

WAIC 2026 摩尔线程展台介绍

主题：“词元时代 万物智能”

WAIC 2026 将于 7 月 17 日在上海拉开帷幕。作为国内领先的全功能 GPU 企业，摩尔线程（688795.SH）将以“词元时代 万物智能”为主题，携“云-边-端”全场景智算产品矩阵与前沿解决方案精彩亮相，首次开启上海世博展览馆与张江科学会堂双展区联动，并于 7 月 18 日举办高规格主题论坛，系统呈现从算力基础设施到智能应用落地的全栈能力与产业实践。

随着 AI 浪潮将“词元（Token）”从技术概念推向千行百业，一个以词元为核心的智能时代正加速到来。万物智能已处于爆发前夜，算力成为决定这一进程速度与深度的基石。面对词元需求的指数级跃升，摩尔线程创始人、董事长兼 CEO 张建中提出建设三大“AI 工厂”：“模型训练工厂”“词元生产工厂”“智能体生产工厂”，旨在以全场景智算能力，加速千行百业智能化升级。

展览信息

展览时间：7 月 17 日-20 日

展览地点 1：上海世博展览馆 H2-C109

展览地点 2：上海张江科学会堂 Z1C-101



WAIC 2026 摩尔线程展区核心亮点

在本届 WAIC 上，摩尔线程将以三大“AI 工厂”为核心主线，全景展出打通大模型“训、推、



用”闭环的硬核亮点与产品矩阵：

▼ 亮点一：走进“模型训练工厂”，见证万卡级集群与“国芯训国模”硬核实力

展区现场将深度解密支撑万卡万亿参数大模型训练的“夸娥”底座，实景呈现大规模 GPU 智算集群的高效运营状态。依托这一强大底座，**摩尔线程现场首秀了多项标杆级“国芯训国模”成果**：包括完成了从零起步完整训练的 MoE-236B 基础模型，以及全球首个 5D 世界模型“北大 EvoPhys-World”的全栈原生训练，充分验证了国产芯片在超大规模、超长连续稳定训练上的全链路闭环能力。

▼ 亮点二：探秘“词元生产工厂”，体验极致吞吐与大模型极致推理部署

摩尔线程集中展示了基于旗舰级 AI 训推一体智算卡 MTT S5000 打造**的高性能推理方案**。现场不仅有针对 GLM、DeepSeek 等主流大模型的极速推理展示，树立国产 GPU 推理新标杆。更重磅推出了 PD（Prefill-Decode）异构分离方案，通过将高算力需求的 Prefill 计算池（MTT S5000）与高带宽的 Decode 生成池（国际主流 GPU）异构分离、分池部署，独家支持 KV Cache FP8 与 1M 超长上下文，重塑了算力成本边界。

▼ 亮点三：解码“智能体生产工厂”，直击全栈具身智能仿真与端侧 AI 生产力

智能体与端侧承载是 AI 走进生产与生活的终极形态。摩尔线程依托强大的通用全功能 GPU 算力，推动“智能体生产工厂”全面运转。在现场展示了全栈国产化具身智能仿真平台 MT Lambda 及面向工业场景的 MT Lambda for Factory 解决方案。此外，集成家庭 AI 中枢的 MTT AICUBE 与“为智能体而生”的 AIBOOK 算力本也开放了现场 Live 体验，其凭借端云协同和独特的内存等效扩容技术，让端侧无需升级硬件即可流畅运行 80B 超大 MoE 模型与智能体协作。

以下是摩尔线程 WAIC 2026 展区完整介绍：

“云-边-端”全场景智算产品矩阵

▼ MTT S5000-AI 训推一体智算卡

专为大模型训练、推理及高性能计算而设计的全功能 GPU 智算卡，基于第四代 MUSA 架构“平湖”打造，完整支持从 FP8 到 FP64 的全精度计算。目前 S5000 正加速规模化量产，基于该产品搭建的万卡级集群已完成部署并上线服务。

▼ MTT SGX5000-AI 大模型训推一体机

摩尔线程 AI 大模型训推一体机 MTT SGX5000，专为大语言及多模态大模型、科学计算和物理仿真等前沿领域打造。该服务器基于摩尔线程全功能 GPU 模组与高速全互联架构，全面释放计算性能，为高性能计算集群的构建提供强大支持，赋能企业在复杂 AI 场景中实现极致算力突破。

▼ 超节点架构

摩尔线程亮相了 MTT C256 超节点，用一层 Scale-up 网络把 256 张 GPU 变成一台数据中心级计算机。它具备三大核心优势：首先是高密度，单机柜 GPU 密度领先行业，两机柜实现 256 卡一层 Scale-up 全互联，节省机房空间；其次是极致互联，用一层 Scale-up 网络实现 256 卡全互联（业界多为一层 64 卡），卡间通信延迟压缩至亚微秒级，带宽利用率更高，为并行策略优化提供更大空间；最后是生态兼容，2U 节点设计兼容现有智算软硬件生态，便于从万卡向十万卡级集群平滑扩展。同时，MTT C256 超节点的工程化优势极为突出，通过高效液冷散热、三盲插设计、完善的 RAS 机制和集群管理系统，保障了超节点在大规模部署下的稳定性和可维护性，为万卡及十万卡级超大规模长稳训推集群打造坚实基础。

▼ MTT E300-AI 计算模组

MTT E300 AI 模组搭载自研智能 SoC“长江”，具备 50TOPS 异构 AI 算力，支持混合精度计算。凭借高算力、全栈 AI 工具链及端云协同架构，可提供高性能、低延迟、强可靠的国产边缘 AI 解决方案，广泛应用于工业、能源、教育、交通、医疗等行业。

▼ MTT AIBOOK-AI Agent 笔记本

MTT AIBOOK AI 算力本是专为 AI 学习与开发者的个人智算平台，在 MTT AIOS 原生 Linux 系统下预装 OpenClaw 智能体，单机可以支持多达 12 个智能体高效并行运行，实现了智能体应用开发、调试与部署的完整闭环。该产品搭载摩尔线程自研智能 SoC 芯片“长江”，提供 50TOPS 异构 AI 算力，支持混合精度计算；具备多系统兼容能力，并预置完整的 AI 开发环境与工具链。

▼ MTT AICUBE-家庭 AI 中枢

MTT AICUBE 是未来 AI 智能家庭的核心中枢，它具备完整的 AIPC 及 AI NAS 功能，搭载自研智能 SoC“长江”，具备 50TOPS 异构 AI 算力，内置基于 Linux 内核的 MTT AIOS 操作系统及智能体“小麦”，可以为用户提供开箱即用的一站式 AI 体验。通过 AICUBE，用户可以和“小麦”智能体进行自然语音交互，轻松完成日常照片及视频等文件的存储管理，畅享娱乐、学习及社交活动，拥抱由 AI 充当数字管家所带来的全新智能生活。

以“三大 AI 工厂”为主题的展区介绍

模型训练工厂

- ▼ 展示内容一：夸娥（KUAE）智算集群解决方案，支持国产全功能 GPU 万卡万亿大模型训练

摩尔线程基于 MTT S5000 打造的夸娥智算集群解决方案，面向超大规模智算中心，具备单集群万卡级 AI 算力，覆盖 FP64 至 FP8 全精度计算，支持 AI、图形、科学计算等全场景加速。夸娥智算集群在 AI 大模型训练性能表现优异，Dense 模型 MFU 达到 60%，MoE 模型 MFU 达到 40%，万卡集群综合算力利用率达到国际领先水准。在训练精度方面，夸娥万卡级智算集群可与国际主流计算集群对齐，覆盖 MoE、多模态、具身智能等前沿架构，训练 Loss 曲线与参照基准基本重合，收敛行为一致。基于 CPT 与 Long CPT 全链路预训练数据，MMLU 评测结果表现优异。这标志着大模型从零开始的完整训练，可在夸娥集群上端到端完成。夸娥智算集群解决方案在扩展性方面同样处于领先水平，其线性扩展效率达 95%，千卡至万卡规模的算力扩展中，单卡有效算力损耗控制在 5% 以内。

▼ 展示内容二：夸娥智算集群运行实景展示，高效管理集群运营

摩尔线程夸娥智算中心解决方案以全功能 GPU 为底座，是软硬一体化的全栈解决方案，包括以夸娥计算集群为核心的基础设施、KUAE Platform 集群管理平台以及 KUAE ModelStudio 大模型平台，旨在以一体化交付的方式解决大规模 GPU 算力的建设和运营管理问题。该方案可实现开箱即用，大大降低传统算力建设、应用开发和运维运营平台搭建的时间成本，实现快速投放市场开展商业化运营

集群建成后，系统管理平台可实时监控算力使用率、节点状态、网络吞吐及存储性能等关键指标。运维人员通过可视化大屏直观掌握 GPU、服务器、网络及存储设备的全栈运行状态，确保集群资源始终处于高负载、稳定运行的最佳效能，为大规模 AI 训练与推理任务提供坚实运维保障。

▼ 展示内容三：现场首秀多项标杆级“国芯训国模”成果

训练成果 1：MoE-236B 基础模型，“国芯国模”万卡训练新突破

依托摩尔线程与合作伙伴联合建设的国产万卡 GPU 集群，**相关基础模型项目已完成 MoE-236B 架构模型从零开始的完整训练流程**。模型基于 25T+ 高质量训练语料，采用混合专家（MoE）架构，覆盖预训练（Pretrain）、持续预训练（CPT）、长窗口持续预训练（Long-CPT）全链路，MMLU 评测精度与国际主流计算卡持平，训练损失曲线高度一致。

训练全程由摩尔线程夸娥 AI 训练套件（KUAE Training Suite）提供底层技术支持，有效训练时长占比超 90%，在万卡规模下展现出与国际同类集群相当的系统稳定性与训练效率。此次突破标志着“国产芯片+国产模型”技术路线的实质性验证——从芯片、集群、训练框架到基础模型，全链路实现自主可控，已开始超万亿参数大模型的训练，同时为各类前沿科学领域提供了可落地的算力支撑。

训练成果 2：北大 EvoPhys-World，国产 GPU 首次全栈支撑世界模型原生训练

北大 EvoPhys 团队推出全球首个 5D 世界模型 EvoPhys-World，一经发布便在斯坦福 WorldScore“世界生成”维度位列全球第一，并连续 37 天蝉联榜首。模型基于约 4 万小时

无标注第一视角手部交互数据训练，首创 World Engine（世界生成与物理交互）与 World Policy（世界预演与动作决策）双形态架构，构建“生成—操控”闭环。

其原生训练全程在基于摩尔线程 MTT S5000 的夸娥智算集群完成，MUSA 软件栈全栈支撑，训练稳定性、扩展性与生成质量均达国际主流水平，为中国原创模型提供坚实算力底座。作为摩尔线程“灯塔计划”代表成果，该项目是“国芯国模”协同创新的标杆实践，为高水平科技自立自强提供了可复制的产学研协同范式。

训练成果 3：智源 RoboBrain 2.5，国产算力集群首次全流程支撑具身大脑训练

RoboBrain 是北京智源研究院面向真实物理场景打造的通用具身大脑，以统一的视觉—语言多模态架构，支撑机器人在感知、认知、推理与决策上的核心能力。RoboBrain 2.5 新增了对动作时序价值评估和三维空间结构的理解推理能力，显著提升下游任务执行成功率。

基于 FlagOS-Robo 框架，依托摩尔线程 MTT S5000 千卡智算集群，智源成功完成 RoboBrain 2.5 的全流程训练，首次验证国产算力集群在具身智能大模型训练中的可用性与高效性。通过 FlagOS 与硬件的协同优化，该方案实现“训得稳、训得快”，为具身智能从实验室走向产业落地提供了坚实底座。

训练成果 4：硬核国产芯算力，高效训练智驾感知模型

自动驾驶 BEV 感知是智能驾驶量产体系的核心环节，其模型训练对算力吞吐与多机扩展稳定性要求极高。本方案基于摩尔线程 MTT S5000 的夸娥智算集群，对行业主流 BEV 感知模型 BEVFormer 完成训练实测，覆盖智驾感知模型训练的真实负载。

实测结果显示，MTT S5000 凭借 MUSA 软件栈，单机与双机训练吞吐均达国际主流训练 GPU 的 1.56–1.67 倍，扩展倍数稳定，充分验证了国产全功能 GPU 高效支撑智能驾驶核心模型训练的能力。作为夸娥集群产业实践的代表场景，该成果证明智驾模型训练全流程国产化切实可行，为 MUSA-X 落地千行百业，服务高水平科技自立自强提供了可复制的国产算力底座方案。

词元生产工厂

▼ 展示内容一：大模型高性能推理方案，极速释放潜能

摩尔线程基于旗舰级 AI 训推一体智算卡 MTT S5000 与全栈自研 MUSA 架构，构建从底层算子库到上层推理框架的完整推理生态。深度适配 SGLang、vLLM、ONNX Runtime、Llama.cpp 等主流推理框架，全面覆盖大语言模型、视觉、语音、DiT 及视频生成等多模态推理场景。

针对大语言模型的 PD 分离部署，通过通信-计算重叠降低资源开销，支持 TP/PP/DP/CP/EP 多并行策略，以及超大规模 EP 的负载均衡；集成 FlashMLA、DeepGEMM 等高性能算子

库，算子效率接近峰值 90%。面向视频生成模型，通过并行策略及推理框架优化，以及融合算子、NDHWC 内存布局、FasterCache、Attention Cache 等算法创新进一步降低推理成本，持续提升生成效率，为不同场景提供高性价比、高吞吐、低延迟的国产推理解决方案。

▼ 展示内容二：打通多元生态，全栈推理赋能大模型

摩尔线程大模型推理方案基于旗舰级 AI 训推一体智算卡 MTT S5000 打造，方案结合 MTT S5000 硬件、MUSA 软件栈、AI 推理套件及夸娥云服务，为用户提供高性能、易部署的推理方案。该方案深度适配 GLM、DeepSeek、Qwen、Kimi、MiniMax 等业界主流大模型，覆盖通用对话、智能编码 Agent、多工具自主任务等复杂应用场景。

MTT S5000 采用先进“平湖”架构，支持 FP8 至 FP64 全精度算力。第四代 MUSA 平台原生适配 PyTorch、SGLang 等框架，实现多模型 Day-0 适配。通过分布式并行与算子优化，实现高吞吐、低延迟、长序列的稳定推理，彰显自主可控算力底座的性能与兼容性。

▼ 展示内容三：PD 异构分离，重塑算力成本边界

长上下文大模型推理场景的性能瓶颈，正由 Decode 生成侧，逐步转向 Prefill 输入侧计算与 KV Cache 生成。Prefill 依赖高算力，Decode 依赖高带宽与大显存，传统同构部署难以兼顾性能与成本。PD 异构分离通过分池部署、分别调度、独立优化，使 Prefill 与 Decode 按不同资源需求灵活扩容，更高效匹配长输入、高吞吐、低时延等复杂推理场景。

在该方案中，MTT S5000 作为 Prefill 计算池，充分发挥其长输入高吞吐优势，承担 Prompt 编码、Attention 计算和 KV Cache 生成；国际主流 GPU BXXX/HXXX/Hxx 作为 Decode 生成池，专注 TPOT、输出吞吐和流式稳定性，通过 KV Cache 高效传输实现 P/D 解耦协同。方案可支撑 L3/L4 级别 Token 服务，精度与官方结果对齐、误差小于 1%，独家支持 KV Cache FP8，并支持 1M 上下文处理能力，为大模型推理提供高性价比的国产算力底座。

▼ 展示内容四：破局场景之困，推理能力落地千行百业

随着 AI 技术加速落地千行百业，场景化推理需求爆发式增长，成为产业发展的核心驱动力。不同行业的推理场景对算力规格、精度、时延及并发能力诉求差异显著，单一算力形态难以满足多元化需求。摩尔线程全功能 GPU 凭借 AI 计算加速、图形渲染、物理仿真和科学计算、视频编解码等核心能力，天然适配多样化推理负载，在不同领域发挥独特优势。

依托 MTT S5000 训推一体智算卡、MTT E300 AI 模组、MTT AIBOOK 算力本等产品矩阵，摩尔线程灵活匹配从云端大规模训练到边缘轻量化推理、从移动端部署到便携式开发等多元场景需求，提供高效、稳定、易用的推理算力支撑。方案可应用于新能源、金融、工业、具身智能、能源热力、财务税务、安全生产、数字资产等多个关键领域，持续赋能行业智能化转型。

智能体生产工厂

MTT AICUBE

▼ 展示内容一：MTT AICUBE-家庭 AI 中枢，懂生活更懂你

MTT AICUBE 以全域智能体“小麦”为核心，深度整合“智能体+AI PC+AI NAS”三大核心能力，通过全场景语音控制颠覆传统 PC 对键鼠交互方式的依赖，具备完整的桌面 AI PC 能力，可高效支撑端侧智能体、超高清视频播放以及 3D 游戏等家庭应用场景；AI NAS 模式下，配备全闪 SSD 高速存储，并支持免拆机扩容，轻松收纳全家人的数字资产。为 AI 时代家庭智能体验提供一站式解决方案。

MTT AIBOOK

▼ 展示内容二：MUSACODE × MusaCoder MUSA 开发新利器

MUSACODE 是摩尔线程自主研发的 AI 编程智能体（AI Coding Agent），支持开发者通过自然语言指令完成代码生成、调试与重构，深度优化 MUSA 生态开发体验。工具覆盖算子生成、通用计算库自动化迁移，以及开发、测试、编译、性能优化全流程，大幅提升研发效率。

其背后的专用代码大模型引擎 MusaCoder，是业内首个基于国产 GPU 算力底座完成全链路训练与验证的开源代码大模型，其完整训练流程均在基于 MTT S5000 构建的夸娥智算集群上完成。MusaCoder 专注 MUSA 原生 Kernel 自动生成与优化，在权威评测中性能达行业领先水平。

通过“工具+大模型”双引擎驱动，MUSACODE 让 MUSA 开发更简单、更高效。

▼ 展示内容三：OPC 一人公司 开启“单人成军”新模式

借助 MTT AIBOOK 预装的 OpenClaw 等智能体，个人创业者可以突破自身技能与精力的限制，以一己之力统筹全局，调度多种 AI Agent 组建全天候运作的数字员工团队。各类智能体协同分工，覆盖产品研发、内容设计、营销推广、客服运维等全链路工作，帮助个人以低成本实现商业闭环，打造“单人成军”的轻量化创业模式。

MTT AIBOOK 基于摩尔线程“长江”SoC 的高并发算力能力，内置 OpenClaw，并支持多智能体协同运行，充分满足 Agent 集群的算力需求。MTT AIBOOK 可打通夸娥云端算力，灵活部署各类大模型，通过端云协同为数字员工团队提供强大的 AI 能力支撑，为单人创业模式筑牢算力基础。

▼ 展示内容四：创新支持多系统并行 MTT AIOS 筑牢应用底座

MTT AIOS 是摩尔线程基于 Linux 深度自研的全功能操作系统，其端侧载体为 MTT AIBOOK。

内置完整 MUSA 开发套件与 AI 智能体，覆盖办公、开发、影音娱乐全场景。系统预装 Chromium 浏览器，自研 AI 邮箱，各类国产文档编辑、即时通信和会议软件，集成 MUSACODE、VSCode 等全套编程工具。

针对跨平台兼容难题，系统创新支持原生 Linux、虚拟化 Windows、容器化 Android 多系统并行，首创 vGPU 硬件直通与 DirectX12 适配，支持高分高刷显示及双系统文件无缝共享，可流畅运行 Windows 办公、设计、剪辑类专业软件，兼顾国产生态与跨平台生产力。

MTT AIOS 打通芯片、系统、应用全栈链路，大幅降低政企与开发者迁移成本，拓宽国产芯片落地场景，支撑云边端布局，推动国产 AI 产业规模化发展。

▼ 展示内容五：端侧内存不再受限 MTT AIBOOK 等效扩容跑通 MoE 大模型

大模型参数越大智能越强，但端侧内存上限限制百亿级 MoE 模型本地部署。MTT AIBOOK 搭载 32GB 高速统一内存，结合 AI 加速分区与智能推理软件系统，通过 AISSD 自适应缓存管理机制，集中调度 MoE 专家权重、长文本 KV 缓存与智能体内存数据，无需升级硬件即可在原生 32GB 内存上流畅运行 80B 大模型推理与智能体编程。

摩尔线程携手合作伙伴基于 Llama.cpp-MUSA-M1000 深度优化，运行 80B MoE 模型可等效扩容 20-25GB 内存，预填充 8.34 Token/s，解码突破 7 Token/s；运行 35B MoE 模型可扩容 9-15GB，解码达 17+ Token/s，同时运行 OpenCode 智能体框架。不依赖物理内存堆叠，不受内存涨价影响，让普通用户的 32GB MTT AIBOOK 轻松运行超大模型。

具身智能

▼ 展示内容六：MT Lambda 全栈国产化具身智能仿真平台

摩尔线程首个全栈具身智能仿真平台 MT Lambda，旨在赋能用户构建数据合成、策略训练、仿真验证的高效工作流。

该平台构建了从底层算力、核心引擎到上层框架及工具的完整解决方案：其底层基于摩尔线程全功能 GPU，实现渲染、物理、AI 计算在同一芯片中完成，数据“零拷贝”；中间层深度融合自研物理、渲染、AI 三大引擎；上层则提供 MTLambda-Lab 具身策略开发与训练平台以及 MT Lambda-Sim 高保真物理仿真与渲染平台。

▼ 展示内容七：MT Lambda for Factory 工业具身智能解决方案

为促进具身智能与工业制造深度融合，摩尔线程推出自研工业具身智能平台 MT Lambda for Factory，集成高保真工厂渲染、高精度物理模拟、域随机化数据生成、轨迹生成与任务规划、产线点位基准评测等核心能力，打造智能工厂专属的“虚拟调试场”与“AI 模型孵化器”：企业在虚拟环境中即可完成产线调试、工艺优化与 AI 模型迭代，大幅降低物理试错成本，缩短新产线从部署到投产的周期。



平台依托摩尔线程全功能 GPU 与 MUSA 统一架构，自研 AlphaCore 物理引擎与光追渲染引擎，物理仿真、渲染、AI 训练三类负载在同一 GPU 内完成、数据零拷贝，实现具身智能混合负载的架构级适配。作为摩尔线程具身智能全栈布局的核心平台，该方案证明智能工厂的训练底座可以全链路自主可控，为新型工业化进程中培育新质生产力提供了可靠的国产化基座。

+++

关于摩尔线程

摩尔线程以全功能 GPU 为核心，致力于向全球提供加速计算的基础设施和一站式解决方案，为各行各业的数智化转型提供强大的 AI 计算支持。我们的目标是成为具备国际竞争力的 GPU 领军企业，为融合人工智能和数字孪生的数智世界打造先进的加速计算平台。我们的愿景是为美好世界加速。

- 国内唯一实现“图形+AI 计算”深度融合的全功能 GPU 厂商
- 国内唯一同时深耕 B 端、C 端市场的全场景 GPU 厂商
- 国内唯一单芯片同时支持 AI 计算加速、图形渲染、物理仿真和科学计算、超高清视频编解码
- 国内首个发布支持 DirectX 12 的图形加速引擎
- 国内极少数支持从 FP8 到 FP64 等全计算精度的国产 GPU 厂商
- 国内唯一功能可对标英伟达的全功能 GPU 厂商
- 国内首个完成万卡集群建设、并高效支持万亿参数的大模型训练的国产 GPU 公司
- 推出国内首个全栈具身智能仿真训练平台 MT Lambda，成为继英伟达之后，全球第二家具备该全栈能力的 GPU 企业。
- 国内首个支撑顶尖世界模型北京大学 EvoPhys-World 完成全栈训练、并登顶斯坦福 WorldScore 世界模型评测榜首的国产 GPU 厂商
- 国内首个自主全栈训练出代码大模型 MusaCoder、且性能超越 Claude Opus 4.7 等国际主流 SOTA 模型的国产 GPU 厂商