下载日志已记录，仅供内部参考

中信建投证券 CHINA SECURITIES

证券研究报告·行业深度

端侧：AI 的下一个浪潮

——端侧行业研究报告

电子

核心观点

当前算力主要部署在云侧，我们认为端侧不仅能成为云侧的入口，更能够提供隐私保护、即时响应、低成本、以及主动响应的差异化价值，从而有望成为 AI 的下一个浪潮。从场景上看，端侧 AI 不仅仅能够应用在如手机、PC、眼镜等大的消费终端，端侧 AI 的价值更在于多个碎片化场景的渗透，这体现在端侧 AI 会赋能传统成熟产品的升级，同时端侧 AI 也会带动全新终端的诞生，最后，从落地上看，成熟产品的升级迭代落地更快，目前已能看到部分产品开始放量，而全新 AI 的终端需要更长的开发周期，需要将模型、场景与算力深度融合，围绕用户核心痛点，打造系统级能力，端侧 AI 从“通用能力”到“行业价值”的最后一公里仍需模型厂与终端相互协同，也会在未来逐步落地，我们认为这些碎片化场景的逐步渗透是国内相关龙头公司业绩能够持续快速增长的核心原因之一。

摘要

端侧 AI 的价值：入口、主动响应、隐私、低延时与成本

当前算力主要部署在云侧，我们认为端侧不仅能成为云侧的入口，更能够提供隐私保护、即时响应、低成本、以及个性化主动式的差异化价值。首先，端侧 AI 可作为 AI 进入物理世界的神经末梢，端侧在本地主要负责感知、唤醒、第一层判断，把需要深度思考的任务脱敏处理后交给云端大模型；其次，端侧 AI 深度融合多维度实时数据，拓宽个性化服务的覆盖边界与体验深度，推动服务逻辑从“被动响应”升级为“主动预判”，云端 AI 无法获取、分析用户全面的本地数据，在实现“主动式 AI”、深度理解用户需求方面，端侧 AI 具备独特的优势；此外，对于工业控制和自动驾驶等场景，云侧算力的延迟会直接影响交互流畅度、任务连续性乃至安全边界，而端侧通过本地化部署能够即时响应，同时对于金融风控、医疗辅助、企业办公等场景，端侧处理能够显著减少敏感数据离开设备或组织边界的次数，具备较强的隐私保护价值。最后，相比云端，端侧 AI 在数据处理需求高或推理请求量大等的高频任务场景具有成本优势，特别长上下文、多轮对话、多工具调用和多智能体协同时，单纯任务的 Token 消耗可能成倍放大，而端侧 AI 采用硬件成本模式，长期成本基本确定。

维持 强于大市

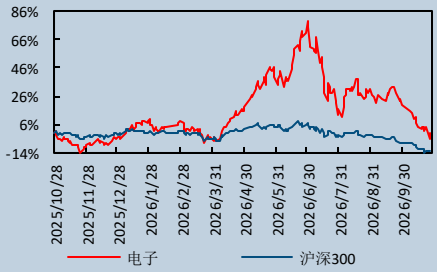
姚康
yaokang@csc.com.cn
SAC 编号:S1440526100001

梁艺
liangyidzz@csc.com.cn
SAC 编号:S1440525080006

何昊
hehaodzz@csc.com.cn
SAC 编号:S1440526030011

发布日期： 2026 年 10 月 29 日

市场表现



相关研究报告

本报告由中信建投证券股份有限公司在中华人民共和国（仅为本报告目的，不包括香港、澳门、台湾）提供。在遵守适用的法律法规情况下，本报告亦可能由中信建投（国际）证券有限公司在香港提供。同时请参阅最后一页的重要声明。

Remove Demo Watermark from http://www.wancaiinfo.com/pdfmaster

下载日志已记录，仅供内部参考

端侧 AI 的场景：AI 赋能千行百业，AIoT 多彩缤纷

端侧 AI 不仅仅能够应用在如手机、PC、眼镜等大的消费终端，端侧 AI 的价值更在于多个碎片化场景的渗透。一方面，端侧 AI 会赋能传统成熟产品的升级，例如智能家居、清洁类机器人、大屏等，这些成熟产品将从过去的自动化、数字化向智能化、AI 化过渡升级，端侧 AI 在这些场景的落地也会更快，目前已能看到部分产品开始放量；另一方面，端侧 AI 也会带动全新终端的诞生，例如各类 AI BOX、服务型机器人、AI NAS、工业质检等，而全新 AI 终端需要更长的开发周期，需要将模型、场景与算力深度融合，围绕用户核心痛点，打造系统级能力，同样会在未来逐步落地；最后，伴随 Personal Agent 的发展成熟、Muse Gadgets 等开源硬件开发工具包的推出，端侧 AI 有望成为云侧 AI 的物理载体，较低成本的外设类硬件有望快速爆发，我们认为这些碎片化场景的逐步渗透是国内相关龙头公司业绩能够持续快速增长的核心原因之一。

端侧 AI 的落地：模型、场景与算力需深度融合，系统级能力是规模化落地的关键

与云端大模型主要围绕算力集群和模型服务展开不同，端侧 AI 对软硬件适配、功耗控制、响应时延、系统集成和场景理解提出了更高要求，需要模型、终端与芯片多方协同，其规模化落地并非单一环节能够独立完成。回顾云端 AI 的发展历程，模型能力的提升扮演关键角色，我们认为端侧同样如此，随着端侧模型能力密度的提升，以及轻量化的发展，端侧通用模型的能力持续突破，例如 30B 参数级别的小模型智能水平正快速逼近千亿级大模型，3-7B 的小模型也能够在本地完成问答、文档处理、视觉语音识别等中难度任务，然而对于行业场景，端侧 AI 从“通用能力”到“行业价值”的最后一公里仍需模型厂与终端相互协同；

此外，算力也是端侧体验的重要一环，一颗好用的端侧算力芯片标称算力只是起点，在实际使用中，工具链、内存带宽与实时功耗等也是影响推理体验的关键，为应对端侧 AI 在终端差异化的部署要求，国内头部 SoC 公司均在加速产品迭代，同时不断完善工具链以及生态，例如瑞芯微推出 RK182X 协处理器，采用 3D 堆叠的创新架构从而突破存储的带宽限制，我们认为在 AIoT 领域，国内头部 SoC 具备全球竞争力，而海外头部厂商更侧重云侧 AI 以及手机、PC 等核心消费终端芯片，伴随产品迭代加速，这些国内头部 SoC 将充分受益端侧 AI 浪潮，保持快速成长态势。

投资建议

我们看好端侧 AI 未来在千行百业的逐步落地，重点关注 SoC 产业链，重点推荐标的：

1）瑞芯微：公司作为 AIoT SoC 领导者，构建“芯片+算法+行业方案”的全栈能力，过去十年在千行百业里沉淀的软件工具和生态积累十分深厚，特别是 AI 协处理器通过 3D 堆叠架构突破内存带宽限制，目前已有数百个客户、几十个行业在深化研发中，首批客户在清扫/陪伴机器人、NAS、智能家居等已进入产品发布、量产阶段；

2）乐鑫科技：公司围绕“处理+连接”的产品战略，持续迭代 AIoT 芯片等产品，S31、P4 等新产品放量有望驱动公司营收加速成长，工控领域、智能农业、新型开发者市场等新兴下游业务快速成长，同时，公司重点关注产品生态建设，通过 ESP-Claw、ESP-Vision 以及对 MCP 等前沿 AI 开发技术的支持，进一步降低 AI 与物理设备结合的门槛，以吸引更多开发者并扩大产品的应用领域；

3）晶晨股份：公司以 TV SoC 和机顶盒 SoC 为基础，积极布局 AIoT 通用算力平台，推出 A9 高算力 6nm 端侧 AI 芯片，由积累的电视屏幕大客户出发，已推进至机器人、边缘计算、智能座舱等领域，AIoT 新品与头部大客户达成合作，预计 26 年下半年量产，有望成为长期增长引擎；

4）恒玄科技：公司是全球领先的低功耗无线计算 SoC 芯片设计企业，2026 年上半年，受存储价格上涨、国补政策退坡影响，下游终端市场景气度下降，公司经营承压，我们认为公司拐点在即，公司新一代旗舰 BES6100 已于 2026 年上半年完成流片，预计下半年将进入测试及客户送样阶段，有望助力公司开启新的成长周期；

Remove Demo Watermark from http://www.wancaiinfo.com/pdfmaster

下载日志已记录，仅供内部参考

目录

一、价值：入口、主动响应、隐私、低延时与成本 1
二、场景：AI 赋能千行百业，AIoT 多彩缤纷 3
2.1 Personal Agent 的硬件载体 3
2.2 机器人：多形态机器人量产落地，数采市场有望快速兴起 6
2.3 AI NAS：构筑智能家庭与企业中枢 9
2.4 办公：AI Agent 重构生产力，端侧落地实现主动式 AI 体验 11
2.5 智能显示：端侧 AI 助力构建新一代多媒体平台 13
2.6 汽车电子：车载 AI Box 赋能座舱智能化升级 14
2.7 工业：AI 重塑机器视觉 15
三、落地：模型、场景与算力需深度融合，系统级能力是规模化落地关键 18
3.1 模型与场景：端侧模型能力持续进步，不同场景适配不同参数模型 18
3.2 算力：国内头部 SoC 厂商充分受益于端侧 AI 浪潮 20
3.2.1 峰值算力、内存带宽、软件工具以及生态效率共同作用于 AIoT 算力芯片 20
3.2.2 国内头部 SoC 公司产品迭代加速，满足多样化算力要求 21
风险分析 27

Remove Demo Watermark from http://www.wancaiinfo.com/pdfmaster

下载日志已记录，仅供内部参考

一、价值：入口、主动响应、隐私、低延时与成本

当前的算力部署主要集中在云端，但云端集中式推理模型在大规模应用中遭遇结构性瓶颈，端侧 AI 的兴起是确定性趋势。随着大模型从“尝鲜式应用”走向“生产级部署”，推理调用量、上下文长度、智能体工具调用次数和多模态数据处理需求同步上升，**纯云端方案正在同时面临成本、时延、隐私合规和服务稳定性四类约束，**而这些约束不能简单地依靠继续增加云端算力供给加以根本化解。云端模型无法突破物理时延边界与敏感数据监管要求，弱网、断网场景下也无法继续应用，而调用量爆发使云端推理经济上持续承压，企业高度担忧云端 AI 成本，云端 AI 也无法为用户提供主动响应、个性化服务。因此，端侧智能的兴起成为解决云端约束的必然结果。

端侧 AI 通过本地化部署，可覆盖隐私保护、即时响应等多项需求。云端推理的体验瓶颈并不只来自模型生成速度，也来自“终端采集—网络传输—云端排队—模型推理—结果回传—终端执行”的完整链路。对于普通文本问答而言，数百毫秒到数秒的响应延迟仍可被用户容忍；但对于语音助手、实时翻译、图像理解、AR 交互、智能座舱、工业控制和自动驾驶等场景，延迟会直接影响交互流畅度、任务连续性乃至安全边界，例如自动驾驶场景通常要求 10~50 毫秒级响应，以保障安全导航；又例如工业机器人、自动化产线等场景百毫秒级延迟足以导致任务中断、执行偏差或安全风险。而端侧模型可以将响应从“等待云端返回”转变为“本地即时处理”，从而显著改善交互体验。同时，在弱网或无网环境下，端侧 AI 可提供稳定性服务，可靠性相比云端更强，适配场景更加多元。

在隐私方面，通过在本地完成语音识别、图像理解、行为识别、个性化推荐和初步推理，终端可以在“不上传原始数据”或“仅上传脱敏结果”的前提下提供智能服务。对于金融风控、医疗辅助、企业办公、智能座舱和个人助理等场景，端侧处理能够显著减少敏感数据离开设备或组织边界的次数，具备较强的隐私保护价值。

相比云端，端侧 AI 在高频使用场景下具有成本优势。云端 AI 主要按 Token/API 计费，初始成本低但是边际推理成本较高，长期成本随 API 调用频率线性上升，而在长上下文、多轮对话、多工具调用和多智能体协同时，单纯任务的 Token 消耗可能成倍放大。而端侧 AI 采用硬件成本模式，即一次性买断，后续推理在本地完成，边际成本几乎只有电费与折旧，无需付出 API 调用费用，长期成本基本确定。因此，在数据处理需求高或推理请求量大等的高频任务场景，如果不刚需云端大模型的超大规模能力，选用端侧 AI 方案在长期成本上具有优势。

端侧 AI 也可作为 AI 进入物理世界的神经末梢，提供低成本硬件 AI Agent 的实现路径。AI 的趋势不再只是电脑和手机上的大模型，而是让家电、穿戴等小设备自带感知与交互。因此，AI 下沉至端侧的低成本路径可以是在端侧部署轻量 AI 模型，端侧主要负责感知、唤醒、第一层判断，把需要深度思考的任务脱敏后交给云端大模型，实现“端—云”协同，既保持端侧 AI 的低时延、高隐私特点，也充分利用云端大模型的高算力

端侧 AI 可以通过深度学习、感知本地数据，提供个性化体验以及主动式服务。端侧 AI 通过分析本地数据，在设备端动态更新知识库与服务策略，实现基于本地数据的深度个性化适配。一方面，端侧 AI 依托本地部署、低延迟的优势深度挖掘实时交互数据，精准匹配场景核心需求；另一方面，端侧 AI 深度融合多维度实时数据，拓宽个性化服务的覆盖边界与体验深度，推动服务逻辑从“被动响应”升级为“主动预判”。云端 AI 无法获取、分析用户全面的本地数据，在实现“主动式 AI”、深度理解用户需求方面，端侧 AI 具备独特的优势。

Remove Demo Watermark from http://www.wancaiinfo.com/pdfmaster

下载日志已记录，仅供内部参考

表 1: 端侧 AI 与云端 AI 对比

维度	云端 AI	端侧 AI
响应速度	受网络影响，一般百毫秒级	10-100 毫秒
隐私安全	数据上传云端，隐私数据无法处理	数据不出设备，低泄露风险
成本	随 Token/API 用量线性提升	基本只有一次性成本
网络需求	依赖网络，受地缘约束	可离线运行
优势任务	复杂计算、大数据存储等	物理感知、个性化适配等

资料来源：中信建投证券

端侧大模型可通过精细化分层架构设计，实现端云融合互补。面对实际应用场景，端侧与云端模型并非二选一，更优解是将任务分层，达成“端-边-云”三级资源的动态优化配置与高效调度，一方面对任务智能分流实现全域计算协同，不同类型任务分别交给不同层级处理，云端负责处理复杂计算及大数据存储，边缘层完成本地实时分析和无线带宽连接，终端层涵盖个人设备与行业终端，完成对物理世界的感知、控制与本地推理，充分利用端侧的响应速度与边云的算力优势；另一方面实现多设备智能互动，弥补单一终端算力、感知局限，各设备功能优势互补。

图 1: 大小模型端云协同架构



数据来源：Frost & Sullivan，中信建投证券

Remove Demo Watermark from http://www.wancaiinfo.com/pdfmaster

下载日志已记录，仅供内部参考

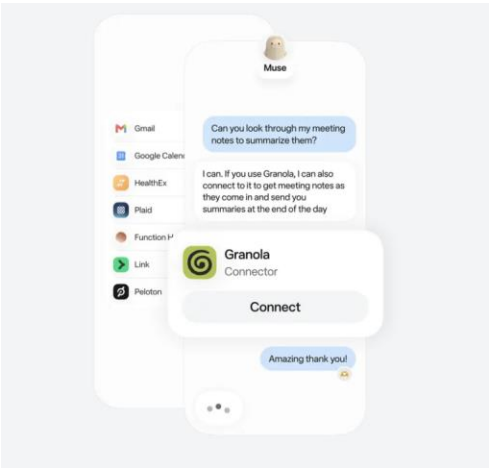
二、场景：AI 赋能千行百业，AIoT 多彩缤纷

端侧 AI 不仅仅能够应用在如手机、PC、眼镜等大的消费终端，端侧 AI 的价值更在于多个碎片化场景的渗透。一方面，端侧 AI 会赋能传统成熟产品的升级，例如智能家居、清洁类机器人、大屏等，这些成熟产品将从过去的自动化、数字化向智能化、AI 化过渡升级，端侧 AI 在这些场景的落地也会更快，目前已能看到部分产品开始放量；另一方面，端侧 AI 也会带动全新终端的诞生，例如各类 AI BOX、服务型机器人、AI NAS、工业质检等，而全新 AI 终端需要更长的开发周期，需要将模型、场景与算力深度融合，围绕用户核心痛点，打造系统级能力，同样会在未来逐步落地；最后，伴随 Personal Agent 的发展成熟、Muse Gadgets 等开源硬件开发工具包的推出，端侧 AI 有望成为云侧 AI 的物理载体，较低成本的外设类硬件有望快速爆发。本篇报告我们更加聚焦于端侧 AI 在 AIoT 中的应用，我们认为这些碎片化场景的逐步渗透是国内相关龙头公司业绩能够持续快速增长的核心原因之一。

2.1 Personal Agent 的硬件载体

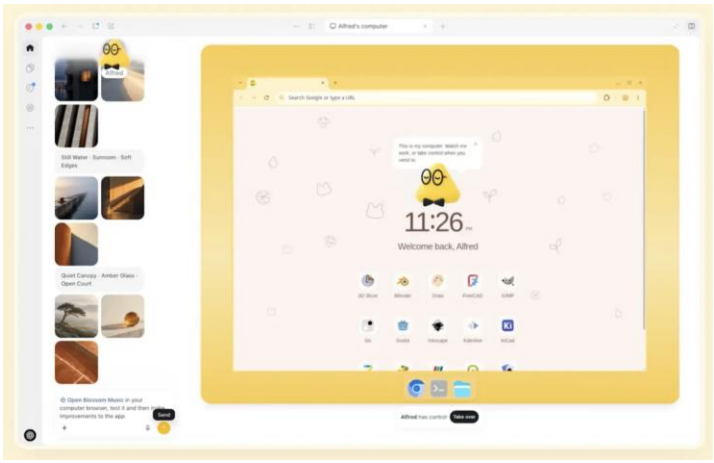
Personal Agent 全球火爆，全面覆盖工作与生活。Meta 与 OpenAI 相继布局 PersonalAgent，AI 产品由单次对话或任务转向长期陪伴服务用户。Personal Agent Muse 于 2026 年 9 月 8 日推出，用户仅需下达目标，Muse 即可根据用户需求分析生成任务流程，可智能完成预定行程、在线购物、收发邮件等日常事务和项目，并且用户关闭应用后仍可在云端继续执行。根据 Apptopia 统计，Muse 上线 12 天累计下载约 280 万次，同期 ChatGPT 是 130 万、Claude 约 40 万，Muse 上线 12 天内在美加移动端下载量已超 ChatGPT 同期水平，并登顶 App Store 榜首。同月 29 日，OpenAI 推出 Dots，同样主打全天候为用户推进工作、分担重要任务，由 GPT-6 Astra 驱动，能从反馈中学习，实现逐渐了解用户偏好与优先级，在执行任务之外参与“发现任务”环节。同时 OpenAI 打造专家型 Dot，可在组织和公司内承担专门职责。国内多家厂商跟进 Personal Agent，例如阿里千问助手升级为千问 GO，具备购物、出行、物流、编程等数十个领域的个人办事能力。

图 2:Meta-Muse



数据来源：Muse，中信建投证券

图 3: OpenAI Dots



数据来源：OpenAI，中信建投证券

端侧能够赋能 Personal Agent 体验升级，成为其重要的硬件载体。从留存率方面来看，在美国市场，Muse 7 天后的留存率为 30%，这一指标与 ChatGPT 上线初期相当，然而 14 天后的留存率为 23%，这一指标低于

Remove Demo Watermark from http://www.wancaiinfo.com/pdfmaster

下载日志已记录，仅供内部参考

ChatGPT 同期数据，这说明纯云侧的个人助理在用户体验上依旧有欠缺，例如用户仍需较长时间与 AI 交互，并且部分关键网页也可能关闭接口权限，部分用户也会担心隐私问题，同时从商业化上来看，Agent 所消耗的 Token 可能更高，然而与 AI Coding 场景不同，Personal Agent 场景用户对于性价比的要求较为极致，若无法为用户提供足够好体验的服务，用户的付费意愿也较低，因此用户体验十分影响 Personal Agent 的商业化和后续发展。然而硬件载体的出现能够较好提升用户体验，与云侧 AI 相互补充，一方面终端硬件通常配有麦克风、摄像头等各类传感器，使得 AI 在终端设备上的实时响应和“主动感知”成为可能，AI 从单纯被动式响应升级为“主动预判”，另一方面，在终端部署的端侧小模型主要负责感知、唤醒、第一层判断，之后再将更复杂的推理任务交给云侧，减少部分云侧的推理压力。此外，从争夺流量入口的角度来看，各大模型厂商也需积极布局消费硬件赛道。

OpenAI 加速布局 AI 硬件产品，实现“模型+终端”的硬件闭环。OpenAI 在 2025 年 5 月宣布以约 65 亿美元收购由苹果前首席设计师 Jony Ive 创立的 AI 硬件公司 IO，同时多位来自苹果的供应链、产品设计、制造、人工智能研究等人才加入，在产品设计理念上，设计团队希望创造一种“始终陪伴、几乎隐形”的 AI 伴侣，强调语音交互和情境感知，颠覆传统的屏幕依赖模式，在具体产品形态上，OpenAI 正在考虑的产品可能包括类似没有屏幕的智能音箱，智能眼镜、数字录音设备以及可穿戴 AI Pin 等多个方向。

OpenAI 消费级硬件即将面世，背靠 ChatGPT 推出全新物理载体。彭博社爆料，OpenAI 与 Jony Ive 合作的首款硬件产品即将面世，产品外形像甜甜圈、体积与冰球相仿。这款 AI 终端目标成为“以 AI 为核心的便携计算机”，没有屏幕、有麦克风，主要通过语音交互，可能配置摄像头帮助产品进行环境感知，配合指示灯可完成交互提示，主体与充电底座分离便于携带。这一物理终端背靠 OpenAI 的 ChatGPT Live，模型能力全球领先，支持语音、图像多模态交互，推动智能设备向理解、决策、执行进化。OpenAI 首批产品计划于 2026 年底或 2027 年初发布，我们也期待 OpenAI 等头部大模型公司的入局能够打造新的 AI 终端。

Muse 开源硬件开发工具 Muse Gadgets 项目，打通从虚拟助手走向真实环境的链路。Meta 选择软件与硬件解耦，将硬件载体的定制权完全让渡给第三方设备与开源社区。Meta 依托 Personal AI Agent—Muse 建立云端智能与连接服务，而 Muse Gadgets 则给它增加了硬件接口，开发者可利用现成的单板计算机、微控制器与显示屏幕组合出适用的物理形态，模块化开源生态赋予云侧 AI 的物理外设无限可能。

图 4: Muse Gadgets 参考设计



数据来源：Github，中信建投证券

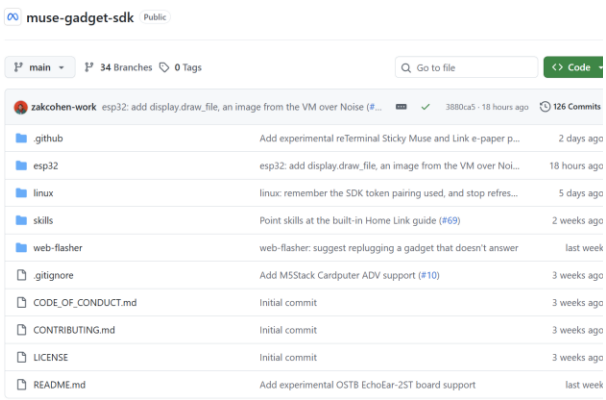
Muse 开源硬件可搭配 ESP32 开发板、树莓派等多种现有硬件。Meta 在 Github 上以 Apache 2.0 许可发布了 ESP32 固件和 Linux 设备 SDK，开发者可把 Muse 连接到显示屏、按钮、传感器、执行器，以及工作台上任何可联网的设备。Linux 设备 SDK 开发套件面向具备完整操作系统的单板计算机，如树莓派、Linux 主机，能够通过系统服务常驻后台托管。ESP32 设备固件则专门面向搭载乐鑫 ESP32 的开发板，支持开发者直接调用显示

Remove Demo Watermark from http://www.wancaiinfo.com/pdfmaster

下载日志已记录，仅供内部参考

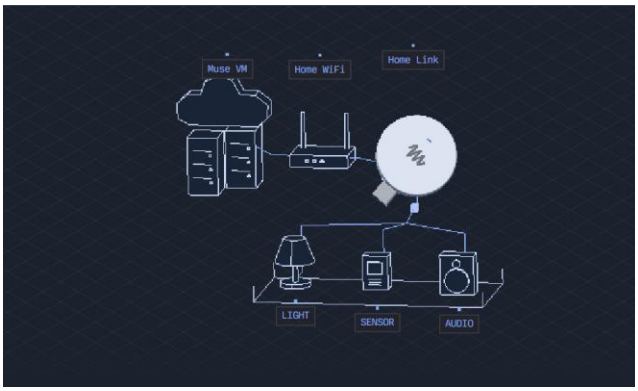
屏展示图像或文本，仅凭几十元的基础芯片赋予硬件语音交互与信息呈现能力。针对不想从零开始开发的用户，Muse 也提供基于 ESP32-C5 的成品 Muse Home Link，部分舍弃开发生态换取更便捷的使用体验。

图 5:Muse Gadget SDK



数据来源：Muse，中信建投证券

图 6: Muse Home Link 示意图



数据来源：Muse，中信建投证券

ESP32 作为 Muse 的端侧入口，极大节约了端侧硬件的成本。在 Muse 开源的 SDK 中，ESP32 主要用于输入与反馈，即在用户触摸采集按钮的时候，感知环境传感器数据、触摸动作以及微型摄像头的拍摄指令，通过 Wi-Fi 传回 Muse 云端智能体，而在 Muse 返回执行决策后，在现实中驱动状态指示灯、小型墨水屏/彩屏显示、蜂鸣器音效等，实现与真实世界的交互。通过将大模型和真实硬件接起来，云端负责理解和生成，端侧负责感知、控制、联网，这种软硬件生态可以取得端侧与云侧的平衡，以极低的成本取得更好的使用体验。

除了 Meta，Anthropic 在其桌面应用 Claude 中向开发者开放了一个面向硬件设备的 Bluetooth API，开发者可借助这一接口构建与 AI 实时交互的物理设备。官方示例中，团队使用基于 ESP32 的开发板构建了一个“桌面宠物”，在 Claude 需要用户授权时主动提醒，并支持在硬件端的交互。ESP32 具备极强的开发者属性，单块仅 20 元出头，凭借其易用性、低价格、低功耗的特性适合在消费者市场普及，相关类型产品有望快速爆发。

图 7: Claude“桌面宠物”



数据来源：乐鑫董办，中信建投证券

Remove Demo Watermark from http://www.wancaiinfo.com/pdfmaster

下载日志已记录，仅供内部参考

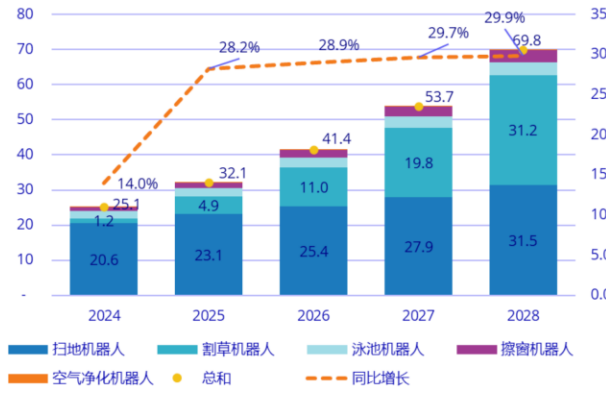
2.2 机器人：多形态机器人量产落地，数采市场有望快速兴起

端侧 AI 赋能消费级机器人，应用场景全域扩张。消费级机器人是 AI 大模型的最佳落地场景之一，具备“硬件+内容+服务”的生态潜力，相比过去主要以概念展示为主的“类人机器人”，面向真实家庭、商用场景需求的消费级机器人开始密集出现，这类机器人能够以其感知、决策、执行能力感知真实世界，能够为用户提供显著价值，是 AI 走向物理世界的重要入口。消费级机器人形态多样，可以分为功能性机器人、服务类机器人以及陪伴类机器人，其中功能性机器人是当前商业化相对成熟的应用场景，能够完成单一任务，替代重复劳动，任务场景覆盖清洁、割草等。

功能性机器人增速稳健，割草、擦窗机器人等细分品类增速亮眼。根据 IDC 的数据，2025 年全球智能家居清洁机器人市场出货 3210 万台，同比增长 28.2%，预计到 2028 年五年复合增长率达 26%，从细分品类情况来看

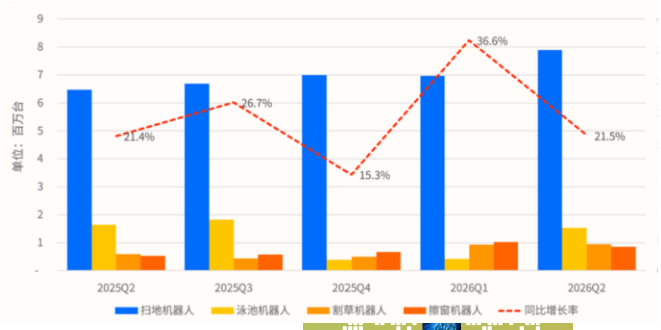
- 扫地机器人 2026 年上半年出货超 1400 万台，同比增速超 20%，扫地机器人的快速增长主要得益于欧洲区域，Q1/Q2 欧洲扫地机器人市场分别同比增长 81%/64%。
- 割草机器人 26H1 出货将近 200 万台，Q1/Q2 同比增速分别为 97%/59.7%，割草机器人高景气度的背后主要得益于 AI 带动的智能化升级，传统割草机器人普遍依赖预埋边界线作业，人工布线费用高昂、施工繁琐，场景适配灵活性偏弱，面对不规则草坪、零散绿化区域、复杂围挡地形时容易出现越界行驶、漏割、重复割草等问题，智能割草机可通过主控多路高清摄像头同步采集画面，开发视觉自主建图、草坪边界识别、全域路径规划等功能，精准区分复杂场景，大幅降低整机硬件与售后施工成本，满足海外庭院适配性及成本需求。
- 擦窗机器人同样保持高景气度，无线化、智能喷水、边角清洁等技术升级显著改善产品体验，叠加欧美等市场的大面积玻璃、高层住宅及人工清洁成本进一步强化需求，同时中国品牌凭借成熟供应链和快速产品迭代加速全球渠道铺设，使擦窗机器人从小众尝鲜品类逐步走向大众化，26 年上半年擦窗机器人出货将近 200 万台，同比增速超 60%。

图 8:家用清洁机器人出货量及预测（百万台）



数据来源：IDC，中信建投证券

图 9:25Q2-26Q2 全球家用清洁机器人出货量



数据来源：IDC，中信建投证券

Remove Demo Watermark from http://www.wancaiinfo.com/pdfmaster

下载日志已记录，仅供内部参考

服务类机器人往往配备更为复杂的传感器及 AI 功能，核心功能包括自主导航与定位、自动感知与识别、人机交互、任务调度、运动控制等，其聚焦于特定场景，实现多重功能。举例而言，在家庭场景，2026 年，科沃斯推出行业内首款搭载 OpenClaw 的具身智能机器人，集成双 RK3588 高阶算力、静音移动底盘、六自由度机械臂和多模态感知系统，覆盖桌面、地面全场景作业，具备突出扩展性与续航能力。八界机器人突破过去家庭服务机器人的立体整理限制，全面覆盖家务需求。通过集成 OpenClaw，八界可以理解口语化指令，同时在面临家务场景中的大量复杂场景与任务时不断学习，持续突破能力边界，最终实现主动服务。

在户外及工业等非标准化场景，受制于崎岖地形、斜坡、楼梯及恶劣天气等因素，传统的机器人在巡检、后勤保障、应急响应等关键任务的部署较为困难，PUDU D5 系列采用 RK3588 协同 NVIDIA Orin，支持导航、目标识别、避障规划等多线程任务，可部署端侧大模型实现 AI 决策，在复杂的动态环境中也能实现稳定的自主运行，PUDU D5 系列可实现百万平方米级别场地的建图与自主导航，覆盖“出发-巡逻-障碍处理-自动回充”全流程自主作业，为机场、地铁、工业园区等大规模场景提供解决方案。

网球机器人同样是近年来正在兴起的新应用场景，与传统发球机相比，网球机器人能够实现实时感知球场环境，分析来球轨迹与速度，作出决策并自主运动完成击球回合，实现与人类的全场对抗或陪练，一台网球机器人能够集球童、教练、陪练等多个功能于一体，为用户创造多重价值，例如卧安机器人旗下的 Acemate，成了赛事指定网球训练机器人供应商，累计出货量已经从年初的 2000 多台增加到了 26 年 9 月的 6000 多台。

母婴育儿同样是端侧 AI 部署的潜力场景，针对“养育压力大、信息繁杂、保姆成本高”三大痛点，母婴育儿机器人能够通过传感器实测监控婴幼儿状态，在端侧/云端分析决策， 驱动硬件执行如自动安抚、智能喂养、睡眠监护等功能，实现感知、判断、执行、反馈的闭环，可以替代部分人工照护职责的“AI 虚拟月嫂”。

图 10: 部分服务机器人应用



数据来源：瑞芯微电子，科沃斯官网，普渡官网，中信建投证券

开源机器人“机器鸭”全球爆火。Pollen Robotics 推出的桌面级双足机器人 Microduck，售价仅 399 美元，开售 24 小时销售额便突破 260 万美元，高峰期平均每 4 秒卖出一台。Microduck 作为被赋予了宠物感的具身智能终端，能完成行走、踢腿、抓握、摔倒自站立等 7 种出厂动作，还支持用户训练专属音色与个性化技能，把具身智能的体验门槛拉到了消费级。Microduck 能在全球开发者社区引爆，依托的是瑞芯微 RK3566 处理器，

Remove Demo Watermark from http://www.wancaiinfo.com/pdfmaster

下载日志已记录，仅供内部参考

四核 A55 架构与 1 TOPS NPU 的 SoC 正好实现端侧 AI“算力—功耗—响应”的平衡。同时，整机仅配 1GB 内存与 32GB 存储，RK3566 表现出极致的性价比。其配套的 RKNN 工具链原生兼容 TensorFlow、PyTorch 等主流框架，叠加全球上亿台终端设备所沉淀的海量开源方案与社区问答，让开发者实现开箱即用。

图 11: 机器鸭 Microduck



数据来源：Pollen Robotics，中信建投证券

除各形态垂类机器人之外，具身智能庞大的训练数据要求带动数采市场正在快速兴起。当前机器人泛化能力推进慢、落地不及预期，其核心原因是真实数据的供给矛盾，预训练阶段数据规模的扩大，能够支撑具身智能认知和行动能力的显著提升，例如 Generalist AI 的新一代模型 GEN-1 的预训练数据较前代 GEN-0 模型多近一倍，平均任务成功率则提升了约 1.5 倍。

然而，具身智能数据积累更加困难，距离实现模型智能仍有较大距离。具身智能所需的视觉、动作等多模态交互数据不像大语言模型所需的文本资料被系统记录过，因此，行业正沿着仿真和真实两条路径缩小数据差距。中国信通院估计，假设实现机器人“ChatGPT”时刻的模型参数量级为 55B，若训练数据全部由真机回流提供，约需 21200 台机器人连续工作一年，产出千万小时级有效数据，达到智能涌现则需要 100 亿小时的数据，而当前全球可用的高质量真实数据仅数十万至百万小时级，远未达到支撑具身基础模型智能涌现的门槛。

在真实数据采集方式上，无本体采集有望从“补充方案”走向“预训练主力”。无本体采集包括 Ego 人类第一视角视频采集、UMI 末端操作采集、动作捕捉采集，其中，UMI 末端操作采集为操作员手持或穿戴 UMI，采集与机器人末端同构的观测和动作数据，Ego 人类第一视角采集为利用头戴式相机采集人类第一视角的行为数据。相较于真机数据，无本体采集数据的优势在于：1）成本显著下降，无本体采集数据的单小时采集成本为小几百元，显著低于真机数据的大几百到上千元；2）无本体采集摆脱对本体与场地的依赖，通过头戴/腕戴/夹爪等可穿戴设备，将采集活动嵌入真实工作与生活过程；全国百余城点位、万人网络与“居家/商超/工厂”分布式采集网络加速覆盖长尾任务与复杂环境。

当前，多个头部玩家正在积极入局数采市场，有望带动数据采集设备出货量大幅增长。头部玩家进入“设备规模化—数据产能化—治理平台化”阶段。觅蜂科技 Mego 无本体数据采集设备正式下线交付京东，同时宣布与腾讯 Robotics X 实验室达成无本体数据业务合作，形成了 Ego 裸手、Ego+Wrist 腕戴、Ego+UMI 夹爪三类数据体系，累计形成 100 万小时无本体数据，涵盖居住、办公、零售、生产、仓储、餐饮、文娱休闲、交通出行等八大高频场景。简智机器人宣布累计百万小时真实数据场景、2000+实操技能，月产能冲刺百万小时目标年累计融资超 2 亿美元。场景侧，京东利用 10 万/50 万内部/外部人员与多场域沉浸式场景，规划两年构建 1000 万小时第一人称数据；地方平台大额招投标聚焦“数据采集系统+训练平台”基础设施，具身智能训练场已覆盖全国过半的省级行政区，行业场景覆盖工业制造、家庭场景、商业零售、医疗康养、特种应急五大类行业场景。

Remove Demo Watermark from http://www.wancaiinfo.com/pdfmaster

下载日志已记录，仅供内部参考

表 2: 不同真实数据采集技术对比

路线	数据采集方法	成本	规模化程度	真机映射难度	数据用途
无本体采集	人类第一视角视频数据采集	低	高	高	支持场景理解、人类行为理解、空间预测
	末端操作数据采集	低	高	高	支持多场景下末端精细交互操作任务
	动作捕捉数据采集	中等	中等	中	支持全身或手部复杂运动交互
真机采集	遥操作数据采集	高	低	低	无域偏移问题，与本体运动、力控、结构一致
	自主作业回流数据	最高	低	低	具备真实环境信息、真实扰动、失败案例等高价值数据

资料来源：中国信息通信研究院，中信建投证券

图 12:觅蜂科技 Mego view 采集视角



数据来源：觅蜂科技，中信建投证券

图 13:UMI 夹爪方案

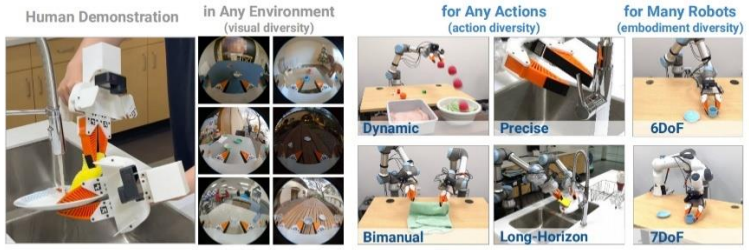


Fig. 1: **Universal Manipulation Interface (UMI)** is a portable, intuitive, low-cost data collection and policy learning framework. This framework allows us to transfer diverse human demonstrations to effective visuomotor policies. We showcase the framework for tasks that would be difficult with traditional teleoperation, such as dynamic, precise, bimanual and long-horizon tasks.

数据来源：高工人形机器人，中信建投证券

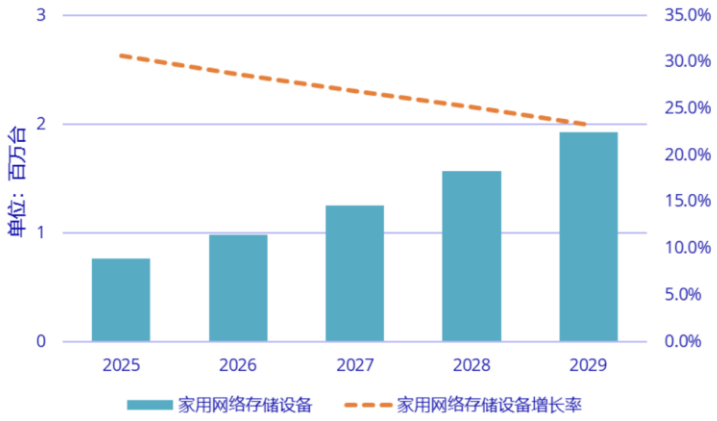
2.3 AI NAS：构筑智能家庭与企业中枢

智能体驱动 NAS 变革，AINAS 实现家用到企业级全面爆发。随着端侧大模型轻量化技术持续突破，隐私存储需求集中爆发，NAS 作为家庭和企业数据的汇聚点，自然地成为了端侧推理需求的落脚点。AI NAS 作为将本地大容量存储能力与端侧 AI 算力深度融合的新一代智能存储设备，本质是将传统 NAS 的存储能力与 AI 的计算、分析与推理深度融合，完成从“数据在哪存/怎么取”到“数据怎么用/如何创造价值”的位阶跃迁。IDC 数据显示，2024 年中国家用 NAS 设备出货量突破 58.5 万台，预计至 2029 年有望突破 192 万台，五年 CAGR 将超过 26%。QYResearch 数据显示，2024 年全球 NAS 市场规模 107.30 亿元，预计 2031 年将达 453.63 亿元，CAGR 为 24.0%。

Remove Demo Watermark from http://www.wancaiinfo.com/pdfmaster

下载日志已记录，仅供内部参考

图 14：中国家用网络存储设备市场预测，2025-2029

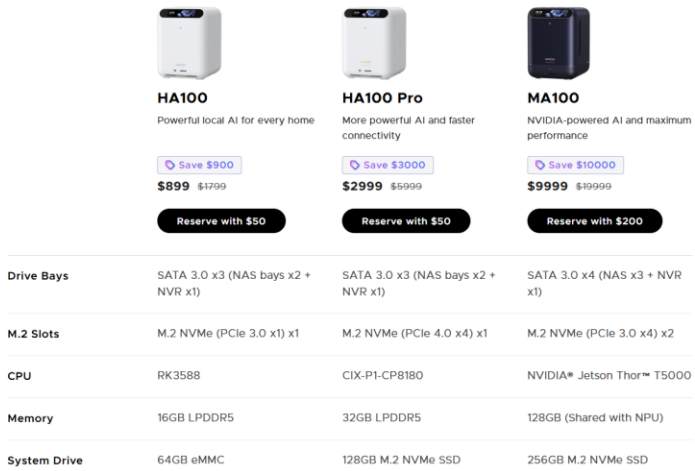


数据来源：IDC，中信建投证券

在家庭场景，AINAS 的价值在于本地隐私 AI 能力与家庭智能中枢。一方面，存储作为 NAS 的基础功能，AI 的应用可以使存储的数据更加易得。传统的 NAS 具有较高的上手门槛，一般作为小众的极客工具存在，AI NAS 提供 AI 自动化配置，有效降低上手门槛。AI NAS 具备智能相册功能，可做到人脸与场景识别、照片修复、视频逐帧色彩调教等功能，同时支持自然语言跨模态、跨生态访问，让本地隐私数据不上传云端的情况下，实现便捷、快速的数据整理归类与搜索使用。

另一方面，AI NAS 可作为**智能家庭管家**，为摄像头、清洁机器人、智能灯具、音箱等提供统一的管理平台，有效解决智能设备数据分散管理导致的使用体验不佳。同时，通过统一存储数据，AI NAS 可通过 NPU 在本地运行 VLM 等模型，智能识别家庭中的异常行为，如宠物与小孩的异常行为，并与智能家居进行联动，而数据本地处理更显著提升对家庭敏感影像与隐私数据的保护能力。例如绿联推出的 HomeAgent 系列，机身集本地存储、本地 AI 计算与全屋管控于一体，其入门款 HA100 采用瑞芯微 RK3588，通过 Matter 接入音箱、灯光、窗帘等不同品牌的智能设备，可实现数据存储、AI 分析与设备控制在本地完成。

图 15：UGREEN 绿联 HomeAgent 产品参数



	HA100	HA100 Pro	MA100
	Powerful local AI for every home	More powerful AI and faster connectivity	NVIDIA-powered AI and maximum performance
	Save $900	Save $3000	Save $10000
	$899 $1799	$2999 $5999	$9999 $19999
	Reserve with $50	Reserve with $50	Reserve with $200
Drive Bays	SATA 3.0 x3 (NAS bays x2 + NVR x1)	SATA 3.0 x3 (NAS bays x2 + NVR x1)	SATA 3.0 x4 (NAS x3 + NVR x1)
M.2 Slots	M.2 NVMe (PCIe 3.0 x1) x1	M.2 NVMe (PCIe 4.0 x4) x1	M.2 NVMe (PCIe 3.0 x4) x2
CPU	RK3588	CIX-P1-CP8180	NVIDIA® Jetson Thor™ T5000
Memory	16GB LPDDR5	32GB LPDDR5	128GB (Shared with NPU)
System Drive	64GB eMMC	128GB M.2 NVMe SSD	256GB M.2 NVMe SSD

数据来源：UGREEN 绿联官网，中信建投证券

Remove Demo Watermark from http://www.wancaiinfo.com/pdfmaster

下载日志已记录，仅供内部参考

在企业场景，AINAS 的主要价值在于效率与安全的双重诉求。效率方面，AINAS 以私有云形态构建跨部门统一知识库，支持 AI 语义搜索、文件问答、自动摘要与思维导图，让企业文档、手册与视频能以 RAG 形态在本地完成检索与生成，效率显著高于传统文件管理。同时 AI NAS 可承载本地 Agent 与自动化流程，支撑低时延响应与长记忆的本地编排。相比 PC、Mac mini 或公有云，NAS 免去数据迁移，更适合长期常开，也规避了云端 Token 消耗与隐私风险，综合成本与安全性更优。协同与治理层面，它支持多人协作、版本管理与快照回滚、精细化权限与多网段隔离，而本地隐私计算可避免敏感数据出网，满足合规审查与数据主权诉求。一次性硬件投入替代云端长期付费，则更契合中小企业对成本与可控性的双重诉求。

图 16: UGREEN 绿联 iDX6011 Pro



数据来源：UGREEN 绿联官网，中信建投证券

图 17: TerraMaster 企业级 U16-725 Plus



数据来源：TerraMaster 官网，中信建投证券

2.4 办公：AI Agent 重构生产力，端侧落地实现主动式 AI 体验

AI Agent 应用多点开花，端侧部署有望突破局限。2026 年 OpenClaw 全球走红，腾讯、阿里、字节跳动等多家厂商争相入场，推出个人与企业级智能体平台，用户可以通过其强大的推理与行动能力，以更低的成本获取更高阶的数字化生产力。2026 年 3 月中国日均 Token 调用量突破 140 万亿，较 2024 年初增长超千倍，应用覆盖金融、工业、医疗等多个行业。但是，云端 AI 的隐私性问题、权限风险、高推理成本已成为阻碍 AI 智能体发展的痛点，端侧 AI 具备本地部署、成本可控等优势，有望突破云端局限，成为未来的趋势。

图 18: 中国企业活跃智能体关键数据预测



数据来源：IDC，中信建投证券

Remove Demo Watermark from http://www.wancaiinfo.com/pdfmaster

下载日志已记录，仅供内部参考

主动式 AI 助手正重塑用户体验，有望成为未来办公场景新趋势。目前主要的 AI Agent 采用 API 模式，需要通过命令调用 AI，而端侧 AI 的部署能够使得 Agent 具备全局感知功能，**人工智能从“被动响应指令”进入“主动理解并参与工作的全新计算形态**。

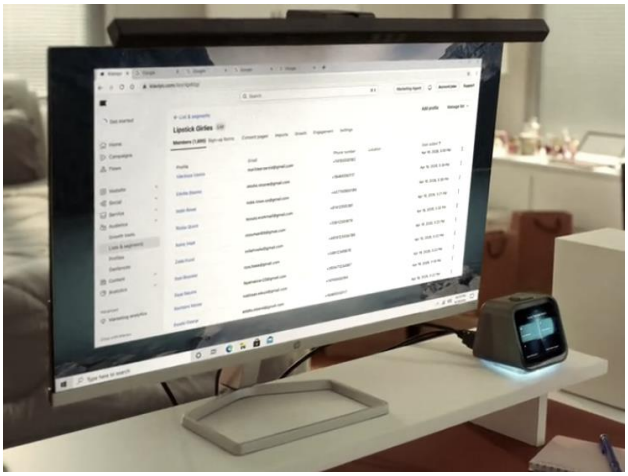
以 Vioploop 为例，Vioploop 是一款专注于 AI PC 自动化硬件的桌面 AI 设备，搭载 RK3576 八核处理器及专用 AI 加速器，支持双屏连接，能在本地跑 10B 参数的模型，处理速度 45 token/秒，大约是 Mac mini 的 2.7 倍。相比于传统纯 AI 软件，Vioploop 的价值在于：1）**兼容性方面**，像微信、剪映等软件没有开放 API，只能通过外接硬件读屏的方式来获取其信息；2）**速度方面**，即使用云端最强的模型，信息上传再回来也要 3.5 秒以上，而 Vioploop 把响应压到 1 秒以内，最长 1.5 秒，真正做到实时响应，实时的响应的意义在于用户在处理多个简单任务时，若等待响应任务时间过久，则会放弃使用 AI，转而自己进行任务，而 Vioploop 的实时响应能力能够进一步加强用户使用 AI 的粘性；**3）个性化体验方面**，Vioploop 可以主动应对一系列场景，例如在微信中提到上次的简历，Vioploop 直接弹出小窗询问是否需要发送简历，用户只需进行“发送”动作。又例如在浏览商品时，AI 主动询问“要不要帮你比价”，用户接受后发回结果。Vioploop CEO 何佳霖将产品定位为电脑的第二个使用者，Vioploop 通过学习电脑屏幕的内容，记住和用户有关的事实，学会用户的工作方式，构建“记忆图谱”，最后把验证过的经验内化到模型权重中，成为高度定制化的模型，最终做到真正懂用户。

图 19:Vioploop 产品图



数据来源：瑞芯微电子，中信建投证券

图 20: Vioploop 产品图



数据来源：Violoop 官网，中信建投证券

端侧 AI 在办公领域的应用还可体现在本地算力拓展、数据不出端的高隐私场景。针对传统 PC 等终端算力不足的情况，面向通用电脑，朗国 RK1828 AIBOX 通过 USB 与网口连接 Windows PC，配合 WorkBuddy 等 Agent，为通用电脑补足 20TOPS 算力，进而实现 PC 的本地 AI 能力升级，无需更换整机即可完成设备智能化改造。**面对政企及对数据安全有较高要求的办公场景**，瑞芯微与亿次网联联合推出离线本地 AI 办公智能体，基于 RK3588 或 RK3576，搭载 RK1828 AI 加速芯片，双芯架构可支撑最高 27B 稠密模型流畅推理，模型、文件与对话数据全部留在本地，并可对接客户的协同文档、私有网盘与 OA 应用，面向设备管理场景还可按项目对接会议大屏、门禁、NAS、OA 及常用物联网协议，目前已在部分省份部署。离线会议方案面向数据安全要求较高的会议场景，语音转写、翻译与纪要生成全流程在本地完成，录音与文本不出设备；支持多种语言的离线识别与双向翻译，会后自动输出结构化纪要。

Remove Demo Watermark from http://www.wancaiinfo.com/pdfmaster

下载日志已记录，仅供内部参考

2.5 智能显示：端侧 AI 助力构建新一代多媒体平台

端侧 AI 赋能显示设备，构建新一代 4K 多媒体平台。视觉作为人最常用的感知觉，屏幕可成为最直观适合用户的端侧 AI 入口。边缘 AI 正在重塑家庭终端，超越传统媒体播放的能力边界。晶晨、瑞芯微等协同显示器厂商，推出新一代多功能 4K 多媒体平台，依托 VLM 多模态模型、LLM 模型等端侧 AI 模型，在实现视听体验大幅突破的前提下，在端侧实现了多种家庭所需功能。

- **视听体验：**传统显示领域中，即使显示设备本身参数足够强大，片源本身的质量也会限制视听体验。海信 RGB-Mini LED 电视 UX 依托 RK1828 对多模态模型的原生算力支撑，可实现实时 2D 转 3D 画面渲染，无需原生 3D 片源即可生成立体观影效果。同时，通过 VLM 多模态画面识别，智能电视实现自动识别画面内容，可一键弹出相关资料，不用切换设备查资料，全面提升用户体验。

- **智能会议与智能翻译：**依托端侧模型，智能显示屏现在已可实现智能会议分析，精确识别会议中的不同发言人与其所说的言论，在会议结束后生成会议纪要，智能总结会议关键内容。同时，智能屏幕还具备智能翻译能力，可处理多种语言的文本，实现实时语音转写，为用户观看其他语言的视频提供辅助，也可为跨国会议纪要提供更深入的辅助。

- **视频内容分析：**智能屏幕搭载 NPU，可通过多模态 VLM 画面识别，实现每隔几帧识别一次屏幕画面并生成对应的文字描述。用户可通过文字描述寻找自己感兴趣的内容，也可利用自动识别的信息，实时分析视频中的行为，或者智能识别屏幕中的广告，根据用户的地区或习惯替换广告内容，在公共场所或工业环境可提高信息处理效率，在家庭环境中则可显著改善用户使用体验。

- **多语言 AI 语音助手：**新一代显示设备同时搭载了多语言 AI 语音助手，在本地环境中部署 AI 模型，支持自然语言交互，实现在断网环境下维持服务，并且保护用户隐私数据不泄露。智能显示 AI 语音助手具备多种功能：**1）智能客服中心**，旨在解决在观影中用户遇到的各种问题，例如在弱网环境下识别视频卡顿原因，同时也可智能识别用户需求，为用户提供各种信息；**2）AI 智能购物**，用户在观看视频时，可随时暂停视频查找视频同款商品，智能显示屏可通过图片识别在网络上寻找同种商品，同时用户可通过自然语言交互特定选定其中某项商品；**3）智能家居中枢，**智能屏幕可作为智能家居的延伸，为用户提供控制不同智能家居的统一入口，搭载端侧 AI 赋予显示屏实时理解和分析用户行为的能力，能够更加理解用户需求，在用户提出要求时更快做出响应，使用户享受便捷全面的体验。

Remove Demo Watermark from http://www.wancaiinfo.com/pdfmaster

下载日志已记录，仅供内部参考

图 21：智能屏幕部分应用



数据来源：晶晨 Amologic，中信建投证券

2.6 汽车电子：车载 AI Box 赋能座舱智能化升级

车载 AI Box 赋能座舱智能化升级，加速端侧 AI 全面落地。车载 AI Box 通过集成汽车级 CPU、GPU/NPU、内存、电源管理、散热系统及高速车载通信接口，为车辆提供端侧 AI 推理、多模态交互、智能座舱 Agent、本地知识服务和个性化用户体验等能力。车载 AIBOX 的单价约在 1500-2000 元，可以通过前装或者后装方式接入，后装市场主要针对低端、中低端车型，无需更换昂贵的座舱域控，即可快速提升车辆的 AI 能力，特别是在面向商用车、物流车、网约车、出租车的企业版 AI 边缘计算终端中，这类场景需要车队管理、行驶数据上传、驾驶员行为分析、AI 识别和云端监管，车载 AI BOX 的商业价值可以更为充分体现。

在前装市场，其主要的功能在于：**1）场景化自主任务拆解**：传统车载语音仅支持单一机械指令，AI Agent 可识别用户模糊场景需求，自动拆解多步骤整车控制任务。例如用户表述“我有点累，想要休息”，系统自动完成关窗、调温、播放音乐、座椅放平全套操作；依托车内车外多模态感知，主动识别用户潜在需求并完成交互。**2）个性化长效记忆功能**：AI 模型具备用户偏好记忆能力，自动记录出行习惯、常用路线、用车需求，无需重复输入基础指令。例如指令“接孩子”，系统自动匹配接送时间、停靠点位、目的地，大幅提升交互粘性与使用体验，复刻手机端智能助手交互逻辑。**3）情感化多模态交互引擎**：模型搭载情绪识别引擎，具备拟人化交互性格，提供情绪价值；联动车内、车外摄像头实现增强哨兵模式，识别车辆异常触碰行为，自动生成预警信息同步车主，拓展安全防护功能。

瑞芯微车载 AI Box 以千元级成本，实现对座舱交互的全面重塑。瑞芯微推出多款基于 RK3576M+RK1828 分布式架构运行和应用端侧大模型的千元级 AI Box 解决方案，部署千问、MiniCPM 等业界主流大模型，支持语音、文本、图片、视频等多模态场景应用。瑞芯微最新的 Omni 全模态感知智能车控方案，具有多传感器输入、低时延全模态感知、超过 80 tokens/s 的连贯输出等优势，实现精准调用各类车控功能。瑞芯微车载 AI Box

Remove Demo Watermark from http://www.wancaiinfo.com/pdfmaster

下载日志已记录，仅供内部参考

可作为独立 AI Box 扩展单元对接新车型或存量车型升级，也可只使用 AI 协处理器灵活搭配各座舱 SoC，实现以千元级成本解决算力和带宽不足难题，已有超 20 个在研项目在业内头部主机厂及 Tier 1 顺利推进，预计未来将规模化全面爆发。

图 22: Omni 全模态智能车控方案



数据来源：瑞芯微电子，中信建投证券

图 23:车载龙虾智能体



数据来源：瑞芯微电子，中信建投证券

2.7 工业：AI 重塑机器视觉

在产业实践中，端侧 AI 在工业视觉领域展现出显著价值。工业现场要求低时延、连续运行、数据安全以及网络异常情况下的可用性，决定了视觉推理不能完全依赖云端。随着 CPU、GPU、NPU 和 FPGA 等异构计算平台持续演进，更多图像预处理、AI 推理、结果输出和设备通信功能将在智能相机、视觉控制器和边缘服务器中完成。端侧视觉的价值不仅在于降低延迟，提升准确度，还在于减少原始图像上传、降低带宽与存储成本、保护客户数据并提高系统独立运行能力。

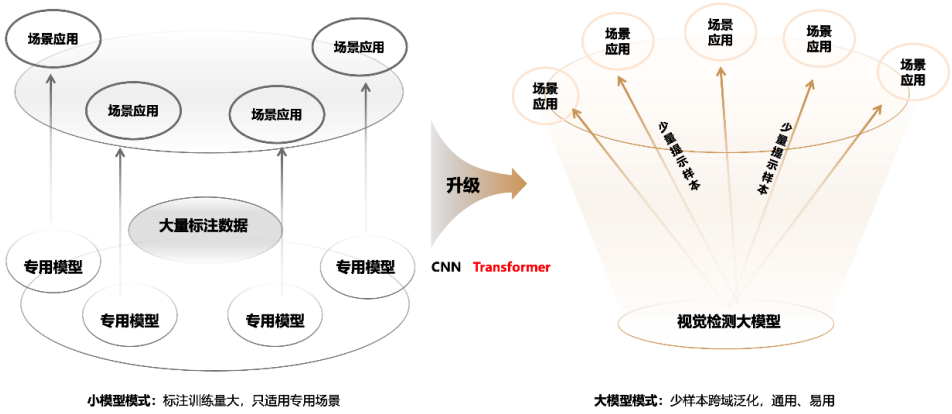
传统机器视觉技术在数据获取与模型训练中存在局限。传统机器视觉技术主要基于卷积神经网络（CNN）架构，依赖大量准确标注，而工业缺陷拥有低频、长尾、变化快和难以穷举等特点。这种技术路径存在样本依赖性强、泛化能力差、维护成本高、行业壁垒深的局限，直接导致了工业视觉市场的碎片化，阻碍了工业视觉技术的规模化应用。

视觉检测大模型推动工业 AI 向自监督、少样本、生成式数据和持续学习演进。视觉检测大模型采用 Transformer 架构，摒弃了传统的 CNN 结构，生成式 AI 则改变了工业视觉的数据准备方式，通过缺陷迁移、局部编辑、三维仿真渲染和物理约束生成，可补充极少见缺陷、危险工况和尚未量产产品的数据。这种架构具备少样本学习能力、跨领域泛化能力、快速适应能力，模型训练将由一次性离线开发转向持续迭代，有效应对传统技术难以解决的痛点。考虑到工业现场对低延迟、高确定性、可验证性、数据安全和成本控制具有更严格要求，在项目开发阶段大模型实现图像检索、预标注、缺陷语义描述等功能后，小模型在生产阶段负责稳定执行，例如检测、分割、分类、测量和定位。

Remove Demo Watermark from http://www.wancaiinfo.com/pdfmaster

下载日志已记录，仅供内部参考

图 24: CNN 框架与 Transformer 架构对比



数据来源：机器视觉产业联盟，中信建投证券

工业视觉产品落地需要成套的软硬件方案。工业 AI 既需要模型能力支持，也需要摄像头、芯片等硬件的协同。以格灵深瞳的 VE² S-GBOX 为例，它面向真实行业场景打造端到端产品体系，其中 MENTOR 面向云端或中心侧、负责数据处理、模型训练、效果评测与算法持续迭代；EXPERT 承接算法全生命周期运营，将通用模型能力转化为适配客户业务的专属算法；GBOX 则进入靠近摄像头、产线和设备的业务现场，完成设备接入、数据采集、边缘推理、事件判断、结果回传与数据回流。三者共同形成一条“从现场运行—数据回流—算法迭代—再部署”的持续进化链路及端到端产品体系。

图 25:格灵深瞳 VE² S-GBOX



数据来源：格灵深瞳，中信建投证券

图 26:格灵深瞳 GBOX



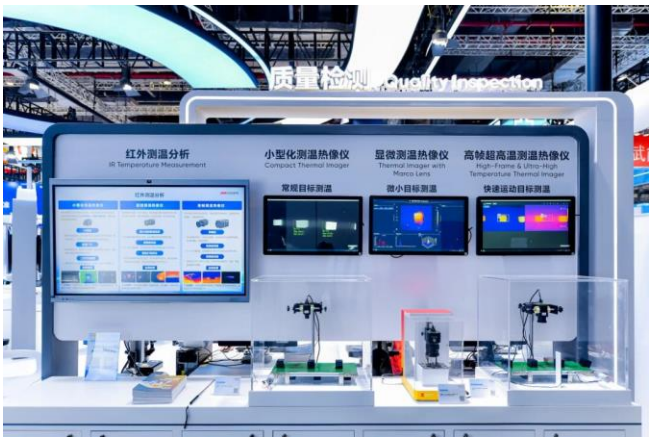
数据来源：瑞芯微电子，中信建投证券

工业视觉端侧 AI 产品形态丰富，覆盖多种工业场景。配合端侧 AI，芯片厂商与终端厂商已合作推出丰富的工业视觉产品，包括工业相机、三维激光扫描仪、智能扫码枪、高精度测绘设备、太阳能摄像头及热成像开发板等。面向多路并发、现场检测响应和柔性换型需求，工业视觉产品已构建起从前端成像、多路检测到多模态分析的分层视觉方案，广泛覆盖工业检测、物流扫描、能源巡检、户外监控等多元化的新型工业场景。

Remove Demo Watermark from http://www.wancaiinfo.com/pdfmaster

下载日志已记录，仅供内部参考

图 27：红外测温分析系列摄像机



数据来源：海康威视，中信建投证券

图 28：三维激光扫描仪



数据来源：瑞芯微电子，中信建投证券

针对工业巡检及产线 SOP 等实际应用场景，芯片厂商与模型厂商协同推出系统级解决方案。瑞芯微与 State Labs 在工业巡检及产线 SOP 合规检测中达成合作，其中工业巡检方案以 sVL 模型为核心，聚焦人工巡检覆盖不足、信息分散、响应滞后等痛点，本地常驻检测与语义分析模型，已在停车场运营平台、工业 AI 眼镜巡检等场景投入使用。针对工厂生产安全与合规问题，多家厂商推出产线 AI-SOP 方案，旨在优化电子制造产线上的操作合规问题，自动识别当前工序与标准流程对比并给出判定，作业现场隐患当场“喊停”，同时管理人员行为与机械臂行为，目前已应用于大型嵌入式模组方案产线。

图 29:工业巡检方案



数据来源：瑞芯微电子，中信建投证券

图 30:端侧 AI 安全生产方案



数据来源：海康威视，中信建投证券

Remove Demo Watermark from http://www.wancaiinfo.com/pdfmaster

下载日志已记录，仅供内部参考

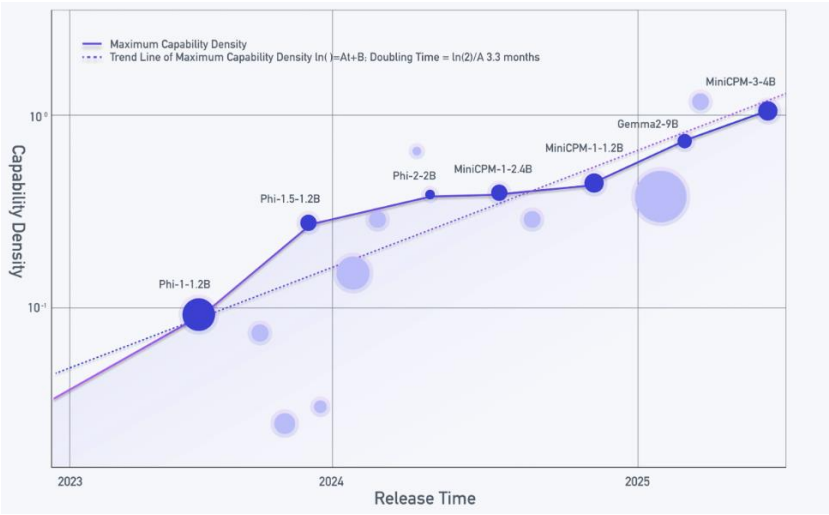
三、落地：模型、场景与算力需深度融合，系统级能力是规模化落地关键

端侧 AI 的规模化落地，并非单一环节能够独立完成。与云端大模型主要围绕算力集群和模型服务展开不同，端侧 AI 对软硬件适配、功耗控制、响应时延、系统集成和场景理解提出了更高要求，需要模型、终端与芯片多方协同。

3.1 模型与场景：端侧模型能力持续进步，不同场景适配不同参数模型

作为端侧 AI 模型的能力基础，端侧模型密度提升和轻量化共同发展。回顾云端 AI 的发展历程，模型能力的提升扮演关键角色，我们认为端侧同样如此，端侧模型发展的核心在于模型密度的提升以及轻量化，在端侧若模型太大，硬件端内存、功耗与散热等限制会使得模型跑不动，若模型太小则会导致能力不够。根据“密度定律”，同样训练工艺下，大约每过 3.5 个月，同等能力的大模型参数量持续减半，可用模型的体量正在快速变小，模型能力密度的提升主要来自于：1）更高效的模型架构等方式，如采用 MoE 架构在调用模型时仅激活部分参数；2）小模型的高效训练与能力迁移，通过高质量的数据配比与训练策略将大模型能力迁移至小模型；3）面向端侧部署的数据治理与评测闭环，用真实场景评测驱动迭代。在模型轻量化方面，蒸馏、剪枝、低比特量化、混合精度、KV Cache 优化和稀疏化等方式可将模型压缩到可部署的范围内。

图 31: 面壁智能模型参数与密度定律



数据来源：面壁智能官网，中信建投证券

我们认为端侧场景众多，不同场景适配不同参数模型。1B 及以下模型在端侧主要用于异常检测与关键词唤醒、轻量分类，也可实现语音交互、家居控制等功能，其主要推理任务均在云端；**3B/4B** 左右的模型能够实现离线文档、实时翻译、视觉理解和本地问答等功能，但是对于复杂推理、多步骤规划、需要广泛世界知识的任务，4B 模型会明显吃力，比如让 4B 模型做一个完整的方案，它可能会遗漏关键细节或逻辑不自洽；**7B** 左右的模型能够更好地进行任务理解与文本创造，适合更复杂对话、端侧指令与一定代码/文档处理，然而在需要高度准确性的场景（医疗诊断、法律建议），7B 模型还不能作为独立决策工具，只能作为辅助参考，同时，长上下文处理

Remove Demo Watermark from http://www.wancaiinfo.com/pdfmaster

下载日志已记录，仅供内部参考

能力有限；**13B** 左右的模型因为参数规模更大，能精准理解复杂指令，理解和处理长上下文，在需要更强推理能力的专业场景中表现突出，例如可用于合同初筛、风险点识别、法律文书分析、论文核心要点总结以及跨文献对比等。未来随着端侧模型能力密度的进一步提升，小参数模型的能力也有望进一步增强。

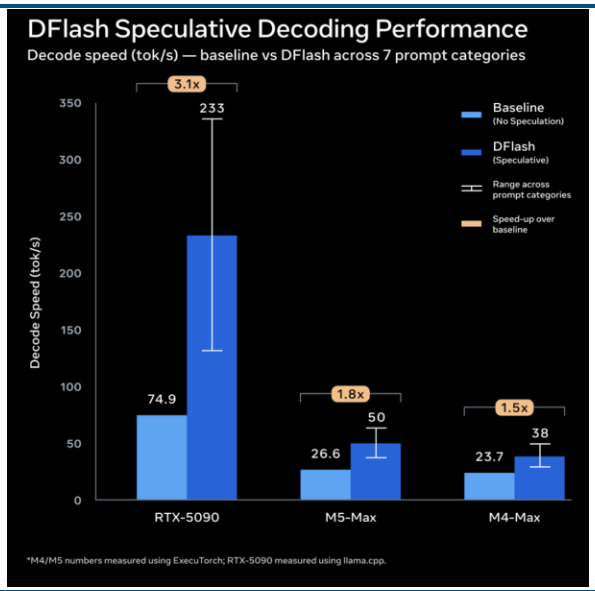
30B 参数级别的小模型智能水平正快速逼近千亿级大模型，其具备较强的多步推理和指令跟随能力，任务的进行更多在端侧，可满足高隐私、高及时响应等场景下的要求。以 2026 年 8 月发布的 Qwen3.8-27B 为例，其 Intelligence Index 与 300B 以上头部模型的差距已缩窄至个位数，在编程、办公及 Agent 任务中甚至超越更大参数模型。Meta 以 Apache 2.0 发布开源模型 Muse Glimmer，具备端到端的任务完成、多步推理、故障恢复、多模态输入与推理等功能，可全面胜任本地代理。Muse Glimmer 搭载基于 DFlash 的轻量级模型，将语言模型缩小至 20GB 以下，适配单张消费级 GPU，实现在 24GB 或 32GB 的显存下同时运行 KV Cache、图像编码器与解码器，真正赋予端侧模型本地代理能力。30B 模型能够处理复杂推理、编程、本地知识库问答、长周期办公（如多步骤任务规划、文档总结）及代码生成，适合部署在本地 AI 一体机上，作为个人或办公的“本地 Agent”， 用于对数据隐私要求高、需结合内部知识库的行业，从而实现低时延、隐私保护且离线可用。

图 32:Qwen3.8 27B 与部分模型文本任务表现对比

	Qwen3.8-27B	Qwen3.6-27B	Qwen3.7-Plus	Muse Glimmer-30B	Opus4.6 Max
Coding					
Agentic terminal coding	73.0	63.4	64.0	51.7	**78.2**
Agentic coding	**61.7**	53.5	57.6	51.2	53.4
Repo-level code generation	42.3	36.2	41.1	--	**47.6**
Agentic coding	**42.2**	13.3	14.2	--	--
Software engineering	**79.0**	49.3	59.2	--	63.8
Agent					
Long-horizon office work	**70.7**	61.0	65.1	--	68.2
Professional job tasks	**33.4**	21.8	27.6	--	--
Frontier agentic tasks	20.4 **42.9**	10.6 27.3	13.2 33.6	--	--
General					
Instruction following	**79.5**	69.1	79.1	77.0	62.5
Scientific reasoning	89.2	87.8	90.3	83.5	**91.3**
Multidisciplinary reasoning	30.8	24.0	34.7	22.0	**40.0**
Competitive coding	**90.3**	83.9	89.6	--	88.8

数据来源：千问大模型，中信建投证券

图 33: RTX-5090、M5-Max、M4-Max 部署 Muse Glimmer



数据来源：Muse，中信建投证券

对于部分行业场景，端侧 AI 从“通用能力”到“行业价值”的最后一公里仍需模型厂与终端相互协同。云侧 AI 的任务类别通常在于问答、Agent 等，而端侧 AI 差异化价值在于能够连接真实的物理世界，基于真实场景，形成“感知—推理—控制/执行”的本地闭环。其规模化落地的难点并不在于通用端侧模型不够聪明，而在于跨过从“通用能力”到“行业价值”的最后一公里需要行业知识、真实场景、产品判断、工程交付和商业化能力共同作用。在这一链条中，终端厂商掌握用户入口、系统权限和场景数据，但其 AI 能力较弱，而模型厂商具备大模型能力、压缩部署经验和多模态推理能力，但是缺少对场景需求的理解以及数据，进而导致基于不同场景的专用模型缺失。因此，端侧 AI 在行业方面真正发挥价值，需场景拆解、模型适配、智能体设计、系统集成、私有化部署、算力适配产品工程化和客户交付等多环节协同发力。

Remove Demo Watermark from http://www.wancaiinfo.com/pdfmaster

下载日志已记录，仅供内部参考

图 34: 终端与模型厂商协同推进端侧 AI 落地



数据来源：中信建投证券

3.2 算力：国内头部 SoC 厂商充分受益于端侧 AI 浪潮

3.2.1 峰值算力、内存带宽、软件工具以及生态效率共同作用于 AIoT 算力芯片

算力芯片是端侧体验的重要一环，然而一颗好用的端侧算力芯片标称算力只是起点，TOPS 指标代表“理论峰值算力”。然而在实际使用中，工具链、内存带宽与实时功耗等才是将标称的 TOPS 转化为实际可用的推理效率关键。

NPU 性能不只受算力约束，端侧 AI 场景下内存带宽才是性能的决定因素。根据 Roofline 模型，一个给定的计算任务在特定硬件上的可达性能上限受算力天花板与内存带宽双重限制。而在 LLM 推理这类极低算术强度的场景下，内存带宽才是推理体验的决定因素。理论生成速度（tokens/s）≈内存带宽/模型权重大小，举例而言，在 50GB/s 有效带宽情况下，3B 模型 INT4 量化后约 1.7GB，理想速度约 30 token/s，这一速度能够被用户接受，而 8B/30B 模型 INT4 量化后约 4.5GB/16GB，同样带宽下只有 11/3 token/s，这样的推理生成速度明显吃力，将大幅影响端侧的推理体验。同时，**功耗是另一重约束峰值算力的限制，**端侧推理时 NPU 和内存系统同时高负载，整机功耗更大，在持续高负载下，终端的温度会明显升高，进而导致推理速度进一步下降。

此外，软件工具即算子和编译器也影响硬件性能上限，卷积和矩阵乘法通常可以高效执行，但 LayerNorm、Softmax、Gather、Scatter、动态 Shape、注意力机制和自定义算子可能受到更多限制，当某个算子无法运行在 NPU 上时，运行时需要将数据转移到 CPU 或 GPU，完成计算后再传回 NPU。一个只占模型计算量很小的算子，也可能因为数据复制、格式转换和同步开销，成为整张网络的性能瓶颈。这正是编译器价值所在，优秀的编译器需要完成模型转换、算子融合、计算图切分、 张量排布、片上存储分配和混合精度选择。卷积、偏置、归一化和激活函数如果能够融合执行，可以减少多次中间结果写回；如果不能融合，每一个算子都可能产生新的访存开销。因此，软件工具决定开发者能否靠近硬件性能上限。

表 3: NPU 主要操作能耗

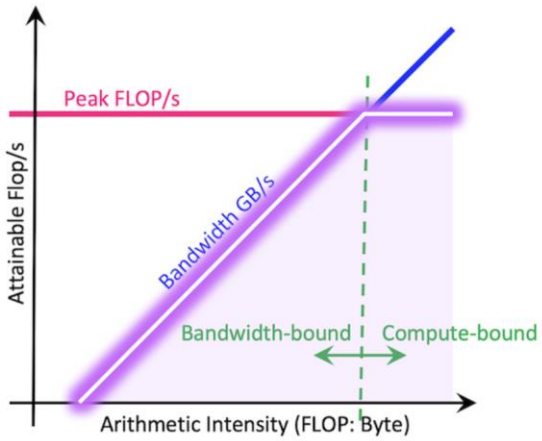
操作	典型能耗（7nm 工艺）
MAC 运算（INT8）	~0.1pJ
读取片上 SRAM（64-bit）	~1pJ
读取 HBM（64-bit，含 I/O）	~20pJ
读取 LPDDR5（64-bit，含 I/O）	~50pJ

资料来源：On-Chip Notes，中信建投证券

Remove Demo Watermark from http://www.wancaiinfo.com/pdfmaster

下载日志已记录，仅供内部参考

图 35: Roofline Model——峰值算力与内存带宽共同决定性能上限



数据来源：NERSC，中信建投证券

生态效率也是影响端侧 AI 部署的重要一环，端侧 AI 面向千行百业，应用场景众多，对于客户而言，某款芯片的社区活跃度和开发板更加完善，相关的文档、量产案例更加丰富，便能降低项目的开发周期、部署难度并减少试错成本。然而端侧 AI 所需要的 NPU 具备碎片化特点，相较于多数均使用 ARM 架构的 CPU/GPU，其 CPU 指令集几十年稳定，软件编译后可适配不同终端，而 NPU 的架构类型相当碎片化，PU 算子、模型架构持续迭代，由 CNN 转向 Transfomer 最后再转向 KV Cache/MoE，同时不同的 NPU 架构适应于不同的应用场景，各家 NPU 设计架构不同，缺乏统一的标准，这会显著提升终端客户的项目开发难度。

3.2.2 国内头部 SoC 公司产品迭代加速，满足多样化算力要求

为应对端侧 AI 在终端差异化的部署要求，国内头部 SoC 公司均在加速产品迭代，同时不断完善工具链以及生态，我们认为在 AIoT 领域，国内头部 SoC 具备全球竞争力，而海外头部厂商更侧重云侧 AI 以及手机、PC 等核心消费终端芯片，伴随产品迭代加速，这些国内头部 SoC 将充分受益端侧 AI 浪潮，保持快速成长态势。

3.2.2.1 瑞芯微：SoC+AI 协处理器双轨并行，AIoT 生态持续完善

瑞芯微推出 AI 协处理器，3D 堆叠架构突破内存带宽限制。2025 年瑞芯微推出全球首颗 3D 架构端侧算力协处理器 RK182X，采用 3D 堆叠封装，把逻辑芯片与高带宽 DRAM 垂直集成，突破平面工艺的限制，垂直互连同时缩短内部传输距离、降低数据搬运功耗，在更小体积内集成更强算力与更大存储。RK1828 理论内存带宽达 1TB/s，充分满足 3B/7B 大模型推理时的数据运力需求。同时，RK182X 采用模块化设计，应用厂商可按需叠加一颗或多颗扩展算力，后续 RK1860（计划 2026Q4）、RK1899 将补足高阶档位。目前已有数百家客户、数十个行业深化研发，覆盖机器人、车载 AI BOX、机器视觉、智慧工业、智慧教育、智能家居等领域，清扫/陪伴机器人、NAS、智能家居等首批项目已率先进入量产。

Remove Demo Watermark from http://www.wancaiinfo.com/pdfmaster

下载日志已记录，仅供内部参考

图 36: RK182X 系列参数

CPU	• 多核RISC-V
NPU	• 支持3B/7B大模型
	• 广泛支持从INT4到FP16的多种计算精度（INT4/INT8/INT16/FP8/FP16/BF16）
存储	• 内置2.5GB/5GB高带宽DRAM，eMMC 4.51
多媒体	• RGA，4K JPEG解码
接口	• USB3.0/2x PCIe2x1/RGMII

数据来源：瑞芯微电子，中信建投证券

图 37: RK1828 部分模型性能表现

RK182X VLM模型核心性能数据

模型	加速芯片	Vision分辨率	Vision(ms)	LLM TTFT(ms)	LLM Decode TPS
FastVLM_1.5B_stage3	RK182X	512 * 512	144.13	47.99	148.47
InternVL3-2B	RK182X	448 * 448	190.80	47.93	148.26
Qwen2.5-VL-7B	RK1828	392 * 392	279.34	159.42	70.02
Qwen3-VL-4B	RK1828	384 * 384	158.89	108.29	89.69
MiniCPM-V_4	RK1828	448 * 448	237.55	94.94	106.62

RK182X LLM模型核心性能数据

模型名称	加速芯片	TTFT(ms)	TPOT(ms)	Decode TPS
Qwen2.5-3B	RK182X	83.44	9.80	102.01
Qwen3-1.7B	RK1828	52.16	7.20	138.88
Qwen3-8B	RK1828	177.87	16.36	61.11

数据来源：瑞芯微电子，中信建投证券

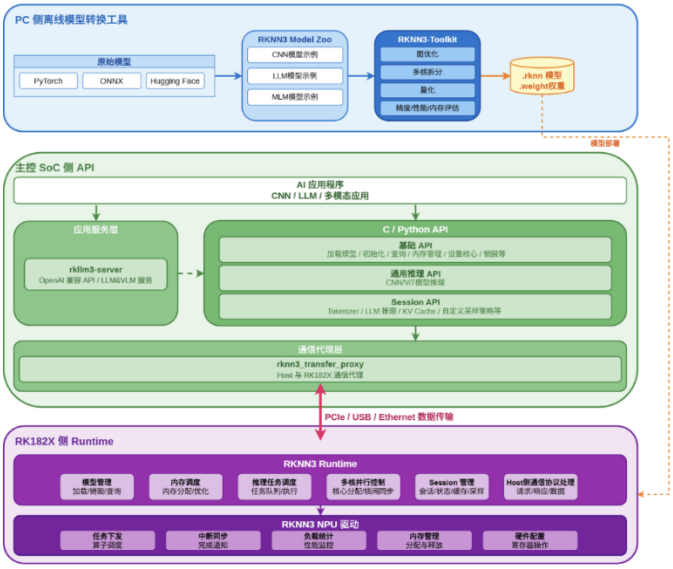
主控 SoC 方面，瑞芯微持续完善“旗舰—高端—中端—入门级”的 AIoTSoC 产品矩阵，新一代旗舰 SoC 正在研发。瑞芯微现旗舰 RK3588 采用 8nm 制程、八核 64 位架构、6TOPS NPU；中高端由 RK3576 承接，2024 年推出，8nm 八核、双核 NPU，接口配置全面，可满足汽车、工业等中高端需求；2026 年 5 月新增 RK3572 下沉中端，同为 8nm 八核，NPU 降至 4TOPS 但保留 8K 解码与多种高速接口，是 AIoT 行业的性价比选择，已快速导入平板与移动终端、AI 摄像头、视频分析、NAS、零售设备及数字标牌。下一代旗舰 RK36 系列正在重点研发，将搭载自研 ISP、NPU、视频编解码等核心 IP，预计内置 32TOPS NPU，综合性能较 RK3588 大幅提升。

瑞芯微自研 AI 开发工具链，帮助客户快速部署模型。面向 SoC，RKNN-Toolkit2 提供模型转换、推理与性能评估的开发套件，并从早期服务传统 CNN 转向 Transformer 架构，引入量化技术提升大模型性能；RKNN LLM 专为大模型优化，进一步降低端侧部署的软件阻碍。面向协处理器与新 SoC，公司自研 RKNN3，兼具高性能、高适配与高能耗比，打通感知—决策—执行链路：感知层覆盖 YOLO 系列与 ASR/TTS（联动科大讯飞、思必驰、大象声科），决策层兼容 0.5B—8B 全规格 LLM 并适配 Qwen3-VL、GLM Edge、MiniCPM 等，执行层支持协处理器自定义后处理。此后持续更新，陆续适配 Qwen3 ASR/TTS、PaddleOCR、SigLIP、DINOv2、Gemma4 等模型，并对不同类型模型采用差异化量化策略，兼顾推理性能与精度。

Remove Demo Watermark from http://www.wancaiinfo.com/pdfmaster

下载日志已记录，仅供内部参考

图 38: RKNN3 软件栈



数据来源：瑞芯微官网，中信建投证券

瑞芯微构建 AIoT 2.0 创新生态圈，持续赋能生态伙伴的算法创新与场景落地。2026 年 1 月，瑞芯微首届 AI 软件生态大会举办，汇聚超 500 位 AI 软件生态伙伴，覆盖基座模型、软件/算法与终端领域。大会中，聚焦实战的 RKNN Workshop 技术培训专场降低开发门槛，持续建设生态。同时，AI 软件及模型公司可通过展示其算法最前沿力量及最新的 AI 算法成果，一方面实现技术交流，另一方面提供 AI 软件伙伴与终端伙伴一对一交流洽谈产品需求的机会，与生态伙伴共同达成硬件与软件的深度协同，推动创新技术落地。

图 39：瑞芯微 AI 软件生态大会



数据来源：瑞芯微电子，中信建投证券

图 40：RKNN Workshop



数据来源：瑞芯微电子，中信建投证券

3.2.2.2 乐鑫科技：云侧 AI 端侧入口，生态建设推动应用边界扩展

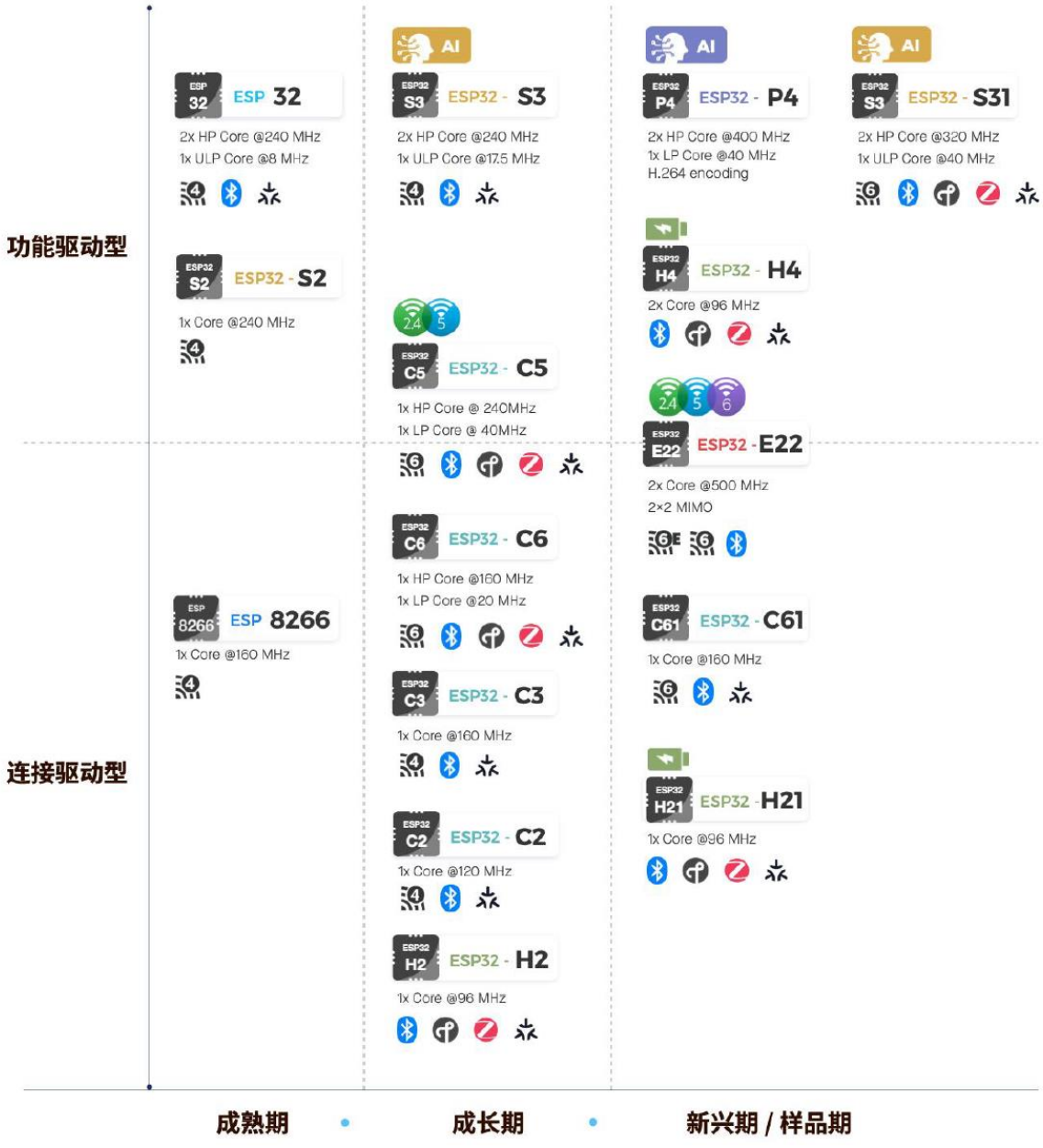
乐鑫持续推出新产品，积极拓展应用场景，可作为云侧 AI 的端侧入口。乐鑫产品以“处理+连接”为战略，从连接驱动周期转向功能驱动周期，相关产品持续迭代。以 ESP32-S 系列为代表的交互式设备芯片集成 AI 加速、多媒体处理及图形显示能力，配置微量化模型，可通过关键词唤醒、第一层感知判断，作为云侧 AI 的端侧入口。ESP32-S31 主要面向嵌入式 AI 和信号处理任务，也已经进入量产阶段，预计未来将实现放量增长。ESP32-

Remove Demo Watermark from http://www.wancaiinfo.com/pdfmaster

下载日志已记录，仅供内部参考

P 系列作为乐鑫向物理 AI 与高交互领域的拓展，高算力的 AIoT ESP32-P4 芯片已经量产，目前正在销售推广的新兴期阶段。同时公司正积极推进下一代基于自研 RISC-V IP 的 AI 端侧芯片的研发，预计采用 12nm 先进制程，大幅提升能效比于本地 AI 推理算力，实现向物理 AI 时代所需的高性能、高交互智能计算平台发展。

图 41：乐鑫产品矩阵



数据来源：乐鑫官网公告，中信建投证券

乐鑫持续迭代升级软件工具链，助力客户快速落地产品。乐鑫以通用开发框架 ESP-IDF 为核心，已针对视觉开发、AI 辅助编程、语音交互等场景推出专用工具链。视觉开发场景中，ESP-Video 主要针对底层采集，ESP-VISION 支持目标检测、颜色追踪、二维码识别等功能，ESP-WHO 则提供人脸检测、人脸识别等 AI 视觉处理

Remove Demo Watermark from http://www.wancaiinfo.com/pdfmaster

下载日志已记录，仅供内部参考

方案。AI 辅助编程场景下，ESP-Claw 作为 Coding Agent 可即插即用，创作者只需直接对话或使用 Skills Lab 中已经封装好的 Skill 分享设备应用包，ESP-Claw 就会结合当前设备和已接模块的能力生成应用，配合其结构化长期记忆系统可在本地保存私密信息。ESP-SR、ESP-Skainet 面向语音识别开发，可实现唤醒词识别、语音命令词识别等功能。乐鑫工具链协同芯片拓展应用边界，覆盖场景将持续拓宽。

开源软件生态原生适配多家云侧开源项目，开发者社群高度活跃。Meta 正式推出 Muse Gadgets 开源项目，全面下放硬件能力，鼓励全球开发者围绕 Muse 打造外围硬件设备。其中 Muse Gadgets 原生适配 ESP32 SDK，开发者基于 ESP32 开发板，在低成本微控制器上部署 Muse 交互逻辑。Anthropic 在其桌面应用 Claude 中，向开发者开放了一个面向硬件设备的 Bluetooth API，其官方示例中则基于乐鑫 M5Stack 平台的 M5StickC Plus 开发板构建了一个“桌面宠物”。同时，乐鑫开发者社群维持高活跃度，ESP32 以高性价比、原生端侧轻 AI 等优势吸引众多开发者设计分享端侧设备软件。乐鑫科技官方数据统计，在 Github 上 ESP32 日均项目快速增长，26Q3 每日新增高达 402 个项目。

图 42: 乐鑫开发者社群关键数据（截至 2026Q3）



数据来源：乐鑫官网公告，中信建投证券

3.2.2.3 晶晨半导体：端侧 AI 芯片等新产品导入市场

晶晨在自身积累的 NPU，大量音视频/模拟接口的 IP 基础上，顺利推出 AI 端侧通用算力平台。高算力通用 AI 端侧平台芯片（A311Y3，即 A9）完成研发并已成功回片，采用台积电 6nm 先进制程工艺，在第二季度业务拓展迅速，公司坚持结合自身擅长的头部大客户战略为主，细分行业生态合作伙伴为辅的双轨拓展模式，聚焦：1）商显；2）机器人（具身及泛机器人领域）；3）平板；4）边缘计算；5）汽车智能座舱这五大细分市场，已立项项目最快有望在 2026 年三季度实现量产，部分项目给出了积极的 2027 年需求预测，有望成为公司成长的新动力。此外，晶晨已上线 Amlogic AI 开发者门户，可集中托管模型、工具、技术文档、性能测试数据，一站式支撑开发者开展端侧 AI 应用开发。

Remove Demo Watermark from http://www.wancaiinfo.com/pdfmaster

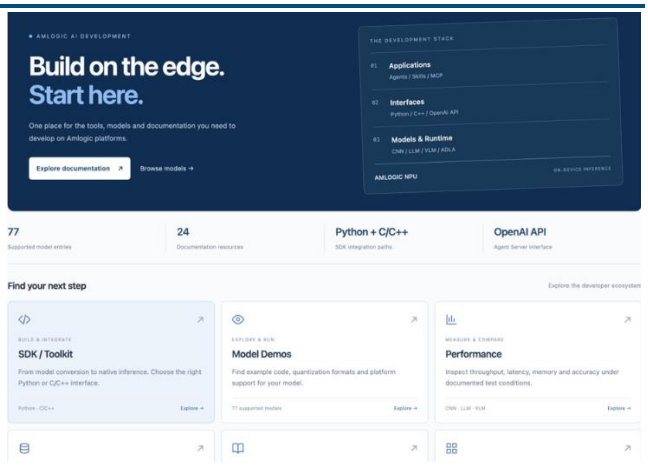
下载日志已记录，仅供内部参考

图 43：A311Y3 机器人关节电机解决方案



数据来源：先楫半导体，中信建投证券

图 44: Amlogic AI 开发者门户



数据来源：晶晨 Amologic，中信建投证券

3.2.2.4 恒玄科技：拐点在即，BES6100 有望打开全新增长曲线

恒玄旗舰芯片 BES6100 顺利流片，有望在 2027 年快速放量。BES2810、BES2710、BES2720 等多款芯片新产品陆续量产推向市场，同时导入雷鸟、Rokid 等智能眼镜领域重点新客户，实现 Wi-Fi 音频、智能伴侣等新终端产品量产落地。同时，恒玄新一代可穿戴芯片 BES6100 顺利流片，下半年将进入测试及客户送样阶段。BES6100 采用 6nm 先进工艺和混合系统架构，兼顾高性能计算与超低功耗需求，实现多模态、多感官与全天候续航一体化，可以有效解决传统智能眼镜、智能手表双芯片方案中存在的芯片间通信复杂、系统设计难度大、PCB 面积大、成本高等问题，可以实现视觉计算与音频计算的深度融合，显著降低系统时延，进一步加强公司在低功耗端侧 AI 市场的领先地位，有望从 2027 年开始为公司积极贡献增量业绩，为公司打开全新的成长空间。

Remove Demo Watermark from http://www.wancaiinfo.com/pdfmaster

下载日志已记录，仅供内部参考

风险分析

AI 终端需求不及预期风险：传统 AI 终端中内存、存储等部件价格持续上涨，智能眼镜、AI NAS、消费级机器人等应用仍处于快速发展阶段，若下游终端需求、资本开支或产品迭代放缓，可能导致端侧 AI 芯片订单及产能利用率低于预期。

端侧模型发展不及预期风险：端侧 AI 产品需要依托端侧模型，端侧模型仍在快速迭代阶段，轻量化、模型能力密度提升均会影响端侧部署效果。若端侧模型发展迭代速度不及预期，可能造成端侧 AI 功能不完整或使用体验不佳，用户需求低于预期。

技术迭代与路线变化风险：NPU 等关键部件生态碎片化，尚未确定统一的不同技术路线，若产品生产路线发生调整，可能造成已有设备、产线及研发投入回报低于预期。

Remove Demo Watermark from http://www.wancaiinfo.com/pdfmaster

下载日志已记录，仅供内部参考

分析师介绍

姚康

电子行业首席分析师，上海交通大学电气工程与自动化学士，集成电路工程硕士，曾任职于东吴证券、兴业证券，从事电子行业研究，2026 年加入中信建投证券研究发展部，任电子行业首席分析师，研究领域包括半导体、海外算力和消费电子等，擅长产业链研究、产业趋势判断和个股挖掘。2021 年所在团队获得新财富第四名。

梁艺

北京大学金融硕士，曾就职于浙商证券，主要覆盖消费电子等领域

何昊

电子团队分析师，复旦大学管理学硕士，2020 年起先后就职于国海证券、光大证券研究所担任电子行业分析师，2026 年 3 月加入中信建投证券研究发展部，主要覆盖消费电子等领域。

Remove Demo Watermark from http://www.wancaiinfo.com/pdfmaster

下载日志已记录，仅供内部参考

评级说明

投资评级标准		评级	说明
报告中投资建议涉及的评级标准为报告发布日后6个月内的相对市场表现，也即报告发布日后的6个月内公司股价（或行业指数）相对同期相关证券市场代表性指数的涨跌幅作为基准。A股市场以沪深300指数作为基准；新三板市场以三板成指为基准；香港市场以恒生指数作为基准；美国市场以标普500指数为基准。	股票评级	买入	相对涨幅15%以上
		增持	相对涨幅5%—15%
		中性	相对涨幅-5%—5%之间
		减持	相对跌幅5%—15%
		卖出	相对跌幅15%以上
	行业评级	强于大市	相对涨幅10%以上
		中性	相对涨幅-10-10%之间
		弱于大市	相对跌幅10%以上

分析师声明

本报告署名分析师在此声明：（i）以勤勉的职业态度、专业审慎的研究方法，使用合法合规的信息，独立、客观地出具本报告，结论不受任何第三方的授意或影响。（ii）本人不曾因，不因，也将不会因本报告中的具体推荐意见或观点而直接或间接收到任何形式的补偿。

法律主体说明

本报告由中信建投证券股份有限公司及/或其附属机构（以下合称“中信建投”）制作，由中信建投证券股份有限公司在中华人民共和国（仅为本报告目的，不包括香港、澳门、台湾）提供。中信建投证券股份有限公司具有中国证监会许可的投资咨询业务资格，本报告署名分析师所持中国证券业协会授予的证券投资咨询执业资格证书编号已披露在报告首页。

在遵守适用的法律法规情况下，本报告亦可能由中信建投（国际）证券有限公司在香港提供。本报告作者所持香港证监会牌照的中央编号已披露在报告首页。

一般性声明

本报告由中信建投制作。发送本报告不构成任何合同或承诺的基础，不因接收者收到本报告而视其为中信建投客户。

本报告的信息均来源于中信建投认为可靠的公开资料，但中信建投对这些信息的准确性及完整性不作任何保证。本报告所载观点、评估和预测仅反映本报告出具日该分析师的判断，该等观点、评估和预测可能在不发出通知的情况下有所变更，亦有可能因使用不同假设和标准或者采用不同分析方法而与中信建投其他部门、人员口头或书面表达的意见不同或相反。本报告所引证券或其他金融工具的过往业绩不代表其未来表现。报告中所含任何具有预测性质的内容皆基于相应的假设条件，而任何假设条件都可能随时发生变化并影响实际投资收益。中信建投不承诺、不保证本报告所含具有预测性质的内容必然得以实现。

本报告内容的全部或部分均不构成投资建议。本报告所包含的观点、建议并未考虑报告接收人在财务状况、投资目的、风险偏好等方面的具体情况，报告接收者应当独立评估本报告所含信息，基于自身投资目标、需求、市场机会、风险及其他因素自主做出决策并自行承担投资风险。中信建投建议所有投资者应就任何潜在投资向其税务、会计或法律顾问咨询。不论报告接收者是否根据本报告做出投资决策，中信建投都不对该等投资决策提供任何形式的担保，亦不以任何形式分享投资收益或者分担投资损失。中信建投不对使用本报告所产生的任何直接或间接损失承担责任。

在法律法规及监管规定允许的范围内，中信建投可能持有并交易本报告中所提公司的股份或其他财产权益，也可能在过去12个月、目前或者将来为本报告中所提公司提供或者争取为其提供投资银行、做市交易、财务顾问或其他金融服务。本报告内容真实、准确、完整地反映了署名分析师的观点，分析师的薪酬无论过去、现在或未来都不会直接或间接与其所撰写报告中的具体观点相联系，分析师亦不会因撰写本报告而获取不当利益。

本报告为中信建投所有。未经中信建投事先书面许可，任何机构和/或个人不得以任何形式转发、翻版、复制、发布或引用本报告全部或部分内容，亦不得从未经中信建投书面授权的任何机构、个人或其运营的媒体平台接收、翻版、复制或引用本报告全部或部分内容。版权所有，违者必究。

中信建投证券研究发展部

北京
朝阳区景辉街16号院1号楼18层
电话：（8610） 56135088
联系人：李祉瑶
邮箱：lizhiyao@csc.com.cn

上海
上海浦东新区浦东南路528号南塔2103室
电话：（8621） 6882-1600
联系人：翁起帆
邮箱：wengqifan@csc.com.cn

深圳
福田区福中三路与鹏程一路交汇处广电金融中心35楼
电话：（86755）8252-1369
联系人：曹莹
邮箱：caoying@csc.com.cn

中信建投（国际）

香港
中环交易广场2期18楼
电话：（852）3465-5600
联系人：刘泓麟
邮箱：charleneliu@csci.hk

Remove Demo Watermark from http://www.wancaiinfo.com/pdfmaster