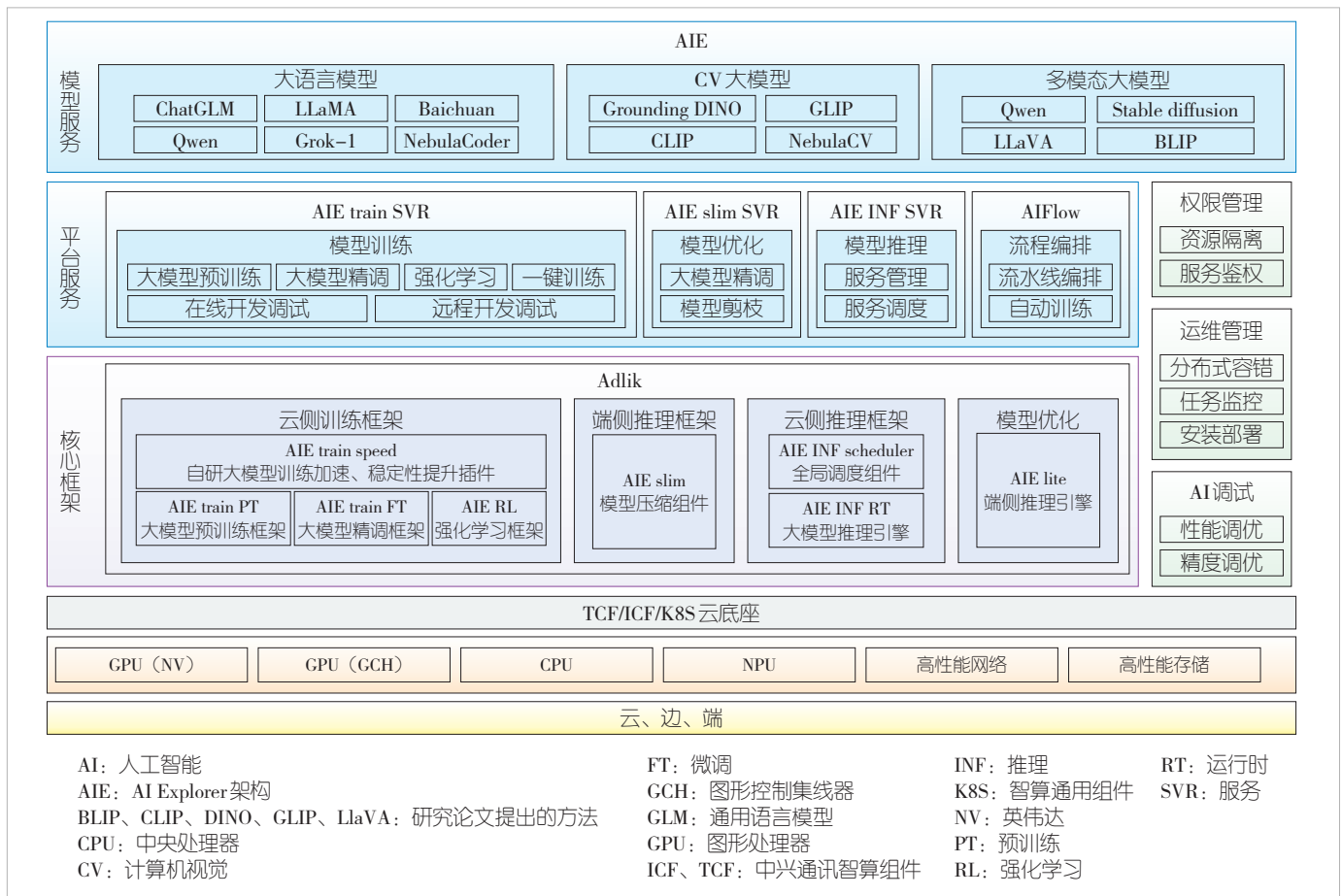


图3 中兴通讯的AI平台架构图

开源项目 Dify^[53] 和 RAGFlow^[54] 自主开发了高效检索系统 NebulaRAG，如图4所示。该系统集成了多类开源生成式大模型、向量化嵌入模型与 Rerank 精排模型，实现语义检索与智能问答任务的高效闭环，构建面向行业知识库的多源智能交互平台。

在智能体应用方面，中兴通讯开源的 Co-Sight 超级智能体采用多智能体 (Multi-Agent) 架构，构建“数字团队”式协同体系。系统通过主管智能体统一调度各执行智能体，在有向无环图 (DAG) 机制支持下，实现并发任务优化与自适应容错，具备动态重规划与跨平台自动化执行能力，通过智能反思模型机制实现能力持续进化。在 GAIA 基准测评中，Co-Sight 以 72.72 分位居开源框架榜首，已应用于行业研究、舆情分析、旅游规划等场景，大幅提升任务自动化水平与入效。

总体来看，中兴通讯在大模型算法架构、智算 AI 平台、智能检索系统与智能体应用等关键技术路径上持续发力，不仅实现了技术创新，更在多行业完成高质量落地。其在推动国产大模型生态建设、加速人工智能产业发展中，正发挥着

日益重要的引领作用。

4 总结与展望

当前，大模型技术正沿着多元融合的路径演进，核心突破集中于算子优化、异构架构协同、强化学习，以及思维链驱动下的模型自学习能力增强，并朝着“理解-生成”一体化的多模态大模型架构不断演进。伴随着微调与后训练范式的日益普及，VeRL、Huggingface Transformers 等开发框架持续演进，为大模型训练、部署与优化提供了更加灵活高效的支持体系。同时，智算服务器与异构算力资源的协同调度能力持续增强，为人工智能研发提供坚实的底层算力保障。在生态方面，以 DeepSeek 为代表的开源技术浪潮，推动了模型能力向更大范围普惠开放演进，促使 AI 技术生态从“闭源竞跑”转向“开源协同”。然而，开源体系下也暴露出潜在的滥用风险与安全漏洞，技术共享与主权控制之间的博弈加剧，亟需构建全球协同治理框架与标准体系，实现“安全可控”与“开放创新”的双重目标。

总体而言，大模型在复杂任务泛化能力、跨模态理解与

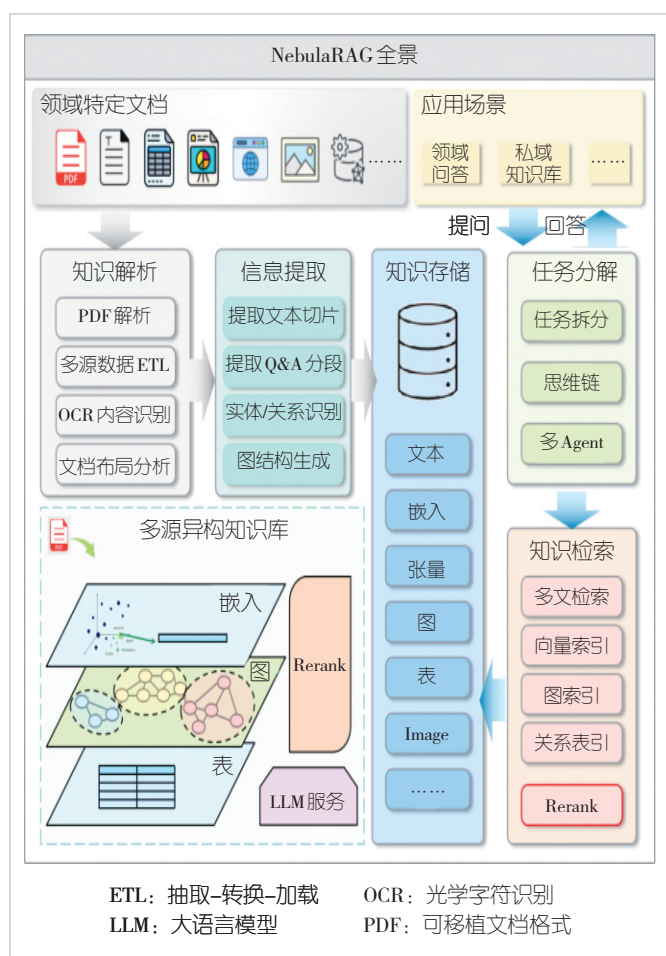


图4 NebulaRAG全景图

生成、推理路径稳定性、系统级集成能力等方面仍存在诸多挑战。硬件兼容性、部署效率与模型规模之间尚未实现理想平衡，多模型协同、Agent系统部署与安全治理体系尚待完善。

未来，人工智能的发展需在技术创新与风险防控之间实现动态平衡，在端侧智能化、具身智能、Agent协同等方向实现更深层次突破。跨学科协同与开源生态建设将成为AI下一阶段发展的主动力，助推其在通信网络、智能制造、医疗健康、教育服务等关键行业的深度融合与规模落地，真正实现从能力突破到价值创造的跃迁。

参考文献

- [1] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [EB/OL]. [2025-04-05]. <https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>
- [2] DAO T. Flashattention-2: faster attention with better parallelism and work partitioning [EB/OL]. (2023-07-17)[2025-04-05]. <https://arxiv.org/abs/2307.08691>
- [3] OHNSTON W A, DARK V J. Selective attention [J]. Annual review

- of psychology, 1986, 37: 43-75
- [4] LIU A, FENG B, WANG B, et al. DeepSeek-V2: a strong, economical, and efficient mixture-of-experts language model [EB/OL]. (2024-05-07)[2025-04-05]. <https://arxiv.org/abs/2405.04434>
- [5] YUAN J, GAO H, DAI D, et al. Native sparse attention: hardware-aligned and natively trainable sparse attention [EB/OL]. (2024-05-07)[2025-04-05]. <https://arxiv.org/abs/2502.11089>
- [6] DAI D, DENG C, ZHAO C, et al. DeepSeekMoE: towards ultimate expert specialization in mixture-of-experts language models [EB/OL]. (2024-01-11)[2025-04-05]. <https://arxiv.org/abs/2401.06066>
- [7] WANG A, SUN X, XIE R, et al. HMoE: heterogeneous mixture of experts for language modeling [EB/OL]. (2024-08-20)[2025-04-05]. <https://arxiv.org/abs/2408.10681>
- [8] ZHU J, CHEN X, HE K, et al. Transformers without normalization [EB/OL]. (2025-03-13) [2025-04-05]. <https://arxiv.org/abs/2503.10622>
- [9] SU J, AHMED M, LU Y, et al. Roformer: enhanced transformer with rotary position embedding [J]. Neurocomputing, 2024, 568: 127063
- [10] PENG B, QUESNELLE J, FAN H, et al. YaRN: efficient context window extension of large language models [EB/OL]. (2023-09-15)[2025-04-05]. <https://arxiv.org/abs/2309.00071>
- [11] BAI S, CHEN K, LIU X, et al. Qwen2.5-VL technical report [EB/OL]. (2025-02-19)[2025-04-05]. <https://arxiv.org/abs/2502.13923>
- [12] BLAKEMAN A, BASANT A, KHATTAR A, et al. Nemotron-H: a family of accurate and efficient hybrid mamba-transformer models [EB/OL]. (2025-04-04) [2025-04-05]. <https://arxiv.org/abs/2504.03624>
- [13] GitHub. LLM Hunyuan [EB/OL]. [2025-04-05]. <https://github.com/Tencent/llm.hunyuan.T1>
- [14] WEI J, WANG X, SCHUURMANS D, et al. Chain-of-thought prompting elicits reasoning in large language models [J]. Advances in neural information processing systems, 2022, 35: 24824-24837
- [15] LI W, LI Y. Process reward model with q-value rankings [EB/OL]. (2024-10-15)[2025-04-05]. <https://arxiv.org/abs/2410.11287>
- [16] JAECH A, KALAI A, LERER A, et al. Openai o1 system card [EB/OL]. (2024-12-21) [2025-04-05]. <https://arxiv.org/abs/2412.16720>
- [17] SHAO Z, WANG P, ZHU Q, et al. Deepseekmath: pushing the limits of mathematical reasoning in open language models [EB/OL]. (2024-02-05) [2025-04-05]. <https://arxiv.org/abs/2402.03300>
- [18] GUO D, YANG D, ZHANG H, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning [EB/OL]. (2024-01-11)[2025-04-05]. <https://arxiv.org/abs/2501.12948>
- [19] YU Q, ZHANG Z, ZHU R, et al. DAPO: an open-source LLM reinforcement learning system at scale [EB/OL]. (2024-01-11) [2025-04-05]. <https://arxiv.org/abs/2503.14476>
- [20] YUAN Y, YU Q, ZUO X, et al. VAPO: efficient and reliable reinforcement learning for advanced reasoning tasks [EB/OL]. (2025-04-16)[2025-04-25]. <https://arxiv.org/abs/2504.05118>
- [21] ALAYRAC J B, DONAHUE J, LUC P, et al. Flamingo: a visual language model for few-shot learning [J]. Advances in neural information processing systems, 2022, 35: 23716-23736
- [22] WU Z, CHEN X, PAN Z, et al. DeepSeek-v2: mixture-of-experts vision-language models for advanced multimodal understanding [EB/OL]. (2024-12-05) [2025-04-05]. <https://arxiv.org/abs/2412.10302>
- [23] Chameleon Team. Chameleon: mixed-modal early-fusion foundation models [EB/OL]. (2024-01-11)[2025-04-05]. <https://arxiv.org/abs/2405.09818>
- [24] WANG X, ZHANG X, LUO Z, et al. Emu3: next-token prediction is all you need [EB/OL]. (2024-09-23) [2025-04-05]. <https://arxiv.org/abs/2409.23100>

- org/abs/2409.18869
- [25] ROMBACH R, BLATTMANN A, LORENZ D, et al. High-resolution image synthesis with latent diffusion models [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. IEEE, 2022: 10684–10695
- [26] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: transformers for image recognition at scale [EB/OL]. (2020-10-22) [2025-04-05]. <https://arxiv.org/abs/2010.11929>
- [27] LI J, LI D, SAVARESE S, et al. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models [C]//International conference on machine learning. PMLR, 2023: 19730–19742
- [28] PEEBLES W, XIE S. Scalable diffusion models with transformers [C]//Proceedings of the IEEE/CVF international conference on computer vision. IEEE, 2023: 4195–4205
- [29] WANG X, WANG Z, GAO X, et al. Searching for best practices in retrieval-augmented generation [EB/OL]. (2024-07-11) [2025-04-05]. <https://arxiv.org/abs/2407.01219>
- [30] HUANG Y, CHENG Y, BAPNA A, et al. Gpipe: efficient training of giant neural networks using pipeline parallelism [EB/OL]. (2018-11-16)[2025-04-05]. <https://arxiv.org/abs/1811.06965>
- [31] QI P, WAN X, HUANG G, et al. Zero bubble pipeline parallelism [EB/OL]. (2024-01-11) [2025-04-05]. <https://arxiv.org/abs/2401.10241>
- [32] JACOBS S A, TANAKA M, ZHANG C, et al. DeepSpeed ullysses: system optimizations for enabling training of extreme long sequence transformer models [EB/OL]. (2023-09-15)[2025-04-05]. <https://arxiv.org/abs/2309.14509>
- [33] LIU H, ZAHARIA M, ABBEEL P. Ring attention with blockwise transformers for near-infinite context [EB/OL]. (2023-10-11) [2025-04-05]. <https://arxiv.org/abs/2310.01889>
- [34] GALE T, NARAYANAN D, YOUNG C, et al. Megablocks: efficient sparse training with mixture-of-experts [J]. Proceedings of machine learning and systems, 2023, 5: 288–304
- [35] MIAO X, WANG Y, JIANG Y, et al. Galvatron: efficient transformer training over multiple gpus using automatic parallelism [EB/OL]. (2022-11-11)[2025-04-05]. <https://arxiv.org/abs/2211.13878>
- [36] MICIKEVICIUS P, STOSIC D, BURGESS N, et al. FP8 formats for deep learning [EB/OL]. (2024-01-11)[2025-04-05]. <https://arxiv.org/abs/2209.05433>
- [37] CHEN C, LI X, ZHU Q, et al. Centauri: enabling efficient scheduling for communication-computation overlap in large model training via communication partitioning [C]//Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems. ACM, 2024: 178–191
- [38] LIU A, FENG B, XUE B, et al. DeepSeek-V3 technical report [EB/OL]. (2024-01-11) [2025-04-05]. <https://arxiv.org/abs/2412.19437>
- [39] GitHub. DeepSeek AI: 3FS [EB/OL]. [2025-04-05]. <https://github.com/deepseek-ai/3FS>
- [40] WANG Q, SANG B, ZHANG H, et al. DLRouter: an elastic deep training extension with auto job resource recommendation [EB/OL]. (2023-04-04) [2025-04-05]. <https://arxiv.org/abs/2304.01468>
- [41] GLOECKLE F, IDRISI B Y, ROZIERE B, et al. Better & faster large language models via multi-token prediction [EB/OL]. (2024-04-30)[2025-04-05]. <https://arxiv.org/abs/2404.19737>
- [42] LU E, JIANG Z, LIU J, et al. MoBA: mixture of block attention for long-context LLMs [EB/OL]. (2024-01-11)[2025-04-05]. <https://arxiv.org/abs/2502.13189>
- [43] CHEN Q, QIN L, LIU J, et al. Towards reasoning era: a survey of long chain-of-thought for reasoning large language models [EB/OL]. (2024-01-11) [2025-04-05]. <https://arxiv.org/abs/2503.09567>
- [44] 韩银俊, 牛家浩, 屠要峰. 数据管理系统发展趋势与挑战 [J]. 中兴通讯技术, 2023, 27(2): 64–71. DOI: 10.12142/ZTETJ.202304012
- [45] 田海东, 张明政, 常锐, 等. 大模型训练技术综述 [J]. 中兴通讯技术, 2024, 30(2): 21–28. DOI: 10.12142/ZTETJ.202402004
- [46] 韩炳涛, 刘涛. 大模型关键技术与应用 [J]. 中兴通讯技术, 2024, 30(2): 76–88. DOI: 10.12142/ZTETJ.202402012
- [47] FENG B Y, FENG M X, WANG M R, et al. Multi-agent hierarchical graph attention reinforcement learning for grid-aware energy management [J]. ZTE Communications, 2023, 21(3): 11–21. DOI: 10.12142/ZTECOM.202303003
- [48] GitHub. Model context protocol [EB/OL]. [2025-04-05]. <https://github.com/modelcontextprotocol>
- [49] GitHub. Agent2Agent (A2A) protocol [EB/OL]. [2025-04-05]. <https://github.com/google/A2A>
- [50] 邵宏, 谢大雄. 具身智能机器人技术 [J]. 中兴通讯技术, 2024, 30(S1): 40–44. DOI: 10.12142/ZTETJ.2024S1006
- [51] KIM M J, PERTSCH K, KARAMCHETI S, et al. Openvla: an open-source vision-language-action model [EB/OL]. (2024-01-11) [2025-04-05]. <https://arxiv.org/abs/2406.09246>
- [52] YANG Z, LI L, LIN K, et al. The dawn of LMMs: preliminary explorations with GPT-4V (ision) [EB/OL]. (2023-09-29) [2025-04-05]. <https://arxiv.org/abs/2309.17421>
- [53] Dify. Build production: ready agentic AI solutions [EB/OL]. [2025-04-05]. <https://dify.ai/>
- [54] RAGFlow. Build generative AI into your business [EB/OL]. [2025-04-05]. <https://ragflow.io/>

作者简介



包义明, 中兴通讯股份有限公司人工智能算法工程师、中兴通讯“蓝剑计划”员工; 主要研究方向为计算机视觉、AI Infrastructure、多模态大模型等。



林阳, 中兴通讯股份有限公司技术预研工程师; 主要研究方向为人工智能、数据库、大模型技术及应用等。



屠要峰, 中兴通讯股份有限公司中心研究院副院长、中国人工智能学会理事、中国计算机学会杰出会员、大数据/数据库/信息存储专委会常务委员; 主要研究方向为大数据、人工智能、数据库及存储。

中兴通讯技术杂志社

促进产学研合作青年专家委员会

主 任 陈 为 (北京交通大学)

副主任 秦晓琦 (北京邮电大学) 卢 丹 (中兴通讯股份有限公司)

委 员

曹 进 西安电子科技大学

陈 力 中国科学技术大学

陈 为 北京交通大学

陈琪美 武汉大学

陈舒怡 哈尔滨工业大学

陈思衡 上海交通大学

官 科 北京交通大学

韩凯峰 中国信息通信研究院

何 姿 南京理工大学

侯天为 北京交通大学

胡 杰 电子科技大学

黄 晨 紫金山实验室

李 昂 西安交通大学

刘 凡 东南大学

刘春森 复旦大学

刘俊宇 西安电子科技大学

卢 丹 中兴通讯股份有限公司

陆游游 清华大学

宁兆龙 重庆邮电大学

祁 亮 上海交通大学

秦晓琦 北京邮电大学

秦志金 清华大学

史颖欢 南京大学

唐万恺 东南大学

王景璟 北京航空航天大学

王兴刚 华中科技大学

王勇强 天津大学

温森文 华南理工大学

吴泳澎 上海交通大学

武庆庆 上海交通大学

夏文超 南京邮电大学

徐梦炜 北京邮电大学

徐天衡 中国科学院上海高等研究院

杨川川 北京大学

尹海帆 华中科技大学

于季弘 北京理工大学

张 娇 北京邮电大学

张宇超 北京邮电大学

章嘉懿 北京交通大学

赵昱达 浙江大学

赵中原 北京邮电大学

周 伊 西南交通大学

朱秉诚 东南大学

刊物相关信息



投稿须知



投稿平台



过刊下载



论文索引与
引用指南

办刊宗旨:

以人为本, 荟萃通信技术领域精英
迎接挑战, 把握世界通信技术动态
立即行动, 求解通信发展疑难课题
励精图治, 促进民族信息产业崛起

产业顾问:

段向阳、高 音、胡留军、华新海、刘新阳、
陆 平、史伟强、屠要峰、王会涛、熊先奎、
赵亚军、赵志勇、朱晓光

双月刊 1995 年创刊

第 31 卷 总第 184 期

2025 年 8 月 第 4 期

主管: 安徽出版集团有限责任公司

主办: 时代出版传媒股份有限公司

深圳航天广宇工业有限公司

出版: 安徽科学技术出版社

编辑、发行: 中兴通讯技术杂志社

总编辑: 王喜瑜

主编: 王利

执行主编: 黄新明

副主编: 卢丹

编辑部主任: 王萍萍

责任编辑: 徐烨

编辑: 杨广西、朱莉、任溪溪

设计排版: 徐莹

发行: 王萍萍

编务: 王坤

《中兴通讯技术》编辑部

地址: 合肥市金寨路 329 号凯旋大厦 1201 室

邮编: 230061

网址: tech.zte.com.cn

投稿平台: tech.zte.com.cn/submission

电子信箱: magazine@zte.com.cn

电话: (0551) 65533356

发行方式: 自办发行

印刷: 合肥添彩包装有限公司

出版日期: 2025 年 8 月 15 日

中国标准连续出版物号: ISSN 1009-6868
CN 34-1228/TN

定价: 每册 20.00 元