东吴证券 SOOCHOW SECURITIES

# 汽车行业深度报告

# AI 系列深度 26 期：云端基座模型如何在毫秒级车端跑起来

2026 年 08 月 07 日

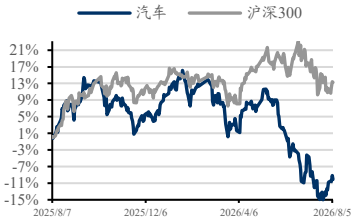
证券分析师 黄细里
执业证书：S0600520010001
021-60199793
huangxl@dwzq.com.cn
研究助理 童明祺
执业证书：S0600125080005
tongmq@dwzq.com.cn

增持（维持）

## 投资要点

- **学术缩放定律与车端工程约束对应不同的优化目标。**Kaplan、Chinchilla 等缩放定律讨论的是给定训练算力下参数量、数据量与损失之间的最优关系，其“最优”指向训练算力前沿，并未把推理成本和毫秒级实时时延纳入目标函数。将推理成本显式纳入后，部署最优点会向更小模型、更多训练数据偏移；智能驾驶则进一步强化了这一部署约束。
- **车端实时性构成大模型上车的核心约束。**自动驾驶模型必须在固定车规算力上满足帧率、端到端时延和功能安全要求，大型视觉 Transformer 叠加大语言模型难以在现有车规算力下持续满足高频推理要求。理想披露 VLA 推理帧率约 10Hz，NVIDIA Alpamayo-R1 论文披露端到端推理时延 99ms，小鹏称第二代 VLA 指令输出时延压缩至 80ms 以内。
- **云端基座模型的效率优化主要可归纳为两类路径：**一是通用模型稀疏化和注意力/KV 缓存压缩，如 MoE、MLA、MTP、FP8、线性注意力等；二是面向驾驶的视觉 Token 剪枝、动态/隐式 Token 和思维链压缩，用更少 Token 承载关键场景信息。其目标是在云端继续探索能力上限，同时降低长序列和多模态推理成本。
- **车端时延压缩更强调可部署性，主流范式是“云端大模型蒸馏上车+快慢双系统”。**Waymo 通过教师—学生蒸馏把大模型能力迁移到实时学生模型，并采用 Think Fast/Think Slow 架构；理想由云端基座蒸馏出车端 VLA，并用扩散流匹配减少去噪步数；小鹏则以基座蒸馏和视觉思维链效率提升，降低车端推理时延。
- **我们归纳，未来云车分工或趋于“云端探上限、车端保实时”。**云端模型负责更大规模、更复杂推理和世界模型训练，车端模型则围绕低时延、稳定性和安全冗余做压缩部署。该判断仍受蒸馏能力上限、统一评测缺失、芯片供给、量产时间表和功能安全验证约束，因此，训练侧的规模扩张逻辑不能直接外推至车端部署。
- **我们认为，数字 AI 底座模型发展路径已逐渐清晰、价值在于确定性，但其难以实现物理世界的建模闭环；我们认为物理 AI 将成为物理世界的 AI 解决方案，从而在 AI 的当前发展阶段成为增量更大的方向，智能驾驶作为最先跑通闭环的代表场景，我们认为应重点关注。**
- **风险提示：**技术迭代不及预期的风险；大模型厂商 ARR 增速放缓的风险；云服务商资本开支不及预期的风险；智能驾驶及机器人商业化落地不及预期的风险。

## 行业走势




## 相关研究

《AI 系列深度 17 期：视觉智能为何引入 Transformer？》

2026-08-06

《AI 系列深度 16 期：语言智能为何从 RNN 转向 Transformer？》

2026-08-06

# 内容目录

# 图表目录

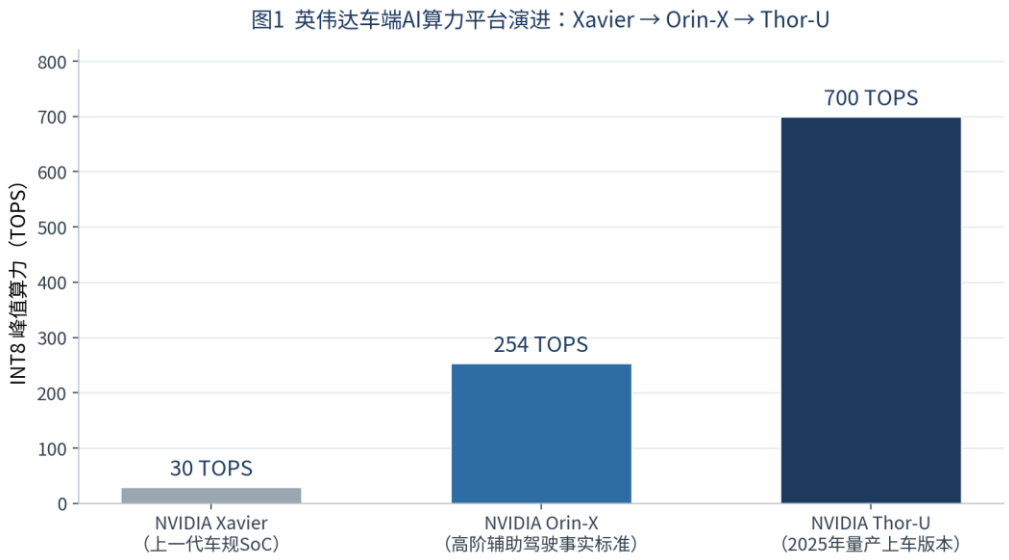
## 1. 缩放定律与车端约束：训练最优不等于部署最优

**经典缩放定律以给定训练算力下的损失最小化为主要目标，“最优”指向训练算力最优前沿，而非部署时延。**Kaplan 等在 2020 年提出语言模型的损失随参数量、数据量与算力呈幂律变化，其结论更偏向扩大参数规模，约七成算力给参数、三成给数据，对应约 1.7 个训练 Token/参数。Hoffmann 等在 2022 年提出 Chinchilla 式算力最优配比，对前述结论作出修正，给出参数与数据应近似等比扩张、约 20 个 Token/参数的“算力最优”配比，并以 70B 参数、1.4 万亿 Token 的 Chinchilla 超越了 280B 参数、约 3000 亿 Token 的 Gopher。

**计入推理与部署成本后，最优点向“更小模型、更多训练数据”偏移，这一变化构成理解车端部署约束的起点。**Sardana 等在 2024 年缩放定律中显式加入推理成本后指出，在高频部署场景下，算力最优点会移向参数更小、训练数据更多的模型，以压低单次推理开销；Llama2、Llama3 分别以 2 万亿、15 万亿 Token 远超 Chinchilla“最优”数据量进行训练，正是这一部署导向的体现。Waymo 在 2025 年 12 月的公开材料中亦采用相近路径：先训练高容量教师模型，再以蒸馏得到紧凑学生模型，并称由此获得更适用于学生模型的缩放规律。

**智能驾驶把部署约束推向更高强度：**在固定车规算力上同时满足硬实时帧率，是大模型上车的首要约束。在车载计算单元上以≥30Hz 运行大型视觉 Transformer 叠加大语言模型，在现有车载算力下实现该目标仍具较高难度。数字 AI 部分所称 Transformer 作为通用底座，主要指云端与通用基础模型口径，并不意味着其在车端低时延场景中仍是部署最优解。与云端算力可弹性扩展不同，车端算力平台相对固定：NVIDIA Orin-X 约 254TOPS INT8 长期被广泛用于高阶辅助驾驶平台，其继任者 Thor 自 2025 年起量产上车，主流落地版本 Thor-U 单颗约 700TOPS，理想 L8、领克 900、小米 YU7、极氪 9X 等约为 Orin-X 的 2.8 倍。在这一算力与时延的双重边界下，云端基座持续扩张与车端毫秒级响应之间形成明显矛盾。

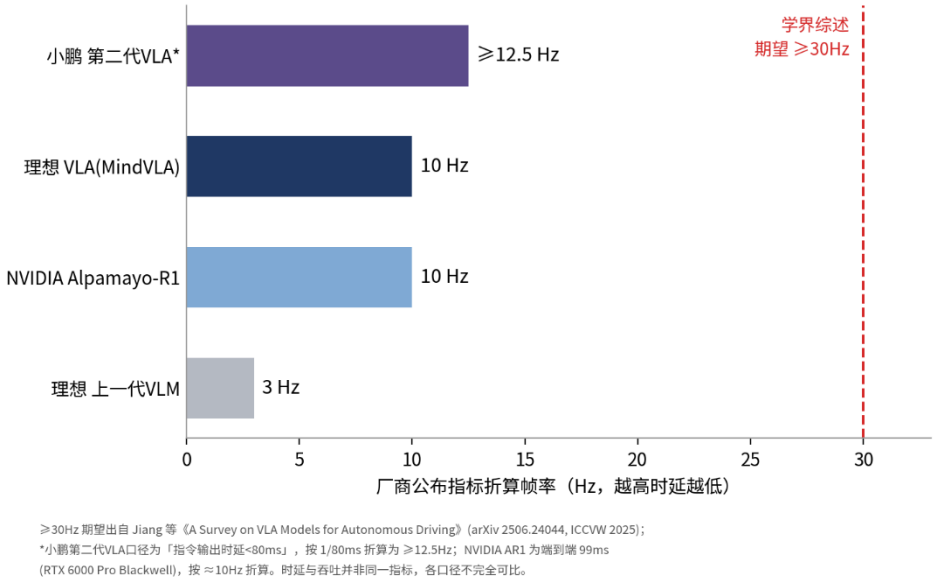
**图1：主流车端 AI 算力平台峰值算力对比**




数据来源：NVIDIA 官网及公开技术解读与量产上车报道，东吴证券研究所整理

**当前量产大模型的实际推理帧率仍低于部分研究提出的 30Hz 参考水平，其中，语言模型和链式推理带来的串行计算是重要时延来源。**理想自动驾驶团队披露，其 VLA 推理帧率约 10Hz，较上一代 VLM 的约 3Hz 提升三倍多，且每一帧都要经过语言模型；NVIDIA 在 Alpamayo-R1 论文中报告端到端推理时延 99ms，约合 10Hz；小鹏公开资料称第二代 VLA 把指令输出时延压缩至 80ms 以内。

**图2：智驾大模型车端推理帧率对比**



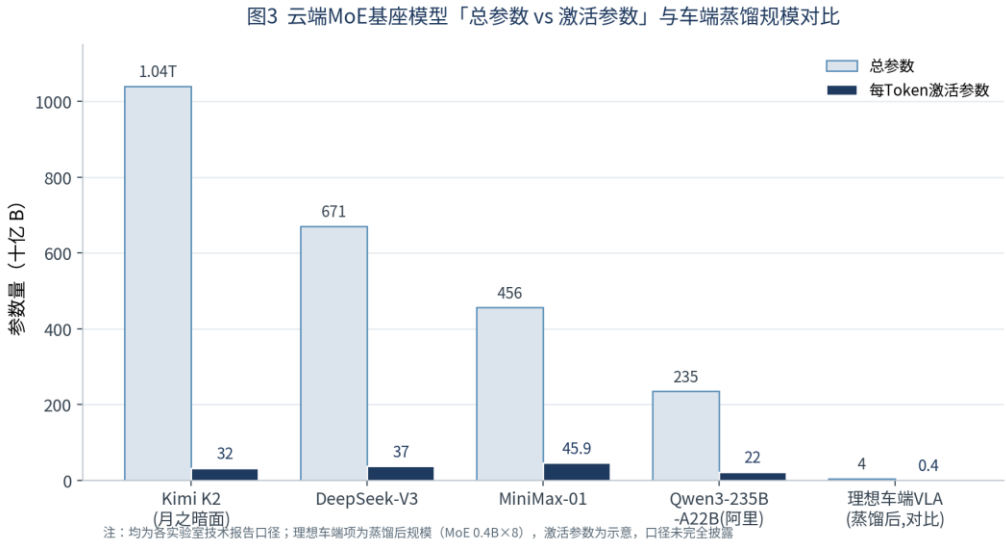

数据来源：《A Survey on Vision-Language-Action Models for Autonomous Driving》，《Alpamayo-R1: Bridging Reasoning and Action Prediction for Generalizable Autonomous Driving in the Long Tail》，理想、小鹏等公司公开资料，东吴证券研究所

## 2. 云端效率优化：稀疏化、注意力优化与 Token 剪枝

**云端压缩与车端压缩的边界在于优化对象不同**：云端侧主要降低模型固有算力与复杂度，重点在算法和架构层面；车端侧则在固定车规算力约束下实现低时延部署，重点在系统和工程层面。云端侧可按压缩对象归为四类：1）降低模型容量与注意力/缓存开销，包括通用基座的 MoE、MLA、线性注意力；2）削减输入侧视觉 Token 冗余，以视觉 Token 剪枝为代表；3）压缩推理过程中的链式思维，以隐式/动态 Token 替代显式长 CoT；4）加速训练与仿真侧的世界模型。视觉 Token 剪枝与 CoT 压缩虽降低模型固有算力、收益最终在车端兑现，但按优化主要发生环节仍归入云端基座侧。

**通用基座侧的效率优化方法以“稀疏激活 + 注意力/缓存压缩 + 低精度 + 投机解码”为主，为车端部署提供可复用的上游技术。**混合专家 MoE 通过每个 Token 只激活一小部分专家，将模型容量与单次推理算力解耦：DeepSeek-V3 以 671B 总参仅激活 37B，256 个路由专家激活 8 个；MiniMax-01 为 456B 总参激活 45.9B，32 专家；阿里 Qwen3-235B-A22B 为 235B 激活 22B；月之暗面 Kimi K2 为约 1.04 万亿总参激活 32B，384 专家激活 8；智谱 GLM-4.5 为 355B 量级 MoE。注意力与缓存侧，DeepSeek 提出的 MLA 以低秩联合压缩把 KV 缓存压至约 70KB/Token，较 GQA 类模型的 192–328KB/Token 降低约 2.7–4.7 倍；MiniMax-01 则以闪电线性注意力把序列复杂度由二次降为线性，其 M1 在生成 10 万 Token 时的 FLOPs 仅为 DeepSeek-R1 的约 25%。此外，DeepSeek-V3 的多 Token 预测 MTP 模块用于投机解码、并以 FP8 进行训练与推理。

**图3：云端 MoE 基座模型总参数与激活参数对比，含车端蒸馏规模**




数据来源：《DeepSeek-V3 Technical Report》，《MiniMax-01: Scaling Foundation Models with Lightning Attention》等其他文献，理想公司公开资料，东吴证券研究所

**物理 AI 尤其是智驾的车端 MoE 则呈现“小而专、按场景/技能路由”的不同形态。**云端 LLM 的 MoE 以总参数百亿至万亿、激活数十亿、专家数以百计为特征，其目标是在数据中心 GPU 上用稀疏性扩大模型容量。智驾车端则相反：受限于 Orin-X、Thor-U 等固定车规算力与内存带宽、以及约 10Hz 的硬实时要求，车端模型规模通常压缩至数十亿参数以内。理想车端 VLA 即由约 32B 云端基座蒸馏为约 4B，并采用为车载平台量身定制的 MoE0.4B×8 结构，在有限算力下平衡多场景适配与推理速度。更关键的差异在路由依据：智驾 MoE 更多按“驾驶场景与技能”而非通用语义 Token 分工，学界 DriveMoE 即设“场景特化视觉 MoE”按驾驶情境动态选取相机视角以压低多视角视觉冗余、并设“技能特化动作 MoE”按驾驶行为激活专用专家以避免对罕见动作的“模式平均”，SAMoE-VLA 等亦走场景自适应路线。当前多数量产智驾 VLA 的规模仍在个位数十亿量级，部分甚至为稠密小模型，MoE 在智驾的主要价值是“专家分工”而非单纯扩容。

**面向驾驶的第一类云端优化方法是视觉 Token 剪枝，其核心动因在于 VLA 输入中的视觉 Token 数量通常高于文本 Token，注意力计算开销因此主要集中在视觉侧。**理想研究团队提出的 LightVLA 以动态查询评估视觉 Token 重要性、并用 Gumbel-Softmax 实现可微分的 Token 选择，论文报告在 LIBERO 基准上 FLOPs 与时延分别下降 59.1%与 38.2%、任务成功率提升 2.6%；LIBERO 为机器人操作基准而非真实驾驶场景，论文将驾驶场景的视觉冗余优化列为后续工作。小鹏与北京大学合作的 FastDriveVLA 专为驾驶设计，采用重建式前景 Token 剪枝器 ReconPruner；该方法借鉴 MAE 像素重建机制保留前景信息。论文称，当视觉 Token 由 3249 降至 812 时，计算量约减少 7.5 倍，CUDA 时延上预填充与解码分别减少约 3.7 倍与 1.3 倍。

**面向驾驶的第二类云端优化方法是以动态或隐式 Token 替代冗长思维链，缩短链式推理带来的串行时延。**其核心是先预测紧凑的世界动态 Token、再生成驾驶动作，以减少文本思维链表达模糊和视觉思维链生成完整未来帧带来的冗余。香港中文大学深圳校区与小鹏汽车联合提出的 FutureX 在隐空间内做思维链：以 Auto-think 开关判断场景难度，复杂场景进入“思考”模式，由隐式世界模型推演未来表征；简单场景直接采用快速前向路径。论文报告在 NAVSIM 上使 TransFuser 的 PDMS 提升 6.2。小鹏自身则在第二代 VLA 中称把视觉推理思维链效率提升约 32 倍，以保障车端实时运行。

**世界模型/仿真侧的推理加速可单独归类，其压缩对象是训练与验证用的生成式视频世界模型，而非行车实时回路。**小鹏于 2026 年 4 月底发布的 X-Cache 是面向其 X-World 世界模型的无需额外训练的加速方法：基于相邻画面段在去噪网络同层、同步上的中间特征高度相似，跨分块复用中间结果、跳过整层计算，官方称块跳过率达 71%、实测推理加速 2.6–2.7 倍且画质损失较小。

**表1：云端/车端时延压缩代表性方法一览**

| 方法 | 机构，出处主体， | 层面 | 核心机制 | 公布收益 |
|---|---|---|---|---|
| LightVLA | 理想 | 云端·VLA | 可微分视觉 Token 剪枝，动态查询+Gumbel-Softmax， | FLOPs−59.1%、时延−38.2%、LIBERO 成功率+2.6%，机器人基准， |
| FastDriveVLA | 小鹏+北京大学 | 云端·VLA | 重建式前景 Token 剪枝 ReconPruner | Token3249→812 时 FLOPs 约÷7.5；预填充÷3.7、解码÷1.3 |
| X-Cache | 小鹏，2026-04-30 发布， | 云端·世界模型仿真 | 无需额外训练的跨分块特征缓存复用 | 块跳过率 71%、实测加速 2.6~2.7×，X-World 侧， |
| FutureX | 香港中文大学深圳校区+小鹏汽车 | 云端·E2E 规划 | 隐式 CoT 世界模型+Auto-think 双模 | TransFuser 在 NAVSIM 上 PDMS+6.2 |
| DynVLA | 华为引望 | 云端·VLA | 动态 Token Dynamics CoT 替代冗长 CoT | 以紧凑动态 Token 替代冗余推理 |
| DriveMoE | 学界 | 云端·VLA | 场景特化视觉 MoE+技能特化动作 MoE | 压低多视角视觉冗余、避免罕见动作模式平均；论文报告在 Bench2Drive 闭环基准上达到 SOTA； |
| MLA+MoE+MTP+FP8 | 深度求索 DeepSeek-V3 | 云端·通用基座 | KV 低秩压缩/稀疏激活/投机解码/低精度 | KV Cache 约 70KB/Token，较典型 GQA 模型降低约 2.7–4.7 倍；671B 总参、激活 37B |
| 闪电线性注意力+MoE | MiniMax-01/M1 | 云端·通用基座 | 线性注意力将复杂度由二次降为线性 | M1 生成 10 万 Token 的 FLOPs 约为 R1 的 25% |

数据来源：《DynVLA: Learning World Dynamics for Action Reasoning in Autonomous Driving》，《FutureX: Enhance End-to-End Autonomous Driving via Latent Chain-of-Thought World Model》等其他文献，小鹏、理想等公司公开资料，东吴证券研究所

## 3. 车端部署优化：蒸馏上车与快慢双系统

**车端侧压缩主要围绕三条路径展开，其共同约束是在既定模型能力目标和车规算力预算下完成部署，核心问题是如何在固定车规芯片上实现低时延、低内存占用和稳定部署。**三条路径依次为：1）把云端大模型蒸馏为可上车的小模型，实现云端能力向车端迁移；2）以快慢双系统在运行时分离快速响应与复杂推理，优化车端运行架构；3）通过量化、投机解码、采样步数压缩、编译优化等底层推理工程降低端到端时延。

**车端部署优化的第一类路径是“云端大模型蒸馏上车”，即云端训练高容量教师、车端运行蒸馏后的学生模型。**Waymo 在 2025 年 12 月公开材料中描述：以 Waymo Foundation Model 适配出 Driver/Simulator/Critic 三类大教师模型，因其规模过大无法在车端实时运行，需安全蒸馏为更小的学生模型；车端架构刻意镜像基座结构，并设独立的车端校验层对生成轨迹做验证。小鹏自动驾驶团队称其面向 L4 的 720 亿参数基座约相当于主流 VLA 的 35 倍，经预训练与强化学习后蒸馏为较小模型部署车端，并计划向 MAX 车主推送蒸馏版第二代 VLA；其云端算力约 10EFLOPS、利用率长期保持 90%以上。理想则在云端以 13EFLOPS，约 3EFLOPS 推理、10EFLOPS 训练，构建约 32B 多模态基座，经强化学习与蒸馏压缩为约 4B 的车端 VLA 部署于 Thor 芯片，采用 INT8/FP8 混合精

度、在 10Hz 帧率下完成视觉语言交互。

**表2：代表性厂商的“云端基座—车端部署”路径**

| 主体 | 云端基座 | 车端部署 | 蒸馏与快慢架构 |
|---|---|---|---|
| 理想 | 约 32B 多模态基座；云端 13EFLOPS | 约 4B，MoE0.4×8；Thor；INT8/FP8；10Hz | 基座→RL+蒸馏→车端 VLA；早期 E2E+VLM 双系统，双 Orin-X→统一 VLA |
| 小鹏 | 720 亿参数物理世界基座；云端约 10EFLOPS | 蒸馏版第二代 VLA 推送 MAX 车型/自研图灵芯片 | 基座蒸馏上车；CoT 效率约 32×保实时；编译效率约 12× |
| Waymo | Waymo Foundation Model，统一 Driver/Simulator/Critic | 学生模型车端实时运行+独立轨迹校验层 | 教师→学生蒸馏；Think Fast/Slow：传感融合编码器+Driving VLM Gemini+World Decoder |
| 华为/引望 | 诺亚前瞻研究+引望量产；DynVLA 以 EMU3 8B 为骨干 | 面向量产落地 | 以动态 Token 替代冗长 CoT 降低推理负担 |

数据来源：《DynVLA: Learning World Dynamics for Action Reasoning in Autonomous Driving》，Waymo、小鹏等公司公开资料，东吴证券研究所

**车端部署优化的第二类路径是快慢双系统，将“快速直觉反应”与“慢速复杂推理”在架构层面分工，以兼顾时延与长尾能力。**理想较早采用基于卡尼曼“思考快与慢”的端到端 + VLM 双系统：系统 1 为端到端模型，直接由传感器输入输出轨迹，运行于一颗 Orin-X，处理约 95%常规场景；系统 2 为 22 亿参数 VLM，以思维链做复杂语义推理后将语义判断结果反馈至系统 1，运行于另一颗 Orin-X。这一双系统后续被统一为单一 VLA 4B MoE。Waymo 在 2025 年 12 月材料中明确其基座采用 Think Fast/Think Slow：传感融合编码器负责快速反应，融合相机、激光雷达和毫米波；Driving VLM 基于 Gemini 负责复杂语义推理，能在“前方车辆起火”等罕见语义场景给出绕行信号；二者汇入 World Decoder 输出预测、地图、轨迹与校验信号。Waymo 早期基于 Gemini 的 EMMA 亦显示，思维链推理使端到端规划提升 6.7%。

**车端部署优化的第三类路径是推理工程手段，覆盖量化、并行/投机解码、采样步数压缩与编译优化。**量化方面，理想以底层推理引擎在 Orin-X 上以 INT4 量化运行 VLM、在 Thor 上以 INT8/FP8 运行 VLA；小鹏通过芯片-编译器-模型联合优化与自研“图灵结构”模型，称把基座编译效率提升约 12 倍。解码与采样方面，理想 MindVLA 采用 Action Token 并行解码、思维链使用小词表叠加投机推理、并以常微分方程 ODE 采样器与扩散流匹配把生成轨迹的去噪步数由约 10 步压缩至 2 步，以在 10Hz 帧率下改善响应时延。这些手段大多源自通用大模型的推理优化方法，如投机解码、低精度量化，经车规适配后用于压低端到端时延。

**产业与投资侧，云端训练算力多由车企与云厂商合作构建，头部基座实验室构成上游技术供给。**国内主机厂在算力中心建设上多与腾讯云、阿里云、字节火山引擎、金山云等云服务商合作，以缓解高算力 AI 芯片的获取约束。上游基座侧，DeepSeek、MiniMax、

阿里 Qwen、月之暗面 Kimi、智谱 GLM、字节豆包、腾讯混元、百度文心等以 MoE 为主的稀疏化路线，持续向产业输出可被驾驶模型借鉴的效率技术，包括 MoE、MLA、线性注意力、投机解码等；小米 LaST-VLA 隐式时空思维链、元戎启行等亦在 VLA 路线上推进。

## 4. 结论：云端探上限，车端保实时

**从已披露路线看，云端基座与车端部署正逐步形成“云端探上限、车端保实时”的协同分工。**云端以 MoE 稀疏化、注意力/KV 压缩、低精度与投机解码持续扩展能力边界，并以视觉 Token 剪枝、隐式/动态 Token 替代冗长思维链来压缩驾驶 VLA 的串行开销；车端则以教师-学生蒸馏获取紧凑学生模型、以快慢双系统分置直觉与推理、并以量化与解码工程把端到端时延压入 10Hz 量级。

## 5. 风险提示

**技术迭代不及预期的风险：**数字 AI、物理 AI 技术进展低于预期，任务或应用场景拓宽速度低于预期。

**大模型厂商 ARR 增速放缓的风险：**若客户增长、续费率或客单价不及预期，模型厂商 ARR 增速可能放缓。

**云服务商资本开支不及预期的风险：**若自由现金流承压，或 AI 算力需求及投资回报低于预期，云服务商可能下调资本开支。

**智能驾驶及机器人商业化落地不及预期的风险：**若产品可靠性、成本下降或用户需求低于预期，相关产品商业化进程可能放缓。

东吴证券 SOOCHOW SECURITIES

## 免责声明

东吴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。

本研究报告仅供东吴证券股份有限公司（以下简称“本公司”）的客户使用。本公司不会因接收人收到本报告而视其为客户。在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议，本公司及作者不对任何人因使用本报告中的内容所导致的任何后果负任何责任。任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。

在法律许可的情况下，东吴证券及其所属关联机构可能会持有报告中提到的公司所发行的证券并进行交易，还可能为这些公司提供投资银行服务或其他服务。

市场有风险，投资需谨慎。本报告是基于本公司分析师认为可靠且已公开的信息，本公司力求但不保证这些信息的准确性和完整性，也不保证文中观点或陈述不会发生任何变更，在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。



## 东吴证券投资评级标准

投资评级基于分析师对报告发布日后 6 至 12 个月内行业或公司回报潜力相对基准表现的预期（A 股市场基准为沪深 300 指数，香港市场基准为恒生指数，美国市场基准为标普 500 指数，新三板基准指数为三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的），北交所基准指数为北证 50 指数），具体如下：

公司投资评级：

买入：预期未来 6 个月个股涨跌幅相对基准在 15%以上；

增持：预期未来 6 个月个股涨跌幅相对基准介于 5%与 15%之间；

中性：预期未来 6 个月个股涨跌幅相对基准介于-5%与 5%之间；

减持：预期未来 6 个月个股涨跌幅相对基准介于-15%与-5%之间；

卖出：预期未来 6 个月个股涨跌幅相对基准在-15%以下。

行业投资评级：

增持： 预期未来 6 个月内，行业指数相对强于基准 5%以上；

中性： 预期未来 6 个月内，行业指数相对基准-5%与 5%；

减持： 预期未来 6 个月内，行业指数相对弱于基准 5%以上。

我们在此提醒您，不同证券研究机构采用不同的评级术语及评级标准。我们采用的是相对评级体系，表示投资的相对比重建议。投资者买入或者卖出证券的决定应当充分考虑自身特定状况，如具体投资目的、财务状况以及特定需求等，并完整理解和使用本报告内容，不应视本报告为做出投资决策的唯一因素。

东吴证券研究所
苏州工业园区星阳街 5 号
邮政编码：215021
传真：（0512）62938527
公司网址： http://www.dwzq.com.cn