

AI Edge

需求、愿景与潜在关键技术白皮书

**Requirements, Vision,
and Potential Key Technologies
for AI Edge White Paper**



AI Edge 联盟
2026.04

目 录

前言	5
1. 背景与需求	5
1.1 DOICT 融合的产业与技术背景	5
1.2. 全球研发现状	6
1.2.1 ITU-R 6G“通智融合”愿景	6
1.2.2 产业界研发现状	7
1.2.3 学术界研发现状	8
1.3 需求与驱动力	11
2. AI EDGE 的技术内涵	12
2.1 AI EDGE 的定义与关键特征	12
2.1.1 定义	12
2.1.2 AI Edge 的三大关键特征	12
2.2 AI EDGE 的技术优势	13
2.3 AI EDGE 的创新理念	14
3. AI EDGE 的典型应用场景和潜在价值	15
3.1 工业机器人与智能制造	15
3.1.1 场景描述	15
3.1.2 潜在价值分析	16
3.2 智慧能源与电网调度	16
3.2.1 场景描述	16
3.2.2 潜在价值分析	17
3.3 智慧农业与无人农机	17
3.3.1 场景描述	17
3.3.2 潜在价值分析	18
3.4 低空无人机通信与监管	18
3.4.1 场景描述	18
3.4.2 潜在价值分析	19

3.5 具身机器人训练场	20
3.5.1 场景描述	20
3.5.2 潜在价值分析	20
3.6 EDGE 增强的沉浸式 XR.....	21
3.6.1 场景描述	21
3.6.2 潜在价值分析	22
3.7 智能驾驶与车路协同	23
3.7.1 场景描述	23
3.7.2 潜在价值分析	23
3.8 应急通信与保障	24
3.8.1 场景描述	24
3.8.2 潜在价值分析	24
3.9 智慧体育	25
3.9.1 场景描述	25
3.9.2 潜在价值分析	26
3.10 机械导盲犬	26
3.10.1 场景描述	26
3.10.2 潜在价值分析	27
4. AI EDGE 的技术方向与主要挑战	28
4.1 系统架构	28
4.2 AI for Edge 技术	30
4.2.1 AI for Edge 的兴起	30
4.2.2 AI for Edge 的核心价值	30
4.2.3 边缘网络对 AI 的核心需求	31
4.2.4 AI for Edge 性能提升的关键路径	32
4.2.5 AI for Edge 模型与算法的测试验证	34
4.3 AI over Edge 技术.....	35
4.3.1 多模态感知与融合处理	36
4.3.2 模型轻量化与低时延推理技术	36
4.3.3 云边端大中小模型协同技术	37

4.3.4 AI Agent 技术	39
4.3.5 面向具身智能的端到端信息服务技术	41
4.3.6 数据安全和隐私	42
4.4 芯片与算力底座	44
4.4.1 通感智算控融合的芯片架构创新	45
4.4.2 全域异构算力智能调度引擎	46
4.4.3 智能算力开放生态体系	48
4.5 AI Edge 系统、平台与测试	50
4.5.1 AI Edge 系统与平台	50
4.5.2 AI Edge 测试	51
5. 总结	53
参考文献	53
缩略词列表	57
白皮书贡献者列表	59

前言

移动通信网络正突破传统的单一连接服务，加速向通信、感知、智能、计算、控制一体化的综合信息服务升级演进。通信网络的功能将进一步下沉，形成通感智算控超融合的超级边缘网络节点，通过网络连接资源、算力资源、存储资源的平台化和开放化，为用户提供低时延、智能化、定制化的服务，满足 5G 和 6G 多场景业务开放的发展需求。在这一背景下，AI Edge 应运而生。AI Edge 是一种综合移动信息服务基础设施，它基于网络内置的异构开放性可编程共享算力基座，在感知信道环境和用户需求的基础上，实现数据技术、运营技术、信息技术、通信技术 (data technology, operation technology, information technology, communication technology, DOICT) 的深度融合，并对移动边缘信息服务、网络功能虚拟化、网络内生 AI 与自治等功能进行按需编排。本白皮书对 AI Edge 的背景、发展驱动力、国内外研究现状、基本技术内涵与核心特征、典型应用场景与潜在价值等进行了系统性阐述，并围绕 AI Edge 系统架构，AI for Edge 技术，AI over Edge 技术，芯片与算力底座，AI Edge 系统、平台与测试等技术方向进行了深入探讨，指出了 AI Edge 可能的研究方向与技术挑战。

1. 背景与需求

1.1 DOICT 融合的产业与技术背景

在数字化浪潮的席卷下，DT（数据技术）、OT（运营技术）、IT（信息技术）、CT（通信技术）正以前所未有的态势深度融合，这一融合趋势是产业变革与技术创新双轮驱动的必然结果。

从产业角度看，全球各行业正加速数字化转型进程。制造业期望借由技术融合实现智能制造升级，提升生产效率、降低成本、提高产品质量与定制化水平。例如汽车制造企业，利用 DT 对生产线上海量数据进行挖掘分析，借助 OT 精准控制生产流程，依靠 IT 搭建智能化管理系统，通过 CT 实现设备间及工厂与外部的高效通信，构建起高度自动化与智能化的生产体系。能源行业同样如此，在智能电网建设中，DT 用于分析电力供需数据，OT 保障电力系统稳定运行，IT 实现能源管理信息化，CT 支持电力数据实时传输与远程控制，提升能源利用效率与供电可靠性。**从技术角度看**，各技术自身发展遭遇瓶颈，亟待融合突破。CT 领域，5G 虽带来显著性能提升，但面对工业互联网、自动驾驶等场景对低时延、高可靠、海量连接的严苛要求，仍显不足，需与其他技术协同；IT 领域，云计算发展促使计算资源集中化，但数据传输时延与隐私安全问题制约其在部分场景应用，需借助边缘计算等融合手段优化；OT 长期专注工业特定场景，在数字化转型中，其系统封闭性、数据处理能力局限凸显，急需引入 DT 与 IT 技术实现开放互联与智能升级。DT 则需依托 CT、OT、IT 获取多源数据，并借助它们实现数据价值落地。

➤ 技术趋势 1：智能化协同演进

AI 作为 DT 核心，将深度嵌入 OT、IT、CT。在 OT 中，AI 助力工业机器人实现更精准灵活操作，基于实时感知数据做出智能决策；IT 领域，云计算平台借助 AI 实现资源智能调度，提升服务质量；CT 方面，通信网络利用 AI 优化网络规划、故障诊断与流量管理，实现网络自优化、自愈合。多技术融合下的智能化，正朝着端到端智能协同方向

发展，从设备层、网络层到应用层，形成统一智能决策体系，如智能工厂中设备运行状态实时感知、网络传输自动优化、生产计划动态调整的协同闭环。

➤ 技术趋势:2: 边缘融合增强

边缘计算成为 DOICT 融合关键枢纽。在网络边缘，CT 提供网络连接，OT 设备产生数据，IT 提供计算与存储资源，DT 进行数据分析处理。通过边缘融合，数据无需全部上传至云端，可在本地快速处理，降低时延、减轻网络负担、保障数据安全。例如在智能交通中，路侧边缘节点实时处理摄像头采集的交通流量数据（DT），结合交通信号灯控制（OT），通过无线通信（CT）反馈至车辆与交通管理中心，同时利用边缘计算（IT）能力实现实时决策，优化交通信号配时。

➤ 技术趋势 3: 统一标准与开源生态构建

随着融合深入，建立统一标准与开源生态至关重要。目前各行业技术标准不一，阻碍 DOICT 融合大规模推广。例如工业通信协议众多，不同厂商设备难以互联互通。为此，产业界正积极推进标准化工作，如制定统一的数据接口、通信协议、安全规范等。开源项目也不断涌现，促进技术共享与创新，降低企业技术研发门槛与成本，吸引更多参与者构建融合技术生态。

➤ 产业价值 1: 提升产业效率与创新能力

DOICT 融合打破行业信息孤岛，实现数据自由流通与深度挖掘，催生新业务模式与应用。以医疗行业为例，通过 CT 实现医疗设备远程通信，IT 搭建医疗信息系统，DT 分析患者病历、影像等数据，OT 保障医疗设备精准运行，可开展远程医疗、智能诊断等创新服务，提升医疗效率与质量，为患者提供更便捷、精准医疗服务。制造业通过融合实现供应链协同优化、生产过程实时监控与智能排产，大幅提升生产效率与资源利用率。

➤ 产业价值 2: 优化用户体验与服务质量

在消费领域，融合技术为用户带来更智能、便捷体验。智能家居借助 DOICT，用户可通过手机远程控制家电（CT 通信），设备自动感知环境与用户习惯（OT 感知、DT 分析），并智能调节运行状态（IT 控制），营造舒适、节能居住环境。智能出行中，基于 CT 通信与 DT 数据分析，导航系统实时规划最优路线，车辆自动驾驶辅助系统（OT 与 IT 结合）保障行车安全，提升出行效率与舒适度。

➤ 产业价值 3: 促进产业升级与经济增长

DOICT 融合推动传统产业向数字化、智能化转型，培育新兴产业，成为经济增长新引擎。传统农业借助融合技术发展智慧农业，实现精准种植、养殖，提高农业生产效益，推动农业现代化。同时，催生工业互联网、智能物流、数字金融等新兴产业，创造新就业岗位与经济增长点，提升国家整体产业竞争力，促进经济可持续发展。

1.2. 全球研发现状

1.2.1 ITU-R 6G “通智融合” 愿景

在信息技术飞速发展的当下，通信与 AI 的融合已成为 6G 发展的关键趋势。2023 年 6 月，国际电信联盟无线电通信部门（ITU-R）发布的《IMT 面向 2030 及未来发展的框架和总体目标建议书》，明确将“人工智能与通信的融合”列为 6G 的六大应用场景之一，为全球 6G 发展锚定方向。

从 ITU-R 相关工作组的输出看，在未来 6G 网络中，AI 不再是通信的辅助，而是深度嵌入通信系统的各个环节。一方面，通信系统为 AI 提供无处不在的连接，使 AI 服务

能触达各类设备与用户，实现普惠智能。比如在分布式 AI 模型训练中，6G 网络可凭借其高容量、低时延、高可靠的通信能力，助力智能终端间高效交互数据与模型，保护用户隐私的同时提升训练效率。

另一方面，AI 赋能通信系统，实现智能化运维与性能优化。借助 AI 算法，6G 网络能对海量数据实时分析，智能调配通信资源，快速响应网络拥塞、信号干扰等问题，提升网络的灵活性与自适应性。以协作机器人在 6G 场景中的应用为例，通过网络原生智能提供的实时模型推理能力，可满足其对低时延、高推理精度 AI 服务的严苛需求。

此外，ITU-R 识别的面向 IMT - 2030 (6G) 的新兴技术趋势中，原生 AI (AI 空口设计和 AI 无线网络) 是重要一项，这意味着从空口设计、网络架构搭建等底层环节，AI 与通信将深度协同，构建全新通信范式，6G 不仅需要承担通信连接型基础设施的角色，还需要在体系结构中原生引入对 AI 的支持。6G 承载的 AI 应用将具备 AI 需求碎片化、覆盖立体化、交互多样化、AI 业务开放与定制化、以及能力一体化的特征。通过在连接、算力、数据和算法等多维度资源上的深度融合与优化，6G 能够有效保障 AI 服务质量，这也将成为推动服务层面变革的重要驱动力。

1.2.2 产业界研发现状

➤ AI-RAN 联盟

AI-RAN 联盟 (AI-RAN Alliance) 是一个专注于推动人工智能 (AI) 与无线接入网络 (RAN) 深度融合的国际合作组织，成立于 2024 年 2 月 26 日，在西班牙巴塞罗那的 GSMA 世界移动通信大会 (MWC 2024) 上正式宣布成立。AI-RAN 联盟的成立旨在通过将 AI 技术融入蜂窝通信网络，提升无线接入网络的性能、效率与灵活性，推动 5G 和即将到来的 6G 网络发展。其核心使命包括：提高移动网络效率，降低功耗，改造现有基础设施，为电信公司在 5G 与 6G 时代利用 AI 创造新商机。AI-RAN 联盟将研究和创新重点聚焦于以下三大领域：1) AI for RAN: 利用 AI 提升无线接入网络的频谱效率、能效和成本效益；2) AI and RAN: 将 AI 与 RAN 流程深度融合，实现资源高效利用和 AI 驱动的收入模式创新；3) AI on RAN: 在 RAN 边缘部署 AI 服务，提升运营效率，并为终端用户提供新型智能服务。在 2025 MWC 巴塞罗那展会上，AI-RAN 联盟展示了涵盖 AI-for-RAN、AI-and-RAN 及 AI-on-RAN 等多项演示，内容涉及空口技术、节能措施、频谱感知以及网络编排等多个关键领域。

➤ Next G Alliance (美国下一代移动网络联盟)

Next G Alliance 将 AI 定义 6G 网络“能力倍增器”，在其发布的《6G 技术路线图》中提出“AI 驱动的网络自治”体系。该体系包含三级智能架构：边缘层聚焦实时决策（如毫秒级干扰抑制），区域层负责协同优化（如跨基站资源调度），核心层承担全局策略（如业务负载预测）。联盟重点验证了 AI 在动态频谱共享中的应用，通过强化学习算法实现授权与非授权频谱的自适应分配，使频谱利用率提升 40% 以上。其观点认为，通信与 AI 的融合需构建“硬件-算法-数据”协同生态，目前正推动芯片级 AI 加速单元标准化，并联合高校开展联邦学习在网络优化中的安全性研究。

➤ 欧盟 HeXa-X 项目

HeXa-X 作为欧盟 6G 旗舰项目，提出“AI 原生网络架构”理念，将 AI 深度嵌入从空口到核心网的全栈设计。其发布的第二阶段技术报告明确三大方向：一是 AI 驱动的通感算融合，通过多任务学习模型实现通信、感知、计算资源的联合调度；二是网络智能编排，基于数字孪生与强化学习实现端到端业务质量保障，在工业场景验证中使服务可用性达 99.999%；三是可信 AI 框架，通过联邦学习与差分隐私技术解决数据安全

全与模型鲁棒性问题。项目强调，6G 需建立“AI 即服务”（AIaaS）平台，目前已完成跨厂商 AI 模型接口规范，为产业链协同提供技术基准。

➤ 6GANA（6G Network AI 联盟）

6GANA 成员包括运营商、设备制造商、互联网服务提供商和高校等 31 个组织，其的使命和目标是从技术和生态系统的角度积极推动 6G 网络的 AIaaS 化。具体而言，6GANA 旨在通过在整个生态系统内开展联合研究，凝聚网络 AI 共识，涵盖信息和通信技术（ICT）设备制造商（例如芯片制造商、网络基础设施提供商和移动网络运营商）、垂直行业、AI 服务提供商、AI 解决方案提供商、AI 学术界和其他利益相关方，推动 AI 成为 6G 网络的一项新能力和服务，加速泛在智能时代的到来。6GANA 以“网络与 AI 双向赋能”为核心命题，构建了“三横三纵”技术体系：横向覆盖网络架构、数据治理、安全可信三大领域，纵向贯穿需求定义、技术研发、产业落地三个阶段。其发布的《6G 内生智能白皮书》系统提出 AI4NET（AI 增强网络）与 NET4AI（网络支撑 AI）协同范式：在 AI4NET 方面，验证了基于图神经网络的网络故障自愈方案，恢复速度提升 50%；在 NET4AI 方面，设计边缘 - 云端协同的模型训练框架，使分布式训练效率提升 3 倍。联盟通过跨行业工作组推动技术落地，目前已在车联网、工业控制等场景形成 12 项技术规范，推动通信与 AI 融合从概念走向产业化。

➤ IMT-2030（6G）推进组

IMT-2030 推进组将“智联万物”作为 6G 核心愿景，在《6G 总体愿景与潜在关键技术》中明确通信与 AI 融合的“双螺旋”发展路径。一方面，推动 AI 重构通信系统，在空口设计中引入深度学习辅助的波形优化，使复杂环境下通信速率提升 25%；另一方面，构建支撑 AI 服务的通信基础设施，提出“智能体通信网络”（ACN）概念，通过超低时延通信保障边缘 AI 实时推理。推进组联合产学研单位建立了 6G AI 测试床，完成智能超表面与 AI 协同传输等关键技术验证，并形成《6G AI 技术白皮书》，为全球 6G 标准制定提供中国方案。其强调，融合发展需平衡技术创新与产业成熟度，目前正在推动 AI 模型轻量化与通信协议简化的协同优化。

1.2.3 学术界研发现状

通感智算融合不仅是产业界的共识，也是学术界关注的前沿方向。在通信与 AI 融合领域，前期已有大量研究工作基于 AI 赋能通信系统性能优化，在物理层增强、频谱效率提升、网络故障诊断、能效优化等方面取得丰硕成果，近一两年来，研究主要聚焦于生成式 AI 技术在通信网络中的应用；在通信与感知融合领域，目前研究的焦点已由通感频率共享共存、通感射频硬件与软件资源共享互惠逐步迈向多站协同无线通感、多模态网络通感融合等全新阶段，并将应用场景由车联网车辆目标感知、低空无人机目标监测等拓展至智能交通、安防监控等；在通信与计算融合领域，学术界主要是沿着四条路径开展研究工作：外挂式算力（如移动边缘计算 MEC）、网络化算力（将算力信息嵌入路由协议）、内生算力（如 C-RAN、O-RAN）和智能算力（如 AI-RAN）。下面对若干主要研究方向的最新进展做一简要介绍。

➤ 电信基础大模型

大语言模型（LLMs）有望彻底改变第六代移动通信网络（6G）的设计范式，然而当前主流的 LLMs 普遍缺乏电信领域的专业知识。在此背景下，来自阿布扎比科学创新研究所以及哈利法大学的研究团队首次提出了一种将通用 LLMs 应用于电信领域的设计框架[1]。评估结果显示，微调后的 LLM TelecomGPT 在电信数学建模基线测试中显著超越了最先进的 LLMs，包括 GPT-4、Llama-3 和 Mistral，并在 TeleQnA、3GPP 技术文档分类、电信代码摘要与生成及代码补全等各种评估基线中表现出色。

➤ AI 辅助的物理层设计

在物理层通信方面，第一个典型的用例是波束成形。AI 模型经过大量波束成形场景数据集的预训练，能够预测可最大化信号强度并最小化干扰的最佳波束。这可以通过利用多模态来提供有关阻塞概率以及用户状态和活动的附加信息来实现。第二，AI 模型可用于上行链路和下行链路传输之间的信道状态信息（CSI）估计目的。通过自注意力机制和 AI 模型的生成能力，设想模型将能够捕捉上行链路和下行链路传输之间的固有关系，并利用 3D 多模态环境数据（包括摄像头、雷达、激光雷达和 GPS）以选择最佳的上行链路和下行链路波束对，在特定用户位置处，使到达角和离开角完美对齐。第三，在毫米波波束预测方面，东南大学团队将毫米波（mmWave）波束预测问题转化为时间序列预测任务，通过跨变量注意力机制聚合历史观察数据，并使用可训练的分词器将其转化为基于文本的表示，借助 Prompt-as-Prefix（PaP）技术进行上下文增强，利用 LLM 的强大能力来预测未来最优波束[2]。第四，在信源信道联合编码领域，目前已经研究将信道和信源编码集成到语义感知的 JSCC 中，生成式 AI 模型有助于实现高效的 JSCC 方案，以改善无线通信性能[3]。

此外，生成式 AI 也被用于增强接收机性能。扩散模型（Diffusion Models, DM）可以逐步学习去除噪声，近年来在人工智能生成内容（AIGC）中得到了广泛的应用。为了验证 DM 是否可以应用于无线通信，以帮助接收机消除信道噪声，上海交通大学研究团队在 GLOBECOM 2023 上提出了无线通信的信道去噪扩散模型（CDDM）[4]。CDDM 可以作为信道均衡后的一个新的物理层模块来学习信道输入信号的分布，然后利用所学到的知识来去除信道噪声。实验结果表明，CDDM 可以进一步降低均方误差（MSE），具有更好的性能。

为实现对无线信道的高效表征，鹏城实验室团队提出多任务无线基础模型 WirelessGPT，该模型采用基于类 BERT 的 Transformer 架构，并通过三轴注意力机制、多尺度数据编码等技术以适配无线应用，可赋能信道估计、信道预测、定位、行为识别等多种通感任务[5]。北京邮电大学团队基于大语言模型微调的思路构建了无线信道基础模型 ChannelGPT 和 ChannelDS，在信道预测等下游任务上性能优异[6]。北京大学团队提出“机器联觉”概念，基于任务驱动的 AI 原生思想，实现通信与多模态感知智能融合。其研发的基于预训练大语言模型的信道预测方案 LLM4CP、无线物理层多任务方案 LLM4WM，以及无线基座模型 WiFo，为 6G 网络通信感知融合提供新思路[7]。

➤ AI 赋能无线感知

深度学习（DL）模型从根本上促进了无线感知方案的发展，其中射频数据可以被获取并映射到二维图像，用于感知应用，包括定位、遥感和资源分配；生成式 AI 模型可以实现高效的多模态定位方案。这些 AI 模型的通用性和自注意力性质，可以成为检测网络用户和节点的上下文和情境信息的关键，捕获多个图像之间的相互位移，并将这些图像及其变化与无线信号的相应电磁行为相互关联。

射频信号生成技术对于无线感知系统有着重要的意义。为了填补目前高质量时序 RF 信号生成模型的缺失，文献[8]建立了一种新型的时频域扩散理论（Time-Frequency Diffusion Theory），提出了首个针对射频信号的生成式扩散模型 RF-Diffusion，实现了时间序列射频信号的多样化、大规模、高精度自动生成，并成功将其应用于 Wi-Fi 感知数据增广、FDD 信道估计等一系列关键任务中。

➤ 生成式 AI 辅助的语义通信

语义通信（SemCom）在构建背景知识库用于训练语义编码模型的过程中面临诸多挑战，最近出现的生成式人工智能（GAI）技术有望协助 SemCom 中的背景知识构建，增强语义编码模型的推理能力。在此背景下，文献[9]提出了一种 GAI 辅助的 SemCom 框架，通过使用 GAI 根据用户上下文信息来辅助生成用于训练语义编码模型的样本。与传统

SemCom 相比, Gen-SC 在原始训练样本不足的情况下具有更高的语义精度。文献[10]提出了一个专用于图像数据的基于 AI 大模型的语义通信框架(LAM-SC)。设计了基于 SAM 的知识库(SKB), 并提出了一种基于注意力的语义集成(attention-based semantic integration, ASI)方法和一种自适应语义压缩(adaptive semantic compression, ASC)编码方法。文献[11]提出了一种基于预训练基础 AI 模型(Foundation Model)的通用生成式语义通信框架。该框架包括三大核心模块: 多模态语义分解与合成、语义感知的多流传输和低延迟的语义功率分配。

➤ 基于大模型的多智能体

大语言模型(LLM)的快速发展为 6G 通信带来了巨大的机遇, 例如允许用户通过自然语言向 LLM 输入任务要求来进行网络优化和管理。然而, 直接将原生 LLM 应用到 6G 会遇到各种挑战, 例如缺乏专业的通信数据和知识, 模型逻辑推理、评估和优化能力有限。为了解决上述挑战, 文献[12]设计了一个用于 6G 通信的 LLM 增强型多智能体系统, 该系统构建了面向 6G 通信的专业知识库和工具, 并拥有超越原始 LLM 的规划、记忆、工具利用和反省能力。

➤ 基于云边端协同的 AIGC 服务

为了提供低时延与定制化的 AIGC 服务, 采用协作式的云边端 AIGC 框架十分必要。部分高性能终端设备可以直接运行 AIGC 模型来为自身提供服务, 相比于边缘计算, 响应实时性以及安全性进一步提高。同时, 移动终端设备可以将 AIGC 任务卸载到边缘或者云端服务器, 实现灵活的服务配置。考虑到终端设备的轻量性, 通常在设备上运行的模型需要进行压缩量化处理以降低计算、存储资源开销。文献[13]提出了一种边缘适配器模型, 通过该模型可以实现推断准确性、时延以及资源消耗三者之间的折衷关系。在架构设计方面, 文献[14]提出了一种自下而上的 BAIM 架构, 最大化用户数据与边缘专家模型提取的知识利用。该架构有效结合 Pathways 和混合专家模型(MoE), 通过云端大模型与边缘小模型的协同工作, 提升了生成式 AI 服务的效率和用户体验。

➤ 面向 AI 任务的无线网络架构设计

当前的无线网络设计为“数据管道”, 不适宜容纳和利用 GenAI 的能力。为此, 文献[15]提出了一种网络架构, 整合 GenAI 能力以管理网络协议和应用程序。通过构建基于语义的 GenAINet, 从多模态原始数据中提取语义概念, 构建表示它们语义关系的知识库, 然后由 GenAI 模型用于规划和推理。在此模式下, 代理能够从其他代理的经验中迅速学习, 以便更好地决策并更高效地通信。文献[16]提出了内生智能网络架构 NetGPT, 利用云边计算中的资源不均衡, 实现了云端和边缘之间不同规模 LLM 的高效协同。与具有解耦通信和计算资源的 AI 外生网络相反, NetGPT 可以利用融合通信和计算为边缘部署较小的 LLM, 为云端部署较大的 LLM, 并有意地实现云边协同计算, 以提供个性化的内容生成服务。

➤ 基于大模型的无线资源管理与优化

由于用户需求的范围不断扩大, 优化各种无线用户任务对网络系统提出了重大挑战。尽管深度强化学习(DRL)取得了进步, 但需要为个人用户定制优化任务, 这使得开发和应用大量的 DRL 模型变得复杂, 导致大量的计算资源和能源消耗, 并可能导致不一致的结果。为了解决这个问题, 文献[17]提出了一种新的方法, 利用混合专家(MoE)框架, 辅以大型语言模型(LLM), 有效地分析用户目标和约束, 选择专业的 DRL 专家, 并权衡参与专家的每个决策。论文提出的方法减少了为每个独特的优化问题训练新的 DRL 模型的需要, 降低了能耗和人工智能模型的实现成本。

➤ 面向个性化需求的服务质量保障技术

个性化服务应当成为未来 6G 网络的重要能力之一。文献[18]提出了多维指标融合的概念来量化和满足高度差异化的用户需求。该工作提出了“服务需求区域”(Service Requirement Zone, SRZ)的概念,用于在用户侧刻画和可视化个人任务的综合服务需求。SRZ 通过一个八维雷达图来定义,涵盖了时延、能耗、存储、速率、安全隐私、可靠性、知识和成本八个关键性能指标,为用户的个性化体验质量(QoE)设定了明确的边界。在此基础上,该工作进一步引入了“用户满意度”(User Satisfaction Ratio, USR)作为系统侧的评估指标,用以衡量系统在满足不同 SRZ 任务上的整体服务能力。这些概念为实现以每个人为中心的定制化服务提供了理论基础和评估框架,对 AI 下沉至边缘,实现精细化、个性化的资源调度与服务保障具有重要的参考价值。

为推进通信与 AI 大模型融合领域的研究和全球合作,IEEE 通信学会于 2024 年初成立了通信大模型新兴技术委员会(GenAINet ETI)。该 ETI 是一个由全球学者和工业界专家参与的学术组织,主要目标是为探讨通信网络大模型技术搭建一个开放的研究平台,去年 5 月,GenAINet ETI 发布了学术界首个通信网络大模型研究论文集,对相关领域的最新研究成果进行了汇总[19]。

1.3 需求与驱动力

跨界融合、不断拓展应用边界是移动通信发展的重要趋势。5G 技术的多场景应用需求,带动了网络功能虚拟化与切片化的发展以及 ICT 技术的深度融合。未来 6G 垂直行业应用,特别是工业互联网应用,将进一步促进 DT、OT、IT 及 CT 的技术超融合。基于网络内置的异构开放性可编程共享算力基座,为垂直行业用户提供智能化、确定性、定制化以及低时延的通感智算控综合移动信息服务将成为可能。作为 DOICT 技术超融合的重要载体,边缘(Edge)网络凭借其靠近用户所带来的低时延优势体现出独特的价值。

一方面,随着人工智能技术的快速发展,催生了对算力的强烈需求。当前,算力资源已不再局限于云计算中心所提供的集中式、规模化处理能力,更广泛的分布于网络边缘侧及各类终端设备中。在这一背景下,如何高效整合利用分布式算力资源,实现算力的“随取随用”,并延伸算力的触角为垂直行业的智能应用提供低延迟、高可靠的算力服务,实现“compute anywhere”的泛在智能愿景,是人工智能发展迫切需要解决的挑战。边缘网络内部集成的分布式算力可以作为一种新型网络资源,通过对碎片化算力的聚合利用,能够提供“最后一公里”的算力送达服务,构建算力资源不可或缺的网络化延伸。

另一方面,自动驾驶、数字低空、机器人、智慧工厂等新兴业务场景对端到端时延、抖动和可靠性提出了极为严格的要求。为实现更贴近用户的高质量服务能力,网络功能正在持续向边缘侧下沉,构建起分布式、层次化的智能服务新范式。通过对移动边缘网络资源的灵活高效共享,能够实现移动网络功能及智能信息服务在网络边缘的就近部署,并借助边缘网络的连接和感知能力与物理世界充分交互,实现感知、推理与执行的快速闭环,赋能新型 AI 智能体终端,如低空无人机、机器人、自动驾驶等多种低时延具身应用。

2. AI EDGE 的技术内涵

2.1 AI EDGE 的定义与关键特征

2.1.1 定义

AI Edge 是面向智能应用的综合移动信息服务基础设施，基于开放性可编程统一算力构架，同时实现三大功能：

- 移动边缘信息服务
- 网络功能虚拟化
- 网络内生 AI 与自治。

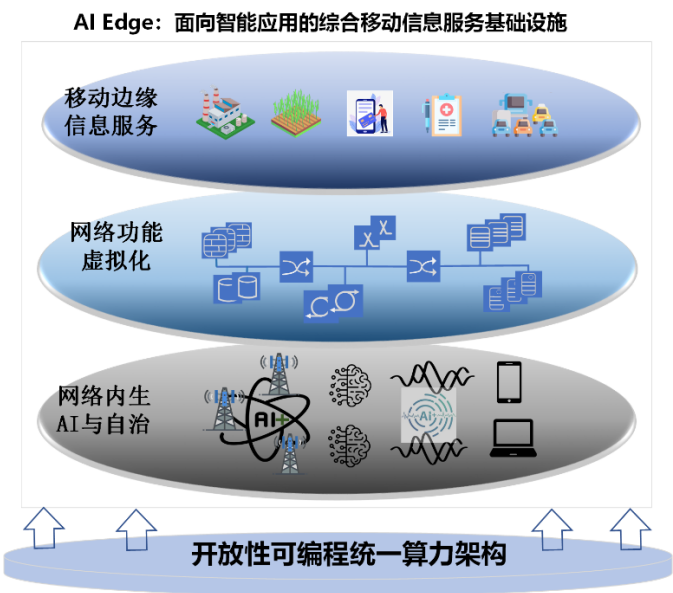


图 1 AI Edge 的定义

2.1.2 AI Edge 的三大关键特征

AI Edge 具备共享化、可扩展、层级化等关键特征，具体论述如下。

共享化：AI Edge 高度兼容 CPU、GPU、NPU、FPGA、SoC 等异构算力，并通过软件可编程的方式实现通信、AI、感知、网络控制与计算服务在同一硬件底座上的功能集成和能力共享。具体的，CPU、GPU、NPU、FPGA、SoC 等多种异构计算硬件构建起统一的算力底座，通过虚拟化与池化技术，将不同硬件的计算资源整合为统一的虚拟算力资源池。该资源池通过设计标准化的统一软件接口，实现对异构算力的集中调用与灵活调度，从而促进算力资源的共享与按需使用。通信信号处理、AI 服务、边缘感知、网络控制等功能无需关注底层硬件差异，可直接从虚拟算力池中获取所需计算资源。借助统一的调度器与协同管理机制，系统能够动态响应业务需求，实现资源的弹性分配与高效管理，提升整体算力利用效率和业务部署的灵活性。

可扩展：AI Edge 不仅在横向维度上跨域整合相邻基站的算力资源，构建弹性可扩展的边缘算力网，也在纵向维度上通过云边端的高效协作，实现跨层级的分布式智能，

支撑移动信息服务在全域网络上的可扩展性。具体的，一方面，AI Edge 构建了支持多节点高速互联的边缘算力网络，依托算力感知、算力路由、算力调度等机制，通过计算任务分解与迁移、模型跨域协同推理、数据并行训练等方法，跨域整合多个边缘节点的算力资源，从而在横向上拓展了 AI Edge 的协同计算能力。另一方面，AI Edge 也在纵向上贯通云、边、端层次化网络的异构算力资源，通过大小模型在云侧、边缘侧、终端侧的自适应部署，实现多尺度算力的动态配置与高效利用。进一步的，通过大小模型的在网协同推理、双向更新、协同进化等机制，有效促进了集中式规模化算力与分布式碎片化算力之间的深度协作，全面提升包括推理时延和设备能耗在内的多维度服务质量。

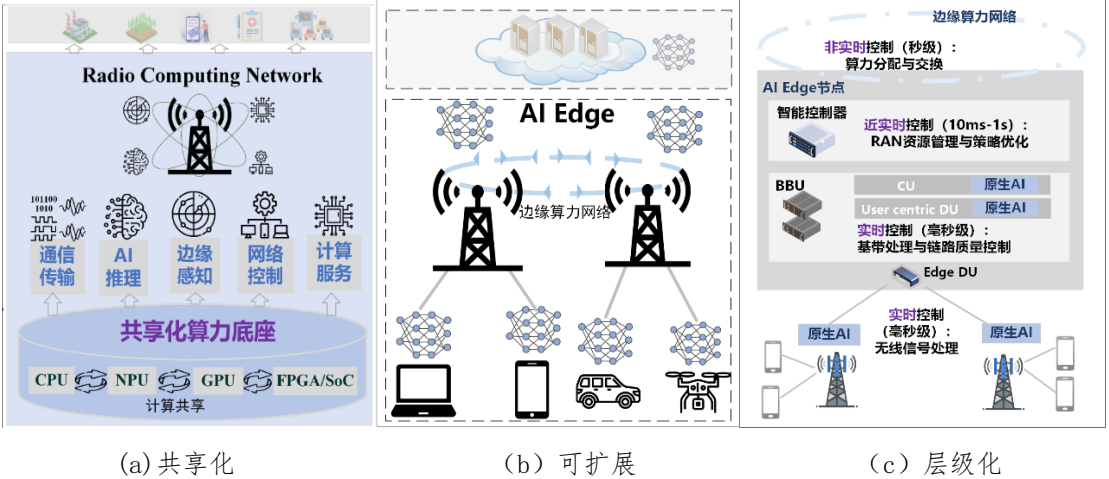


图 2 AI Edge 的核心特征

层级化：AI Edge 充分融合 AI 领域的最新理念和技术，通过信道基础模型（如 WirelessGPT）表征复杂无线环境，有效利用共享底座，并基于 Agentic AI 技术将用户需求直接映射为通信网络的基础能力，进而调用相应功能，按需实现从毫秒级到秒级的实时、近实时、非实时的分层次网络自主管控。具体的，AI Edge 基于信道基础模型，对无线信道及信号的时间域、空间域、频率域等多维度之间的关系进行刻画表征，并在此基础上，通过内生于无线接入单元和基带处理单元的多种下游任务模型完成信道估计、信道预测、波束管理、干扰抑制、解码优化等通信子任务，实现对于射频及基带信号的毫秒级的实时优化处理。同时，AI Edge 借助部署于边缘节点中的智能优化器，实时感知并解析用户需求，动态生成近实时的网络智能调控策略，从而显著提升用户体验。此外，AI Edge 通过边缘算力网络的协同计算，生成面向跨节点通信、感知与计算资源的非实时、高性能全局管理方案，实现多边缘节点资源的高效协同与统一编排。

2.2 AI Edge 的技术优势

基于以上特征，AI Edge 展现出显著的优势与价值，具体如下所述。

首先，AI Edge 基于共享底座，实现 DOICT 技术能力互通和算力资源共享，并原生支持感知、控制、转发、路由、数据管理等新的网络功能，从而显著地提升网络资源的利用效率。基于通信、感知、计算、控制、智能五位一体深度融合与高效协作，实现了环境感知、数据传输、智能分析到精准控制的全链路闭环支撑，为智能应用提供了端到端的融合服务支持。



图 3 AI Edge 的重要优势

其次，通过将云、基站与终端在内的海量分布式算力汇集，不仅实现了算力的“开源”，更是将众多终端由单纯的算力消耗者转变为算力的供给者，从而实现可用算力和智能服务能力随网络规模扩大而同步递增，使能大规模 AI 应用。在 AI Edge 中，人工智能能力不再局限于某个遥远的“云端大脑”的集中化处理，而是通过深度嵌入亿万终端与网络设备，构建起一个与物理世界紧密耦合、持续感知与计算的泛在神经网络，构建了无处不在、无缝衔接的普适智能服务新范式。

再次，通过原生 AI 能力的构筑，主动感知复杂无线环境和差异化用户服务需求，促进通信网络的服务范式由“被动响应”向“主动预测”转变，从而实现从“尽力而为”到“按需保障”的跨越，为传统“管道”注入新的价值增长点。通信网络不再局限于传统的数据传输与连接功能，而是通过嵌入感知、认知与决策能力，基于对业务意图、用户行为及环境状态的超前预测与智能编排，自主调配算力、存储及通信资源，动态优化服务路径与质量，为用户提供精准、自适应、个性化的智能服务体验。

2.3 AI Edge 的创新理念

AI Edge 的创新理念将体现在四个方面，具体如下所述。

第一，革新网络技术：AI Edge 打破了传统 5G 系统中无线接入网（RAN）、用户面功能（UPF）、网络控制、移动边缘计算（MEC）分立实施的模式，支持 RAN、AI 推理、边缘感知与控制执行的 DOICT 技术深度集成，推动 RAN 发展成 RCN（Radio Computing Network），从而赋予移动通信网络新的内涵。此外，AI Edge 还有望通过对全域环境的认知获得对于物理环境、业务特性、网络负载、用户习惯等的精准刻画，从而实现网络的自学习、自优化、自演进。

第二，拓展边缘概念：在传统移动网络中，Edge 是单一边缘网络功能实体，而 AI Edge 将构建分布式、云边端协同的共享算力底座，实现 AI 服务能力的分布式部署和跨层跨域扩展。在 AI Edge 中，终端设备不再视为孤立的数据采集点或单纯的服务消费端，而是内生为 Edge 的一部分，增强了边缘层的感知、计算与响应能力，构建了协同化、智能化、全域联动的泛在边缘节点网络。

第三，增强 AI 能力：AI Edge 所支持的 AI 应用并非单纯的信息检索、内容生成、任务规划等，而是借助网络的连接和感知能力与物理世界充分交互，实现“感知、推理与执行”的快速闭环，促进网络内生 AI 的实际应用落地；通过实际数据反哺 AI 模型，推动 AI 发展范式迈向可实时感知与推理的 AI。AI 模型不再是固定的神经网络参数集合，而是成为具备主动感知、动态适应、持续更新、自主进化的新一代智能体，为构建下一代可持续、自适应人工智能奠定了关键技术基础。

第四，重构交互范式：AI Edge 基于 Agent 技术实现用户意图理解和自动化的编排调度，从而打破了传统的基于固定协议和流程的交互方式，实现基于意图的智能交互。具体的，通过对用户语音、文字输入的分析理解，识别用户对于网络服务质量的个性化

需求，自主生成动态的网络配置策略，驱动底层物理功能单元进行高效配置，从而实现“用户表达”到“系统响应”的闭环智能服务。

3. AI EDGE 的典型应用场景和潜在价值

AI Edge 通过“通感智算控”一体化能力，正在重塑产业数字化的底层逻辑。从技术层面看，其突破了传统通信网络“连接为本”的局限，将通信（泛在互联）、感知（环境洞察）、智能（实时决策）、计算（边缘算力）、控制（精准执行）深度耦合，形成闭环协同的技术体系。这种融合不仅带来性能跃升——如工业场景中设备响应时延从秒级压缩至毫秒级，低空场景中空域管控精度提升至米级，更重构了产业价值分配模式。

在商业维度，AI Edge 催生了“能力即服务”的新范式：硬件厂商从销售设备转向提供“终端 + 边缘节点”的长期服务，算法企业通过模型订阅实现持续收益，行业客户则按实际价值付费（如按每降低 1% 故障率结算）。这种模式推动产业链从“一次性交易”向“共生增值”转型，预计到 2030 年，仅在智能制造、智能交通、低空经济三大领域，AI Edge 相关市场规模将突破 5000 亿美元，成为数字经济增长的核心引擎。

3.1 工业机器人与智能制造

3.1.1 场景描述

在工业生产环境中，AI Edge 的“通感智算控”技术体系为工业机器人打造了“精准感知 - 智能决策 - 高效执行”的工作闭环。工业机器人配备多种传感器，如视觉相机、力传感器、激光雷达等，通过 5G 或工业以太网等通信技术，与边缘计算节点实时交互数据，对周围环境中的物料位置、形状、装配精度要求，以及自身关节状态、运行轨迹等信息进行全方位感知（通感）。

车间内边缘节点通过工业无线（如 5G-Advanced）与物联网传感器，实时采集设备振动、温度、能耗等数据（通感）；边缘 AI 模型（如基于联邦学习的电机故障诊断算法）在本地化算力支撑下（算），实现设备异常的毫秒级识别与根因分析（智）；进而联动 PLC 控制系统自动调整机床参数、触发停机预警或调度维修机器人（控），构建无人化生产体系。

边缘 AI 算法利用本地化的算力资源，对感知数据进行快速分析与处理。例如，基于深度学习的视觉识别算法，能精准识别流水线上不同型号的零部件，确定其抓取位置与姿态；通过对力传感器数据的实时解析，调整机器人抓取力度，避免损坏精密部件。同时，边缘系统结合生产任务与实时工况，运用智能调度算法生成机器人最优行动策略，如在多机器人协同作业场景中，规划各机器人的作业顺序与路径，避免碰撞冲突（智算）。最终，将控制指令快速传输至机器人的电机驱动、关节控制等执行单元，实现对机器人动作的精准控制，完成物料搬运、零件装配、产品检测等复杂生产任务（控）。该场景广泛应用于汽车制造、电子加工、物流仓储等行业，如汽车总装车间中，机器人借助 AI Edge 技术高效完成车门安装、零部件焊接等工序；3C 产品生产线上，实现芯片的高精度贴装与检测。

3.1.2 潜在价值分析

AI Edge 通过 “通感智算控” 的深度协同，为工业机器人赋予了更强的环境适应性、任务执行能力与智能决策水平，其核心价值不仅在于提升生产效率与产品质量，更通过创新商业模式，为制造业带来显著经济效益与社会效益，推动行业迈向智能化、柔性化生产新时代。

➤ 技术价值

突破传统工业机器人 “感知局限、响应迟缓” 的困境：AI Edge 技术体系使工业机器人的感知精度提升至亚毫米级，相较于传统方案，对微小零部件的识别准确率从 85% 提高到 98%；边缘 AI 的实时决策能力将机器人响应延迟从传统云端处理的 200-300ms 缩短至 50ms 以内，满足高速生产线上对 “即时响应、精准操作” 的严苛要求，大幅减少生产误差与次品率。

实现 “复杂工况适应 - 柔性生产” 的技术飞跃：通过多模态感知数据融合与边缘智能算法，工业机器人能够在复杂光照、振动等恶劣环境下稳定作业，且可快速切换生产任务，适应小批量、多品种的柔性生产需求。例如，在电子制造中，能在短时间内重新编程并调整操作流程，生产不同型号电子产品，产线切换时间从数小时缩短至半小时以内。

➤ 商业价值

设备升级与服务订阅模式：制造企业可采购搭载 AI Edge 技术的新型工业机器人，或对现有设备进行升级改造，设备供应商收取一次性升级费用。同时，企业还可按需订阅边缘 AI 算法服务，如针对特定工艺的优化算法，按使用时长或调用次数计费，为供应商创造持续收入流，也降低了企业技术研发成本。

生产效率提升与成本节约：AI Edge 赋能的工业机器人可使企业生产效率提升 30%-50%，人力成本减少 20%-40%，同时降低原材料浪费与设备损耗。例如，汽车制造企业通过引入相关技术，单车生产时间缩短 2-3 小时，年节约成本数千万元，增强了企业市场竞争力。

➤ 社会价值

推动制造业高端化转型：助力传统制造业向智能制造升级，提升国家制造业整体水平，增强国际竞争力，吸引高端制造业回流，促进产业结构优化。

缓解劳动力短缺与技能鸿沟：在劳动力成本上升与专业技术工人短缺背景下，工业机器人智能化升级可降低企业对大量重复性劳动人力的依赖，同时减少新员工培训周期与难度，从 3-6 个月缩短至 1-2 个月，促进制造业可持续发展。

3.2 智慧能源与电网调度

3.2.1 场景描述

智慧能源与电网调度场景中，AI Edge 的 “通感智算控” 技术体系构建 “全域感知 - 智能决策 - 动态调控” 的能源管理闭环。电网边缘节点集成物联网传感器（如智能电表、光伏逆变器、储能电池监测模块）与毫米波通信模块，实时采集分布式能源出力（光伏、风电）、用户负荷波动、输电线路状态（温度、电流）等多维数据（通感）；边缘 AI 引擎通过时空序列预测算法（如 LSTM 模型）预测新能源发电功率与用电负荷，结合边缘算力快速完成供需平衡计算、最优潮流分析（智算）；进而向储能系

统、可调负荷（如充电桩、工业电机）及变电站控制系统发送指令，动态调整充放电策略、负荷优先级与输电线路功率分配（控），实现源网荷储协同优化。

该场景覆盖多元能源场景：分布式光伏电站中，边缘系统实时调整逆变器输出，平抑功率波动；城市配电网中，通过 AI Edge 实现充电桩错峰充电调度，避免台区过载；工业园区内，边缘节点联动微电网与主网，优化自备电厂与外购电比例，降低能源成本。

3.2.2 潜在价值分析

AI Edge 通过“通感智算控”的深度协同，将电网从“被动调度”升级为“主动感知、智能响应”的智慧能源网络，其核心价值不仅在于技术性能的跃升，更在于通过商业化模式创新，平衡新能源消纳、电网安全与用能成本，为构建新型电力系统提供关键技术支撑，推动能源行业向高效、清洁、可持续方向转型。

➤ 技术价值

突破传统电网“响应滞后、新能源消纳难”的瓶颈：AI Edge 的毫秒级感知与决策能力，使电网对负荷波动的响应时间从秒级压缩至 50ms 以内，新能源发电预测精度提升至 90%（传统方法约 70%），弃风弃光率降低 15 个百分点；通过边缘节点分布式调控，配电网线损率从 8% 降至 5% 以下，解决集中式调度“算力不足、时延高”的问题。

实现“复杂场景自适应 - 安全冗余”：边缘 AI 通过多源数据融合（如气象数据、历史负荷），在极端天气（如台风、寒潮）下提前 12 小时预判电网风险，自动启动负荷转供方案，使供电可靠性提升至 99.99%，较传统模式减少 50% 的停电时长。

➤ 商业价值

“平台 + 增值服务”模式：电网企业搭建 AI Edge 能源管理平台，向新能源场站收取数据接入与调度服务费（按发电量比例计费）；向工业用户提供“负荷优化套餐”，按节能收益分成（如通过错峰用电降低电费，平台抽取 10%-20% 分成）。

设备厂商生态模式：储能设备厂商嵌入边缘 AI 控制模块，按“设备销售 + 算法订阅”收费（如提供动态充放电策略，单套系统年服务费超万元）；新能源车企通过车网互动（V2G）技术，联合边缘平台为车主提供“低谷充电折扣 + 电网辅助服务收益”，形成用户与电网的双赢闭环。

➤ 社会价值

推动能源结构绿色转型：AI Edge 技术可支撑高比例新能源并网（如风电、光伏占比提升至 40% 以上），年减少碳排放超亿吨，助力“双碳”目标落地；分布式能源消纳率提升 20%，相当于新增 1000 万千瓦装机容量的清洁电力供给。

降低社会用能成本：工业用户通过边缘优化可降低 10%-15% 的电费支出，年节约成本超百亿元；居民用户通过智慧电表与边缘调度，享受分时电价优惠，年均电费减少 8%-10%，同时提升极端天气下的供电韧性，减少灾害导致的社会经济损失。

3.3 智慧农业与无人农机

3.3.1 场景描述

智慧农业场景中，AI Edge 的“通感智算控”技术体系构建“全域监测 - 智能决策 - 精准执行”的农业生产闭环。田间部署的边缘节点集成土壤传感器、无人机遥感设备与 LoRa/5G 通信模块，实时采集土壤墒情、作物长势、病虫害迹象及气象数据（通感）；边缘 AI 引擎通过图像识别（如叶片病害分类）、生长模型预测（如作物需

水量计算)等算法,结合本地化算力快速生成灌溉、施肥、植保等精准管理策略(智算);随后通过边缘系统控制水肥一体机、无人播种机、无人机喷雾器等设备,实现变量灌溉、定向施肥与病虫害精准防治(控)。

该场景可覆盖多类农业生产需求:如大田种植中,边缘节点根据小麦生长阶段动态调整灌溉量;设施农业(温室大棚)中,通过感知温湿度、光照强度,自动调控遮阳网与通风设备;畜牧养殖中,通过穿戴设备采集牲畜健康数据,边缘 AI 实时预警疫病风险并触发隔离指令。

3.3.2 潜在价值分析

AI Edge 通过“通感智算控”的深度融合,将农业生产从经验驱动转向数据驱动,其核心价值不仅在于产量与效率的提升,更在于通过技术普惠性缩小城乡数字鸿沟,同时实现农业绿色可持续发展,为保障粮食安全、推动乡村振兴提供了可落地的技术路径。

➤ 技术价值

突破传统农业“粗放管理、靠天吃饭”的局限:AI Edge 的分布式感知能力使土壤、作物、环境数据采集精度提升至 90% 以上,较传统人工巡检效率提升 50 倍;边缘 AI 的智能决策将灌溉量、施肥量控制误差缩小至 5%,解决“过量投入”问题,同时通过病虫害早期识别(准确率达 95%),减少农药使用量。

实现“小农户-规模化”的技术普惠:轻量化边缘设备(如低成本土壤传感器)适配小农户生产场景,边缘算力下沉使单地块管理成本降低 60%,打破智慧农业“高门槛”壁垒,推动技术向散户渗透。

➤ 商业价值

“硬件+SaaS”服务模式:农业科技企业提供边缘感知设备(如智能传感器)与云端管理平台,农户按亩数订阅 AI 种植方案(如小麦精准灌溉模型),单亩年服务费仅需数十元,降低使用门槛。

产业链协同模式:农资企业(化肥、农药厂商)与 AI Edge 服务商合作,基于边缘采集的作物数据推送定制化农资套餐,按实际用量收费;电商平台通过边缘数据预判收成,提供“预售+物流”一体化服务,实现产销精准对接。

➤ 社会价值

提升农业生产效率与可持续性:在粮食种植中,AI Edge 技术可使水资源利用率提升 40%、化肥农药使用量减少 30%,亩产增加 10%-15%;在设施农业中,通过精准环境调控,使蔬菜采收周期缩短 20%,年增产超 2000 万吨。

助力乡村振兴与食品安全:小农户通过技术赋能提升收益(如某试点村农户年均增收超 3000 元);边缘系统记录的生产数据可追溯,为农产品质量认证提供依据,推动“从田间到餐桌”的安全管控,增强消费者信任。

3.4 低空无人机通信与监管

3.4.1 场景描述

基于无人机的低空应用中,AI Edge 的“通感智算控”技术体系构建“全域监管-智能协同-精准作业”的低空服务闭环。地面边缘节点与无人机搭载的多模态传感器(毫米波雷达、高清摄像头、北斗定位)协同,实时感知无人机位置、飞行状态、空域障碍物及地面目标(如物流包裹、巡检设备)(通感);边缘 AI 引擎通过空域冲突预测算法、路径规划模型(智),结合边缘节点本地化算力(算),动态生成无人机

避障指令、任务调度方案（如多机协同配送路线）；最终通过空地一体化通信（如 5G-Advanced、LTE-M）控制无人机起降、航迹修正及作业执行（如精准投送、设备巡检）（控）。

该场景覆盖多元低空需求：物流领域，边缘系统调度无人机群完成“3 公里半径 15 分钟达”的即时配送；电力巡检中，无人机通过边缘 AI 识别输电线路缺陷（如绝缘子破损），同步生成维修坐标；应急救援时，边缘节点快速规划无人机搜救路径，结合热成像感知锁定受困人员位置。

3.4.2 潜在价值分析

AI Edge 通过“通感智算控”的深度协同，为低空应用提供了“安全可控、高效经济”的技术底座，其核心价值不仅在于突破空域资源约束，更通过商业化模式创新推动低空经济从“单点试点”走向“规模化运营”，预计到 2030 年将带动相关产业创造超万亿级经济价值，成为数字经济的新增长极。

➤ 技术价值

突破低空应用“监管难、效率低”的瓶颈：AI Edge 的空域感知精度达 0.5 米级，可实现每平方公里 500 架无人机的高密度协同管控，冲突预警准确率提升至 99%，较传统人工调度效率提升 10 倍；边缘 AI 的实时决策使无人机应急响应时延从云端处理的 300ms 降至 50ms，满足抢险救灾“秒级响应”需求。

实现“复杂环境适配 - 成本优化”：通过边缘算力分流无人机数据处理压力，单机传感器成本降低 40%（无需搭载高端芯片）；在雨雪、雾霾等恶劣天气下，多模态感知融合技术使目标识别准确率保持 85% 以上，解决传统无人机“看不远、辨不清”的问题。

➤ 商业价值

“基础设施 + 运营服务”模式：地方政府或企业建设边缘低空管控基站，向物流、巡检等企业收取无人机接入费（按飞行时长 / 架次计费）；第三方服务商提供 AI 算法订阅（如电力缺陷识别模型），单行业年均服务收入可达亿元级。

场景化解决方案模式：针对农业植保场景，提供“无人机 + 边缘 AI”的精准施药方案，按防治面积收费（每亩成本较人工降低 60%）；在城市安防领域，通过边缘系统联动无人机与地面监控，为物业、园区提供“空中 + 地面”一体化安防服务，年订阅费用超千万元。

构建“空域服务 + 算力租赁”生态：物流企业按飞行里程付费使用边缘算力与通信资源，成本较传统卫星定位方案降低 60%；地方政府通过低空管控系统实现合规化管理，可带动无人机物流、应急救援等产业规模突破万亿，创造超 50 万个就业岗位。

➤ 社会价值

激活低空经济潜力：支撑无人机物流规模化落地，使同城即时配送成本降低 50%，惠及生鲜、医药等民生领域；电力巡检效率提升 80%，线路故障检出率从 70% 升至 95%，减少停电事故带来的社会损失。

提升应急响应能力：在森林火灾、洪涝灾害中，无人机群通过边缘协同实现全域监测，救援力量部署效率提升 3 倍，受困人员搜救时间缩短 60%，显著降低生命财产损失。

3.5 具身机器人训练场

3.5.1 场景描述

具身机器人训练场依托 AI Edge 的“通感智算控”技术体系，构建“环境感知 - 智能决策 - 动作执行 - 反馈优化”的闭环训练生态。训练场内部署毫米波雷达、视觉传感器与 5G/6G 边缘节点，实时采集机器人的关节角度、运动轨迹、力反馈数据及周围环境的三维空间信息（通感）；边缘节点集成高算力 AI 芯片，通过强化学习、数字孪生等算法构建虚拟训练场景，模拟极端天气、复杂地形等真实世界挑战，同时实时优化机器人的运动控制策略（智算）；基于边缘系统的低时延特性，控制指令可在 10ms 内传输至机器人执行机构，实现虚拟训练与物理动作的精准映射，同时通过多机器人协同算法调度训练资源，避免设备冲突（控）。

该场景可支撑多类型具身机器人训练：如工业协作机器人通过边缘 AI 学习精密装配动作，家庭服务机器人在虚拟场景中模拟家具避障与人机交互，救援机器人则在边缘生成的地震、火灾等极端环境中训练应急响应能力。

3.5.2 潜在价值分析

具身机器人训练场通过 AI Edge 技术重构了机器人的“学习 - 进化”路径，其核心价值不仅在于训练效率与成本的优化，更在于通过“通感智算控”的深度协同，降低机器人技术的应用门槛，推动具身智能从实验室走向千行百业，成为智能制造、民生服务等领域的核心生产力工具。

➤ 技术价值

突破传统机器人训练的“高成本、高风险”瓶颈：通过边缘数字孪生技术，将物理样机损耗降低 70%，极端场景训练成本减少 60%；边缘 AI 的实时决策能力使机器人动作修正延迟从秒级压缩至毫秒级，训练效率提升 3 倍。

实现“数据闭环 - 泛化能力”的智能化提升：边缘系统聚合多机器人训练数据，通过联邦学习训练通用动作模型，使机器人在未知场景中的任务成功率从 50% 提升至 85%，解决“场景迁移难”问题。

➤ 商业价值

B2B 订阅模式：机器人厂商按训练时长或场景复杂度付费使用边缘训练平台，单台机器人年均训练成本降低 2 万美元；第三方算法公司提供定制化训练模型（如精密操作强化学习算法），按模型调用次数分成。

能力输出模式：训练场运营方将边缘训练技术打包为“训练即服务（TaaS）”，向高校、科研机构开放 API 接口，支持机器人算法验证与人才培养，年服务收入可达千万级。

➤ 社会价值

加速具身机器人产业化落地：在工业领域，可使协作机器人部署周期从 6 个月缩短至 1 个月；在服务业，家庭机器人通过边缘训练实现 90% 以上的日常任务自主完成，推动老龄化社会的照护压力缓解。

构建机器人技术普惠生态：中小厂商无需自建昂贵训练设施，通过边缘平台即可获得尖端训练能力，促进行业技术均衡发展，预计到 2030 年可带动全球机器人市场规模增长超 3000 亿美元。

3.6 EDGE 增强的沉浸式 XR

3.6.1 场景描述

沉浸式 XR 包括虚拟现实 (Virtual Reality, VR)、增强现实 (Augmented Reality, AR) 和混合现实 (Mixed Reality, MR) 等。Edge 增强的沉浸式 XR 被定位为能体现网络智能与业务协同价值的代表性场景之一。图 4 以 AR/VR 游戏场景的八个关键维度勾勒沉浸式体验所需的综合能力边界，可概括为时延与抖动、吞吐与分辨率、可靠性与连续性、存储、能效与终端热管理、安全与隐私、智能协同与个性化、成本与规模化相互制衡的指标簇。以 VR 为例，端到端交互需保持 20ms 级感知-渲染闭环、稳定的高帧率与视场角、精确的头手部追踪与多用户一致性，同时控制能耗与终端发热。这要求将渲染分片、注视点感知编码、场景语义压缩、资源预取与资产缓存等能力前移至边缘，配合链路层速率自适应与跨接入的无缝切换，按需扩展体验八边形的各个“顶点”。

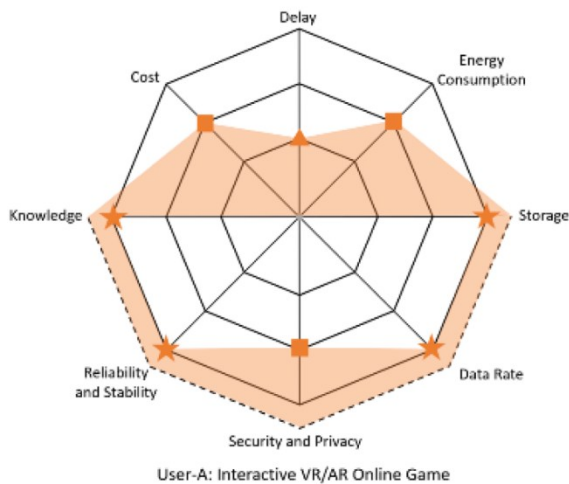


图 4 AR/VR 游戏用户需求示意图

鉴于 AR/VR 体验受限于终端设备的计算能力、电池续航以及云端数据返回的延迟影响，AI Edge 技术可以在边缘侧扩展原生支持 AI 和计算能力，通过将复杂的渲染、计算和 AI 处理任务从终端设备卸载到附近的边缘服务器，为用户提供轻量化、高保真和强交互沉浸式体验，如图 5 所示。

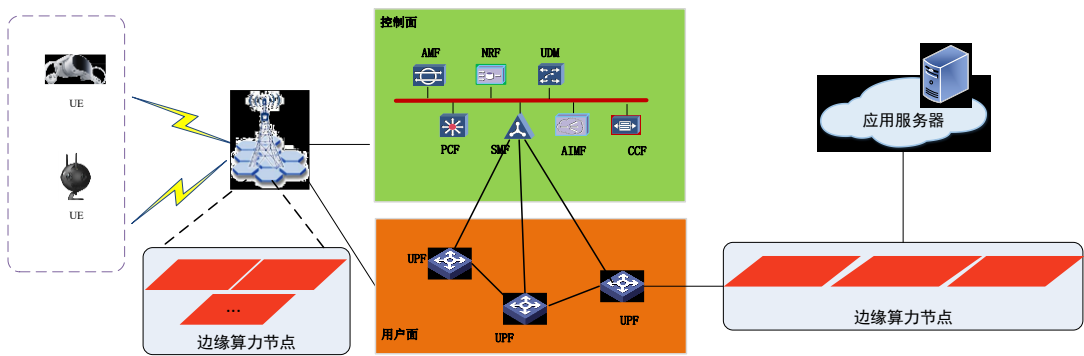


图 5 Edge 增强的 AR/VR

一方面，诸如图像渲染、音视频流特征解析、视频编码等算力任务，可以将其部分或全部任务从云端卸载至边缘算力节点，通过云边协同处理降低传输时延并减少对带宽资源的占用。另一方面，将终端设备的计算任务（如视频解码、跟踪定位等）卸载至边

缘算力节点,实现端边协同处理,解决终端在图像渲染、移动性及互动体验方面的不足,降低对终端电池续航、体积及存储能力的要求,降低终端设备成本。再者,在 AI Edge 增强的 AR/VR 业务场景中,还可以利用 AI 应用分析用户的行为和偏好,根据用户在虚拟环境中的行为调整内容推荐,或者利用 AI 技术助力 XR 内容创作或动画与虚拟形象生成等场景,将部分 AI 增强能力卸载至边缘算力节点,高效协同端边算力,为用户提供更好的沉浸式业务体验。

在具体场景中,该技术可支撑工业 AR 远程运维(如工程师通过 AR 眼镜查看设备内部结构,边缘 AI 实时标注故障点并推送维修指引)、VR 多人协同办公(如虚拟会议室中,边缘节点同步处理 10+ 用户的动作捕捉数据,确保虚拟形象交互延迟 < 20ms)、沉浸式教育(如解剖学 VR 课程中,边缘系统根据学生手势实时渲染器官细节,AI 算法动态调整讲解内容难度)等多样化需求。

3.6.2 潜在价值分析

Edge 增强的 AR/VR 通过 AI Edge 技术重构了“终端 - 边缘 - 云端”的协同范式,不仅解决了用户体验的核心痛点,更通过技术普惠性降低产业准入门槛,加速 AR/VR 从概念走向规模化商用,成为元宇宙基础设施的核心支撑。

➤ 技术价值

突破传统 AR/VR 的“算力瓶颈”与“延迟痛点”:边缘节点将终端算力需求降低 60%,使轻量化 AR 眼镜(重量 < 100g)可实现旗舰级体验;通过边缘 AI 的动态码率调整与预测性渲染,将端到端延迟控制在 15ms 以内,解决眩晕问题,用户体验满意度提升 40%。

实现“环境感知 - 内容适配”的智能化:边缘系统通过多模态感知数据(如用户视线聚焦点、环境光照),自动优化渲染精度(如聚焦区域 4K 分辨率、边缘区域 1080P),在带宽占用减少 30% 的同时,保证核心内容清晰度。

降本增效,促进产业生态发展:通过计算卸载,可使终端设备可以向“轻量化、长续航、低成本”方向演进,极大降低用户购置门槛和硬件迭代成本。同时,海量原始数据在本地边缘处理,可大幅节约回传带宽和中心云的计算成本,为大规模部署创造条件,从而推动产业生态健康发展。

➤ 商业价值

B2B2C 服务模式:硬件厂商(如 AR 眼镜厂商)预装边缘适配模块,按设备激活量向边缘服务商付费;企业客户(如制造业、教育机构)订阅边缘算力与 AI 模型服务(如工业 AR 标注算法),按使用时长计费,单用户年均付费可达数百美元。

内容生态分成模式:边缘平台聚合 AR/VR 内容创作者,通过 AI 推荐算法提升内容曝光率,平台按流量分成(如 10%-20%),同时为创作者提供低代码开发工具(基于边缘 AI 的自动建模功能),降低内容生产门槛。

➤ 社会价值

推动 AR/VR 从“娱乐向”向“生产向”转型:在工业领域,可使设备维修效率提升 50%,培训成本降低 60%;在远程医疗领域,支持外科医生通过 AR 指导基层手术,优质医疗资源可及性提升 30%。

拓展数字经济新场景:预计到 2027 年,Edge 增强的 AR/VR 将带动虚拟办公、数字孪生城市等领域市场规模突破 5000 亿美元,创造超百万个新型就业岗位(如边缘 AR 内容设计师、虚拟空间运维师)。

3.7 智能驾驶与车路协同

3.7.1 场景描述

在智能驾驶场景中，AI Edge 的“通感智算控”技术体系构建起“环境感知 - 决策规划 - 精准控制”的行车闭环。车辆搭载的摄像头、毫米波雷达、激光雷达等传感器，借助 5G/6G 通信与边缘计算融合，实时采集车辆周边 360 度全方位环境数据，包括其他车辆位置、速度、行人动态、交通信号灯状态以及道路状况等（通感）。车辆边缘节点集成 AI 芯片，通过目标检测、语义分割、多目标跟踪等 AI 算法，结合本地化算力对海量感知数据进行快速处理，识别出各类物体、预测其运动轨迹，并根据交通规则和驾驶意图生成最优行驶路径与决策指令（智算）。最终，这些指令在毫秒级时间内传输至车辆动力、转向、制动等执行系统，精准控制车辆加减速、转弯、避让等动作（控）。

该场景能够覆盖多种驾驶场景：在高速公路场景下，边缘 AI 辅助车辆保持安全车距、自动变道；城市道路中，可实现路口智能通行、应对复杂交通状况；停车场场景中，助力车辆自主寻位泊车。此外，车路协同模式下，路侧边缘节点还可与车辆交互信息，进一步提升驾驶安全性与通行效率。

3.7.2 潜在价值分析

AI Edge 通过“通感智算控”的深度协同，将智能驾驶从概念逐步推向现实，其核心价值不仅体现在技术性能的飞跃，更在于通过全新的商业模式，创造显著的社会经济效益，推动交通出行领域的深刻变革。

➤ 技术价值

突破传统智能驾驶“感知局限、决策延迟”的难题：AI Edge 的多传感器融合感知技术使车辆对目标物体的识别准确率提升至 98% 以上，较传统单一传感器方案大幅提高；边缘 AI 的实时决策能力将车辆决策延迟从云端处理的 200-300ms 压缩至 50ms 以内，满足智能驾驶对“低时延、高可靠”的严苛要求，有效避免碰撞事故。

实现“复杂场景适应 - 智能协同”的技术升级：通过边缘计算与 AI 结合，车辆能够在雨、雪、雾等恶劣天气及隧道、城市峡谷等复杂环境下，仍保持稳定可靠的感知与决策能力；车路协同场景下，边缘系统支持车辆与基础设施、其他车辆的实时信息交互，实现协同驾驶，提升道路整体通行效率 30% 以上。

➤ 商业价值

车企采购与服务订阅模式：汽车制造商采购 AI Edge 智能驾驶解决方案，按车辆搭载量向技术供应商付费，以提升车型智能化水平与市场竞争力；车主可按需订阅更高级别的智能驾驶功能服务（如特定场景下的全自动驾驶），增加车企售后收入来源。

数据服务与广告模式：基于 AI Edge 收集的驾驶数据（脱敏处理后），可向保险公司提供精准风险评估数据，实现差异化车险定价；同时，在合规前提下，根据用户驾驶习惯与偏好推送个性化广告服务，创造新的商业价值增长点。

➤ 社会价值

提升交通安全与出行效率：智能驾驶技术可减少人为驾驶失误导致的交通事故，预计可使交通事故死亡率降低 80%；通过优化交通流，缓解城市拥堵，人们的日常通勤时间有望缩短 20%-30%。

释放社会生产力：自动驾驶在物流运输领域的应用，可使货车司机劳动强度大幅降低，同时提升运输效率，降低物流成本 15%-20%；此外，为老年人、残障人士等特殊群体提供出行便利，拓展其社会活动范围。

3.8 应急通信与保障

3.8.1 场景描述

应急通信场景中，AI Edge 的“通感智算控”技术体系构建“全域感知 - 智能组网 - 动态调度 - 精准响应”的应急保障闭环。当地震、洪水等灾害导致传统通信基础设施瘫痪时，可快速部署无人机基站、便携边缘节点等应急设备，通过多频段通信（如卫星窄带、5G 专网）与环境感知（红外成像、振动传感器）融合，实时获取灾区人员位置、建筑损毁情况及信号覆盖盲区（通感）；边缘 AI 引擎基于实时数据生成最优组网策略（如自组织 mesh 网络拓扑），并通过边缘算力快速完成信道资源分配、负载均衡计算（智算）；同时，系统动态控制设备发射功率、切换通信频段，优先保障救援指令传输与生命探测信号，实现“救援人员 - 指挥中心 - 受灾群众”的低时延通信（控）。

该场景可支撑多样化应急需求：如地震救援中，边缘节点通过 AI 识别被困人员手机信号特征，引导救援队精准定位；洪水灾区通过无人机基站与地面边缘节点协同，构建临时通信覆盖网，保障救灾物资调度指令实时传达；疫情封控区通过边缘算力分流健康码核验数据，避免网络拥塞影响应急医疗通信。

3.8.2 潜在价值分析

AI Edge 通过“通感智算控”的深度协同，将应急通信从“被动抢修”升级为“主动感知、智能响应”的现代化体系，其核心价值不仅在于技术性能的突破，更在于通过高效通信保障挽救生命、减少损失，同时构建“政府主导、市场参与”的可持续运营模式，为国家应急管理体系现代化提供关键技术支撑。

➤ 技术价值

突破传统应急通信“响应慢、覆盖弱”的痛点：AI Edge 的自组织网络技术使应急通信部署时间从数小时缩短至 15 分钟，信号覆盖半径扩展至 5 公里，在复杂地形下通信中断率降低 80%；边缘 AI 的智能流量调度可将救援指令优先级提升至 99%，确保关键信息零延迟传输。

实现“资源自适应 - 风险预判”的智能化：边缘系统通过分析环境感知数据（如余震频率、水位涨幅），提前调整通信参数（如增强抗干扰编码），使极端条件下通信可靠性提升至 95%，解决传统应急通信“被动应对”的局限。

➤ 商业价值

政府采购 + 服务外包模式：应急管理部门通过招标采购 AI Edge 应急设备及年度运维服务，按设备部署次数与通信时长付费；电信运营商提供“应急通信专网 + 边缘算力”服务，承接政府与企业的定制化保障需求（如大型活动备用通信）。

设备租赁 + 技术授权模式：设备厂商向消防、武警等专业救援队伍出租便携式边缘基站，同时向行业客户授权 AI 组网算法，按设备型号与授权范围收费，单套系统年均收益可达数十万元。

➤ 社会价值

提升灾害救援效率与生存率：在地震救援中，AI Edge 技术可使被困人员定位时间从平均 6 小时缩短至 1 小时，救援成功率提升 40%；在突发公共卫生事件中，保障医疗物资调度通信畅通，使应急响应速度提升 30%。

降低社会损失与治理成本：通过快速恢复关键通信，减少灾害导致的信息孤岛问题，间接降低经济损失(如 2023 年某省洪灾中，应急通信保障使农业损失减少超 2 亿元)；同时为基层政府提供“平急结合”的通信能力，平时可用于森林防火、地质监测等日常管理，提升资源利用率。

3.9 智慧体育

3.9.1 场景描述

近年来，智能穿戴设备、5G 赛事直播、AI 运动处方等新兴技术不断推动体育与科技的深度融合，“智慧体育”正逐步成为全民健身、竞技体育和体育产业发展的重要引擎。以篮球运动为例，场馆内的运动员、裁判、教练等智能体（Agent）需要在高速动态环境中进行实时互动和协同决策，这对网络的低时延传输和边缘 AI 的快速推理提出了极高要求。

在智慧体育场景中，传统云 AI 架构由于链路较长、端到端时延大，难以满足该场景实时性需求。例如，职业运动员的平均反应时间需控制在 150ms 以内，普通人约为 250ms，如果采用云 AI 的架构，裁判的判罚、教练的战术调整或运动员的动作信息传递存在过大延迟，将直接影响比赛公平性和战术效果。

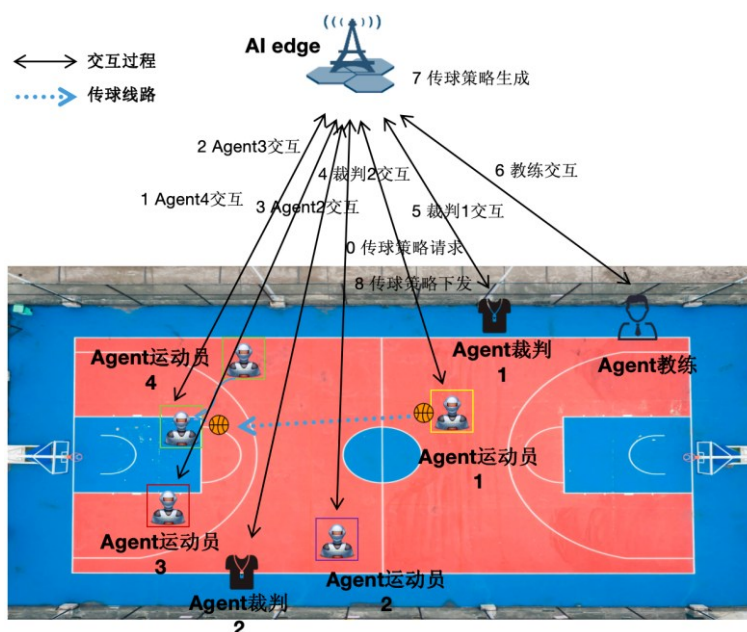


图 6 AI edge 赋能智慧体育的一个示例

AI Edge 通过将 AI 推理与感知能力下沉至 RAN 侧，可以实现通信与计算的深度融合，是应对该类场景的一项有潜力的技术。RAN 不仅承担数据转发，还能够直接调用场馆环境中的实时感知信息，并与多个终端/Agent 进行高频交互，相当于一个集群决策“大脑”。如图 6 所示，运动员之间的传球、跑位信息可通过快速共享；裁判的判罚信

息可通过 AI edge 即时传递给运动员和教练；教练的战术调整可通过 AI Edge 在毫秒级时延上发至场上多个运动员。在此过程中，单次交互数据量可达数百 KB，而 RAN 与 MEC 或云端之间的传输往返延迟可能在数十至上百毫秒。通过 AI Edge 在 RAN 侧直接完成多 Agent 之间的交互，可避免信息多跳转发造成的额外延迟。

3.9.2 潜在价值分析

AI Edge 通过在 RAN 侧部署推理和控制能力，显著缩短了通信路径，降低端到端时延，实现比传统 MEC 和 Cloud AI 更快的响应速度。由于 AI 模型和计算资源更接近终端，通信链路更短，时延更低；同时将通信、AI 推理与控制逻辑一体化，有效降低了系统总时延。比如采用边缘计算架构可大幅减少传输延迟，在真实场景中可将体验延迟控制到几十毫秒以内，远优于云端或远端 MEC 的数百毫秒级延迟，能够满足智慧体育等场景对实时性的严格要求。运动员、教练、裁判之间可以在极低时延下共享场馆视图和状态信息，实现快速协同决策和精细化控制。例如，高速运动过程中可通过 AI edge 实时识别并预测动作轨迹，使裁判判罚和教练指令更加及时准确。

➤ 技术价值

AI Edge 的技术价值主要体现在如下几个方面：1) 端到端时延显著降低：云端往返延迟约 80ms，RAN 侧 AI Edge 可降至 20ms，节省约 60ms，满足运动员 150ms 反应阈值要求；2) 带宽占用优化：以 16 个 Agent、每秒 10 次交互、每次 200KB 计算，每秒流量约 31.25MB/s（约 250Mbps），单场比赛约 88GB，通过 RAN 侧本地处理可将回传流量减少 10 - 100 倍；3) 鲁棒性提升：减少对远端链路的依赖，提升关键判罚与控制功能在弱网环境下的可靠性；4) 实时协同能力：RAN 侧 AI Edge 能快速融合多 Agent 数据，实现毫秒级传球推荐、危险预警和裁判辅助判定。

➤ 商业价值

针对智慧体育场景，AI Edge 可以从如下几个方面带来商业价值：1) 赛事服务增值：观众付费观看低延迟战术视角，假设 1000 人付费 2 美元/场，每场可增收 2000 美元；2) 成本优化：AI Edge 可显著减少带宽和云计算支出；3) B2B 订阅服务：为教练、裁判、俱乐部提供低延迟决策系统；4) 广告与 AR 转播增值：低时延处理支持更丰富的可视化内容与广告嵌入；5) 系统销售与运维服务：RAN/Edge 软硬件部署及长期 SLA 可形成稳定收入。

➤ 社会价值

借助 AI Edge，低延迟辅助裁判系统可减少 40% - 80% 关键误判，从而实现比赛公平性提升；实时监测心率与轨迹，降低碰撞与伤害风险，运动安全性能显著提高；智慧体育下沉公众场馆，提升运动体验与参与率，有利于全民健身普及；此外，通过本地化处理与差分隐私机制能有效保障数据安全。

3.10 机械导盲犬

3.10.1 场景描述

机械导盲犬应用场景依托 AI Edge 的“通感智算控”技术体系，构建“环境感知-智能规划-动作执行-反馈优化”的闭环控制，聚焦视障群体日常出行、跨场景导航等核心需求。

我国约有 1700 万名视障人士，却仅有约 400 只导盲犬，视障人士独立出行仍是亟待解决的社会课题之一。虽然机械导盲犬等设备在研究中展示出潜力，但是成本高、续航短以及在安全方面受到质疑。

得益于 AI Edge 的多维度环境感知，弹性可拓展的边缘算力网以及云边端的高效协作，可促进机械导盲犬设备本身的轻量化，经济性，延长终端电池寿命，提升安全性，使其大规模落地成为可能。具体来说，机械导盲犬终端集成摄像头、激光雷达与声感传感器，融合 AI Edge 以及 5G-A/6G 的环境感知技术，建立全方位多层次的环境感知，构建全景环境视野，远超单一设备的视觉局限，提高其安全可靠。机械导盲犬终端动态向 AI Edge 边端卸载计算任务，持续对数据进行分析与风险判断，例如自动区分如静止护栏与移动的自行车、识别绿灯倒计时，边端的分析结果可快速回传至机械导盲犬执行机构（驱动轮、语音提示模块），实现精准转向、紧急制动与实时语音播报，如“前方 3 米有积水，建议绕行”。通过向 AI Edge 卸载这些数据和任务，大幅削减终端本地算力负载，从而降低终端的成本，有望使更多的残障人士能够以低廉的价格得到可靠的出行辅助；让设备更轻量化、低功耗，避免传统导盲设备因高功耗导致的续航短板。

AI Edge 可支撑机械导盲犬全场景出行需求：基础场景覆盖城市人行道行走、红绿灯识别与过街引导；进阶场景包括商超内货架定位、地铁无障碍电梯指引；特殊场景针对雨天积水、雾霾低光等恶劣环境优化感知能力，补充视觉识别盲区，全面适配视障群体日常出行场景。

3.10.2 潜在价值分析

AI Edge 赋能机械导盲犬重构视障群体智慧出行服务路径，其核心价值不仅在于提升导盲设备的智能化与可靠性，更在于通过网络赋能终端的模式打破传统辅助工具在使用门槛，续航上的局限，推动智能服务走向大众普惠。

➤ 技术价值

突破传统导盲设备的算力瓶颈和高功耗痛点：通过向边缘算力网络侧动态卸载复杂计算任务，机械导盲犬终端本地算力负载削减超 70% 以上，从而避免在终端配备高昂的计算芯片，预计可将机械导盲犬成本降低 30% 以上。

续航时长从传统设备的 1-2 小时延长一倍，将一次充电的巡航距离从 2km 扩大到 5km，活动半径扩大一倍，可以满足视障人士长时间外出需求，如购物，从家到地铁站、再到工作地点。

依托 AI Edge 的通感算协同能力，机械导盲犬可实时根据所处环境需求，实时分析并自动接入基站侧更新的 AI 模型，如新增的“共享单车道避让”“施工路段绕行”模型，无需用户手动升级终端硬件或软件，即可持续获得更全面的安全保障能力，让设备智慧水平随网络进化而提升。

➤ 商业价值

借鉴终端的商业模式，可形成以下两种商业模式：1) 运营商向购买设备的用户交付机器人导盲犬并提供免费通信与计算套餐，模式类似于签约手机；2) 用户直接从制造商购买机器人导盲犬，运营商提供设备运营支持，并向用户销售对应通信与计算服务。多样化的商业模式能够满足不同用户需求，同时降低终端用户的使用门槛，进一步提升方案的可扩展性。对运营商而言，通过 AI 模型运营、算力租用及增值服务，构建长期稳定的收益体系；对终端制造，可通过开放的网络能力降低终端成本，打破设备能力局限；对于公益机构，通过运营商和设备商的合作实现能力互补，可实现技术的快速迭代，可靠的身份认证等，促进落地公益帮扶项目。最终形成“运营商-终端厂商-公益机构”协同共赢的产业格局。

拓展多元化商业生态：基于开放架构，第三方服务商可开发增值功能，例如家人实时位置共享功能，通过运营商平台接入机械导盲犬服务，形成“基础服务+增值服务”的商业生态。

➤ 社会价值

探索视障群体借助科技手段实现独立出行的可能性：借助 AI Edge 能力实现低成本，高续航，轻量化的机械导盲犬，促进机械导盲犬能够真正意义上的落地和推广。AI Edge 赋能机械导盲犬有望大幅度提高视障群体独立出行率，帮助他们自主完成购物、就医、通勤等日常活动，减少对家人、志愿者的依赖，显著增强社会参与感与个人自信心。

实现智能民生服务的普惠化落地：相比全国仅约 400 只的传统导盲犬，机械导盲犬可通过规模化生产与低成本套餐模式快速普及，预计未来 5 年可惠及数十万名视障用户，有效填补传统导盲工具的供给缺口；同时，其轻量化、低功耗特性适配老年、儿童等不同年龄层视障用户，进一步扩大服务覆盖范围，助力无障碍社会建设。

4. AI EDGE 的技术方向与主要挑战

4.1 系统架构

AI Edge 网络的系统架构划分为四个部分，包括：分布式节点、超级边缘节点、核心节点，以及边缘智能算力编排与管理。该体系架构遵循“边缘自智、分层部署、全域协同”的理念，确保在算力分布、智能调度和服务承载等方面形成高效、灵活、韧性且弹性可扩展的体系，满足未来终端用户多样化边缘智能应用的需求。同时，该体系还兼容特定的边缘网络能力开放，无线算力资源共享，以满足垂直行业用户多样化智能应用需求。

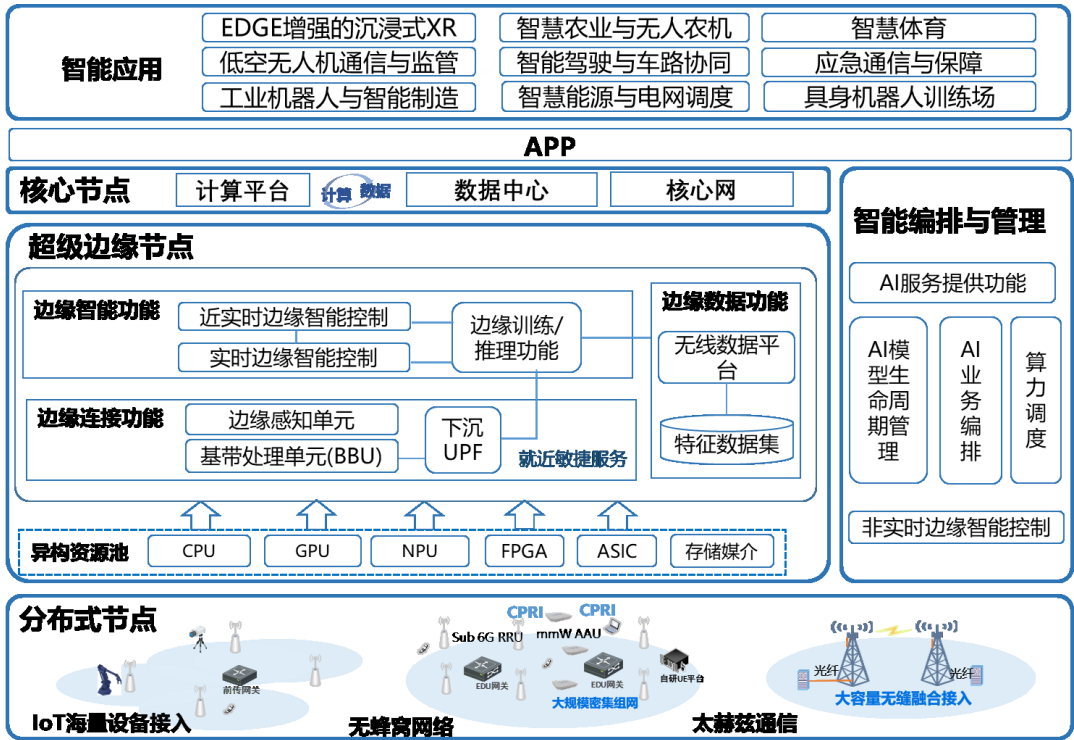


图 7 面向 DOICT 融合的 AI Edge 系统架构

- **分布式单元**位于无线接入网的最前端，是 AI 能力下沉到“最后一跳”的关键载体。该单元通常与分布式处理单元(DU)、小基站、远端射频单元(RRU/AAU)深度集成，具备近端数据采集、信号预处理和轻量级 AI 推理能力。通过将部分 AI 功能下沉至接入侧前端，无线系统能够实现毫秒级的响应速度，支撑可穿戴设备、车联网等对超低时延敏感的应用场景，从而有效降低链路开销，提升用户体验。
- **超级边缘节点**覆盖传统无线接入网络和部分核心网下沉功能，是承接来自终端与分布式单元数据处理的重要计算平台。该节点可基于 CPU、GPU、NPU、FPGA 等异构算力资源，构建边缘连接功能、边缘智能功能以及边缘数据功能，实现在本地完成多模态数据的实时处理与智能推理。在运行方式上，超级边缘节点具备本地边缘自治与跨域协同能力，既能在本地自适应调度算力和模型资源，在数据不出域的前提下完成就近敏捷服务，又能通过跨域协作机制与其他节点协同交互，形成层级化智能服务网络。这一层确保了业务在复杂环境中的连续性与稳定性，同时满足大规模智能应用对实时性和可靠性的严苛要求。
- **核心节点**通常部署在大规模数据中心或云平台，具备更强大的计算与存储能力。核心节点主要承担 AI（大）模型训练与优化、跨区域推理和全局数据汇聚分析等任务。它能够基于来自边缘节点的业务反馈，周期性地完成模型的迭代与优化，并通过统一接口将优化后的模型下发至超级边缘节点和分布式单元，从而形成边缘—核心的模型迭代闭环。同时，核心节点负责网内各种 AI 任务的执行策略协同和模型数据拉通，对来自终端用户和行业用户的各种 AI 服务做统一的认证纳管计费等。
- **智能编排与管理**贯穿整个 AI Edge 体系架构，为网络提供智能化的控制与调度能力。它主要包括：非实时智能控制、模型管理、业务编排、算力调度以及智能服务提供功能。在算力层面，编排系统能够在分布式单元、超级边缘节点和核心节点之间进行网络和算力资源动态调度。在服务层面，编排机制支持 AI 模型的快速部署、在线优化与跨场景迁移，使不同业务灵活匹配网络 AI 能力。在数据层面，编排系统通过边缘采集与预处理、核心优化与反馈的闭环机制，推动 AI 服务持续进化与迭代。管理与编排功能不仅显著提升了 AI Edge 网络的运行效率，还确保了多层级智能之间的有机协同，共同赋能 6G AI 应用。

上述架构具有如下特色和优势：

（1）边缘自智：不同于传统依赖中心云的被动调度模式，AI Edge 能够在边缘节点实现就近感知、就近决策和就近执行，从而显著降低时延并提高服务的确定性。依托边缘节点具备自感知、自诊断和自优化的能力，可以在数据不出域的同时，在本地独立处理智能任务，实现就近敏捷服务。这种边缘自智的方式不仅提升了网络的韧性和容错性，也为关键行业应用，例如工业控制、智能可穿戴设备、低空无人机等场景，提供了低时延、高可靠性、高数据安全性保障。

（2）跨域协同：AI Edge 通过统一的智能管理与编排系统，实现异构网络资源和计算资源的协同调度，能够同时支撑传统通信服务和新兴 AI 应用，如工业互联网、具身智能、低空智联网等。其核心在于对全域环境的动态感知和数据驱动的智能优化。AI Edge 能够综合无线网络环境、业务特性、网络负载以及用户习惯等，对数据进行全生命周期管理和利用，并动态匹配分布式节点、超级边缘节点以及核心节点不同层次的算力与 AI 模型。实现全局资源优化、提升网络算力资源利用率，通过端-网-云协同，赋能多样化智能应用。

（3）异构算力融合：AI Edge 架构在算力层面具有高度的开放性与包容性，能够无缝融合 CPU、GPU、NPU、FPGA、SoC 等多种异构计算资源。通过软件可编程和虚拟化技术，AI Edge 将通信、AI、感知、网络控制和计算服务统一在同一硬件底座上，实现算力与网络能力的

深度共享。这种异构算力融合不仅提升了资源利用效率，还推动了无线接入网（RAN）向无线计算网络（RCN, Radio Computing Network）的演进，使得网络从单一的通信基础设施逐步演变为融合计算与智能服务的综合性平台。（4）**智能可编程**：通过开放的接口以及 AI 驱动的功能框架，AI Edge 能够支持多样化业务需求的快速定制与部署。网络功能和应用可以通过软硬件解耦、模块化设计以及可重构机制，进行按需组合与动态优化。这种智能可编程能力为创新应用的孵化与落地提供了强有力的支撑，例如 ToB 智慧工厂场景下的定制化控制、智慧医疗的实时诊断服务，乃至未来 6G 时代的沉浸式交互和具身智能应用等。

4.2 AI for Edge 技术

4.2.1 AI for Edge 的兴起

随着 5G 和 6G 通信技术的快速演进与物联网（Internet of Things, IoT）设备的爆发式增长，边缘侧数据量呈指数级上升，实时性业务需求日益凸显，推动着边缘计算与人工智能的深度融合。在此背景下，AI for Edge 应运而生，成为实现智能边缘的关键路径[20]。AI for Edge 是指将 AI 技术深度嵌入到边缘环境中，通过在边缘节点上进行智能感知、实时建模、决策优化与资源调度，从而实现低时延、高可靠、高效率的智能服务的一种技术范式。它旨在应对边缘侧数据量激增、业务实时性增强、网络环境复杂多变等挑战，通过 AI 赋能边缘计算，实现对通信、计算、存储等多维资源的智能协同与优化[21]。下文将从技术驱动力、核心需求与性能提升的关键路径三个维度，系统阐述 AI for Edge 的技术挑战与实现方法。

4.2.2 AI for Edge 的核心价值

➤ 动态复杂环境下的实时建模能力

在边缘智能场景中，环境的动态性和复杂性是普遍存在的挑战。无线信道会受到多径效应、用户高速移动、遮挡和外部干扰的共同影响，导致信道特性呈现出强时变和高度不确定性。同时，业务层的流量也存在显著波动，例如短视频、在线游戏、虚拟现实（Virtual Reality, VR）、增强现实（Augmented Reality, AR）等业务的突发流量需求，使得系统的负载在短时间内发生剧烈变化。传统基于数学模型或静态假设的方法往往难以在这种快速演化的环境下保持有效性。而 AI 模型凭借强泛化和在线学习能力，能够实时捕捉复杂环境下的变化规律，实现高效的建模与自适应优化。

➤ 非线性器件补偿与信号保真能力

无线空口性能直接决定边缘网络的吞吐率、时延与稳定性。在高频宽带、多天线和复杂电磁环境下，信号传输不仅受多径衰落与干扰影响，还受到射频器件（如功率放大器、ADC/DAC、天线阵列等）非线性的限制，导致系统性能下降。传统基于解析模型的补偿方法难以应对硬件特性变化及环境动态。AI 技术通过数据驱动的建模与预测能力，为空口性能优化提供了新的手段。AI 模型可从实测信号中学习射频器件的非线性特征，实现数字预失真（DPD）、IQ 失衡补偿与量化噪声抑制等功能，并能在温度漂移或器件老化等条件下实现自适应修正。此外，AI 基于时间序列建模与强化学习方法，可对信道状态信息（CSI）与硬件性能进行联合预测，提前识别链路退化趋势，动态调整调制编码、功率控制与波束赋形策略，实现空口层的前瞻优化。AI 赋能的非线性补偿与空口优化能力能够在复杂硬件与信道条件下保持链路稳定性与高效性，推动无线系统向自感知、自优化和自愈方向演进。

➤ 复杂高维优化问题的求解能力

在边缘智能系统优化中，优化目标通常是多维度的，既需要满足低时延和高可靠性的服务需求，又要兼顾能耗效率、频谱利用率等多重因素。这类优化问题往往具有非凸性、组合爆炸和约束条件复杂等特征，使得传统解析方法或单一启发式算法难以在有限时间内求解。尤其在面向大规模用户、多种业务类型及多样化资源的场景下，问题维度呈指数级增长，导致传统方法陷入维度灾难。而 AI 技术通过强化学习与基于神经网络的函数逼近方法[22]，能够实现高维空间中的高效探索，快速生成接近最优的解决方案，并通过持续学习提升求解性能。

➤ 网络状态预测与主动运维能力

网络运行过程中的状态变化具有显著的时序相关性与突发性，例如链路质量因干扰或用户移动而波动，业务流量因热点事件或应用行为而出现瞬时拥塞。传统的网络运维依赖被动监测与事后调整，往往存在响应滞后，难以满足高可靠、低时延的业务需求。而 AI 技术凭借其时间序列建模优势，能够利用长短期记忆网络（Long Short-Term Memory, LSTM）、门控循环单元网络（Gated Recurrent Unit, GRU）、Transformer 等深度学习模型对网络状态进行精准预测。同时，结合强化学习[23]等方法，AI 能够实现主动运维，即在预测到链路劣化或流量突增之前，提前执行资源重构、路径切换或负载均衡操作。

➤ 快速策略生成与部署能力

在边缘网络中，资源调度、切片编排、任务卸载和干扰管理等任务往往需要在秒级甚至毫秒级内决策与下发。随着用户移动、业务突发和干扰环境的不断变化，传统依赖静态规则或离线优化的方法难以及时响应，容易造成资源利用不足和服务质量下降。此外，边缘部署对跨场景的适配能力也提出更高要求。传统算法通常依赖场景特定的信道模型与参数调优，一旦环境或业务类型变化，性能便会显著下降。而 AI 技术具有对复杂动态环境的持续学习能力，能够学习关键特征并在新场景中迅速迁移经验。这样的泛化与自适应特性，不仅大幅降低了系统迭代与维护成本，也为大规模异构边缘网络的智能演进提供了核心支撑。

4.2.3 边缘网络对 AI 的核心需求

➤ 具备轻量化与高效推理能力

边缘节点因计算能力、内存容量和功耗预算受限，要求部署的 AI 模型具备轻量化与高效推理能力。为此，业界常采用模型剪枝、模型量化、知识蒸馏和稀疏化等方法。其中，模型剪枝技术可采用结构化剪枝，移除卷积核或通道等完整组件；或采用非结构化剪枝，以单个权重或神经元为单位进行裁剪[24]。模型量化通过降低权重和激活值的数值精度，减少存储需求并提升计算效率，常见实施阶段包括训练后量化、量化感知训练及微调等。知识蒸馏则利用复杂教师模型指导轻量学生模型，通过学习其输出软标签或中间表示，在保持模型轻量的同时提升推理精度。稀疏化方法通过训练或推理阶段引入零值权重或零值激活减少无效计算。此外，针对分布式协作场景，多客户端的拆分联邦学习框架[25]通过模型拆分与优先级调度机制，在降低客户端计算负载的同时实现异构设备的协同训练，进一步扩展了轻量化技术的应用边界。

➤ 支持大小模型协同范式

在边缘智能网络中，算力、能耗和存储等资源受限和不同任务对模型的需求差异显著，导致单一的 AI 模型难以同时兼顾所有需求，因此需要构建“通信通用基础大模型 + 下游任务小模型”的协同范式，允许在不同应用场景中灵活部署和更新下游专属小模型，而无需重新训练或大幅修改通信通用基础模型。首先，利用大规模通信数据训练通

信领域的预训练大模型。在此基础上,通过如低秩自适应(Low-Rank Adaptation, LoRA) [26]、适配器模块添加、提示调优等参数高效微调方法,快速派生出面向特定任务的轻量小模型。进一步借助容器化和虚拟化技术,还可以推动模型即服务(Model as a Service, MaaS)在边缘智能生态的落地,为不同应用场景动态提供所需的智能能力,加速边缘智能的普及与规模化应用。

➤ 具备多模态感知与跨层融合能力

边缘网络需要同时处理来自物理层的信号数据、业务层的服务质量(Quality of Service, QoS)指标,以及外部环境感知信息(如雷达、摄像头)等多源异构数据。因此,AI模型必须具备多模态感知与跨层融合能力。该类模型能够对信道状态、业务负载、环境感知等多维度信息进行统一建模与联合分析,实现跨层级的数据融合与智能决策。例如,在通信与感知一体化场景中,通过融合通信信号与感知数据,可有效提升定位精度、链路稳定性及资源调度效率,从而增强整体网络性能[27]。在如图8所示AI驱动的数字孪生平台中,针对无线场景应用,通过融合3D地图、射频测量等数据,可大幅降低真实环境中信道探测的开销,支撑AI模型的高效训练与验证。



图 8 元图工坊数字孪生使能平台

➤ 满足可解释性与可测试性要求

边缘网络需满足高可靠要求,尤其是在运维和安全等关键场景。因此,AI模型必须具备可解释性,能够提供决策逻辑的可视化与因果分析,支持运维人员理解和追溯模型行为。同时,AI模型需具备可测试性,通过系统化验证框架对性能、鲁棒性和安全性进行量化评估。在实验室环境中,可通过信道扰动、业务突发和攻击模拟等方式检验AI模型的稳定性和容错性,实现可信与可靠落地。

4.2.4 AI for Edge 性能提升的关键路径

➤ 无线信道高效表征

为支持低时延具身智能应用,AI Edge网络必须具备超高可靠通信、实时信道预测与链路自适应、高精度定位、快速的环境感知与重构等关键能力,而这些能力的达成都高度依赖于以低成本的方式获取实时的信道信息和实现精准的信道预测。传统做法基于测量获取信道状态信息,再根据预设的先验假设实现内插或外推,其主要问题在于:随着系统复杂度的提升(例如天线端口数的激增),测量多带来的开销将无法承受;另一方面,预设的先验假设往往难以适配多样化的动态场景。近年来兴起的数据驱动方法利

用预先采集的收发信号对训练深度神经网络模型，通过数据拟合信道变化规律，能够在一定程度上解决上述问题。但已有的技术方案大都基于面向特定任务的专属 AI 模型，缺乏多任务和跨场景泛化能力，此外，也未充分融入多模态信息，难以有效支撑感知类应用。为此，在 AI Edge 网络中，可以构建多模态多任务无线信道基础 AI 模型，通过在大规模、多模态的数据集上（包括信道测量数据和环境感知数据，后者又可包含激光点云、毫米波雷达、电子地图、基站感知数据、图像/视频、GPS 等）进行预训练，从而形成对无线信道的统一高效特征表达，再通过微调等技术将其与下游任务模型或算法进行适配或设计面向下游任务的独立任务适配头，以实现跨任务跨场景的泛化[28]。

➤ 无线空口智能优化

无线空口作为连接用户与网络的首要环节，其链路质量直接决定系统的容量、时延和可靠性，是提升无线网络性能的关键途径。在智能波束管理与赋形上，通过动态调控天线阵列的相位与幅度，实现信号定向传输。系统结合用户位置与信道环境的实时感知，优化波束指向与宽度，支持单用户高效覆盖和多用户并行调度，已广泛应用于毫米波通信和超密集组网等场景。在 AI 赋能的信道编码和信号检测上，利用深度学习自适应生成匹配信道环境的编码方案，降低误码并简化复杂度[29]；学习信道失真特征，提升信号恢复能力，已在车联网、卫星通信等动态场景中得到验证。在自适应调制编码策略优化上，通过实时预测信道质量，动态调整调制方式与编码速率，实现信道状态与传输效率的匹配优化。相比传统策略，融合信道预测、QoS 驱动的资源调度及多用户协同分配机制，可以提升调度灵活性与系统效率，已成为多业务场景下关键通信技术。同时，需探索目标导向的 CSI 压缩反馈、机理与数据双驱动接收机设计及无导频传输方案，以适配高动态场景的空口联合优化需求[30]–[31]。更进一步，可以考虑使用端到端训练的神经网络模型替代传统的信道估计、均衡、解调等模块，使系统从原始接收信号直接预测发送符号，实现接收机多模块乃至整体的智能化。

➤ 资源联合调度

在如图 9 的边云协同场景下，终端、边缘与云计算融合带来多维异构资源调度挑战。需构建预测驱动、感知协同、策略智能的资源联合调度机制，实现频谱、计算、存储、通信等资源的按需优化分配。基于 AI 预测模型，提前感知用户位置、信道状态与负载，动态预分配频谱资源，提升效率与通信连续性。通过资源感知与 AI 预测，构建计算、存储、通信一体化调度模型，支持动态切片与资源隔离，满足多样化业务需求。基于强化学习实现智能负载均衡，融合边缘与中心节点状态，自适应调度任务与流量，提升服务质量和系统弹性。

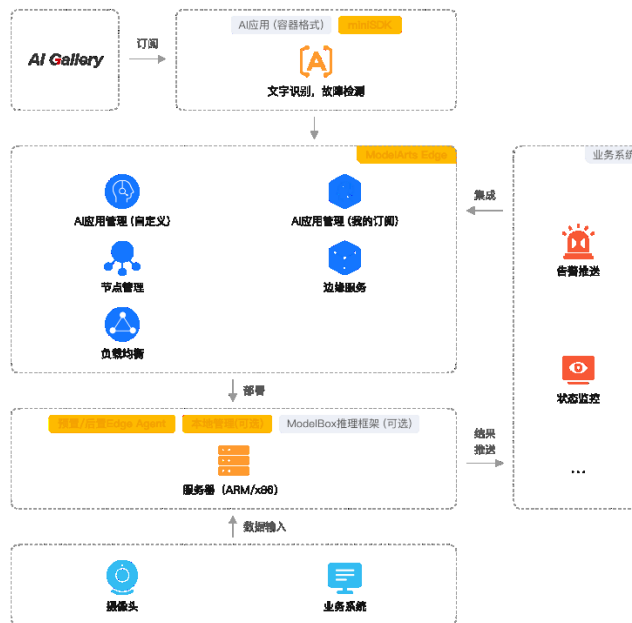


图 9 ModelArts Edge 使能的边云协同场景

➤ 智能运维与安全

随着边缘智能网络日益复杂，传统依赖人工和静态规则的运维模式难以应对动态场景与复杂威胁。AI 的引入正推动网络运维向智能化、自适应和闭环化演进。在网络安全方面，AI 可通过深度学习快速识别异常流量，提升检测精度与响应速度。面对新型攻击，AI 模型相比传统方法更具适应性。同时，对抗训练等鲁棒学习技术可增强模型抗攻击能力，保障智能运维系统的稳定性[32]。在能效管理方面，AI 可基于流量预测动态调整设备状态，实现节能与性能的平衡。例如，通过智能休眠机制，边缘节点可在低负载时降低能耗，提升绿色运维水平。在运维层面，AI 可通过多源数据建模实现故障自动检测与根因分析，并结合自适应恢复机制，实现自愈式闭环运维，显著提升网络鲁棒性与可用性。

4.2.5 AI for Edge 模型与算法的测试验证

为验证 AI for Edge 技术在边缘环境中的有效性、可靠性与效率，需建立系统化测试体系，重点评估模型轻量化效能、动态鲁棒性及组件级性能增益。

➤ 模型轻量化与效能基准测试

在标准化硬件平台上评估模型轻量化水平，核心指标包括剪枝/量化后的精度-复杂度权衡（以 FLOPs、模型大小衡量），及其在边缘芯片上的推理时延与能效比，筛选满足资源约束的高效模型。

➤ 动态环境鲁棒性验证

通过注入信道噪声、数据漂移及对抗样本，系统评估模型在动态扰动下的性能保持度，并结合资源扰动模拟（如 CPU/内存波动），检验其在算力不稳定环境中的稳定性。

➤ 组件级算法性能增益评估

在实验室环境中，通过空口模拟器与网络损伤仪构建高保真无线环境与端到端业务链路，验证 AI 算法性能。具体包括：a) 在高速移动场景中，通过空口模拟器验证智能波束管理算法在频谱效率与跟踪精度上的增益；b) 通过网络损伤仪注入时延、抖动和丢包，结合业务流量生成器模拟 VR/AR、工业互联网等多类业务，评估 AI 调度策略在时延与资源利用率上的优化效果。

➤ 在线学习与协同能力测试

针对持续学习算法，验证其在线自适应速度与抗遗忘能力；在联邦学习等协同场景中，评估多智能体间的策略一致性与通信效率。

该模型级测试以“可重现、可量化、可追溯”为原则，为算法选型提供客观依据，是集成至 AI Edge 系统平台的前置条件，其结论直接支撑全栈系统级测试（4.5.2 节详述）。

4.3 AI over Edge 技术

AI over Edge 技术旨在复用移动通信网络的通信、计算和存储资源，实现 AI 服务在边缘侧的就近部署。与云侧提供的 AI 服务相比，Edge AI 的独特之处不仅在于其低时延和有效保护用户数据隐私的能力，更在于其所能提供的 AI 服务的类型与云侧有本质不同。具体来讲，AI Edge 所支持的 AI 应用并非单纯的信息检索、内容生成、任务规划等，而是借助网络的连接和感知能力与物理世界充分交互，实现“感知、推理与执行”的快速闭环，从而原生地支持具身 AI 应用；此外，AI Edge 能够通过实际数据反哺 AI 模型，推动 AI 发展范式迈向可实时感知与推理的 AI。

目前的移动通信网络尚不具备原生支持 AI 应用的能力，主要存在三大缺失。**第一，数据缺失：**移动通信网络虽然数据量巨大，但输入模态单一，主要是各种 RF 测量结果，例如 CQI/PMI/RI、CSI/DMRS、移动性测量等，但缺乏多模态数据，例如：移动轨迹、业务特性、流量分布、高精地图、气象信息、用户画像、摄像头和传感器数据等，这会严重限制部署于 Edge 的 AI 模型的表征、理解与生成能力；**第二，智能缺失：**移动通信网络是基于规则的，其运转的基本逻辑是，根据测量值，基于预设的规则和静态的策略进行决策（例如：符号判决、MCS 自适应、信道估计，等等），但 AI 应用都是基于智能的，它们的运转逻辑是，基于感知实现理解和生成，这意味着，必须在现有移动通信网络中植入智能化的处理模块；**第三，记忆缺失：**移动通信网络缺乏短期记忆（例如：A 小区的所有行为与 B 小区的所有行为完全独立，每次新的 RRC 请求甚至新的会话请求之间完全独立）和长期记忆（例如：潮汐效应、上下班所带来的周期性变化，以及赛事、演唱会、直播等所带来的稳态趋势），这导致无法在边缘侧完成感知、分析、决策与行动的完整闭环，也不可能通过和环境的互动来实现反思及迭代优化。

基于上述分析，为了实现 AI over Edge 的愿景，需要首先解决多模态感知与融合问题；在此基础上，在边缘引入 AI 模型，并充分利用单点优化技术（如模型轻量化）和云边端协同架构实现边缘侧 AI 的高效训练和推理；更进一步地，为了打通从感知到决策再到执行的闭环，研发 AI Agent 技术，通过赋予边缘网络以记忆、学习与协作能力为 AI 应用提供支撑；最后，围绕具身智能应用，解决信息服务的编排开放、生命周期管理、跨层资源优化以及数据安全和隐私等问题，实现端到端的服务质量保障。

下文将分别介绍 AI over Edge 所涉及的关键技术以及所面临的挑战。

4.3.1 多模态感知与融合处理

为了在边缘侧支持 AI 应用，需要引入多模态感知和处理能力，首先应集成无线感知与通信，利用跨设备的联合调度实现更高质量的数据采集；此外，还应该能够本地处理多源感知数据。对自动驾驶等场景的分析显示，单车安装的多摄像机、毫米波雷达和激光雷达每秒可产生约 2.3 GB 数据，传统做法将这些数据送往云端处理，需要几十毫秒以上的延迟，而边缘计算能够在本地完成计算和决策，借助轻量级传感器融合算法与嵌入式数据预处理技术，可在资源受限条件下提取高质量特征，提升感知的鲁棒性与准确性。更进一步地，在数据产生之初就进行滤波、降噪、压缩等处理，仅上传或处理有价值的信息，极大减少后续计算和传输的压力，满足 20 毫秒甚至更低的时延要求。在多模态融合方面，重点是解决多模态数据特征的对齐问题；此外，如何保证多模态感知与融合的鲁棒性，在部分能力失效时，系统仍能安全、有效地运行，也是一个重要的问题。

4.3.2 模型轻量化与低时延推理技术

在边缘设备中，计算和存储资源的受限使得大规模深度神经网络难以直接部署，因此模型轻量化技术成为关键路径。通过模型压缩、剪枝、量化与结构稀疏化，可以在保证精度基本不受影响的前提下显著减少模型体量。而知识蒸馏则通过将大模型的知识迁移至轻量模型，从而在计算效率和预测精度之间取得平衡。与此同时，结合硬件友好的算子优化与异构加速芯片，可以进一步提升推理性能，确保模型能够在毫秒级响应用户需求。此外，为了满足不同业务场景下的低时延需求，还需引入动态批处理、自适应算力分配与优先级调度等机制，使得模型推理既能处理突发请求，又能稳定支撑连续任务。下面对模型轻量化涉及到的部分关键技术做一介绍[33]。

➤ 剪枝

剪枝通过移除冗余的神经元连接来减小模型规模。剪枝后的模型参数更少、计算量降低，在理想情况下可提升推理速度。然而，非结构化剪枝产生的不规则稀疏权重矩阵可能难以在通用硬件上高效加速，因此剪枝未必总能转化为实际延迟的下降。为获得真正的低延迟优势，往往需要结合结构化剪枝（例如整层或整滤波器剪除）以及针对稀疏的硬件优化，以确保剪枝后的模型计算可以高效并行执行。尽管如此，剪枝在不显著牺牲精度的情况下大幅减少参数量，对于边缘部署仍然是有效途径之一。

➤ 量化

量化通过降低模型权重和激活的数值精度（如从 32 位浮点降至 8 位整数）来实现模型轻量化。低比特宽度不仅压缩模型所需的存储空间，还利用硬件中高效的定点运算提升计算速度。例如，将模型从 FP32 量化为 INT8 后，模型大小和内存占用可显著降低，同时推理延迟大幅减少，而精度仅有微小影响。实际案例表明，相较于浮点模型，INT8 量化可将模型内存占用降低约 4 倍、推理延迟缩短至原来的三分之一以内，且精度损失可以忽略不计。因此，量化技术非常适合资源受限的边缘设备，在提高能效的同时满足实时性的要求。

➤ 高效模型架构设计

高效模型架构设计也是降低延迟的关键手段。手工设计或利用神经架构搜索(NAS)得到的轻量模型（如 MobileNet 系列、EfficientNet 等）在移动端进行了专门优化，以较少算力实现较高精度，适配边缘硬件的实时推理需求。

➤ 分布式知识蒸馏

分布式知识蒸馏（Distributed Knowledge Distillation, DKD）技术通过在算力充裕的中心侧（如：云）训练强大的教师（Teacher）大模型，并采用蒸馏策略，将大模

型蕴含的知识迁移至轻量化的学生（Student）模型，在各个分布式边缘节点高效部署，实现边缘智能协同。具体的，分布式知识蒸馏有以下两种主流框架：

中心侧蒸馏：即在云/中心侧训练教师大模型，并利用全局数据或代表性数据进行知识蒸馏，得到全局学生模型，最后将全局学生模型分发至边缘节点进行部署与推理。在模型微调更新阶段，边缘节点上传本地采集的数据到中心侧，利用教师大模型制作软标签（soft label），微调更新全局学生模型并下发至边缘节点。下发更新后的全局学生模型时，可采用低秩自适应（LoRA）等模型增量适配技术，仅同步参数增量而非全量参数，实现模型高效同步与个性化，提升边缘部署的灵活性与系统效率。该框架的优势在于能够充分利用中心侧算力与低秩自适应高效传输技术，面临的挑战包括数据隐私保护以及模型个性化问题。

边缘侧蒸馏：即在云/中心侧训练教师大模型，并分发至边缘节点，边缘节点利用本地数据，基于教师大模型蒸馏出学生模型。在模型微调更新阶段，可结合联邦学习技术，实现多节点协同蒸馏，具体实现如下。边缘节点定期将知识摘要（如学生模型的输出分布、logits、少量“知识表示”等）上传至中心侧，该步骤中可根据用户上传知识的具体形态采用相应的数据压缩以及空中计算等高效接入方案，大幅降低数据传输量。进一步的，中心侧对上传的知识摘要进行平均或对软标签等进行集成，生成“聚合知识”并下发至边缘节点，各节点可继续利用聚合知识与本地数据进行本地训练。该框架的优势在于能够保护用户隐私，数据上行传输延迟低，模型个性化程度高，面临的挑战包括设备异构、边缘侧通信计算资源受限等。

4.3.3 云边端大中小模型协同技术

可扩展性是 AI Edge 的核心特征之一。AI Edge 并非单一边缘网络功能实体，它不仅能够在横向维度上跨域整合相邻基站的算力资源，构建弹性可拓展的边缘算力网，也能够纵向维度上通过云边端的高效协作，实现跨层级的分布式 AI，从而支撑移动信息服务在全域网络上的可扩展性。根据具体协作模式的不同，云边端大中小模型协同技术可以分为如下几类：

➤ “决策大模型+执行小模型”的大小脑协同

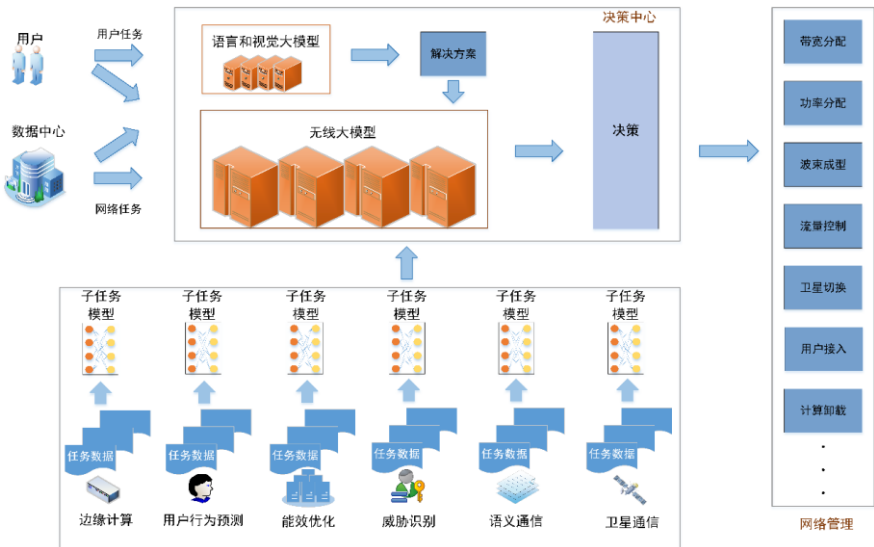


图 10 决策大模型与执行小模型协同示意图

如图 10 所示，在该模式中，大型人工智能模型负责中央决策和用户交互，而无线网络中分布式部署的大量边缘轻量级模型来负责执行具体任务。此外，预计未来大型视觉

和语言模型也会被嵌入到大型人工智能模型中，以增强对用户和运营商需求的理解；同样，无线通信运营商可以根据预定义的功能向大型人工智能模型发送命令，或者利用语音和其他方法实时定制指令。中央大模型能够根据用户输入进行意图理解和任务分解，进而编排调度边缘节点的各种子任务人工智能模型，每个模型都专注于特定的领域，例如边缘计算中的资源优化、语义通信中的内容生成以及卫星通信中的智能调度。因此，未来无线通信中的大型人工智能模型本质上将充当人工智能集群的集合体，根据从各个子任务中获取的信息和用户的交互信息做出决策[34]–[35]。

➤ 切分推理与切分学习

对于较为复杂的推理任务，可以将其在终端设备和边缘节点之间合理划分。例如，使用模型分段推理或早退机制，让终端设备执行前几层网络并在中间层输出达到置信度阈值时直接给出结果，或将中间特征发送至边缘节点完成剩余推理步骤。这种终端与边缘协同的方式能在保证精度的同时减少原始数据传输，降低端到端延迟，实现更快的响应。需要注意的是，切分推理需平衡通信开销和计算节省，并考虑网络状况以选择最佳切分点。为有效地实现切分推理，动态卸载是一项关键技术，它能够实时感知网络带宽、边缘节点负载以及任务本身的计算复杂度等多维因素，利用图优化、强化学习等方法动态决定任务在不同节点间的分配和模型的最优切分方式，从而降低单节点的计算压力并提升整体处理效率。

切分学习的提出是为了通过将模型切分成多个部分，并在客户端和服务端上分别进行训练来减少客户端的存储和计算负载。切分联邦学习是切分学习的一种重要形式，它通过在多个客户端上并行计算学习任务，提高了切分学习框架的可扩展性。PipeSFL 是一种细粒度的切分联邦学习框架。如图 11 所示，PipeSFL 包含两个关键机制：1) 云服务器端的优先级调度机制，该机制将优先处理来自性能最差客户端的切分层激活值，以减轻性能不佳的设备在执行本地反向传播时导致的服务器和其他边缘设备资源的闲置；2) 混合训练模型，允许同一轮次内实现异步训练，而在轮次间实现同步训练，以避免一个轮次内服务器在同步接收所有边缘设备的切分层激活值时产生的资源闲置[25]。

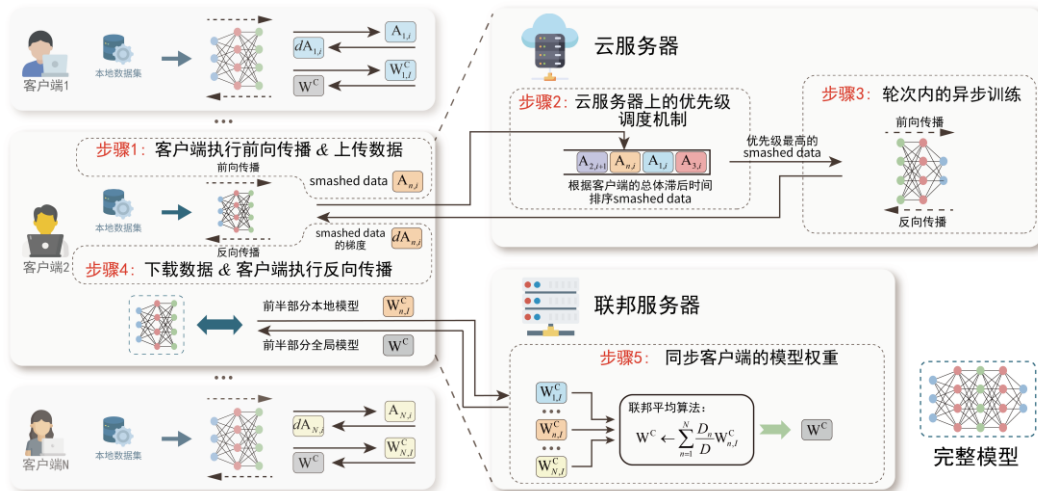


图 11 PipeSFL 框架的工作流

➤ 协作推理

协作推理通过在终端侧运行轻量模型实现快速响应，在遇到复杂或高精度需求的任务时，再由边缘侧的大模型接管，形成“快响应+高精度”互补机制。通过引入资源感知的自适应调度策略，系统能够在不同设备算力、网络带宽和任务复杂度的动态变化下保持稳定与高效，从而显著提升边缘智能的鲁棒性与扩展性。

作为协作推理的一种重要实现方式，投机采样（speculative decoding）在近年来受到越来越多的关注。其核心思想是利用一个轻量级的“草稿小模型”快速生成候选序列，并由一个更精确但参数量更大的“目标大模型”进行验证与重采样修正，从而在保证生成质量的同时显著提升推理效率并降低响应时延[36]–[37]。结合 AI Edge 的分布式计算特性，分布式投机采样的框架被提出[38]：将草稿小模型下沉至用户终端，而目标大模型则部署在边缘节点。如图 12 所示，终端在本地快速生成多候选序列，并将其发送至边缘节点，由目标大模型执行最终的验证与重采样。该架构不仅减少了边缘节点在解码阶段的冗余计算开销，还大幅提升了推理吞吐量，使其能够更好地满足未来无线网络中对算力高效利用与用户体验优化的需求。

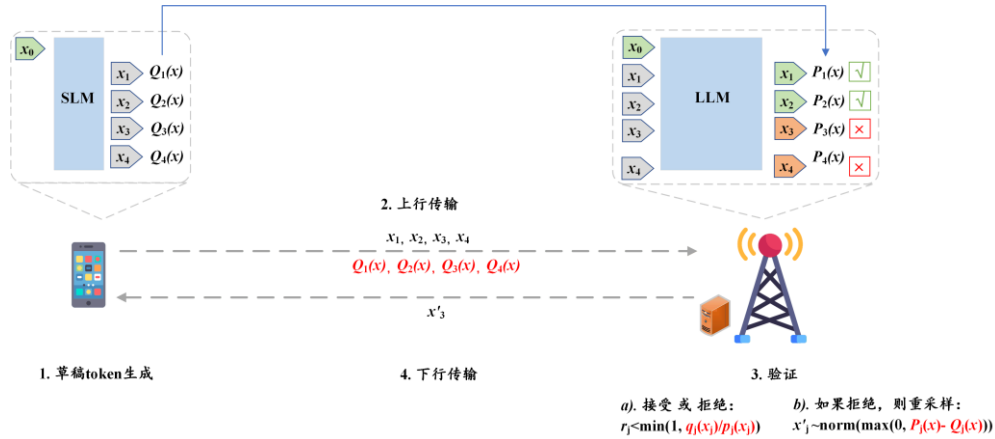


图 12 分布式投机采样

然而，这种分布式部署仍受制于通信开销：每生成一个草稿 token，都需要将其对应的词表概率分布上传至基站或边缘服务器进行验证，因此传输数据量会随着词表规模线性增长。例如，对于大小为 32k 的词表，使用 FP16 表示时，每个 token 的传输量约为 500kbit。可通过建立基于置信度的按需协作机制、验证与重采样分离、稀疏化传输等方法解决上述问题[39]。

➤ 协同训练与模型更新

以上所讨论的各种模式主要针对模型的推理。除推理之外，云边端的协同也有助于实现更高效的模型训练与持续更新。例如，在训练阶段，由云侧强大的算力集群负责复杂大型模型的训练与迭代，再将优化后的轻量化模型下发至边缘节点进行就近部署。此外，边缘侧或终端侧可利用本地数据进行部分模型参数更新，并通过联邦学习或分布式优化方法与云端进行参数聚合，既能提升模型的泛化能力，又能有效保护用户隐私。

为了赋予边缘 AI 持续的适应与进化能力，增量学习与持续学习技术使得部署在边缘的模型能够在不脱离生产环境的前提下，利用本地产生的少量新数据进行微调与迭代更新，从而快速适应数据分布的变化与新出现的场景，大幅降低对云侧再训练的频繁依赖，最终构建出一个既高效敏捷又具备自我进化能力的分布式智能网络。

4.3.4 AI Agent 技术

在传统的移动通信网络中，系统的决策往往依赖于预定义的规则与静态策略，缺乏对环境变化的感知记忆与动态响应能力。AI Agent 是能够在特定环境中感知、理解目标并自主采取行动以达成任务的系统[40]。通过在边缘侧引入 AI Agent，无线系统有望弥补“记忆缺失”和“动态决策”的难题，赋予边缘网络以持续学习、反思与协作的能力。

图 13 给出了 AI Agent 的核心组成和部署，及其在边缘侧的工作模式。AI Agent 主要由感知、记忆、行动和规划四个模块组成，核心功能包括[41]：

- 1) 环境感知与记忆机制：AI Agent 能够持续收集并存储网络状态信息（如信道质量、用户移动性、业务负载等），形成短期与长期记忆，支持上下文感知的智能决策；
- 2) 自主行动与动态规划：基于强化学习与在线学习算法，Agent 可在无需人工干预的情况下，实时调整网络参数（如功率控制、资源分配），优化网络性能与能效；
- 3) 多智能体协同（Multi-Agent）：通过分布式协商与协同学习，多个 Agents 可在边缘节点间实现联合优化，避免冲突、提升整体系统效率，尤其适用于超密集网络与移动性管理场景。这些能力在边缘侧以“工作流”形式编排运行（如感知-规划-行动，并结合记忆），支持多个并行流程在本地数据与用户环境中执行。

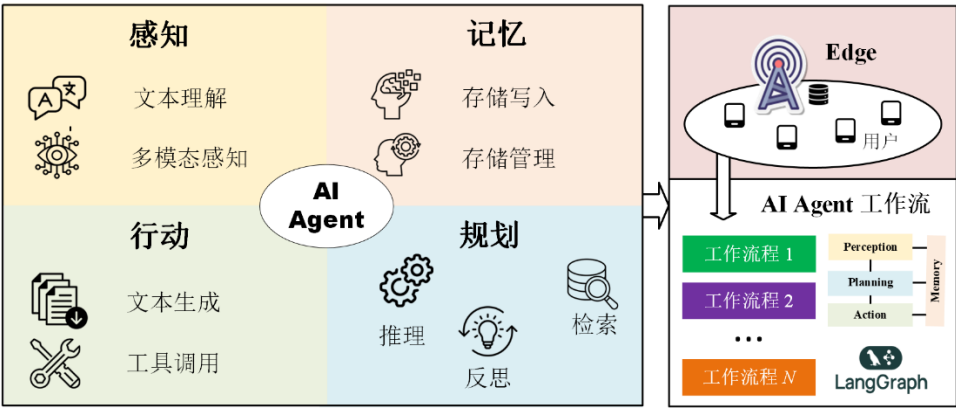


图 13 AI Agent 的核心组成及其在边缘侧的工作模式

AI Agent 在边缘侧设计和部署的关键技术主要包括：

- 1) 记忆与反思机制：构建具备历史数据分析与经验回放能力的 Agent 架构，支持策略迭代与长期网络行为建模（如潮汐效应、事件驱动型流量预测）。
- 2) Multi-Agent 协同算法与协议：搭建 AI Agent 编排框架，管理多个边缘 AI Agents 之间的任务分配、通信与协作；研究基于博弈论、联邦学习、共识机制的多 Agent 协作框架，确保在去中心化环境下的高效、安全与可扩展协同，例如，可采用基于联邦学习的协同训练模式，允许多个 AI Agents 在本地数据不共享的前提下联合优化模型，既保障隐私又提升群体智能。
- 3) 轻量化 Agent 设计与部署：面对边缘侧严峻的算力、存储和功耗约束，需要综合运用模型架构设计、模型压缩与加速、持续学习与自适应能力等技术，在效率与性能之间取得平衡，针对边缘设备资源受限的特点，开发低开销、高并发的 Agent 推理与训练技术，实现“小而灵”的智能体部署。
- 4) 边缘优化的决策与规划：AI Agent 的决策与规划的核心目标是在有限算力与能耗约束下，仍能做出可靠、高效的自主决策。为此，边缘 AI Agent 通常采用层次化与轻量化的决策架构。在高层，AI Agent 基于轻量级强化学习方法，将复杂任务分解为可执行的子任务序列。在低层，依赖本地优化算法等，实现对子任务的快速响应与动作执行。同时，为适应边缘环境的动态性与不确定性，AI Agent 需具备在线学习与自适应能力，能够依据反馈持续调整策略，避免因数据分布变化导致的性能退化，使 AI Agent 能够在多重约束下进行局部滚动优化，实现对动态环境的实时响应和自主决策。
- 5) 高效的资源管理与协同计算：AI Agent 的高效资源管理与协同计算是其实现自主智能的关键前提。面对边缘节点算力有限、能耗约束严格且网络条件多变的挑战，资源管理机制可有效实现动态自适应的调度策略。AI Agent 实时监测本地的计算、存储、能耗、网络资源状态，并基于任务优先级、延迟敏感度和能耗指标等做出智能决策：对于轻量级实时任务，优先在本地完成推理以保障低延迟；对于计算密集型任务，则通过

计算卸载技术将计算任务拆分并分发至边缘服务器或云端，形成云-边-端三级协同的计算范式。

因此，边缘侧 AI Agent 需通过记忆与反思、多智能体协同，并通过轻量化设计以适应受限算力，实现高效、安全、可扩展的部署[42]。

AI Agent 的应用场景与价值包括：

1) 智能无线资源管理：Agent 可动态调整频谱、功率与天线参数，提升网络容量与覆盖；

2) 服务编排与生命周期管理：在边缘计算环境中，Agent 可实现服务的按需部署、迁移与终结，保障 QoAIS (Quality of AI Service)；

3) 通感一体化智能服务：结合多模态感知数据（如 RF 测量、视觉、定位），Agent 可提供端到端的情境感知服务，如智能交通调度、工业物联网监控等。

综上所述，AI Agent 技术是实现“AI over Edge”愿景的核心使能技术之一。通过赋予边缘网络以记忆、学习与协作能力，它不仅解决了传统通信系统中“机械执行”的局限性，也为构建具备持续进化能力的下一代智能边缘网络奠定了坚实基础。

4.3.5 面向具身智能的端到端信息服务技术

随着人工智能与机器人技术的深度融合，具身智能 (Embodied AI) 正成为下一代信息服务的重要载体，同时也是 AI over Edge 的重要应用场景之一。其通过智能体（如机器人、自动驾驶车、XR 设备等）在物理世界中的感知、交互与行动，为用户提供前所未有的沉浸式和主动式服务体验。

具身智能信息服务是指由拥有物理“本体”的智能体，通过多模态感知理解环境，并执行具身化行动，从而向用户提供智能化、情境化服务的技术体系。这要求底层的信息服务技术必须实现从感知、计算到执行的端到端闭环，并能动态适应复杂多变的物理环境。涉及到的关键技术包括：

➤ 与通感深度耦合的端到端智能化服务

“通感”即通信与感知的深度融合，是具身智能服务的基础。端到端智能化旨在将“感知-通信-计算-执行”的传统线性串行架构，重构为一个从原始传感器数据到最终服务动作的协同优化整体，通信不再仅仅是传输数据的管道，其本身也成为感知和控制系统的一部分。通信和感知资源根据当前任务的优先级进行动态分配和调整，并能利用深度学习等技术，联合优化感知模块、控制策略等，在资源受限时做出最优的权衡，实现端到端智能化服务。

➤ 服务的智能编排与开放

单个具身智能体的能力有限，未来必然是多个智能体协同、云边端协同的生态系统。服务的智能编排与开放是实现这一愿景的核心。将各类具身智能服务（如导航、识别等）原子化、模块化，将智能体的能力封装成可被统一发现和调用的网络服务，构建统一的动态编排引擎进行编排管理。编排引擎通过理解任务逻辑、服务间的依赖关系，实时感知环境状态，根据用户的高层指令，自动发现、组合、编排、调度并执行一系列原子服务。同时支持结合知识图谱、强化学习等 AI 技术，动态调整服务链和任务分配策略。通过构建开放服务框架，实现能力服务化封装与安全可信调用，并建立开放平台，允许第三方开发者注册和发布新的智能服务，形成丰富的服务生态。

➤ 边缘计算服务的按需灵活提供及生命周期管理

具身应用对延迟极其敏感，且计算负载波动大。集中式的云计算无法满足需求，必须依赖边缘计算。在网络边缘构建分布式计算资源池，能够根据具身智能体的实时需求，动态地创建、迁移、扩缩容和释放计算服务。依托算力感知路由和网络协议，通过无线

算力网络实现计算任务的智能调度与分发。设计无线算力网络管理框架，对边缘服务实例进行全生命周期管理，在智能体移动过程中，实现服务的无缝迁移，保证任务执行的连续性。

➤ QoAIS 保障技术

QoAIS 是衡量智能信息服务质量的核心指标，是一个多维度的度量体系，超越了传统的网络 QoS（如带宽、延迟），涉及感知、通信、计算和控制多个层面，涵盖了任务成功率、任务完成时间、能耗效率等多个维度，通过实时监测多维度指标，打破各层之间的壁垒，进行跨层的联合资源调度与优化。

4.3.6 数据安全与隐私

边缘人工智能的部署环境从高度受控的云端数据中心转向开放、离散的现实世界，这一根本性转变使其安全范式面临前所未有的严峻挑战。因此，对安全性的关注并非一种补充，而是边缘智能能否成功落地的首要前提。其独特性源于多重因素：（1）物理安全边界变得模糊，边缘设备可能部署在无人值守的公共场所，极易遭受物理接触与篡改；（2）网络连接不可靠且多变，不稳定的网络不仅影响性能，更大幅增加了通信被监听或中间人攻击的风险；（3）设备本身资源高度受限，难以承载复杂的传统安全软件，使其自身成为安全链条中的薄弱环节；（4）异构环境极其复杂，来自不同厂商、架构各异的硬件与软件堆栈共存，极大地扩大了攻击面。基于这些挑战，亟需构建出相应的威胁模型，其攻击面覆盖了系统的每一个层面：（1）在数据层面，攻击者可通过数据投毒污染训练集，或窃取传输中的隐私数据；（2）在模型层面，存在模型窃取、逆向工程以及通过对抗样本攻击误导推理结果的威胁；（3）在基础设施层面，边缘节点或终端设备可能被劫持而沦为僵尸网络的一员；（4）在通信渠道，中间人攻击可篡改或中断关键指令与数据流。系统性识别这些威胁是构建有效安全防护体系的前提。



图 14 端边分离与端边一体模式

在 AI Over Edge 时代，无论端边一体还是端边分离（如图 14 所示），只要数据跨出本地，安全与隐私便是必须直面的底线命题。在端边分离模式下，终端是“零算力”或“弱算力”的设备，所有 AI 推理统一在边缘完成，典型如安防 IPC、工业传感节点。在端边一体模式下，终端自带“强算力”，可本地闭环绝大多数 AI 任务，仅在特殊需求时向边缘/云端请求增量能力，典型如智能机器人、自动驾驶车辆。

与 AI Edge 相关的数据安全与隐私保护技术包括但不限于：

➤ 分布式信任

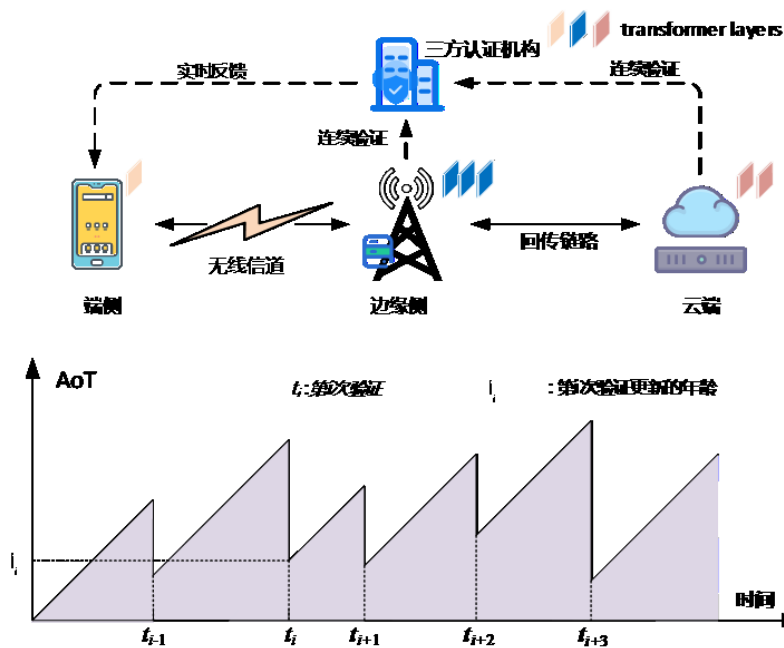


图 15 持续认证云端与边缘端身份

在边缘计算开放、分散且边界模糊的环境中，无论是用户、设备还是边缘工作负载，在访问任何资源前都必须经过严格的、基于身份的认证与授权，并通过微隔离技术实现精细化的访问控制，确保即使单个节点被攻破，攻击者也难以横向移动。更进一步，应当将认证设计为一个持续的过程，其动态访问控制机制会持续评估访问主体的安全状态，如设备指纹、模型版本、地理位置、行为异常等，可结合信任年龄 (AoT) 概念[43]–[44]，实时计算风险等级，并动态调整访问权限，从而形成一种自适应的、基于风险的安全策略。

➤ 数据脱敏

通过对敏感数据进行变形、模糊或替换，在不影响数据分析或业务逻辑目的的前提下，最大限度地降低数据的敏感度，从而保护个人隐私和商业机密。其目标是让数据“可用但不可见”（敏感部分）。

➤ 同态加密 (Homomorphic Encryption, HE)

边缘设备可以将敏感数据（如个人健康数据、视频片段）使用同态加密后发送到边缘服务器或云端。服务器可以在不知道数据内容（数据仍处于加密状态）的情况下，直接执行 AI 模型推理或训练计算，并将加密的结果返回。只有拥有密钥的边缘设备才能解密最终结果。

➤ 差分隐私 (Differential Privacy, DP)

在数据离开边缘设备之前，或是在上传模型更新时，向其中注入适量的随机噪声。这个噪声足够大，可以掩盖任何单个个体的数据贡献，防止从输出结果中推断出特定个体的信息；但又足够小，不会对整个数据集的分析结果产生统计学上的显著影响。

➤ 安全多方计算 (Secure Multi-party Computation, MPC)

通过密码学协议将各方的输入数据拆分、混淆，使得在计算过程中，任何一方都无法看到其他方的原始数据，但最终能获得正确的计算结果。实现多方安全计算的协议主要有基于混淆电路 (Garbled Circuit, GC)，秘密分享 (Secret Sharing, SS) 和同态加密 (Homomorphic Encryption, HE) 方式。

➤ 可信执行环境 (Trusted Execution Environment, TEE)

在边缘设备（如手机、平板电脑的芯片）上创建一个硬件隔离的“安全飞地”。敏感的 AI 模型和数据进行可以在 TEE 内部被处理和存储，与设备上运行的操作系统和其他应用程序完全隔离。即使是拥有根权限的攻击者也无法访问 TEE 内的内容。当前业界主流的可信硬件技术路线包括：ARM Trust Zone, Intel SGX, AMD SEV, RISC-V keystone 等。

➤ 区块链技术

在边缘人工智能分布式、去中心化的部署环境中，建立跨主体的信任与确保运行过程的透明可审计，对其规模化应用具备重要参考价值。区块链技术凭借其分布式账本所固有的不可篡改与可追溯特性，为构建可信边缘 AI 提供了至关重要的基础能力。首先，在模型可信方面，区块链被用于记录模型全生命周期的溯源信息，从训练数据的来源与哈希、训练过程的超参数与环境、到版本更新迭代记录以及最终在边缘节点的分发部署日志，均被永久且防篡改地记录在链，为模型的可信性提供了可验证的证据，有效杜绝了恶意模型或数据污染模型的传播，尤其在多方协作的场景中至关重要。其次，对于 AI 推理的审计与问责，关键推理决策的日志，包括输入数据的哈希、模型版本、推理结果与时间戳等，可被实时上链存证。这不仅为事后的责任界定、纠纷仲裁提供了数据基础，同时也为模型优化迭代提供了高质量、可信的反馈数据闭环。最后，区块链的分布式共识机制本身就在边缘节点间建立了一种无需中心权威的信任网络，任何对日志或数据的篡改企图都会被网络拒绝，有效防止了单个或多个边缘节点因被恶意攻破而伪造数据或作恶的行为。

➤ 模型安全技术

模型安全层面，对抗样本防御技术通过在训练中引入对抗性样本或推理时部署检测过滤器，主动识别并抵御精心构造的恶意输入，确保边缘 AI 模型的判断不被误导，维护其决策的可靠性。同时，为保护投入巨大资源研发的 AI 模型资产，模型水印技术被嵌入其中，将唯一标识信息隐藏在模型参数或结构中，为模型确权与追踪泄露源头提供了有力证据，有效遏制模型的非法复制与滥用。

4.4 芯片与算力底座

AI Edge 芯片与算力底座作为全域智能化的核心基础设施，通过多层次技术突破构建“通感智算控”深度融合的异构计算基座。其核心采用跨维度异构计算架构，在单芯片空间内实现通用计算单元、领域专用加速器及多模态接口的硬件级集成，依托三维堆叠与存算一体技术突破内存墙限制，显著降低数据搬运能耗与时延。动态可配置架构通过标量-向量-矩阵三维融合机制，支持通信与 AI 算力的弹性适配，实现毫秒级实时控制与分布式认知协作的复合需求。全域算力调度引擎构建跨框架协同枢纽，结合绿色调度策略与可信隔离机制，驱动异构资源全局优化与安全流转；内生安全体系集成硬件可信根与多级防护架构，保障物理暴露环境中的功能可靠性与数据隐私。尤其值得注意的是，该架构深度融合 RISC-V 开源指令集所带来的模块化、可扩展及开放协同等核心优势，充分发挥其在定制化计算单元设计、多精度算力支持以及软硬件生态统一方面的潜力，推动实现从芯片物理层到分布式操作系统的全栈支撑能力，为千行百业提供弹性可扩展的智能底座。

4.4.1 通感智算控融合的芯片架构创新

面向全域智能化 AI Edge 芯片设计的核心在于构建支撑未来边缘智能发展的动态能力矩阵。未来边缘智能的发展高度依赖于芯片架构能否实现物理层资源与数字层能力的深度协同，这要求 AI Edge 芯片突破传统刚性约束，建立适应业务弹性演进的“通感智算控”多维融合算力基座，通过异构计算资源的多维度集成与动态调度，满足毫秒级实时控制、分布式认知协作及跨域自治决策的复合需求。在这一过程中，引入 RISC-V 指令集生态尤为重要，其开源性、可扩展性及模块化优势为芯片架构提供了底层技术支撑，使得通信、感知、计算和控制功能可在统一的指令集框架下实现高效协同与动态适配。为实现这一目标并确保持续创新，亟需构建开放协同的技术底座，通过模块化设计标准实现功能单元灵活组合封装、跨协议硬件抽象层确保多域设备无缝互操作、以及存算一体与三维集成工艺的深度融合，最终形成从器件物理、电路架构到分布式操作系统的全栈协同体系，使 AI Edge 芯片成为驱动千行百业智能化转型的隐形引擎。

➤ 跨维度异构计算引擎

AI Edge 芯片需在单芯片空间内完成通用计算单元、领域专用加速器、多模态接口与控制引擎的四维集成。通过三维堆叠与硅通孔垂直互联技术，实现计算核心、高带宽存储器及通信基带单元的微米级近存集成，消除数据搬运导致的时延与能效损耗。在架构层面采用时空复用流水线机制，使工业协议解析、张量加速计算与射频信号处理等异构任务能够在统一硬件资源池上并发执行。借助 RISC-V 开放指令集所带来的高度灵活性和可扩展性，各类处理单元能够根据任务需求实现指令级动态优化与功能重组，进一步增强算力融合的效率与适应性。这种深度集成不仅将端到端处理时延压缩至纳秒量级，更使芯片在突发负载场景下的算力密度较传统架构具备指数级提升，为高并发 AI Edge 智能服务提供底层支撑。

➤ 认知驱动型动态可配置架构

应对通信与 AI 混合负载的核心在于构建模块化可裁剪的异构融合架构。AI Edge 芯片采用标量-向量-矩阵三维计算融合机制：通过“标量处理器核心+一维计算阵列+二维计算阵列”的领域专用架构（DSA），实现计算资源的多维协同与存储-计算高效联动。该架构充分结合 RISC-V 指令集模块化与可扩展特性，实现从通用计算到领域专用加速的无缝衔接与指令集层面的深度融合。层次化存储与高速互连网络支持不同计算单元承担差异化任务，既保障通信向量运算与 AI 矩阵运算的专业效率，又在负载波动时通过计算阵列双向能力拓展实现资源互补，大幅提升整体算力利用率。这种架构打破传统通信与 AI 算力壁垒，形成可扩展的硬件资源池。

面对边缘场景差异化需求，芯片需具备硬件资源动态裁剪与软件算法灵活适配的双重能力。高 AI 需求场景（如自动驾驶、智慧城市）通过激活全二维计算阵列与优化推理算法，提升并行计算能力并保障低延迟数据传输；高通信需求场景（如无人机集群、应急通信）优先配置一维计算阵列与高速通信模块，优化信号处理与多节点协同算法；低功耗需求场景（如智慧农业监测）通过裁剪计算规模与启用低功耗模式，优化数据采集与轻量推理流程。借助 RISC-V 生态所提供的统一而灵活的指令集支持，该架构可有效避免传统“一芯片一场景”导致的生态碎片化问题，显著提升研发效率与跨场景兼容性，为 AI Edge 大规模落地提供兼具灵活性与高效性的硬件基础。

➤ 多重优化的能效自适应体系

针对 AI Edge 边缘场景的极端能效约束，芯片需构建从晶体管级到系统级的跨层优化体系。电压频率岛群技术将芯片划分为多个独立供电域，支持微秒级精度的动态电压频率调节，使芯片在静默态与峰值负载间的切换能耗下降超 90%。突发负载预测器通过

预分析指令流特征提前加载计算资源，消除传统开关机过程的能量空窗期。更关键的是未来芯片可引入光电子互连，在增加数据传输通量的同时进一步降低各系统间数据传输功耗。结合 RISC-V 指令集精简高效的特性和开源社区持续优化的能效控制机制，芯片能够在指令执行层面进一步降低功耗，这种多重能效优化体系保证芯片能在微瓦级待机与满载运行状态间无缝切换，使 AI Edge 芯片能够高效、可靠且长时间地运行，真正推动人工智能在物理世界中的规模化落地。

➤ 内生安全与可信执行架构

随着边缘智能设备的普及和其承载的数据价值日益提升，芯片正面临着更为严峻的物理暴露性威胁，这使得在硬件层面增强防护能力变得至关重要，并要求芯片嵌入硬件级可信根与动态防御机制。四重防护体系将贯穿芯片设计全栈：物理不可克隆功能单元为每个计算任务生成不可复制的加密身份；同态加速核确保敏感数据在计算全周期保持密态；形式化验证硬核对工业控制指令流进行数学证明级校验；抗辐照容错设计通过动态冗余切换维持太空与工业强干扰环境的功能连续性。依托 RISC-V 开源指令集的可审计性与透明性，结合其模块化安全扩展机制，可有效实现硬件木马检测、旁路信道防护与故障自愈能力的深度融合，该架构将硬件级安全与功能安全原生统一，为自动驾驶、远程医疗、工业控制等关键应用领域提供功能安全与信息安全融合保障体系。

➤ 分布式智能协同加速单元

为支撑 AI Edge 的云边端协同架构，芯片需内置跨层协同加速单元。联邦学习硬件引擎支持本地参数聚合与梯度加密交换，使分布式节点能在隐私保护前提下完成模型协同进化；知识蒸馏加速器将全局模型压缩为边缘可承载的轻量化表达，降低协同通信开销 90% 以上；时空对齐接口则对多源感知数据实施硬件级时间戳校准与空间坐标映射，为车路协同、无人机集群等场景提供微秒级精度的环境态势共识。借助 RISC-V 开放指令集标准所带来的软硬件兼容性，不同设备与平台之间可实现更高效的指令集级协同与任务调度，这些机制使边缘节点规模每扩展一倍，协同推理吞吐量即提升近三倍，实现智能能力的超线性增长。

➤ 异构算力统一指令集框架

通信 AI 算力适配的指令集需构建“开源架构为基、标准协同为纲、场景智配为要”的技术体系。以开源指令集架构的模块化扩展能力为核心，打造兼具通用性与定制化的柔性指令框架，通过可扩展指令簇支持多精度计算、实时响应等边缘核心需求，支撑从通用计算到领域专用加速的无缝切换。重点推进跨场景标准体系建设，推动指令集规范与边缘设备接口、安全协议的协同统一，解决异构硬件互操作性难题，形成开放兼容的技术生态。强化场景感知的动态适配能力，依托软硬件协同优化机制实现算力资源的智能调度与能效平衡，融合硬件级安全与分布式协同指令设计，为边缘节点的可信互联与协同决策提供底层支撑，助力构建弹性可扩展的边缘算力底座。

4.4.2 全域异构算力智能调度引擎

通信 AI 算力适配的指令集需构建“开源架构为基、标准协同为纲、场景智配为要”的技术体系。以 RISC-V 开源指令集架构的模块化扩展能力为核心，打造兼具通用性与定制化的柔性指令框架，通过可扩展指令簇支持多精度计算、实时响应等边缘核心需求，支撑从通用计算到领域专用加速的无缝切换。RISC-V 指令集由于其开放、简洁、模块化的设计，具有显著的可扩展性和定制能力，能够针对 AI 与通信融合计算场景进行指令级优化与功能扩展，显著提升能效与实时性。重点推进跨场景标准体系建设，推动指令集规范与边缘设备接口、安全协议的协同统一，解决异构硬件互操作性难题，形成开放兼容的技术生态。强化场景感知的动态适配能力，依托软硬件协同优化机制实现算力资

源的智能调度与能效平衡，融合硬件级安全与分布式协同指令设计，为边缘节点的可信互联与协同决策提供底层支撑，助力构建弹性可扩展的边缘算力底座。

➤ 融合架构驱动的算力共享机制

在 AI Edge 算力平台中，算力调度与共享是实现资源高效利用、保障任务低延迟执行的关键环节。依托一体化算力载体与全栈协同架构，算力调度与共享以算法层面、配置层面、硬件层面为输入，开展全链路的算力调配与共享。算法层面支持多格式任务输入，AI 任务以 ONNX 格式模型为载体，通信任务用类 C 的定制语言编写，为算力调度提供多样任务源；配置层面综合硬件资源状态（标量单元、一维计算阵列、二维计算阵列的负载与可用资源量）与软件参数（任务优先级、时延要求、算力需求等），为调度决策提供约束条件。此外，算力调度与共享具备动态资源调整能力，突破传统架构中硬件资源静态分配的局限，当某类计算单元负载饱和时（如通信高峰期一维计算阵列满负荷），可通过跨单元负载卸载实现资源互补，将部分通信任务的计算负载卸载到空闲的二维计算阵列，或在 AI 推理任务并发量上升时，调配空闲的一维计算阵列承接轻量级的 AI 任务，确保整体算力资源始终处于高效利用状态，避免“忙闲不均”的资源浪费。同时，借助统一的中间表示与任务特征识别，在 AI Edge 的通信-AI 融合应用中，能够支持跨类型任务的流水线式调度，让不同计算阵列交替处理不同阶段的任务，提升整体任务的执行效率。

➤ 绿色算力调度策略

紧扣边缘基础设施可持续发展需求，该策略以“算力与能源协同优化”为核心导向，将边缘节点的算力需求与分布式能源（如光伏、储能）的供应特性深度绑定，形成“能源波动感知-算力动态迁移-能效优先级管控”的闭环体系。在实际应用中，它既能通过错峰调度将非实时任务（如视频数据回溯分析、模型轻量化训练）转移至能源充裕时段，又能优先保障关键服务（如应急监测、民生服务）的算力供给，同时降低边缘集群整体能耗与碳足迹。这不仅缓解了边缘节点的能源供给约束，更推动数字基础设施与绿色能源体系深度融合，为千行百业边缘智能部署提供低碳化、低成本的实现路径。

➤ 可信算力隔离与动态鉴权机制

针对边缘设备物理暴露性高、跨域协同风险突出的安全挑战，该机制以“内生安全+动态防御”为核心，构建从硬件根信任到全链路防护的可信算力体系。一方面，通过算力隔离技术为不同场景任务（如工业控制指令执行、用户隐私数据处理）划分独立运行空间，避免数据泄露或任务干扰；另一方面，建立动态鉴权机制，对边缘节点身份、数据完整性、任务合法性进行实时校验，即便在设备被物理接触或网络环境复杂的情况下，也能保障算力调用的安全性。这一机制为自动驾驶、远程医疗、工业互联网等高安全需求场景筑牢防护屏障，解决了边缘智能规模化应用中的信任难题，支撑跨节点、跨层级的安全协同。

➤ 跨框架算力协同枢纽

聚焦边缘算力生态中“框架碎片化、硬件异构化”导致的资源孤岛问题，该策略的核心是打造统一的算力协同枢纽——通过构建标准化接口与自适应适配引擎，打破 TensorFlow Lite、PyTorch Edge 等不同 AI 框架，以及 ARM、RISC-V、x86 等异构硬件平台之间的协同壁垒。它能实现不同框架模型的自动转换与优化部署，让同一边缘节点可灵活承接多框架训练的推理任务；同时推动“云-边-端”异构算力资源池化，允许跨设备、跨层级的算力弹性拆借。这不仅降低了 AI 模型跨场景部署的技术门槛，避免企业因框架差异重复投入算力资源，更能释放碎片化边缘算力的聚合价值，加速算法创新向产业应用的转化，支撑零售、农业、物流等领域边缘智能的灵活化、差异化落地。

4.4.3 智能算力开放生态体系

智能算力开放生态体系以开发者赋能为核心，构建覆盖全流程的技术支撑链。通信-AI 融合计算架构通过统一编程语言实现通信与 AI 任务的硬件无感描述，结合编译器的智能映射机制，将任务指令动态分配至标量处理器、一维/二维计算阵列等异构单元，突破传统架构的硬件适配壁垒。全栈开发支撑体系提供从功能仿真到原型落地的全生命周期工具链，内置通信-AI 融合验证环境与模块化算子接口，显著降低复杂边缘应用的开发门槛。统一编程范式平台借鉴 CUDA 生态核心优势，建立通信与 AI 无缝集成的开发框架，支持多格式模型迁移与跨精度计算需求，为零售、农业、物流等差异化场景提供弹性可扩展的部署能力。

➤ 通信-AI 融合计算架构 (AI-Unified Radio Architecture, AURA)

AURA 计算架构由 Venus 编程语言、Zoozve 编译器与基础算子库组成，解决传统架构中硬件适配复杂、任务描述割裂的痛点，为硬件层与开发平台层搭建起高效的衔接桥梁。其中，Venus 编程语言支持对通信与 AI 任务进行统一化描述，开发者无需区分任务所属类型（通信或 AI）及对应硬件需求（如基带、NPU 等），仅需使用 Venus 编写任务逻辑，即可实现跨硬件的统一表达，极大简化任务描述的复杂度。Zoozve 编译器作为代码编译与硬件映射核心，具备两大关键能力：一是将 Venus 编写的统一任务代码，编译为 AIEdge 芯片可执行的机器指令；二是通过智能映射机制，将编译后的指令自动分配至芯片内对应计算单元（如通信类指令映射至一维计算阵列、AI 类指令映射至二维计算阵列），完成通信与 AI 算子的硬件适配。尤其针对长向量任务，Zoozve 突破 RISC-V 向量扩展（RVV）静态寄存器数量与 2 的幂次分组限制，通过无条带挖掘（strip-mining-free）设计与数据自适应寄存器分配策略，支持任意长度向量与寄存器分组配置，可减少快速傅里叶变换（FFT）等通信任务的动态指令数至少 10 倍，同时仅增加 5.2% 的芯片面积，在提升编译效率的同时，避免传统编译导致的性能损耗与硬件资源浪费。基础算子库聚焦“硬件级算子封装”，覆盖通信与 AI 领域核心基础算子，通信领域包含 FFT、调制解调、信道编译码等，可直接匹配一维计算阵列硬件特性；AI 领域涵盖 Conv2D/3D、全连接、激活函数（GELU/SiLU）等，适配二维计算阵列并行能力。同时支持算子硬件映射优化，确保每类算子均能匹配对应计算单元的架构优势，为 Echo 平台提供高性能底层算子支撑。

➤ 全栈式开发支撑体系

Echo 平台围绕“降低开发门槛、加速应用落地”目标，对 AURA 架构的底层能力进行“开发者友好型封装”，构建覆盖“开发-验证-部署”全流程的支撑体系，提供从功能仿真到原型落地的一站式服务。在开发验证阶段，平台内置应用功能与性能仿真工具，支持开发者在硬件实物开发前，对通信-AI 融合任务（如“5G 信号处理+ AI 信道估计”）进行全流程验证，既可校验任务逻辑的功能正确性（如通信协议合规性、AI 推理结果准确性），也可量化评估性能指标（如任务端到端时延、多维计算单元利用率、算力资源消耗），提前识别硬件适配冲突、算力瓶颈等潜在问题，避免后期开发返工；在算子调用层面，平台对 AURA 基础算子库进行二次封装，提供通信-AI 常用计算模块（如 5G 物理层处理模块、轻量 AI 推理模块），通过标准化接口向开发者开放。开发者无需关注算子底层硬件实现，可直接调用模块完成基础功能开发，也可基于接口扩展自定义模块（如 6G 语义通信编码模块），平衡易用性与扩展性；在工具链支撑方面，平台提供从代码编写、编译、调试到部署的全流程工具包，代码编写环节兼容主流开发习惯，采用类 C 和类 Python 的编程语法，编译环节集成 Zoozve 编译器实现 Venus 代码到硬件指令的自动映射，调试环节支持计算单元级细粒度监测（如一维/二维计算阵列负载、数据

流转路径），部署环节可一键将应用适配至单芯片或多芯片集群，确保开发链路顺畅高效；在原型落地层面，平台内置 5G/LTE、AI 信道估计等典型场景的应用原型 Demo，开发者可基于 Demo 快速修改参数（如通信频段、AI 模型精度）或扩展功能（如新增多终端协同逻辑），大幅缩短从方案设计到实际应用落地的周期，为 AI Edge 的规模化推广提供便捷支撑。

➤ 通信 AI 统一编程范式平台

传统分立架构下，通信与 AI 任务因硬件单元功能边界固化，需适配不同专属开发工具，导致编程模型碎片化，极大提升了 AI Edge 应用的开发门槛与学习成本以及因开发链路割裂导致的低效率任务协同。Echo 平台聚焦“一套工具链覆盖 AI Edge 全场景开发”的目标，借鉴 CUDA 生态统一编程范式与丰富算子支撑的核心优势，构建通信与 AI 统一的编程生态，从模型迁移、任务开发、算子支撑三方面，彻底解决传统 AI Edge 算力底座“多计算单元编程语言异构、开发链路割裂”的痛点，如图 16 所示。Echo 平台为通信与 AI 任务打造了统一的编程模型，开发者无需区分任务类型（通信或 AI）及对应硬件需求，可采用统一的编程范式开展开发。无论是通信领域的信号调制解调、FFT 运算，还是 AI 领域的卷积、注意力机制推理，均能在同一套编程框架下描述任务逻辑。

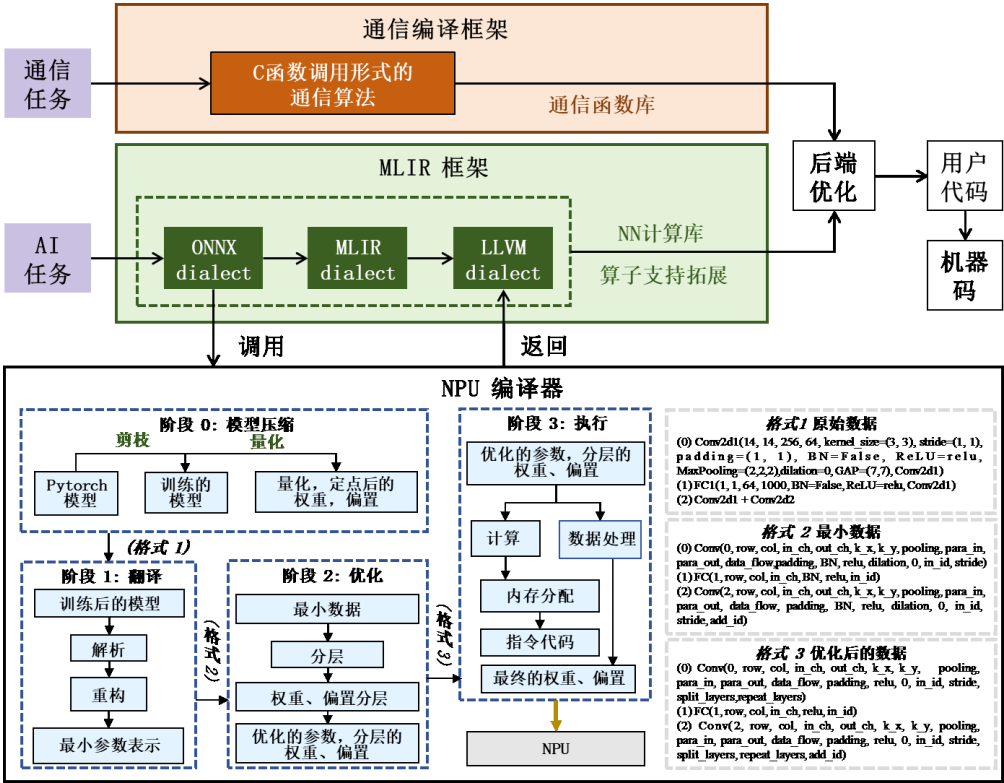


图 16 通信 AI 融合任务转化为统一中间表示的编程框架

➤ 高性能算子库体系

当前 AI Edge 场景面临工业控制、车路协同等核心应用中通信低时延与 AI 高算力难以协同的瓶颈，传统算力架构因任务割裂导致资源浪费严重。该体系构建“基础算子层-融合算子层-优化技术层”架构：基础算子层通过硬件感知封装，将通信算法（FFT/LDPC 编解码）与 AI 计算（卷积/全连接）分别适配异构计算单元特性，破解任务割裂痛点；融合算子层创新开发通信-AI 混合算子（如 AI 信道估计与预编码联合计算），通过减少跨模块数据搬运突破时延瓶颈；优化技术层采用混合精度计算、算子融合及零

拷贝机制，在维持通信精度的同时提升能效比。该体系为边缘侧实时协同计算提供高效算力底座，推动通感算一体化演进，支撑自动驾驶、工业互联网等场景的智能化升级。

➤ 开放算子生态赋能开发者

开发者面临多平台硬件适配复杂、调试效率低下、生态碎片化三大瓶颈，导致 AI Edge 创新应用开发周期长且难以规模化。该生态通过"多框架兼容接口层"构建跨平台能力：CUDA 兼容层实现 PyTorch/TensorFlow 模型零代码迁移，OAI 接口封装 3GPP 与 O-RAN 协议栈使传统基站软件无缝集成，领域专用语言（DSL）扩展以类 Python 语法和预定义模板提升复杂通信-AI 融合任务开发效率；结合"全周期开发工具链"，可视化性能剖析器自动定位算子瓶颈，自动优化引擎智能推荐混合精度切换等策略，一键部署工具通过容器化技术实现开发到部署全流程自动化。开发者可封装自定义算法实现"一次开发、多场景复用"，显著降低边缘侧复杂算法实现门槛，加速通信、工业、交通等领域 AI 应用落地，推动 AI Edge 技术从单点创新走向产业规模化协同发展。

4.5 AI Edge 系统、平台与测试

4.5.1 AI Edge 系统与平台

AI Edge 系统作为一种面向智能应用的综合信息基础设施，包含四个核心层次：硬件基础设施层、操作系统层、软件功能层和应用层。

硬件基础设施层构成系统的物理基石，涵盖感知与智能终端、异构算力与存储资源、接入网与传输网基础设施等核心要素。该层通过整合计算、存储与网络资源，提供异构算力供给、高速互联通信、多模态接入能力及资源池化服务，为上层操作系统与软件功能提供统一的硬件抽象与资源支撑。

操作系统层作为直接部署于硬件之上的系统软件核心，负责对底层异构物理资源进行抽象化与统一管控。该层深度融合云原生技术范式，依托轻量化内核架构、确定性实时调度机制与内生安全框架，实现对 CPU、GPU、NPU 等异构算力资源的统一调度与高可靠运行保障，为上层的边缘应用提供稳定、安全且具备确定性的执行环境。

软件功能层构建于基础设施与操作系统层之上，通过构建分布式数据汇聚、联邦学习协同、边缘推理服务及异构算力感知调度体系，为场景化应用提供数据聚合、协同训练、推理部署的全链路支撑，实现云边端算力的全局协同与 AI 任务的高效可扩展运行。一是依托下层的抽象硬件资源与实时运行管理，形成以边缘计算功能、数据功能为核心的能力中台，为无线网、核心网、智能化和感知等基础功能提供计算和数据服务；二是基础功能模块为应用层提供连接、AI、感知、计算、数据服务等能力；三是由业务编排、网络编排、跨域管理、资源编排构成的管理编排模块以“内生智能”为核心机制，依赖对网络、计算等资源的一体化调度，按需实现实时、近实时、非实时的分层次网络自主管控，并支撑移动信息边缘服务。

应用层基于 Agentic AI 技术将用户需求直接映射为通信网络的基础能力，通过软件功能层实现按需编排，进而调用相应功能，最终将服务状态反馈至应用层，形成“感知-决策-执行”闭环。该过程的实现首先依托于 APP 的开放接口与灵活架构，使上层应用能够无缝调用底层通信与感知能力。在此基础上，借助 AI Edge 所具备的 DOICT 融合技术，实现对差异化用户需求的细粒度定制，从而支撑垂直行业 AI 应用的快速部署与规模化推广。最终，逐步构建起以 AI Edge 为核心的 AI 即服务（AIaaS）“应用商店”生态，为各行业提供可订阅、可组合的智能服务模块，推动网络与智能在能力与价值层面的深度融合。

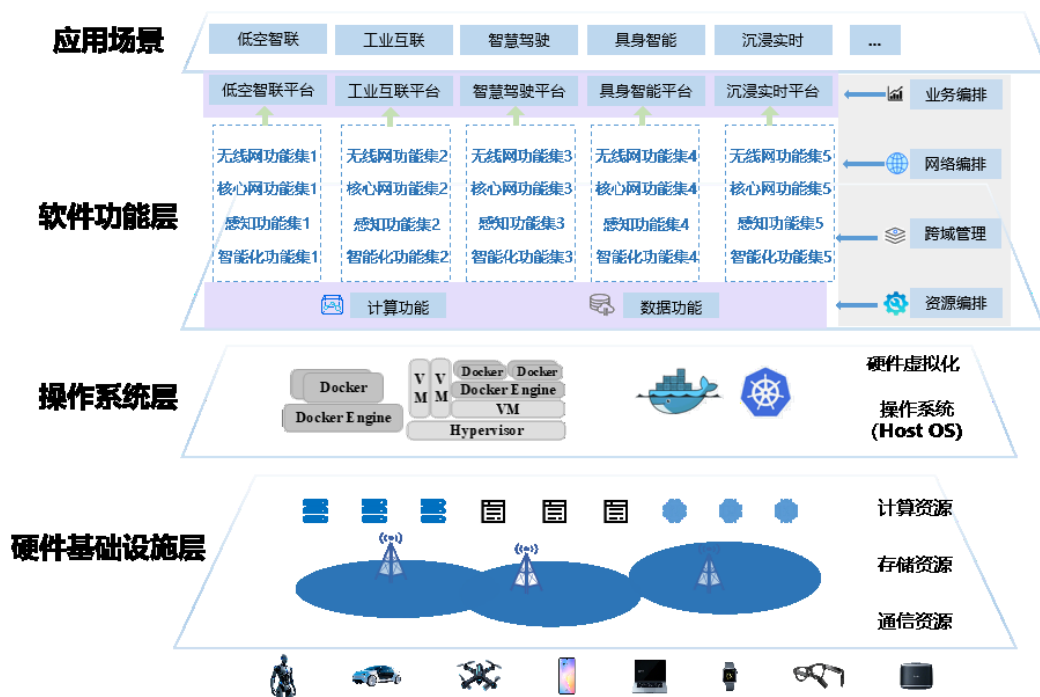


图 17 AI Edge 平台的核心层次

AI Edge 系统平台的技术方向核心是围绕“融合”与“协同”，即通过硬件池化、OS 抽象和软件智能，实现算力、网络与智能的三者深度融合。当前面临的技术挑战主要包括以下几点。

系统面临显著的硬件异构性挑战，表现为计算架构多元（如 CPU、GPU、NPU 及领域专用芯片）、算力形态差异显著，以及终端设备通信协议与接入技术的多样性。这种深度的异构性导致软硬件耦合复杂，驱动适配与系统兼容性要求极高，进而引发严重的生态碎片化问题，对系统资源的统一纳管、应用服务的规模化部署与跨平台移植构成实质性障碍。

在高性能需求场景下的极致性能挑战，一是在共享异构资源环境中，通过硬件虚拟化、实时性内核调度及确定性网络传输，为关键任务提供纳秒级响应精度与 99.9999% 的可靠性保障，确保在极端负载下仍能满足工业级应用的苛刻性能边界。二是需基于实时网络状态、多层次资源可用性及业务服务质量要求，动态判定任务在终端、边缘节点或云端的执行位置，实现高效的任务卸载策略。

在移动边缘场景（如智能网联车、无人机集群）中，终端的高速移动性与动态拓扑变化对业务连续性提出严苛要求。需保障跨基站切换、跨域路由过程中会话状态零中断与业务无感知迁移，这对接入网与核心网的功能下沉、无线资源协同调度以及分布式会话管理机制提出了极高要求，需实现控制面与用户面的近场优化与毫秒级无缝切换能力。

构建全域覆盖的安全可信体系是边缘智能系统的核心要求。需采用轻量化密码学协议与硬件加速加密机制，在保证低延迟与高吞吐的前提下实现传输数据的端到端保密性与完整性；同时，依托隐私计算技术（如联邦学习、安全多方计算等），在原始数据无需集中归集的情况下，实现分布式节点间的安全协同与价值挖掘，最终达成“数据不动价值动”的合规性目标。

4.5.2 AI Edge 测试

随着 AI 模型不断下沉至边缘侧，AI Edge 系统的复杂性与异构性显著提升，如何在多层次、多场景下验证其性能、安全与稳定性，成为规模化部署的关键前提。测试不仅

要覆盖 AI Edge 的功能正确性，还需验证其在动态网络环境、跨域算力调度及复杂应用场景下的持续稳定服务能力。因此，需建立与系统架构相匹配的全栈测试体系，实现从硬件基础设施层到应用层的端到端验证，为高可靠 AI Edge 系统落地提供方法论与工具支撑。

➤ 分层测试框架

硬件基础设施层验证：重点检验终端、基站和服务器等异构算力与通信资源在 AI 驱动下的性能表现。通过终端模拟器、信道仿真仪等工具，重现多径衰落、邻区干扰和频偏失真等复杂信道特性，评估 AI 在信道估计、信号均衡和非线性补偿等 RAN 功能优化中的效果。测试目标是确认 AI 能否在高度动态环境中实现物理层的稳健性与增益。

操作系统层验证：关注异构算力抽象、实时调度与内生安全框架。通过算力负载生成器和调度监测平台，模拟多任务并发与突发计算需求，检验 AI 对 CPU/GPU/NPU 等异构资源调度的效率与公平性。同时，通过引入加密传输和虚拟化隔离测试，验证操作系统层的安全防护与低时延能力。

软件功能层验证：测试对象包括网络编排、跨域管理、资源调度与多模态感知。通过业务流量发生器和多用户业务模拟器，构建语音、视频、物联网等混合业务负载场景，评估 AI 在流量调度、拥塞控制、QoS/QoE 保障中的策略泛化与动态自适应能力。对于边缘-云协同能力，可在实验环境中建立边缘推理节点与云端训练平台，验证任务卸载、参数同步和分布式调度的时延与一致性。

应用层验证：应用层承载典型行业平台，包括低空智联、工业互联、智慧驾驶、具身智能和沉浸实时等。测试重点在于场景化端到端验证：

- 在低空智联中，模拟无人机集群跨基站切换与链路波动，检验 AI 在飞行路径规划与资源调度中的稳定性；
- 在工业互联中，注入高并发控制报文与突发负载，验证 AI 对毫秒级时延和产线稳定性的保障；
- 在智慧驾驶中，重现高速移动与多普勒效应，检验 AI 在 V2X 协同与任务卸载中的实时性；
- 在具身智能中，模拟机器人多模态感知与动作控制流量，评估 AI 在感知-决策-执行闭环中的响应与鲁棒性；
- 在沉浸实时中，生成高带宽 AR/VR 业务流，叠加时延抖动与带宽波动，检验 AI 对 QoS/QoE 的动态优化能力。

➤ 关键测试要求

AI Edge 测试需满足以下四个核心要求：

跨层一致性验证：实现从物理层空口、网络层业务到应用层平台的端到端整体评估，确保 AI 优化效果贯通全栈。

多维度场景复现能力：支持信道时变特性、多用户干扰、跨域切换和高速移动等复杂环境的精准模拟。

动态适应性与闭环优化验证：不仅评估 AI 初始性能，还需测试其在线学习与自我优化能力，形成“测试-评估-优化-再验证”的闭环迭代。

鲁棒性与安全性评估：通过网络损伤仪与攻击仿真平台，注入高噪声、异常流量和对抗样本，检验 AI 在极端条件下的容错能力与安全边界。

➤ 工具与平台支撑

为落实上述测试场景，需要构建标准化的工具体系与集成化平台：

- 实验室仿真环境：依托终端模拟器、信道仿真仪、网络损伤仪和业务流量发生器，形成可控可复现的验证环境；

- 跨层监测与分析系统：实现物理层、业务层和应用层指标的同步采集与联合分析；
- 闭环验证机制：通过实时反馈驱动 AI 模型在线调整，形成动态优化的迭代流程；
- 可视化测试平台：提供多维指标的直观展示，并开放 API 接口对接第三方工具与自动化运维系统。

综上，AI Edge 测试作为系统与平台的重要组成部分，不仅覆盖底层硬件与操作系统的算力与安全验证，更扩展到软件功能与应用层的端到端场景化验证。通过构建跨层一致、可复现、可量化的测试体系，能够系统性评估 AI 在边缘网络中的赋能价值，为 AI Edge 系统的规模化部署和可信应用奠定坚实基础。

5. 总结

借助边缘网络节点靠近用户的低时延独特优势，AI Edge 将实现无线网络内生算力的开放和高效利用，赋能低时延、高可靠通感智算控综合移动信息服务。如果说算力网是信息时代的“大动脉”，那么 AI Edge 就像无数的“毛细血管”，提供边缘分布式算力的随取随用。依托边缘分布式算力的共享底座，AI Edge 不仅能构建出支撑新型 DOICT 技术超融合的移动信息综合服务基础设施，还将能创造出开放包容的移动信息网络垂直行业应用生态，加速千行百业智能应用落地。本白皮书对 AI Edge 的产业与技术背景、技术驱动力、核心技术特征与优势、潜在场景与需求、可能的技术方向进行了较为细致的探讨，希望对业界起到抛砖引玉的作用。在未来，AI Edge 将解决一系列的难点与挑战，包括兼容多种异构算力的开放性计算平台的构建、新型网络架构设计和网络功能定义、高效的边缘 AI 技术、算子库和工具链的开发、边缘侧 APP 的开放、算力的交易与计费机制、数据安全和隐私保护机制等。

参考文献

- [1] H. Zou, Q. Zhao, Y. Tian, L. Bariah, F. Bader, T. Lestable, and M. Debbah, “TelecomGPT: A framework to build telecom-Specific large language models”, <https://arxiv.org/abs/2407.09424>, Jul. 2024.
- [2] Y. Sheng, K. Huang, L. Liang, P. Liu, S. Jin, and Y. Li, “Beam prediction based on large language models,” IEEE Wireless Communications Letters, vol. 14, no. 5, pp. 1406-1410, May 2025.
- [3] L. Bariah, Q. Zhao, H. Zou, Y. Tian, F. Bader, and M. Debbah, “Large generative AI models for telecom: The next big thing?”, IEEE Communications Magazine, vol. 62, no. 11, pp. 84-90, Nov. 2024.
- [4] T. Wu, Z. Chen, D. He, L. Qian, L. Xu, M. Tao, W. Zhang, “CDDM: Channel denoising diffusion models for wireless communications”, in Proc. IEEE GLOBECOM 2023, Kuala Lumpur, Malaysia, Dec. 2023, pp. 7429-7434.
- [5] T. Yang, P. Zhang, M. Zheng, Y. Shi, L. Jing, and J. Huang, “WirelessGPT: A generative pre-trained multi-task learning framework for wireless communication,” IEEE Network,

early access, Jun. 2025.

- [6] L. Yu, L. Shi, J. Zhang, J. Wang, Z. Zhang, Y. Zhang, and G. Liu, "ChannelGPT: A large model to generate digital twin channel for 6G environment intelligence," <https://arxiv.org/abs/2410.13379>, Oct. 2024.
- [7] B. Liu, S. Gao, X. Liu, X. Cheng, and L. Yang, "WiFo: Wireless foundation model for channel prediction," *Science China Information Science*, vol. 68, no. 6, Jun. 2025.
- [8] G. Chi, Z. Yang, C. Wu, J. Xu, Y. Gao, Y. Liu, and T. Han, "RF-Diffusion: Radio signal generation via time-frequency diffusion", in *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking (ACM MobiCom'24)*, Washington D.C., USA, Nov. 2024, pp. 77-92.
- [9] F. Zhao, Y. Sun, L. Feng, L. Zhang, and D. Zhao, "Enhancing reasoning ability in semantic communication through generative AI-assisted knowledge construction", *IEEE Communications Letters*, vol. 28, no. 4, pp. 832-836, 2024.
- [10] F. Jiang, Y. Peng, L. Dong, K. Wang, K. Yang, C. Pan, X. You, "Large AI model-based semantic communications", <https://arxiv.org/abs/2307.03492>, Jul. 2023.
- [11] L. Qiao, M. B. Mashhadi, Z. Gao, C. H. Foh, P. Xiao, and M. Bennis, "Latency-aware generative semantic communications with pre-trained diffusion models", *IEEE Wireless Communications Letters*, vol. 13, no. 10, pp. 2652-2656, Oct. 2024.
- [12] F. Jiang, L. Dong, Y. Peng, K. Wang, K. Yang, C. Pan, D. Niyato, O. A. Dobre, "Large language model enhanced multi-agent systems for 6G communications", *IEEE Wireless Communications*, vol. 31, no. 6, pp. 48-55, Dec. 2024.
- [13] M. Xu, H. Du, D. Niyato, J. Kang, Z. Xiong, S. Mao, Z. Han, A. Jamalipour, "Unleashing the power of edge-cloud generative AI in mobile networks: A survey of AIGC services," *IEEE Communications Surveys & Tutorials*, vol. 26, no. 2, pp. 1127-1170, Second Quarter, 2024.
- [14] Y. Tian, Z. Zhang, Y. Yang, Z. Chen, Z. Yang, R. Jin, T. Q. S. Quek, and K. K. Wong, "An edge-cloud collaboration framework for generative AI service provision with synergetic big cloud model and small edge models", *IEEE Network*, vol. 38, no. 5, pp. 37-46, Sep. 2024.
- [15] H. Zou, Q. Zhao, L. Bariah, Y. Tian, M. Bennis, S. Lasaulce, M. Debbah, "GenAINet: Enabling wireless collective intelligence via knowledge transfer and reasoning," <https://arxiv.org/abs/2402.16631>, Feb. 2024.
- [16] Y. Chen, R. Li, Z. Zhao, C. Peng, J. Wu, E. Hossain, H. Zhang, "NetGPT: An AI-native network architecture for provisioning beyond personalized generative services", *IEEE Network*, vol. 38, no. 6, pp. 404-413, Nov. 2024.
- [17] H. Du, G. Liu, Y. Lin, D. Niyato, J. Kang, Z. Xiong, D. Kim, "Mixture of experts for network optimization: A large language model-enabled approach", <https://arxiv.org/abs/2402.09756>, Feb. 2024.

- [18] Y. Yang, et al., “6G network AI architecture for everyone-centric customized services,” IEEE Network, vol. 37, no. 5, pp. 71-80, Sept. 2023.
- [19] IEEE GenAINet ETI, website: <https://genainet.committees.comsoc.org/home-2/>.
- [20] IMT-2030(6G)推进组. 6GAI 即服务 (AIaaS) 需求研究. 2023.
- [21] R. Singh, S. Gill, “Edge AI: A survey”, Internet of Things and Cyber-Physical Systems, vol.3, pp. 71-79, 2023.
- [22] Q. Hu, Y. Cai, Q. Shi, et al., “Iterative algorithm induced deep-unfolding neural networks: Precoding design for multiuser MIMO systems”, IEEE Transactions on Wireless Communications, vol. 20, no. 2, pp. 1394-1410, Feb. 2021.
- [23] Q. Luo, J. Zhang, S. Hu, and et al., “Joint task migration and resource allocation in vehicular edge computing: A deep reinforcement learning-based approach,” IEEE Transactions on Vehicular Technology, vol. 74, no. 6, pp. 9476-9490, Jun. 2025.
- [24] S. Liu, G. Yu, R. Yin, and et al., “Joint model pruning and device selection for communication-efficient federated edge learning,” IEEE Transactions on Communications, vol. 70, no. 1, pp. 231-244, Jan. 2022.
- [25] Y. Gao, B. Hu, M. B. Mashhadi, and et al., “PipeSFL: A fine-grained parallelization framework for split federated learning on heterogeneous clients,” IEEE Transactions on Mobile Computing, vol. 24, no. 3, pp. 1774-1791, Mar. 2025.
- [26] E. J. Hu, Y. Shen, P. Wallis, and et al., “LoRA: Low-rank adaptation of large language models,” <https://arxiv.org/abs/2106.09685>, Jun. 2021.
- [27] F. Liu, Y. Cui, C. Masouros, and et al., “Integrated sensing and communications: Toward dual-functional wireless networks for 6G and beyond,” IEEE Journal on Selected Areas in Communications, vol. 40, no. 6, pp. 1728-1767, Jun. 2022.
- [28] T. Jiao et al., “Addressing the Curse of Scenario and Task Generalization in AI-6G: A Multi-Modal Paradigm,” IEEE Transactions on Wireless Communications, vol. 24, no. 9, pp. 7377-7391, Sept. 2025.
- [29] G. Zhang, H. Li, Y. Cai, and et al., “Progressive learned image transmission for semantic communication using hierarchical VAE,” IEEE Transactions on Cognitive Communications and Networking, early access, Feb. 2025.
- [30] L. Sun, Y. Wang, A. Swindlehurst, and X. Tang, “Generative-adversarial-network enabled signal detection for communication systems with unknown channel models,” IEEE Journal on Selected Areas in Communications, vol. 39, no. 1, pp. 47-60, Jan. 2021.
- [31] Y. Wang, L. Sun, and A. Swindlehurst, “Knowledge-driven signal detector for uplink transmission in IoT networks with unknown channel models,” IEEE Internet of Things Journal, vol. 11, no. 15, pp. 25839-25852, Aug. 2024.
- [32] B. D. Son, N. T. Hoa, T. V. Chien, and et al., “Adversarial attacks and defenses in 6G network-assisted IoT systems,” IEEE Internet of Things Journal, vol. 11, no. 11, pp.

19168-19187, Nov. 2024.

- [33] J. Xiong, M. Wang, D. Zhou, and et al., “Edge intelligence: A review of deep neural network inference in resource-limited environments,” *Electronics*, vol. 14, no. 12, p. 2495, 2025.
- [34] G. Zhu, D. Liu, Y. Du, C. You, J. Zhang, and K. Huang, “Toward an intelligent edge: Wireless communication meets machine learning,” *IEEE Communications Magazine*, vol. 58, no. 1, pp. 19–25, Jan. 2020.
- [35] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y.-J. A. Zhang, “The roadmap to 6G: AI empowered wireless networks,” *IEEE Communications Magazine*, vol. 57, no. 8, pp. 84–90, Aug. 2019.
- [36] W. Zhao, W. Jing, Z. Lu, and X. Wen, “Edge and terminal cooperation enabled LLM deployment optimization in wireless network,” In *2024 IEEE/CIC International Conference on Communications in China (ICCC Workshops)*, Hangzhou, China, Aug. 2024, pp. 220-225.
- [37] S. Oh, J. Kim, J. Park, S. Ko, T. Q. S. Quek, and S. Kim, “Uncertainty-aware hybrid inference with on-device small and remote large language models,” <https://arxiv.org/abs/2412.12687>, Dec. 2024.
- [38] J. Ning, C. Zheng, and T. Yang, “DSSD: Efficient edge-device deployment and collaborative inference via distributed split speculative decoding,” in *Proceedings of the ICML Workshop on Machine Learning for Wireless Communication and Networks (ML4Wireless)*, Vancouver, Canada, Jul. 2025.
- [39] C. Zheng and T. Yang, “Communication-efficient collaborative LLM inference via distributed speculative decoding,” <https://arxiv.org/abs/2509.04576>, Sep. 2025.
- [40] R. Sapkota, K. I. Roumeliotis, and M. Karkee, “AI agents vs. agentic AI: A conceptual taxonomy, applications and challenges,” <https://arxiv.org/abs/2505.10468>, May 2025.
- [41] J. Tong, W. Guo, J. Shao, and et al., “WirelessAgent: Large language model agents for intelligent wireless networks,” <https://arxiv.org/abs/2505.01074>, May 2025.
- [42] Y. Shen, J. Shao, X. Zhang, and et al., “Large language models empowered autonomous edge AI for connected intelligence,” *IEEE Communications Magazine*, vol. 62, no. 10, pp. 140-146, 2024.
- [43] Y. Xiao, Q. Du, W. Cheng, P. D. Diamantoulakis, and G. K. Karagiannidis, “Age of trust (AoT): A continuous verification framework for wireless networks,” <https://arxiv.org/abs/2406.02190>, Jun. 2024.
- [44] Y. Xiao, Q. Du, W. Cheng, G. K. Karagiannidis, A. Nallanathan, and M. Guizani, “Redefining information freshness: AoGI for generative AI in 6G networks,” <https://arxiv.org/abs/2504.04414>, Apr. 2025.

缩略词列表

3GPP: 3rd Generation Partnership Project, 第三代合作伙伴项目计划

6GANA: 6G Alliance of Network AI, 6G 网络 AI 联盟

AAU: Active Antenna Unit, 有源天线单元

AI: Artificial Intelligence, 人工智能

AIaaS: AI as a Service, 人工智能即服务

AoT: Age of Trust, 信任年龄

API: Application Programming Interface, 应用程序编程接口

AR: Augmented Reality, 增强现实

ASC: Adaptive Semantic Compression, 自适应语义压缩

ASI: Attention-based Semantic Integration, 基于注意力的语义集成方法

B2B2C: Business to Business to Consumer, 企业通过第三方平台向消费者提供商品或服务

BAIM: Big Artificial Intelligence Model, 大型人工智能模型

BERT: Bidirectional Encoder Representations from Transformers, 基于 Transformer 的双向编码器表示模型

CDDM: Conditional Denoising Diffusion Model, 条件去噪扩散模型

CPU: Central Processing Unit, 中央处理器

CQI: Channel Quality Indicator, 信道质量指示

CSI: Channel State Information, 信道状态信息

DM: Diffusion Models, 扩散模型

DMRS: Demodulation Reference Signal, 解调参考信号

DOICT: Data Technology, Operation Technology, Information Technology, Communication Technology, 数据运营信息通信技术

DP: Differential Privacy, 差分隐私

DRL: Deep Reinforcement learning, 深度强化学习

DSA: Domain Specific Architecture, 领域专用架构

FDD: Frequency Division Duplex, 频分双工

FP32: Floating-Point 32, 32 位浮点数

FPGA: Field-Programmable Gate Array, 现场可编程门阵列

GAI: Generative Artificial Intelligence, 生成式人工智能

GenAI: Generative Artificial Intelligence, 生成式人工智能

GenAINet ETI: Large Generative AI Models in Telecom Emerging Technology Initiative, 生成式大模型新兴技术倡议委员会

Gen-SC: Generative Artificial Intelligence based Semantic Communication, 基于生成式 AI 的语义通信

GLOBECOM: Global Communications Conference, 全球通信大会

GPU: Graphics Processing Unit, 图形处理单元

GRU: Gated Recurrent Unit, 门控循环单元网络

HE: Homomorphic Encryption, 同态加密

INT8: Integer 8, 8 位整数

IoT: Internet of Things, 物联网

ITU-R: International Telecommunication Union-Radiocommunication Sector, 国际电信联盟无线电通信部门

JSCC: Joint Source-Channel Coding, 信源信道联合编码

LAM-SC: Large AI Model-Based Semantic Communications, 基于 AI 大模型的语义通信

LLM4CP: Large Language Model-empowered Channel Prediction, 大语言模型赋能的无线通信信道预测

LLM4WM: Large Language Model for Wireless Multi-Tasking, 大语言模型赋能的无线多任务处理

LoRa: Long Range Radio, 远距离无线电

LoRA: Low-Rank Adaptation, 低秩自适应

LSTM: Long Short-Term Memory, 长短期记忆网络

MaaS: Model as a Service, 模型即服务

MEC: Mobile Edge Computing, 移动边缘计算

MoE: Mixture of Experts, 混合专家模型

MPC: Secure Multi-party Computation, 安全多方计算

MR: Mixed Reality, 混合现实

MSE: Mean-Square Error, 均方误差

MWC: Mobile World Congress, 世界移动通信大会

NAS: Neural Architecture Search, 神经网络架构搜索

NPU: Neural Processing Unit, 神经网络处理器

PaP: Prompt-as-Prefix, 提示作为前缀

PLC: Programmable Logic Controller, 可编程逻辑控制器

PMI: Precoding Matrix Indicator 预编码矩阵指示

QoAIS: Quality of AI Service, 智能信息服务质量

QoE: Quality of Experience, 用户体验质量

QoS: Quality of Service, 服务质量

RAN: Radio Access Network, 无线接入网

RCN: Radio Computing Network, 无线计算网络

RF: Radio Frequency, 射频

RI: Rank Indicator, 秩指示

RISC-V: Reduced Instruction Set Computer - V, 第五代精简指令集

RRU: Remote Radio Unit, 射频拉远单元

SaaS: Software as a Service, 软件即服务

SAM: Segment Anything Model, 图像分割模型

SemCom: Semantic Communication, 语义通信

SKB: Segment Anything Model based Knowledge Base, 基于 SAM 的知识库

SKT: SK telecom, SK 电信集团

SoC: System on a Chip, 片上系统

SRZ: Service Requirement Zone, 服务需求区域

TaaS: Training as a Service, 训练即服务

TEE: Trusted Execution Environment, 可信执行环境

UPF: User Plane Function, 用户面功能

USR: User Satisfaction Ratio, 用户满意度

V2G: Vehicle-to-Grid, 电动汽车与电网互动

VR: Virtual Reality, 虚拟现实

WiFo: Wireless Foundation Model, 无线基础模型

白皮书贡献者列表

本白皮书由鹏城实验室牵头组织撰写，国内外 20 家企业、高校、科研机构的 50 余位专家学者共同完成。鹏城实验室负责材料的汇总和统稿。各章节的贡献单位及贡献人列表如下（排名不分先后）。

章节		贡献单位	贡献人
1. 背景与需求		中国移动	崔莹萍, 王新尧, 陈天骄
		紫金星宇(南京)科技有限公司	陈雷, 彭劲东
		鹏城实验室	孙黎, 黄宁
		叠拓信息技术有限公司	季清
		香港科技大学	杨旻, 马牧雷
2. AI Edge 的技术内涵		鹏城实验室	杨婷婷, 黄宁, 孙黎
3. AI Edge 的典型应用场景和潜在价值		紫金星宇(南京)科技有限公司	陈雷, 彭劲东
		中信科移动	王亚鹏
		香港科技大学	杨旻, 马牧雷
		中国移动	崔莹萍, 王新尧, 陈天骄
		中国电信	吴超, 王晴天
4. AI Edge 的技术方向与主要挑战	4.1 系统架构	中国电信	王越, 王晴天, 李泽旭, 王靖壹
		中兴通讯	杨立, 孙文文, 谢峰
		香港科技大学	杨旻, 马牧雷
	4.2 AI for Edge 技术	华为技术有限公司	王坚, 高慧国
		上海大学	周婷, 刘胜利
		中科院计算所	王聪聪, 齐彦丽, 于含笑
		西安交通大学	王熠晨

		香港中文大学（深圳）	崔曙光，任金科，张泽中，许杰
		英国萨里大学	Chong Huang, Pei Xiao
		鹏城实验室	杨婷婷，孙黎
		北京触点互动信息技术有限公司	王航，毛俊
		紫金星宇（南京）科技有限公司	张珍兵，陈雷
		上海诺基亚贝尔股份有限公司	陶涛，蔡立羽
	4.3 AI over Edge 技术	鹏城实验室	孙黎，郑策
		中信科移动	王亚鹏
		香港科技大学	Jun Zhang
		香港中文大学（深圳）	崔曙光，任金科，张泽中，许杰
		西安交通大学	杜清河，肖玉权，张朝阳
		上海大学	周婷，刘胜利
		英国萨里大学	Chong Huang, Pei Xiao
		叠拓信息技术有限公司	季清
	4.4 芯片与算力底座	中科睿芯	杨子欣
		上海大学	周婷，姜之源，刘胜利
		壁仞科技	段柳成，刘泽
	4.5 AI Edge 系统、平台与测试	中国联通	李福昌,张涛,马艳君
		紫金山实验室	黄永明,刘泽宁,尤建洁
		北京触点互动信息技术有限公司	王航,毛俊
		紫金星宇（南京）科技有限公司	顾亮，彭劲东，郭文婷
	前言、总结及其他		鹏城实验室

在白皮书策划及撰写过程中，AI Edge 联盟专家委员会主席尤肖虎院士及专家委员会全体委员杨婷婷、吴建军、刘光毅、王越、朱尔霓、李福昌、段向阳、李俨、Shuguang Cui、李革、张朝阳、杨旻、周一青、周婷、孙程君、孙韶辉、常晓涛、Pei Xiao、Lin Cai、Sumei Sun、Mérrouane Debbah 针对白皮书内容提出了大量宝贵的意见和建议。在此谨对上述专家的辛勤付出和认真指导致以崇高敬意和衷心感谢！