



电子行业研究

买入（维持评级）

行业深度研究

证券研究报告

电子组

分析师：樊志远（执业 S1130518070003）

fanzhiyuan@gjzq.com.cn

AI 走进现实：Physical AI 驱动端侧硬件进入升级周期

行业观点：

我们认为，本轮 Physical AI 的核心变化并非机器人机械结构本身发生突破，而是通用模型开始由感知和理解进一步进入规划、决策与动作空间，使 AI 由数字世界向物理世界延伸。随着模型输出由语言、图像逐步扩展至真实动作，机器人对硬件的要求也由单一 AI 算力进一步扩展至多模态感知、本地计算、Memory/Storage、实时控制及功能安全等完整端侧系统能力。与此同时，Physical AI 并非一条孤立的机器人硬件路线：手机、PC 已率先验证本地大模型和 Agent 的规模化部署，机器人则进一步将同一趋势由数字交互延伸至物理执行。我们认为，模型能力跃迁是本轮 Physical AI 的技术起点，机器人是最典型的新终端，而端侧计算平台升级则是电子行业更广泛的产业落点。

投资逻辑：

模型进入物理世界，推动机器人端侧硬件由单一算力升级向系统级扩容。随着机器人任务由固定工位、单一操作向开放环境、全身运动和灵巧操作扩展，模型需要持续获取视觉、姿态、运动和接触等多维信息，并在本地完成感知融合、推理、规划和控制。对应硬件来看，Physical AI 带来的增量并非单纯提升 TOPS，而是“大模型+多传感器+实时控制”共同推动整条端侧数据链路扩容：模型规模提升带动内存容量与带宽需求，多路感知和数据记录增加本地存储及高速 I/O 需求，而动作输出进一步增加 EtherCAT、CAN-FD 等实时控制接口及功能安全资源，机器人电子硬件价值量由 SoC 向 Memory、Storage、互联及控制环节扩散。

机器人进入小批量产业验证，Training 需求有望先于本体大规模放量释放。2026H1 全球人形机器人出货已接近 2.5 万台，半年出货量超过 2025 全年，产业由几千台级研发市场逐步迈入数万台级商业化探索；与此同时，应用结构也开始由科研、展示和数据采集向制造、物流和服务等真实作业场景扩散，电子零部件需求有望由项目型采购向标准化、批量化采购迁移。更值得关注的是，模型训练需要提前积累大量真实交互数据，无本体采集能够降低对机器人数量的依赖，EGO、UMI、WristCam、Hub 等产品形态已经开始由实验室自建工具向标准化训练硬件演进。电子硬件需求不必完全等待机器人本体大规模出货，可能形成“训练需求先行、本体需求承接”的释放节奏。

模型加速下沉，Physical AI 进一步融入更广泛的端侧计算升级周期。MLPerf 标准测试对象已由传统视觉 CNN 进一步扩展至 LLM、生成式 AI 及 Agent，手机和 PC 则率先验证本地大模型进入规模化终端；机器人 VLA 同步向端侧迁移，本地模型开始直接承担动作输出任务。随着 LLM、VLM 和 Agent 本地运行，模型权重、上下文、KV Cache 及本地知识库持续增加，端侧硬件约束由计算能力进一步扩展至 Memory 容量、带宽及 Storage；同时，小模型、量化压缩及部署工具链持续成熟，又在降低复杂 AI 进入终端的算力、功耗和存储门槛，使 AI 能力有望由高性能 PC、智能手机和机器人进一步向工业设备、AIoT、可穿戴及智能家居扩散。

投资建议与估值

Physical AI 本体及产业化相关标的：宇树科技、优必选等；数据采集及感知硬件相关标的：奥比中光、思特威、韦尔股份、芯动联科、安培龙、必易微等；端侧 AI 芯片相关标的：瑞芯微、星宸科技、乐鑫科技、恒玄科技等；整机制造相关标的：立讯精密、工业富联等。

风险提示

Physical AI 模型能力提升不及预期；机器人产业化及终端出货不及预期；Physical AI 数据采集及训练硬件需求不及预期；端侧 AI 渗透及模型本地部署进度不及预期；端侧硬件配置升级幅度不及预期；技术路线快速变化及产业竞争加剧。



内容目录

一、模型走进现实，Physical AI 硬件需求同步升级	4
1.1 通用模型跃迁，物理任务边界持续拓宽	4
1.2 感知、推理与控制融合，端侧平台成为 Physical AI 关键底座	6
二、产业化进程提速，硬件需求开始兑现	8
2.1 应用场景迁移，机器人硬件进入批量验证	8
2.2 训练数据需求前置，训练硬件率先释放	9
三、模型加速下沉，端侧计算进入平台升级周期	11
3.1 工作负载持续升级，端侧 AI 由感知走向智能体与动作执行	11
3.2 模型本地化深化，端侧硬件需求持续扩散	13
四、相关标的	14
五、风险提示	14

图表目录

图表 1: Astra Agentic 策略提升机器人任务成功率	4
图表 2: Gemini Robotics 2 实现机器人全身操作能力扩展	5
图表 3: Gemini Robotics 2 具备多指灵巧操作能力	5
图表 4: 人形机器人由单摄像头向多视角、多模态感知升级（以 NVIDIA Isaac GROOT Reference Humanoid 为例）	5
图表 5: AI 厂商积极布局机器人硬件环节	6
图表 6: 机器人端侧平台由单一计算芯片升级为“计算+感知+控制+连接”完整系统（以高通 Dragonwing IQ10 Robotics Reference Design 为例）	7
图表 7: 机器人端侧计算平台向大内存、高带宽与多传感器接口演进	8
图表 8: 机器人端侧计算由 AI 推理进一步扩展至实时控制与功能安全	8
图表 9: 2026H1 全球人形机器人出货量同比接近 300%	9
图表 10: 人形机器人应用结构加速从验证型场景向真实作业场景扩散	9
图表 11: EGO、UMI 与 WristCam 等多路径扩展真实交互数据入口	10
图表 12: 以奥比中光为例，已覆盖 EGO、UMI、WristCam 及 Hub 等多采集路径产品	10
图表 13: 训练侧硬件需求有望先于机器人本体大规模出货释放	11
图表 14: LLM、生成式 AI 与 Agent 逐步成为端侧标准工作负载	12
图表 15: GenAI 手机出货占比迅速提升	12
图表 16: 机器人 VLA 由云端推理向本地动作生成下沉	13



图表 17: 复杂 AI 模型本地运行提升 Memory 与 Storage 配置要求	14
图表 18: 模型轻量化推动端侧 AI 向更广泛设备扩散	14

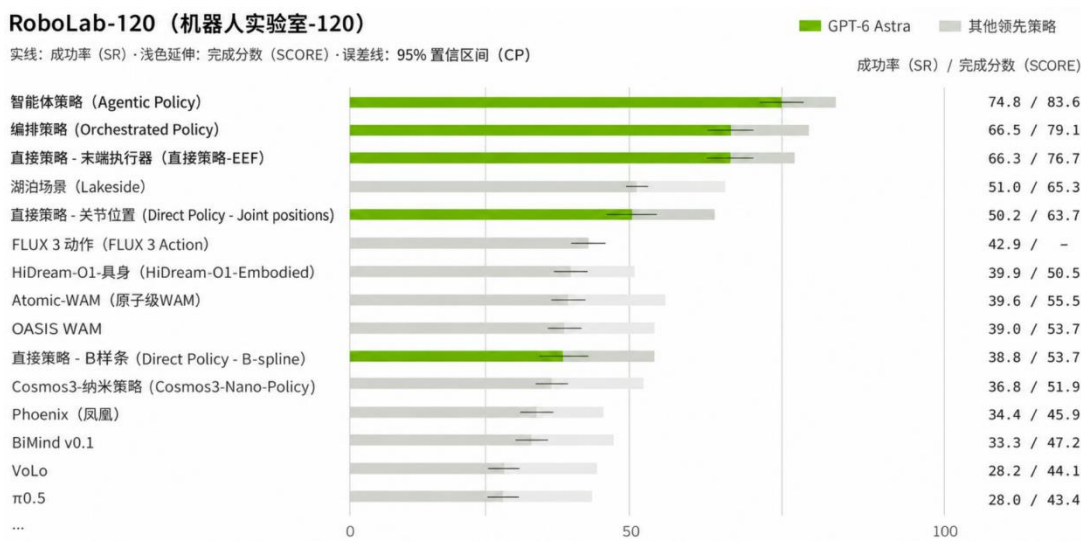


一、模型走进现实，Physical AI 硬件需求同步升级

1.1 通用模型跃迁，物理任务边界持续拓宽

通用模型正由机器人感知与语义理解环节进一步向动作空间延伸，推动机器人模型从“理解环境”向“规划并执行任务”演进。传统 VLM 主要承担图像理解、语言理解与场景识别等感知任务，而新一代机器人模型开始建立 Vision、Language 与 Action 之间的直接连接，使通用模型不仅能够理解机器人所处环境及用户指令，还可进一步参与任务拆解、动作规划、执行反馈及失败纠错等控制环节。NVIDIA Astranomy 的 RoboLab-120 测试显示，GPT-6 Astra 已经能够直接作为 Robot Policy 参与机器人操作任务，其中 Agentic Policy 的任务成功率达到 74.8%，高于 Direct Policy-EEF 的 66.3%，表明通用模型通过引入任务规划与自主决策能力，可以在直接动作策略之外进一步提升复杂操作任务的完成效果。随着模型能力由感知、理解持续向规划和控制环节延伸，通用模型可介入的机器人任务链条进一步拉长，为机器人覆盖更多长流程、多步骤及需要动态调整的任务提供了能力基础。

图表1: Astra Agentic 策略提升机器人任务成功率



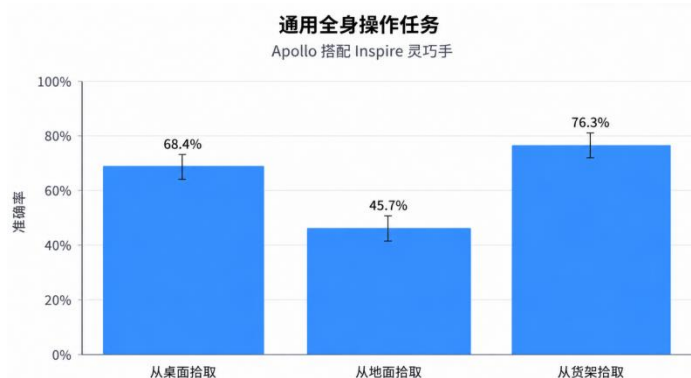
来源: 英伟达、国金证券研究所

通用模型的能力边界正进一步由环境理解向全身控制和灵巧操作延伸，机器人可执行任务的复杂度与动作空间持续拓展。除 NVIDIA Astranomy 展示通用模型直接参与任务拆解、动作规划和执行控制外，Google Gemini Robotics 2 进一步将模型能力从以桌面抓取为代表的局部操作扩展至行走、下蹲、伸展、双臂协同以及多指灵巧操作等更复杂的全身动作。在 Apollo 机器人测试中，模型在桌面、地面和货架取物任务中的准确率分别达到 68.4%、45.7% 和 76.3%，表明模型已能够结合不同空间位置调动全身完成操作；在多指灵巧操作测试中，拧出灯泡任务准确率达到 92%，同时在拧入灯泡、系垃圾袋、使用簸箕和操作密封袋等任务中分别达到 36%、44%、32% 和 40%，显示模型已经开始处理需要手指协调和精细动作控制的任务。

我们认为，从当前前沿的 Astranomy 与 Gemini Robotics 2 实验结果来看，通用模型正在由“识别和理解物理世界”进一步向“规划并作用于物理世界”延伸，其输出对象由语言和视觉表征逐步扩展至机器人动作，为后续覆盖更复杂的全身协同和灵巧操作任务提供了能力基础。



图表2: Gemini Robotics 2 实现机器人全身操作能力扩展



来源: Google、国金证券研究所

图表3: Gemini Robotics 2 具备多指灵巧操作能力

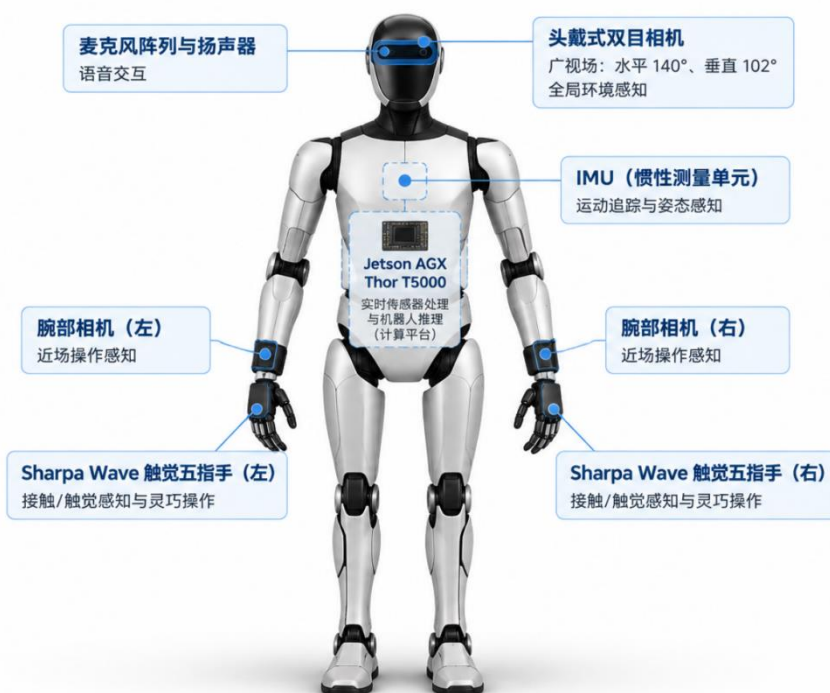


来源: Google、国金证券研究所

机器人由固定工位走向开放场景后,任务复杂度提升正在同步抬升对感知硬件的要求。传统工业机器人在固定相机、固定场景和固定程序下,依靠有限环境信息即可完成重复性任务;而当机器人进入家庭、仓储、工厂及商业服务等动态环境,并执行长流程、全身运动和灵巧操作时,模型需要持续感知目标位置、自身姿态、接触状态及环境变化,并据此实时调整动作,感知由任务开始前的“识别”逐步转向贯穿执行过程的闭环反馈。NVIDIA 的人形机器人参考设计已体现这一趋势:通过头戴式双目相机与腕部相机获取全局及近场视觉信息,同时结合惯性测量单元(IMU)、触觉五指手及麦克风阵列补充运动、接触与语音信息,并由 Jetson AGX Thor 完成实时传感器处理与机器人推理。我们认为,物理 AI 模型能力提升并非对传感器的替代,而是提高了模型对多维物理信息的利用能力,推动机器人感知由单一视觉向多视角、多模态融合演进;随着机器人任务复杂度提升,视觉、运动及交互类传感器有望成为感知系统升级的核心方向。

图表4: 人形机器人由单摄像头向多视角、多模态感知升级(以 NVIDIA Isaac GROOT Reference Humanoid 为例)

多视角视觉、触觉、惯性与语音感知协同



来源: 英伟达、国金证券研究所

随着模型能力由视觉、语言进一步向动作与多模态感知延伸,头部 AI 厂商的机器人布局也开始由模型和数据层向本体硬件拓展。OpenAI 已单列 Robotics 团队,并持续招聘控制



系统、电气工程、电机、齿轮及执行器等岗位，其中控制系统岗位覆盖实时系统、执行器、动力学与底层硬件交互，电机与执行器岗位涉及定制电机、执行器制造及验证，计算电气工程岗位则明确面向 OpenAI 人形机器人开发 AI 推理计算系统。Physical AI 的发展正在推动 AI 厂商由单纯提供“大脑”能力，进一步介入计算平台、控制系统和执行器等机器人硬件环节，软硬件协同的重要性持续提升。

图表5: AI 厂商积极布局机器人硬件环节

OpenAI			Research	Products	Business	Developers	Company	Foundation	Q	Log in	Try ChatGPT
robotics			Showing 15 jobs out of 802			All teams			All locations		
Compute Electrical Engineer, Robotics			Robotics			San Francisco			Apply now		
Control Systems Software Engineer, Robotics			Robotics			San Francisco			Apply now		
Electrical Engineer, PCBA Test Infrastructure, Robotics			Robotics			San Francisco			Apply now		
Electrical Engineer, Robotics			Robotics			San Francisco			Apply now		
Field Engineer			Robotics			San Francisco			Apply now		
Lead Safety Engineer, Robotics			Robotics			San Francisco			Apply now		
Manufacturing Engineer, Motors & Actuators			Robotics			San Francisco			Apply now		
Mechanical Design Engineer, Motor Design & Prototyping			Robotics			San Francisco			Apply now		
Mechanical Engineer, Dynamometer and Actuator Testing Infrastructure			Robotics			San Francisco			Apply now		

来源：OpenAI、国金证券研究所

1.2 感知、推理与控制融合，端侧平台成为 Physical AI 关键底座

多模态感知持续增加后，机器人端侧计算正由单一处理器向系统化计算平台演进。机器人运行过程中需要持续接收多路相机、激光雷达（LiDAR）、飞行时间（ToF）及惯性测量单元（IMU）等传感器数据，并进一步完成感知融合、AI 推理、任务规划与运动控制，高带宽、连续输入及低时延的工作特征使关键计算与控制环节需要更多在本地完成。高通 2026 年发布的 Dragonwing IQ10 机器人参考设计即体现了这一趋势：平台将异构计算、传感器接口、网络通信、实时控制及软件栈集成于统一系统，最高提供 700 TOPS AI 算力，并原生支持最多 12 路 GMSL2 相机以及 LiDAR、ToF、IMU 等多模态传感器；同时通过 EtherCAT、CAN-FD 等高速确定性接口连接执行机构，支持端侧感知、规划、推理与运动控制。随着机器人感知维度和自主能力持续提升，端侧计算的核心需求将不再局限于单颗 SoC 算力，而是进一步扩展至传感器接入、数据同步、网络通信和实时控制等系统能力，机器人计算硬件有望逐步形成以 SoC 为核心、配套存储与多类高速接口的完整计算平台，单机电子硬件的复杂度随之提升。



图表6: 机器人端侧平台由单一计算芯片升级为“计算+感知+控制+连接”完整系统（以高通 Dragonwing IQ10 Robotics Reference Design 为例）

将计算、传感、网络 and 软件集成于单一、可直接部署的系统中



来源: 高通、国金证券研究所

物理 AI 端侧硬件升级并不止于 AI 算力提升, 大模型运行与多传感器数据处理正在同步推高内存、存储及高速 I/O 配置。NVIDIA Jetson AGX Thor 和高通 Dragonwing IQ10 机器人参考设计均体现出这一趋势: 前者在高 AI 算力之外配置大容量 LPDDR5X 内存、NVMe 存储及高速以太网, 后者则进一步整合大容量内存、本地存储、多路相机接入以及 EtherCAT、CAN-FD 等机器人控制接口, 面向 AI、感知、控制和系统任务的一体化运行。

我们判断, 物理 AI 带来的硬件增量并非集中于单颗 AI 芯片, 而是“大模型+多传感器”共同推动整条端侧数据链路扩容: 模型规模提升带动内存容量与带宽需求, 持续感知和数据记录增加本地存储需求, 多路传感器与执行机构接入则提升高速通信及实时控制接口的重要性。对应产业链来看, 机器人电子硬件的价值增量有望由 SoC 进一步向 LPDDR 等内存、UFS/NVMe 等存储以及高速网络与控制接口等环节扩散, 物理 AI 因此不仅是算力升级逻辑, 也有望成为机器人端侧存储与互联硬件升级的重要驱动力。


图表7：机器人端侧计算平台向大内存、高带宽与多传感器接口演进

配置	NVIDIA Jetson AGX Thor 开发套件 (T5000)	Qualcomm Dragonwing IQ10 RRD
AI 算力	2070 TFLOPS (FP4 Sparse)	最高 700 TOPS
CPU	14 核 Arm Neoverse V3AE	18 核 Qualcomm Oryon CPU
计算架构	Blackwell GPU, 2560 个 CUDA Core、96 个第五代 Tensor Core	CPU + NPU + GPU 异构计算
内存	128GB LPDDR5X	64GB LPDDR5X
内存带宽	273GB/s	未披露
本地存储	1TB NVMe	512GB UFS 4.0
存储扩展	PCIe Gen5 NVMe	PCIe Gen5 NVMe
摄像头/视觉接入	HSB 高速摄像头、USB 摄像头	最高 12 路 GMSL2 摄像头
其他传感器支持	通过高速 I/O 扩展多类传感器	LiDAR、ToF、IMU 等
以太网	1×5GbE + 4×25GbE	2×10GbE + 1×2.5GbE
实时控制接口	CAN 等	EtherCAT、CAN-FD、TSN-ready Ethernet
主要定位	Physical AI / 通用机器人端侧 AI 计算	人形机器人 / AMR / 工业机器人一体化参考设计

来源：英伟达、高通、国金证券研究所

随着 AI 由感知与理解进一步进入运动决策环节，机器人端侧计算开始呈现出区别于手机、PC 等通用 AI 终端的新增硬件需求。手机和 PC 运行端侧 AI 时，核心瓶颈更多集中于算力、内存和存储；机器人则需要将 AI 输出直接转化为真实动作，因此还需增加运动控制与功能安全能力。高通 Dragonwing IQ10 机器人参考设计已将 EtherCAT、CAN-FD 及 TSN 以太网等机器人控制接口纳入平台，其中配置 4 路 EtherCAT 和 8 路 CAN-FD，并通过安全岛 (Safety Island)、域隔离及独立网络域支持安全层与应用层之间的隔离和受控交互；NVIDIA 最新 Isaac 平台亦强调确定性运行，并计划进一步集成 Halos 全栈安全系统。

我们研判，相较其他 AI 终端，Physical AI 更具增量意义的并非单纯增加 TOPS，而是在通用 AI 计算之外增加了连接执行机构的实时通信与安全控制硬件：前者对应 EtherCAT、CAN-FD、TSN 以太网等控制接口及相关 I/O，后者则对应安全岛、独立安全域及网络隔离等系统资源。由此，机器人电子价值量有望由主控、内存和存储进一步向工业通信、控制 I/O 及功能安全相关硬件扩散，形成区别于手机、PC 等 AI 终端的新增硬件需求。

图表8：机器人端侧计算由 AI 推理进一步扩展至实时控制与功能安全

功能域	主要任务	NVIDIA 方案	Qualcomm Dragonwing IQ10 RRD
AI 计算	视觉感知、VLM/VLA 推理	Jetson Thor: 2070 FP4 TFLOPS、128GB 统一内存	CPU + NPU + GPU 异构计算，最高 700 TOPS
实时控制	运动控制、传感器与执行器协同	Isaac OS: 连接 AI、传感器及控制系统，支持确定性实时运行	EtherCAT、CAN-FD、TSN-ready Ethernet
功能安全	故障监测、安全隔离与保护	NVIDIA Halos 全栈安全平台，基于 IGX Thor	Safety Island + Domain Separation

来源：英伟达、高通、国金证券研究所

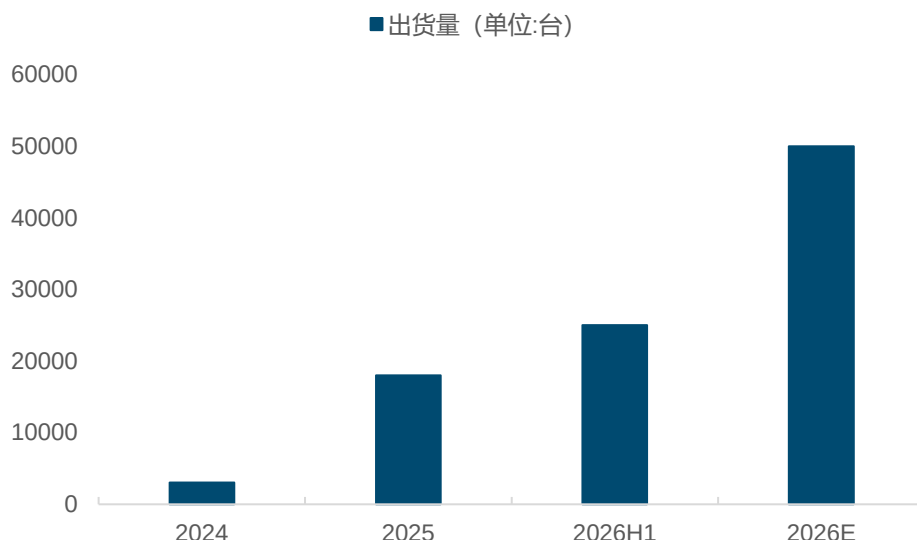
二、产业化进程提速，硬件需求开始兑现

2.1 应用场景迁移，机器人硬件进入批量验证

人形机器人出货正在由几千台级研发市场迈入数万台级商业化探索阶段。根据 IDC，2025 年全球人形机器人出货量接近 1.8 万台，同比增长约 508%，对应 2024 年出货量约 0.3 万台；2026H1 出货进一步接近 2.5 万台，同比增长 432.1%，半年出货量已超过 2025 全年。相比远期百万台级市场空间，我们认为当前更值得关注的是产业已经跨过单纯依赖研发样机和技术验证的几千台阶段，开始进入数万台级的小批量交付周期。随着整机出货基数扩大，摄像头、传感器、计算模组及其他电子零部件的需求也将逐步由项目型、研发型采购向更稳定的批量采购迁移，产品成本、可靠性及供应链交付能力的重要性随之提升。



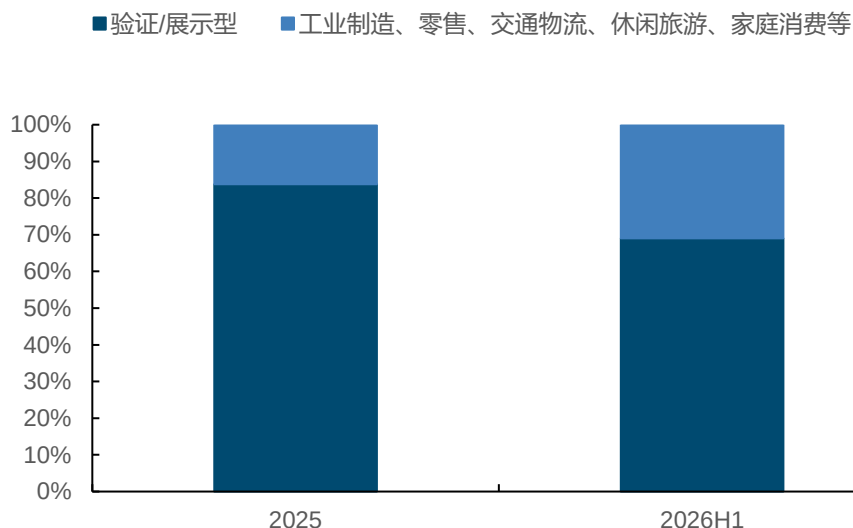
图表9：2026H1 全球人形机器人出货量同比接近 300%



来源：IDC、counterpoint、国金证券研究所

相比出货量增长，应用结构变化更能反映人形机器人商业化进程。2026H1，娱乐表演、数据生产及科研仍占全球出货的 60%以上，但服务引导、智能制造和仓储物流占比已分别达到约 19%、13%和 5%，机器人需求开始由 Demo、科研和数据采集向工厂、物流及商业服务等真实运行场景扩展。对于电子硬件而言，这一变化意味着评价标准也在发生迁移：研发阶段更关注性能、扩展性和功能实现，而规模部署后则更加重视 BOM 成本、可靠性、功耗、温度适应性、体积、生命周期及供应链稳定性。我们判断，随着真实场景渗透率提升，电子零部件需求有望由一次性、定制化的研发订单，逐步向标准化、批量化采购演进，具备成本控制、稳定供货和长期可靠性能力的供应商将更具优势。

图表10：人形机器人应用结构加速从验证型场景向真实作业场景扩散



来源：IDC、Counterpoint、国金证券研究所

2.2 训练数据需求前置，训练硬件率先释放

机器人本体尚未进入大规模出货阶段，但模型训练已经需要大量真实交互数据，因此数据采集硬件有望先于整机放量。当前机器人训练数据主要来自三类路径：真机遥操作获得的数据与目标机器人匹配度最高，但采集规模受机器人数量和使用效率限制；无本体采集则先记录人的真实操作，再通过动作映射等方式转化为机器人可用数据，从而降低对机器人本体数量的依赖；仿真数据则更多用于扩展长尾任务和组合场景。在此背景下，EGO、UMI 等训练侧硬件开始率先搭载相机、IMU 等传感器，用于获取第一视角、手部运动及操作过



程数据。我们认为，这意味着部分感知硬件的需求释放不必完全等待机器人整机出货，训练数据生产本身即可形成独立的硬件需求，并与后续真机采集、仿真数据共同构成机器人训练的数据来源。

图表11: EGO、UMI 与 WristCam 等多路径扩展真实交互数据入口

数据采集形态	是否占用机器人本体	主要采集信息	主要优势	主要局限
真机遥操作	是	图像、Robot Action、Joint State、Force 等	数据与目标机器人本体直接匹配，可直接用于策略训练	受机器人数量、场地及遥操效率约束
EGO 第一视角	否	场景视觉、人体行为、任务流程	设备轻量，人员和场景易扩展，适合大规模真实世界数据采集	缺少可直接执行的 Robot Action
Wrist-view 腕部采集	否	手-物体近场视觉、局部运动信息	补充头戴视角难以观察的接触、工具操作等细节	仍需完成 Human-to-Robot 视角及动作映射
UMI 手持式采集	否	图像、末端位姿 /IMU、夹爪状态等	在不占用机器人本体的情况下采集更接近机器人 Action 的操作数据	与目标机器人仍存在 Embodiment Gap，需要 Retargeting
无本体 HIL 采集	否	人类示教 + 同步 Policy 推理结果	可针对模型薄弱状态定向补充数据，进一步将数据采集与机器人部署解耦	尚处研究阶段，仍需最终真机验证

来源：UMI、WARPED、HIL-UMI、奥比中光、国金证券研究所

无本体数据采集正在由实验室自建工具向标准化训练硬件演进。过去相关设备多由科研团队围绕具体任务临时搭建，产品形态和接口体系较为分散；随着 Physical AI 训练对真实交互数据的需求提升，产业端开始将第一视角采集、手持操作采集、腕部近场采集以及多设备同步等功能拆分为独立产品，并形成可重复部署的标准化硬件组合。以奥比中光为例，其已推出 EGO、UMI、WristCam 及 Hub 等产品形态，分别覆盖不同采集位置与系统同步需求。这一变化说明训练侧硬件正在从科研项目配套工具逐步转向独立产品类别，其需求释放不必完全等待机器人本体放量；随着数据采集流程进一步标准化，相关相机、IMU、同步模组及边缘计算等电子硬件也有望由零散项目需求向可复制的批量采购演进。

图表12: 以奥比中光为例，已覆盖 EGO、UMI、WristCam 及 Hub 等多采集路径产品



来源：奥比中光官网、国金证券研究所

训练侧需求有望先于机器人本体大规模出货，形成一轮电子硬件的前置采购。当前人形机器人出货仍处数万台级，但模型训练和任务适配需要提前积累大量真实交互数据，因此产业节奏更可能表现为“模型能力提升—数据需求增加—数采设备采购—模型训练与适配—机器人量产—同类感知硬件进入本体”。从受益环节看，我们认为前置需求最直接的是视觉与运动感知硬件：EGO、WristCam 等设备首先增加相机模组需求，UMI 等操作采集设备



则进一步引入 IMU 及操作状态采集；随着多设备、多视角数据同步，数据采集 Hub、同步控制及高速传输环节的重要性也随之提升，同时大量视频和动作数据的持续记录还会带动本地计算与存储配置。相比之下，触觉、力觉等硬件更多与具体机器人本体和执行器结构绑定，其规模需求或更依赖后续整机部署。

图表13：训练侧硬件需求有望先于机器人本体大规模出货释放

模型训练与任务适配前置，电子硬件需求先于机器人本体规模量产释放



来源：国金证券研究所

三、模型加速下沉，端侧计算进入平台升级周期

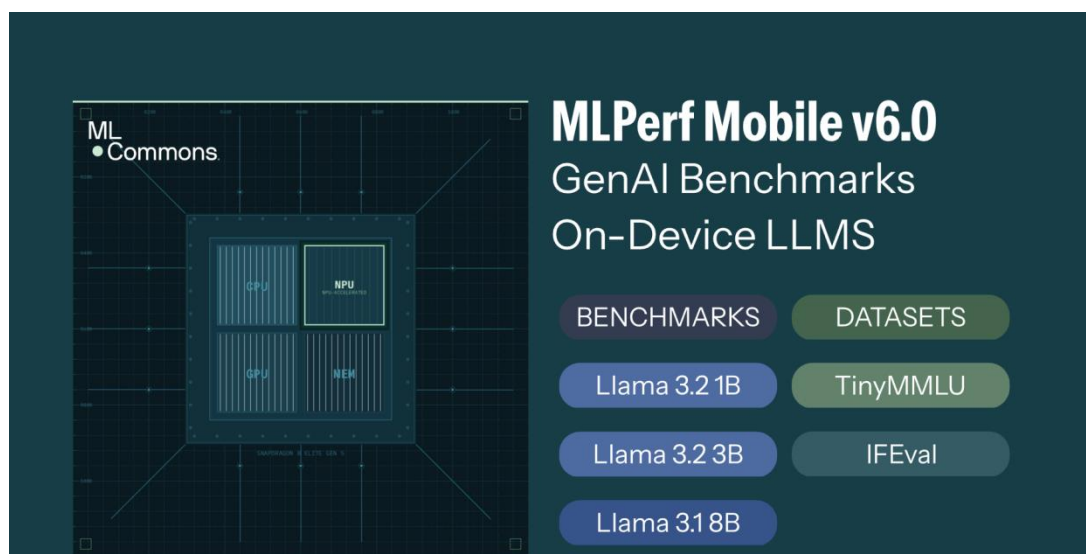
3.1 工作负载持续升级，端侧 AI 由感知走向智能体与动作执行

端侧 AI 的标准化测试对象正在由传统视觉任务向大模型、生成式 AI 及 Agent 扩展，这一变化本身反映了端侧硬件评价体系的升级。MLCommons 在 2026 年发布的 MLPerf Mobile v6.0 首次加入端侧 LLM 测试，覆盖 Llama 3.2 1B、3B 及 Llama 3.1 8B，并开始测试 8B 模型的 NPU 加速能力；随后 MLPerf Client v2.0 进一步纳入 Qwen3 8B、Phi 4 Reasoning 14B、图像生成及 Agent 端到端执行等工作负载。

相比过去以图像分类、目标检测等 CNN 任务为主，Benchmark 从“单次视觉推理”转向“大模型持续推理+生成式计算+多步骤 Agent 执行”，意味着端侧性能评价已不再只看峰值 AI 算力，而开始同时考察模型装载、持续数据搬运和长时间推理能力。对应硬件来看，更大的模型规模直接提高内存容量与带宽需求，生成式和 Agent 工作负载则进一步增加持续计算、存储访问及系统级调度压力。我们认为，Benchmark 对象的变化说明端侧 AI 竞争正由单点 NPU 性能向“算力+内存+存储+软件栈”的系统能力迁移，相关硬件升级也将从单纯提升 TOPS 进一步扩展至内存容量与带宽、存储及整机数据通路。



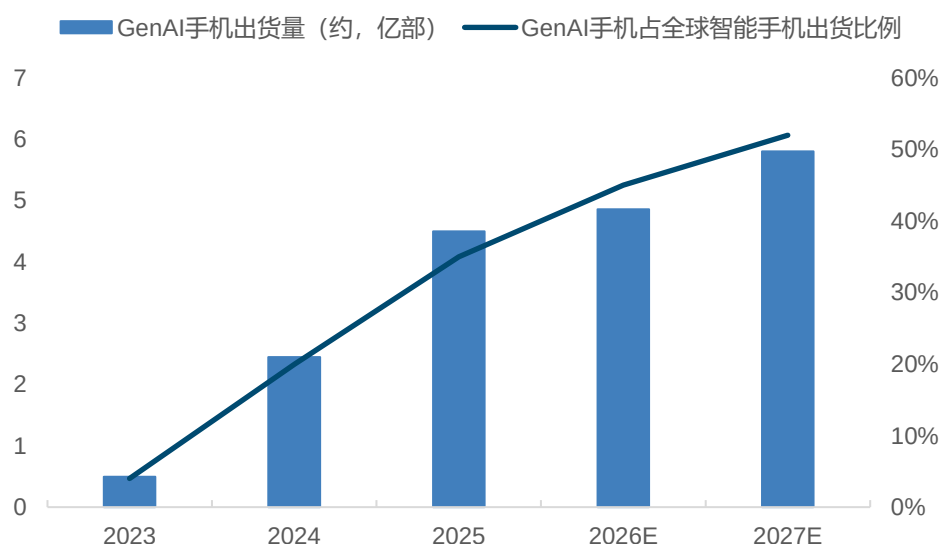
图表14: LLM、生成式AI与Agent逐步成为端侧标准工作负载



来源: MLCommons、国金证券研究所

手机与PC已经率先验证端侧AI由实验性功能走向规模化部署。Counterpoint预计,2026年GenAI手机将占全球智能手机出货量的45%,Gartner此前预计2026年AI PC占PC市场比重将超过50%,说明本地大模型正在进入亿级终端。其意义不仅在于新增AI功能,更在于端侧硬件开始围绕大模型重新配置算力、内存、存储及功耗管理等资源。Physical AI可以看作这一趋势向物理世界的进一步延伸:手机和PC主要在本地完成理解、生成和Agent任务,而机器人等物理终端还需要在此基础上进一步接入多模态传感器、实时控制与执行机构,将模型输出真正转化为物理动作。我们认为,机器人并不是一条孤立的硬件升级路线,而是端侧AI从“数字交互”走向“物理执行”的更高阶形态,手机、PC先完成本地大模型的规模化验证,Physical AI进一步把同一套趋势延伸至感知、控制和执行链条。

图表15: GenAI手机出货占比迅速提升



来源: Counterpoint、国金证券研究所

机器人VLA正在同步向端侧下沉,本地模型开始直接承担动作输出任务。Google DeepMind于2026年7月发布Gemini Robotics On-Device 2,将Vision-Language-Action模型直接部署至机器人本地设备,模型输入包括图像、文本和动作信息,输出则直接为机器人动作;官方明确强调其针对网络时延和无网络连接场景进行优化,能够在本地运行,并可较快适配新的机器人本体。对Physical AI而言,机器人所需的核心模型不再只是部署在云端或作为辅助模块存在,而是开始被专门压缩、适配并落地到本地计算平台,进一步抬升



对端侧算力、内存、存储和系统协同能力的要求。

图表16: 机器人 VLA 由云端推理向本地动作生成下沉

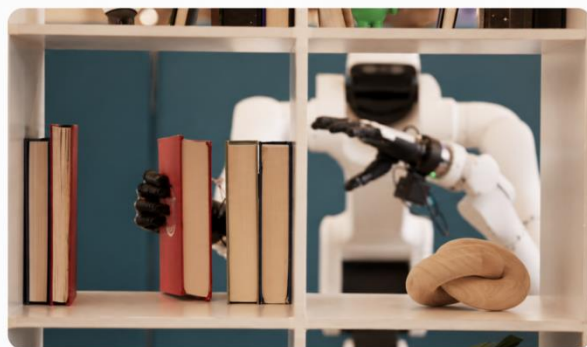


快速的硬件适配

原生支持多种机身形态，能够在仅需数小时的训练时间内，快速适配全新的机器人本体、形状和传感器。

支持本地运行，优化实时性能

许多机器人需要在无网络延迟或无互联网连接的环境下运行。Gemini Robotics 端侧模型 2 针对这些约束进行了专门优化。



来源：Google、国金证券研究所

3.2 模型本地化深化，端侧硬件需求持续扩散

大模型由云端向本地部署后，端侧硬件升级开始从 AI 算力进一步扩散至内存与存储。LLM、VLM 及 Agent 在本地运行时，需要同时容纳模型权重、上下文、KV Cache 及本地知识库等数据，对内存容量、带宽以及本地存储空间提出更高要求。MLPerf Client v2.0 的系统配置已经体现出这一趋势：不同平台最低内存要求普遍在 16GB 以上，部分混合计算及 GPU 模式进一步提升至 32GB 甚至 48GB，同时完整 Benchmark 运行需要数百 GB 可用磁盘空间；MLCommons 亦明确指出，更大模型和更长上下文可能因系统内存或计算资源不足而无法在部分客户端运行。我们认为，随着复杂模型持续向端侧下沉，终端 AI 硬件升级将不再局限于 SoC 或 NPU 性能提升，LPDDR 等内存容量与带宽、UFS/NVMe 等本地存储也将成为影响模型可运行规模和实际体验的重要配置约束，端侧 AI 带来的价值增量有望进一步向 Memory 与 Storage 环节扩散。


图表17: 复杂 AI 模型本地运行提升 Memory 与 Storage 配置要求

平台 / 运行方式	MLPerf Client v2.0 最低/推荐内存要求	主要特征
AMD Ryzen AI, NPU 模式	≥16GB	基础端侧 AI 推理
AMD Ryzen AI, Hybrid 模式	推荐≥32GB	CPU/GPU/NPU 混合运行
AMD Ryzen AI, GPU-only 模式	≥48GB	内存要求进一步提升
Qualcomm Snapdragon X 系列	≥32GB	本地大模型运行
Intel Core Ultra 系列	≥16GB, 扩展工作负载≥32GB	长上下文等场景提高内存需求
Apple M 系列 Mac	≥16GB	统一内存架构
Apple M2 及以上 iPad	≥16GB	移动端本地 AI
本地存储空间	至少 200GB, 完整 Benchmark 推荐 400GB	用于模型、数据集及运行文件

来源: MLCommons、国金证券研究所

模型轻量化推动端侧 AI 向更广泛设备扩散。随着小模型、量化压缩及端侧部署工具链持续成熟,生成式 AI 和复杂模型对终端算力、存储及功耗的要求逐步下降,端侧 AI 的应用范围有望由高性能 PC、智能手机和机器人进一步扩展至工业设备、AIoT、可穿戴设备及智能家居等更广泛终端。我们认为,端侧 AI 产业的增长动力将不仅来自单台设备算力及 Memory 配置提升,也来自 AI 能力向更多终端品类渗透带来的芯片需求扩容;不同终端根据算力、功耗和成本约束承载不同复杂度的模型,有望推动端侧 AI 芯片市场由高性能计算平台向更广泛的 Edge 设备持续延伸。

图表18: 模型轻量化推动端侧 AI 向更广泛设备扩散

设备层级	典型 AI 负载	代表场景
高性能客户端/边缘端	大语言模型 / 视觉语言模型 / 智能体	PC、高性能边缘计算
嵌入式边缘端	小语言模型 / 视觉 AI / 智能体	工业、机器人、智能设备
低功耗边缘端	视觉 AI / 语音 AI / 轻量 AI	IoT、智能家居
MCU 级边缘端	传感器 AI / 常时在线 AI	传感、唤醒、控制

来源: Google、瑞芯微、NXP、国金证券研究所

四、相关标的

- 1) **Physical AI 本体及产业化:** 宇树科技、优必选等。随着通用模型由感知理解进一步进入动作空间,机器人在任务规划、全身运动和灵巧操作等环节的能力持续提升,本体厂商有望率先受益于 Physical AI 能力升级及后续商业化落地。
- 2) **数据采集及感知硬件:** 奥比中光、思特威、韦尔股份、芯动联科、安培龙、必易微等。Physical AI 训练对真实交互数据的需求持续增加,同时机器人进入开放环境后,对视觉、惯性、触觉/力觉等多模态感知以及配套信号链、供电与控制硬件提出更高要求,相关电子零部件需求有望由研发验证逐步向标准化、批量化采购演进。
- 3) **端侧 AI 芯片:** 瑞芯微、星宸科技、乐鑫科技、恒玄科技等。随着端侧 AI 工作负载由传统视觉识别向 LLM、VLM 及 Agent 扩展,同时模型轻量化与部署工具链逐步成熟,AI 能力有望由高性能终端向机器人、工业、AIoT、智能家居及可穿戴等更广泛设备渗透,带动端侧 AI SoC 算力、内存接口及系统集成能力持续升级。
- 4) **整机制造:** 立讯精密、工业富联等。OpenAI 等头部 AI 厂商正在由模型和数据能力进一步向机器人控制、计算平台、电机及执行器等硬件环节延伸,Physical AI 产业竞争逐步从单一模型能力走向软硬件一体化。随着头部厂商后续推进机器人产品工程化与规模制造,具备复杂电子产品组装、系统集成及全球供应链管理能力的制造平台有望迎来新的增量需求。

五、风险提示

Physical AI 模型能力提升不及预期的风险。



当前 VLA、世界模型及机器人 Agent 仍处于快速迭代阶段，复杂场景泛化、长任务执行及失败恢复等能力仍有待验证，若模型能力提升速度不及预期，可能影响 Physical AI 应用落地节奏。

机器人产业化及终端出货不及预期的风险。

当前人形机器人仍处于小批量验证及商业化早期阶段，产品成本、可靠性、应用场景及商业模式尚未完全成熟，若终端需求或量产进度不及预期，相关感知、计算及控制硬件需求可能低于预期。

Physical AI 数据采集及训练硬件需求不及预期的风险。

真实世界数据采集路径仍在演进，无本体采集、遥操作、仿真及合成数据等方案之间存在替代和互补关系，若真实数据需求低于预期或数据采集方式发生明显变化，相关训练硬件需求可能不及预期。

端侧 AI 渗透及模型本地部署进度不及预期的风险。

端侧大模型仍受到算力、功耗、Memory 容量及应用体验等因素约束，云端与端侧的任务分工尚未定型；若模型轻量化、部署工具链或终端应用发展不及预期，端侧 AI 芯片及相关硬件升级节奏可能放缓。

端侧硬件配置升级幅度不及预期的风险。

端侧 AI 对 SoC、Memory 及 Storage 的增量需求受到模型压缩效率、软硬件协同及成本约束影响，若模型量化及压缩技术进步快于预期，或终端厂商优先通过软件优化降低硬件需求，相关硬件单机价值量提升幅度可能低于预期。

技术路线快速变化及产业竞争加剧的风险。

Physical AI 及端侧 AI 仍处于技术快速迭代阶段，模型架构、芯片方案、数据采集方式及软硬件生态均可能发生变化，若相关公司技术路线选择失误、产品迭代不及时或市场竞争加剧，可能影响其市场份额及盈利能力。



行业投资评级的说明：

买入：预期未来 3—6 个月内该行业上涨幅度超过大盘在 15%以上；

增持：预期未来 3—6 个月内该行业上涨幅度超过大盘在 5%—15%；

中性：预期未来 3—6 个月内该行业变动幅度相对大盘在 -5%—5%；

减持：预期未来 3—6 个月内该行业下跌幅度超过大盘在 5%以上。



特别声明：

国金证券股份有限公司经中国证券监督管理委员会批准，已具备证券投资咨询业务资格。

形式的复制、转发、转载、引用、修改、仿制、刊发，或以任何侵犯本公司版权的其他方式使用。经过书面授权的引用、刊发，需注明出处为“国金证券股份有限公司”，且不得对本报告进行任何有悖原意的删节和修改。

本报告的产生基于国金证券及其研究人员认为可信的公开资料或实地调研资料，但国金证券及其研究人员对这些信息的准确性和完整性不作任何保证。本报告反映撰写研究人员的不同设想、见解及分析方法，故本报告所载观点可能与其他类似研究报告的观点及市场实际情况不一致，国金证券不对使用本报告所包含的材料产生的任何直接或间接损失或与此有关的其他任何损失承担任何责任。且本报告中的资料、意见、预测均反映报告初次公开发布时的判断，在不作事先通知的情况下，可能会随时调整，亦可因使用不同假设和标准、采用不同观点和分析方法而与国金证券其它业务部门、单位或附属机构在制作类似的其他材料时所给出的意见不同或者相反。

本报告仅为参考之用，在任何地区均不应被视为买卖任何证券、金融工具的要约或要约邀请。本报告提及的任何证券或金融工具均可能含有重大的风险，可能不易变卖以及不适合所有投资者。本报告所提及的证券或金融工具的价格、价值及收益可能会受汇率影响而波动。过往的业绩并不能代表未来的表现。

客户应当考虑到国金证券存在可能影响本报告客观性的利益冲突，而不应视本报告为作出投资决策的唯一因素。证券研究报告是用于服务具备专业知识的投资者和投资顾问的专业产品，使用时必须经专业人士进行解读。国金证券建议获取报告人员应考虑本报告的任何意见或建议是否符合其特定状况，以及（若有必要）咨询独立投资顾问。报告本身、报告中的信息或所表达意见也不构成投资、法律、会计或税务的最终操作建议，国金证券不就报告中的内容对最终操作建议做出任何担保，在任何时候均不构成对任何人的个人推荐。

在法律允许的情况下，国金证券的关联机构可能会持有报告中涉及的公司所发行的证券并进行交易，并可能为这些公司正在提供或争取提供多种金融服务。

本报告并非意图发送、发布给在当地法律或监管规则下不允许向其发送、发布该研究报告的人员。国金证券并不因收件人收到本报告而视其为国金证券的客户。本报告对于收件人而言属高度机密，只有符合条件的收件人才能使用。根据《证券期货投资者适当性管理办法》，本报告仅供国金证券股份有限公司客户中风险评级高于 C3 级（含 C3 级）的投资者使用；本报告所包含的观点及建议并未考虑个别客户的特殊状况、目标或需要，不应被视为对特定客户关于特定证券或金融工具的建议或策略。对于本报告中提及的任何证券或金融工具，本报告的收件人须保持自身的独立判断。使用国金证券研究报告进行投资，遭受任何损失，国金证券不承担相关法律责任。

若国金证券以外的任何机构或个人发送本报告，则由该机构或个人为此发送行为承担全部责任。本报告不构成国金证券向发送本报告机构或个人的收件人提供投资建议，国金证券不为此承担任何责任。

此报告仅限于中国境内使用。国金证券版权所有，保留一切权利。

上海

电话：021-80234211

邮箱：researchsh@gjzq.com.cn

邮编：201204

地址：上海浦东新区芳甸路 1088 号

紫竹国际大厦 5 楼

北京

电话：010-85950438

邮箱：researchbj@gjzq.com.cn

邮编：100005

地址：北京市东城区建国门内大街 26 号

新闻大厦 8 层南侧

深圳

电话：0755-86695353

邮箱：researchsz@gjzq.com.cn

邮编：518000

地址：深圳市福田区金田路 2028 号皇岗商务中心

18 楼 1806



【小程序】
国金证券研究服务



【公众号】
国金证券研究