

AI 系列深度 12 期：从模块化到端到端 增持（维持）

2026 年 08 月 04 日

证券分析师 黄细里

执业证书：S0600520010001

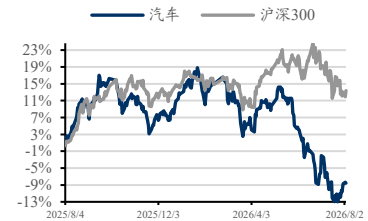
021-60199793

huangxl@dwzq.com.cn

投资要点

- **端到端的核心**，是以原始传感器输入直接映射规划或控制输出，并通过整体可微、全局优化的方式训练系统。该概念最早来自网络工程语境，进入机器学习后逐步转化为输入端到输出端整体训练的方法论；在自动驾驶中，ALVINN、DARPA DAVE、NVIDIA DAVE-2、UniAD 和特斯拉 FSD v12 共同构成其从学术探索到产业落地的演进脉络。我们将端到端理解为训练方法、系统架构和技术思想三重属性的组合。
- **端到端与模块化流水线的差异**，在于是否将感知、预测、规划和控制拆成多个人工定义模块。模块化系统可解释、可调试，但信息容易在模块边界损失，且规则维护成本较高；端到端系统可让模型直接围绕最终任务优化，信息传递更完整，但可解释性、可控性和安全验证难度更高。2024 年以来，一段式与二段式成为中国智驾行业讨论焦点。
- **端到端并不等于完全取消规则**，更准确的表述是替代人工规则在主决策回路中的主导地位。特斯拉 FSD v12 以神经网络替代大量手写 C++ 规则，是该趋势的重要标志；但规则仍会保留在安全兜底、形式化校验、冗余备份和合规约束中。由此看，端到端普及是主决策范式从人工规则转向数据驱动模型，而非规则系统消失。
- **从应用进展看**，自动驾驶由特斯拉率先量产推动，中国厂商 2024 年集中跟进，并在 2025 年前后进一步向 VLA 和世界模型演进；机器人领域自 RT-2 起即以 VLA 形态采用端到端；数字 AI 中端到端已是默认范式，通常不再作为单独技术标签。例外和分歧仍存在，如 Mobileye 复合 AI 系统、Zoox 模块化路线等。
- **端到端、VLA 与世界模型不是替代关系，而是处于不同能力层级和功能维度**。VLA 可视作端到端在多模态理解与动作生成方向上的延伸；世界模型关注环境状态预测和场景生成，可作为云端数据引擎、仿真训练场或车端模型组件。产业实践中，三者更可能形成组合方案，而非线性替代关系。
- **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

- 《智能汽车主线周报：曹操出行开放 Robotaxi 无安全员测试，看好智能化》
2026-08-03
- 《AI 系列深度 11 期：大模型如何进行模仿与强化学习？》
2026-08-03

内容目录

1. 端到端：从输入到动作的整体优化框架4

1.1. 概念溯源：从网络工程术语到机器学习范式4

1.2. 定义梳理：整体可微、全局优化是公认内核4

1.3. 发展路径：学术奠基、产品催化与产业跟进5

1.4. 属性辨析：架构、方法论还是思想6

1.5. 技术原理：端到端与模块化对比6

2. 一段式与二段式：端到端替代主决策回路规则7

2.1. 一段式与二段式的界定与阵营7

2.2. 核心议题：端到端是否是对"规则算法时代"的替代8

3. 端到端应用：自动驾驶率先量产，机器人以 VLA 承接.....9

3.1. 自动驾驶：特斯拉率先，中国厂商 2024 年集中跟进9

3.2. 机器人：自 RT-2 起即以 VLA 形态采用端到端10

3.3. 数字 AI：端到端是默认范式，通常不再单独标注10

4. 端到端与 VLA：演进与包含，而非二选一.....11

4.1. VLA 的概念起源与技术原理11

4.2. 关系辨析：技术原理与产业口径的双重验证11

4.3. 从快慢系统看：VLA 强化端到端的推理与泛化11

5. 端到端与世界模型：一个出动作，一个预测世界12

6. 风险提示12

图表目录

表 1: “端到端”代表性定义对照 5

表 2: 端到端发展关键时间线..... 6

表 3: 一段式与二段式端到端对比..... 8

表 4: 主要玩家端到端采用时间一览..... 10

1. 端到端：从输入到动作的整体优化框架

在表征、训练范式与 LLM/VLM/VLA 之后，端到端是理解智能驾驶和机器人落地路径的关键工程范式。其核心在于以可整体训练的神经网络压缩传统模块化流水线，使系统从原始输入直接生成规划或控制输出。端到端强调训练链路和系统架构的整体优化，VLA 强调视觉、语言与动作的统一建模，世界模型则强调对环境动态的预测与生成。

1.1. 概念溯源：从网络工程术语到机器学习范式

“端到端”的词形最早成型于计算机网络领域，与机器学习语境并不相同。Saltzer、Reed 与 Clark 在 1981 年会议论文和 1984 年 ACM TOCS 修订版中提出端到端论证，核心含义是某些功能只有依靠通信系统两端应用才能完整实现，不宜由网络中间层承担。该语境下的端到端强调功能位置，机器学习语境下的端到端强调输入到输出的整体训练，两者仅共享减少中间环节的思想隐喻，不应直接混同。

机器学习语境下，端到端驾驶的早期代表是 1988 年 CMU 的 ALVINN，但“端到端”一词属于后世追认。Pomerleau 在 NIPS 1988 发表的 ALVINN 论文提出三层反向传播网络，以摄像头图像和激光测距数据为输入，直接输出车辆沿道路行驶方向。ALVINN 原文并未使用 end-to-end 表述，NVIDIA 在 2016 年官方博客中将其追认为端到端训练神经网络驾驶汽车的早期概念验证。

“端到端学习”作为驾驶任务标题级术语，可考至 2004 至 2005 年。LeCun 等在 1998 年提出图变换网络，使多模块系统能够通过梯度方法全局训练并最小化整体性能指标，为端到端方法论奠定基础。2004 年 DARPA 资助的 DAVE 项目最终报告以 Autonomous Off-Road Vehicle Control Using End-to-End Learning 为题，2005 年 LeCun 等在 NIPS 论文中进一步明确将原始输入图像映射为转向角。

2006 至 2015 年，端到端范式在语音、翻译和强化学习等领域加速落地。Graves 等在 2006 年提出 CTC，使语音识别可在无需人工对齐的情况下训练；Sutskever 等在 2014 年提出 seq2seq，将其表述为通用端到端序列学习方法；百度 Deep Speech 同年以端到端语音识别为标题；DeepMind DQN 在 2015 年以屏幕像素为输入、直接输出游戏动作。至 2016 年前后，端到端已成为深度学习的通用范式。

1.2. 定义梳理：整体可微、全局优化是公认内核

学界与产业界对端到端的定义表述各有侧重，核心要素一致。NVIDIA 在 2016 年给出操作性定义：“与把问题显式分解为车道线检测、路径规划和控制的方法相反，我们的端到端系统同时优化所有处理步骤”。学界目前最常引用的定义来自 IEEE TPAMI 综述：端到端框架“利用原始传感器输入生成车辆运动规划，而非专注于检测、运动预测等单项任务；与模块化流水线相比，端到端系统受益于感知与规划的特征联合优化”。中国业界则更强调，端到端能够减少模块间的信息损耗，并实现感知、决策与控制的全局联合优

化。

表1：“端到端”代表性定义对照

定义主体	时间	定义核心表述
Saltzer/Reed/Clark，网络领域，词源对照	1981/1984	功能只有依靠通信两端的应用才能完整正确实现，不应在中间层实现，“端到端论证”，与机器学习语境含义不同，
LeCun 等 Proc. IEEE	1998	多模块系统用基于梯度的方法全局训练，以最小化整体性能指标 GTN
LeCun 等 NIPS	2005	系统从端到端训练，将原始输入图像映射为转向角
NVIDIA，Bojarski 等	2016	端到端系统同时优化所有处理步骤，而非显式分解为检测、规划、控制
Wayve AV2.0	2021 起	用单一深度神经网络在大量人类驾驶演示上联合学习场景表征与运动规划，替代手工规则栈与高精地图
Chen 等	2023/2024	利用原始传感器输入生成运动规划的算法框架，受益于感知与规划的特征联合优化
中国业界	2024	减少模块间的信息损耗，并实现感知、决策与控制的全局联合优化

数据来源：《End-to-End Arguments in System Design》，《End-to-end Autonomous Driving: Challenges and Frontiers》等其他文献，东吴证券研究所

综合上述定义，端到端系统通常具备三项核心特征：一是输入端直接接收原始或近原始传感器数据；二是输出端直接生成运动规划或控制指令；三是主体处理链路具备可微性，可围绕最终任务目标进行整体优化。

1.3. 发展路径：学术奠基、产品催化与产业跟进

2016 年 NVIDIA DAVE-2 使端到端驾驶路线重新受到关注。NVIDIA 在 2016 年 4 月发布 End to End Learning for Self-Driving Cars，DAVE-2 系统使用 9 层 CNN 和约 72 小时驾驶数据训练，仅以人类转向角为监督信号，并在未显式训练道路轮廓检测的情况下实现演示级自主行驶。2017 年 Wayve 在英国剑桥成立，提出区别于规则栈和高精地图 AV1.0 的自动驾驶技术路线，推动端到端学习成为自动驾驶路线讨论的重要变量。

2023 年 UniAD 获 CVPR 最佳论文，端到端登上学术制高点。上海人工智能实验室 OpenDriveLab 与武汉大学、商汤研究院合作的 UniAD（《Planning-oriented Autonomous Driving》）将感知、预测、规划全栈任务统一进一个以规划为导向、全局可微的网络，获 CVPR 2023 最佳论文奖，属于保留中间任务输出的“模块化端到端”路线。

特斯拉 FSD v12 量产推送是产业侧关键催化。2023 年 8 月马斯克直播试驾 FSD v12 早期版本，提出“光子进、控制出”的表述，并称控制栈由大量 C++规则代码转向神经网络；2023 年 11 月 v12 推送员工，12 月下旬开始向用户推送，2024 年 3 至 4 月 v12.3 在北美大范围铺开并更名 FSD Supervised。特斯拉官方版本说明显示，FSD v12 将城市道路驾驶栈升级为单一端到端神经网络，由数百万视频片段训练而成，取代逾 30 万行显式 C++代码。

中国市场自 2024 年起集中跟进，2025 年前后讨论重心进一步迁移至 VLA 与世界模型。小鹏在 2024 年 5 月宣布国内首个量产上车的端到端大模型；2024 年下半年，行

业公开分化为一段式和两段式两个阵营；2025 年起，理想 MindVLA、小鹏第二代 VLA、元戎启行 IO 2.0、华为 WEWA 等相继发布，产业讨论从端到端本身转向大模型化、多模态化和世界模型化。

表2：端到端发展关键时间线

时间	事件	出处
1981/1984	Saltzer 等发表《End-to-End Arguments in System Design》，网络领域“端到端论证”，词源对照	ACM TOCS
1988	CMU ALVINN: 三层 BP 网络从摄像头/测距输入直接输出行驶方向，原文未用“端到端”一词	NIPS 1988
1998	LeCun 等 GTN: 多模块系统全局梯度训练	Proc. IEEE
2004—2005	DARPA DAVE 项目报告与 LeCun 等 NIPS 论文，“端到端学习”首次标题级用于驾驶任务	Net-Scale 报告；NIPS 2005
2006—2015	CTC，语音、seq2seq，翻译、Deep Speech、DQN，像素到动作，相继采用端到端	ICML/NIPS/Nature
2016	NVIDIA DAVE-2《End to End Learning for Self-Driving Cars》	arXiv:1604.07316
2017	Wayve 成立，主张端到端学习驱动的自动驾驶技术路线	Wayve 官网
2022.12	UniAD 论文提交 arXiv	arXiv:2212.10156
2023.6	UniAD 获 CVPR 2023 最佳论文	CVPR 官方
2023.8	马斯克直播 FSD V12，“光子进、控制出”	直播及媒体报道
2023.12 — 2024.03	FSD v12 向用户推送并大范围铺开，更名 FSD Supervised	特斯拉官方版本说明
2024.5	小鹏宣布“国内首个量产上车的端到端大模型”	小鹏汽车官网
2024.12	中国业界“一段式 vs 两段式”路线分化公开化	每日经济新闻等
2025	理想 VLA、小鹏二代 VLA、华为 WEWA 等发布，讨论重心迁移	各公司官方发布

数据来源：《End to End Learning for Self-Driving Cars》，《Planning-oriented Autonomous Driving》等其他文献，特斯拉、小鹏等公司官方网站，东吴证券研究所

1.4. 属性辨析：架构、方法论还是思想

各方对端到端的属性表述并不一致。学术综述更多称其为端到端算法框架，NVIDIA 将其表述为方法或系统，特斯拉在财报电话会上将 FSD v12 称为彻底的架构重写和端到端人工智能，Wayve 则将其上升为区别于 AV1.0 的代际范式。行业讨论中，端到端还常与大模型混用，前者强调梯度可传导和全局优化，后者强调参数规模、数据规模与涌现能力，两者并非同一维度。

我们认为，端到端在不同语境下承担三重属性。作为训练方法，端到端指整体可微、梯度全局回传的联合优化方式；作为系统架构，端到端指以单一或级联神经网络替代感知、预测、规划模块化流水线的工程形态；作为技术思想，端到端强调让数据学习中间表征，而非由人工先验显式定义全部接口。三重属性分别对应训练机制、系统形态和设计哲学，是行业概念争论的主要来源。

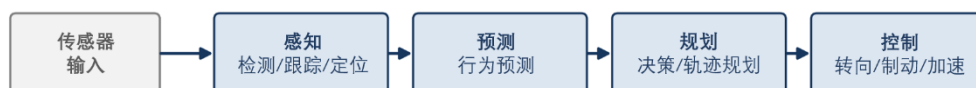
1.5. 技术原理：端到端与模块化对比

模块化流水线将驾驶任务显式拆分，并以人工接口串联各环节。传统自动驾驶栈通常将感知、预测、规划、控制分别开发，通过 3D 检测框、预测轨迹等结构化中间产物

传递信息。该路线优势在于分工清晰、单点故障可定位；约束在于各模块往往优化自身指标，人工设定的中间指标并不一定对应系统整体目标，且信息在接口处会发生压缩和误差累积。NVIDIA 在 2016 年已指出，人为选定的中间指标不自动保证系统整体性能最大化。

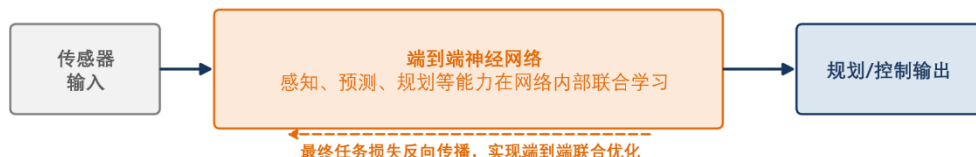
图1：模块化架构与端到端架构的对比示意

模块化架构：通过人工定义中间接口串联不同功能模块



各模块独立设计与训练，模块之间依赖人工定义接口，典型代表包括 NVIDIA 2016 年早期方案。

端到端架构：由单一神经网络直接学习从传感器输入到规划/控制输出的映射



数据来源：《End to End Learning for Self-Driving Cars》，《End-to-end Autonomous Driving: Challenges and Frontiers》，东吴证券研究所绘制

端到端的技术内核是全局可微与数据驱动。端到端模型以最终驾驶任务的损失函数统一驱动参数更新，训练梯度由输出端回传至输入端，内部表征由数据自动习得。该路线的能力与数据规模强相关：NVIDIA DAVE-2 在 2016 年使用约 72 小时数据实现演示级自主；特斯拉 v12 基于超百万车队的数百万视频片段训练；小鹏公开口径提到折算 10 亿公里以上视频训练和较快模型迭代；华为 ADS 3.0 公开口径提到训练数据每日新增 3000 万公里、模型 5 天迭代一次。量产车队数据闭环是端到端在产业侧复兴的重要基础。

端到端的局限也较为清晰。NVIDIA 在 2016 年即指出，端到端训练后难以清晰区分网络内部哪些部分负责特征提取、哪些部分负责控制，仍需提升鲁棒性和内部处理过程可视化；TPAMI 综述将多模态融合、可解释性、因果混淆、鲁棒性与世界模型列为端到端关键挑战。UniAD 采用模块化端到端路线，保留可解释中间任务输出并进行全局可微联合训练，本质上是对高不透明度方案的一种折中。

2. 一段式与二段式：端到端替代主决策回路规则

一段式与两段式之争，核心在于端到端是否正在替代规则算法时代的主决策逻辑。我们将讨论边界限定在主决策回路，而非安全兜底、形式化校验与冗余系统的完全取消；在这一限定下，端到端替代的是人工规则在驾驶策略生成中的主导地位。

2.1. 一段式与二段式的界定与阵营

两类路线的差异在于信息处理环节数量，以及中间表征是否显式保留。两段式端到端通常将感知模型与规控模型级联，第一阶段把原始感知数据转化为语义地图、轨迹点、可行驶区域等中间表征，第二阶段再生成轨迹或控制指令；一段式则将感知、预测决策和规划纳入同一模型，由传感器输入直接输出车辆行驶轨迹。一段式的信息传递更完整、延迟更低，但可解释性和调试难度相应上升。

一段式与两段式的取舍，可概括为“能力上限”与“工程可控性”的权衡。Momenta CEO 曹旭东支持一段式的代表性论证，是未定义物体问题：传统两段式感知更依赖人类预先定义目标类别，对山石、水坑等长尾障碍物可能缺少稳定输出，导致规划环节缺乏有效输入；一段式通过驾驶行为监督学习，有机会直接形成绕行策略。百度王亮则从工程角度强调，两段式保留中间结果，有利于系统透明度、问题定位和决策逻辑调试。由此看，一段式更强调数据驱动上限，两段式更强调工程验证和可控性。

表3：一段式与二段式端到端对比

维度	两段式，两段式端到端，	一段式端到端
信息处理环节	两个：感知→中间表征，中间表征→轨迹/控制	一个：传感器输入直接输出轨迹
中间显式输出	保留，语义地图、轨迹点、可行驶区域等，	不保留显式中间产物，联合训练
信息传递	经中间表征，存在一定信息损失	更接近信息无损
可解释性/调试	模块化，易诊断与定位	可解释性较弱，调试与安全验证较难
代表阵营（2024 末）	华为、百度 Apollo、小鹏、卓驭、极氪	特斯拉、商汤、Momenta
特点概括	工程下限稳，透明度高	上限高、下限低，可学“未定义物体”

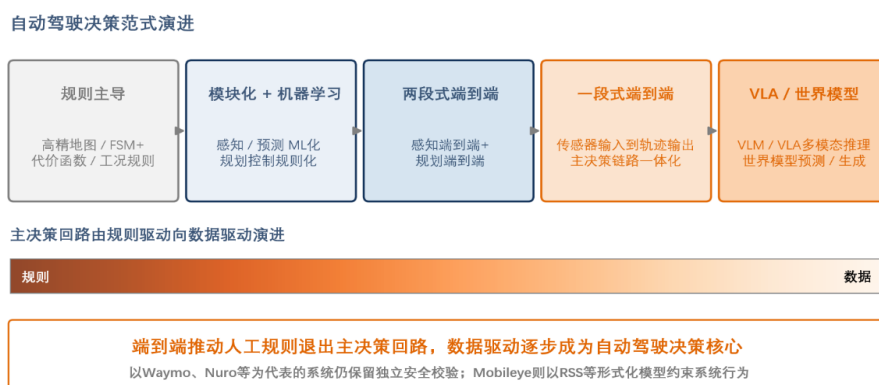
数据来源：Momenta、百度等公开信息，东吴证券研究所整理

2.2. 核心议题：端到端是否是对"规则算法时代"的替代

从决策范式演进看，端到端对应人工规则在主决策回路中的退场。传统规控长期依赖有限状态机、代价函数和大量工况规则；理想汽车将智能驾驶演进概括为从规则算法与高精地图，到端到端+VLM，再到 VLA，并给出规则系统约 1 秒、端到端约 7 秒的推理时长对比。毫末智行在 DriveGPT 发布中亦提出，车端认知模块需要从手工规则向 AI 模型化演进。综合公开材料，决策范式可概括为“规则主导—模块化机器学习—两段式端到端—一段式端到端—VLA/世界模型”，人工规则在主链路中的权重逐步下降。

图2：自动驾驶决策范式演进：从人工规则主导到数据驱动端到端

图3：



数据来源：理想汽车、毫末智行等，东吴证券研究所整理

支持“替代”命题的核心证据，来自特斯拉以神经网络替代大量规则代码。马斯克在2023年8月直播中称，FSD v11 控制栈包含超过 30 万行 C++ 代码，而 v12 基本不再依赖这类显式规则；特斯拉官方版本说明也显示，FSD v12 由单一端到端神经网络取代逾 30 万行显式 C++ 代码。其底层逻辑在于：长尾工况近乎无穷，人工规则难以穷举，而数据驱动端到端可从海量人类驾驶示范中学习策略。未定义物体等案例进一步说明，规则或显式定义之外的场景，是人工规则体系的主要约束。

制约“替代”命题的证据：规则在安全兜底、验证与冗余层面仍不可或缺。“替代”并非彻底。在安全攸关层面，主流 L4 玩家与部分供应商仍保留规则或形式化校验：Waymo 采用“分层安全”，架构层+行为层+运营层，强调“既稳健又冗余”，其行为层显式评估“遵守适用驾驶规则”的能力，并由第三方独立审计——其 2024 年 10 月发布的端到端模型 EMMA 被官方明确定位为“研究原型、面临实际部署挑战”，而非可直接商用的架构（Waymo, 2024/2025）。

端到端替代了人工规则在主决策回路中的主导地位，这一替代在特斯拉 FSD v12 等案例中已经发生，并构成端到端普及的核心驱动力；但规则并未完全消失，而是更多转向安全兜底、形式化校验、冗余备份和合规约束。

3. 端到端应用：自动驾驶率先量产，机器人以 VLA 承接

从应用领域看，端到端首先在自动驾驶侧形成量产催化，随后在机器人领域以 VLA 形态延伸，并在数字 AI 中成为默认建模方式。

3.1. 自动驾驶：特斯拉率先，中国厂商 2024 年集中跟进

特斯拉以 FSD v12 率先推动端到端量产。2023 年 8 月马斯克直播 FSD v12 并提出“光子进、控制出”，2023 年 12 月开始向用户推送，2024 年一季度大范围铺开，成为产业侧端到端路线的重要催化。后续 v12.5 在 2024 年下半年将高速场景并入端到端网络，v13 在 2024 年底继续迭代。

中国厂商在 2024 年集中跟进，并在 2025 年进一步向 VLA 与世界模型演进。小鹏 2024 年 5 月宣布国内首个量产上车的端到端大模型；2024 年下半年至年底，华为 ADS 3.0、理想端到端+VLM、蔚来基于世界模型的端到端、智己联合 Momenta 方案等密集上车；供应商侧 Momenta、商汤绝影、元戎启行、百度 Apollo、地平线、卓驭等在 2024 年陆续量产或发布端到端方案。进入 2025 年后，理想 VLA、小鹏第二代 VLA、华为 WEWA、小米 HAD、比亚迪天神之眼等相继落地，产业讨论重心从端到端迁移至 VLA 与世界模型。

例外路线同样值得关注。Mobileye 两位创始人在 2024 年 5 月撰文质疑仅押注数据管线的单一端到端路线，主推复合 AI 系统；Zoox 官方技术口径仍为感知、预测、规划的模块化架构，并未发布单一端到端方案；通用汽车在 2024 年 12 月停止注资 Cruise robotaxi 业务。即便采用端到端的玩家，其含义也并不统一：既有特斯拉式单一网络，也有华为 GOD+PDP 式分段架构，还有 Waymo、Nuro 式端到端模型叠加独立安全冗余。

表4：主要玩家端到端采用时间一览

公司	首发/量产时间	路线与当前演进
特斯拉	2023.12	FSD v12/v13 单一端到端神经网络，纯视觉
小鹏	2024.5	“国内首个量产端到端大模型”→2025.11 第二代 VLA
华为	2024.4	ADS 3.0 两段式 GOD+PDP→2025.4 ADS 4 WEWA
理想	2024.10	端到端+VLM 双系统→2025.9 VLA 司机大模型
蔚来	2024.7	世界模型 NWM+基于世界模型的端到端→2025 NWM 2.0
Momenta	2023—2024	两段式→一段式端到端→2025 R6 强化学习大模型
商汤绝影	2023	UniAD→2025.2 R-UniAD
元戎启行	2024	一段式无图端到端 DeepRoute IO→2025.8 转 VLA
百度 Apollo	2024.5	感知/规划大模型联合训练，两段式，
卓驭/极氪	2024—2025	两段式→一段式端到端
小米/比亚迪	2025.2	端到端全场景智驾；比亚迪下探至 7 万元级
Waymo（美）	2024.10	端到端模型 EMMA，研究，+独立安全校验
Wayve（英）	2017 起	以端到端自动驾驶技术路线为公司定位
Mobileye（以）	—	质疑单一端到端，主推复合 AI 系统 CAIS，反方
Zoox（美）	—	官方口径仍为模块化架构

数据来源：各公司官网，东吴证券研究所整理

3.2. 机器人：自 RT-2 起即以 VLA 形态采用端到端

机器人领域的端到端通常以 VLA 形态呈现。谷歌 DeepMind 在 2023 年 7 月提出 RT-2，将 VLA 定义为单一端到端训练的模型；此后 Physical Intelligence 的 π_0 、Figure AI 的 Helix、特斯拉 Optimus 以及中国侧智元 GO-1、银河通用 GraspVLA、星动纪元 ERA-42、宇树 UniFoLM、星海图 G0 等均沿 VLA 方向推进。

3.3. 数字 AI：端到端是默认范式，通常不再单独标注

纯软件 AI 中，端到端已成为深度学习后的默认范式，通常不再单独作为技术标签。语音识别、机器翻译、强化学习等领域在 2014 至 2015 年前后已普遍采用端到端训练；当产品从多模型拼接回归单一模型时，端到端才会重新成为表述重点。例如 OpenAI 在 2024 年 GPT-4o 发布中强调训练了横跨文本、视觉和音频的单一新模型，所有输入输出由同一个神经网络处理。与此同时，复合 AI 系统亦在前沿研究中受到关注，UC Berkeley BAIR 在 2024 年指出部分顶尖 AI 成果来自模型、工具和流程的系统组合，这也为 Mobileye 的 CAIS 路线提供了方法论参照。

4. 端到端与 VLA：演进与包含，而非二选一

4.1. VLA 的概念起源与技术原理

VLA 一词由谷歌 DeepMind RT-2 论文在 2023 年 7 月正式提出。RT-2 将互联网规模数据训练的 VLM 纳入机器人控制，把动作表示为文本 token 并纳入同一训练集，由此形成视觉、语言、动作统一建模的 VLA 范式。后续代表性工作大体沿用“预训练 VLM 骨干+动作输出头”的结构：OpenVLA 以 Llama 2 为骨干输出离散动作 token， π_0 通过流匹配为预训练 VLM 增加连续动作输出；自动驾驶侧，Wayve LINGO-2、理想 MindVLA 和小鹏第二代 VLA 则体现 VLA 概念在驾驶场景中的扩展与分化。

相较于朴素端到端，VLA 的增量主要体现在知识注入、语义推理与跨场景泛化。RT-2 论文显示，互联网规模预训练有助于提升新物体泛化、未见指令理解和多阶段语义推理能力。元戎启行公开表述亦指出，传统端到端在盲区感知、逻辑推理和泛化能力上存在约束，VLA 通过思维链推理、知识库和多模态融合补足部分能力。因此，VLA 并非否定端到端，而是在端到端动作生成链路中进一步引入语言、常识和推理能力。

4.2. 关系辨析：技术原理与产业口径的双重验证

从技术原理看，VLA 满足端到端的主要定义要素，属于端到端的子类与延伸。VLA 以原始视觉和语言指令为输入，以动作为输出，并通过单一模型或主体模型整体训练；提出 VLA 的 RT-2 论文也将目标表述为单一端到端训练模型，并强调将 VLM 直接纳入端到端机器人控制。学界综述进一步将自动驾驶 VLA 方法划分为端到端 VLA 和双系统 VLA 两类，前者整合感知、推理、规划，后者以 VLM 慢思考叠加规划器快执行。

产业口径同样将 VLA 表述为端到端的下一阶段，而非完全替代品。理想汽车将端到端视为 VLA 的动作基础，并将技术演进划分为规则算法+高精地图、端到端+VLM、VLA 三个阶段；其自动驾驶研发负责人亦指出，跳过端到端+VLM 直接做 VLA 并不可行。元戎启行则将现有一段式、两段式方案称为端到端 1.0，将 VLA 称为端到端 2.0。因此，VLA 更适理解解为端到端在多模态推理与动作生成方向上的升级，而非对端到端路线的否定。

4.3. 从快慢系统看：VLA 强化端到端的推理与泛化

快慢系统框架有助于解释端到端与 VLA 的差异。理想汽车将端到端对应 System 1，即基于经验的快速反应，将 VLM/VLA 引入的语言与推理能力对应 System 2，即慢思考；DriveVLM 论文也采用类似双系统设计。基于该框架，朴素端到端偏向快速反应和轨迹生成，优势是低延迟，约束是解释性与长时推理不足；VLA 在端到端链路中引入常识、交通规则、语言指令和因果推理，有望提升复杂场景泛化。理想公开口径中，规则系统、端到端 1.0 与 VLA 分别对应约 1 秒、7 秒和几十秒路况推理能力，体现其对长时域决策能力的定位差异。

5. 端到端与世界模型：一个出动作，一个预测世界

从技术原理看，端到端与世界模型分别对应动作映射和环境建模，属于不同维度。端到端回答输入如何映射为动作并通过何种方式优化，世界模型回答环境状态如何演化、未来场景如何生成，两者可以在同一系统中组合。两类特殊形态值得区分：1）世界模型直接上车成为行车模型本体，如蔚来 NWM 2.0，此时世界模型可作为端到端模型的一种实现方式；2）世界模型作为对纯模仿学习的补充，弥补其在空间认知、时间推演和分布外场景中的不足。华为 WEWA 与理想云端世界模型训练 VLA，体现世界模型参与端到端体系的不同路径。

从功能分工看，端到端负责动作生成，世界模型负责环境预测。端到端解决的是在当前观测下如何输出规划或控制；世界模型解决的是未来场景如何变化，以及模型如何在仿真环境中低成本获得训练信号。Ha 与 Schmidhuber 在 2018 年提出智能体可在内部生成环境中训练，蔚来、华为等公司也将世界模型与合成场景、平行世界或 AI 训练 AI 联系起来。我们认为，端到端与世界模型以互补关系为主；只有当世界模型直接参与车端驾驶决策时，两者才会在具体实现中融合。

6. 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。

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。

本报告的版权归本公司所有，未经书面许可，任何机构和个人不得以任何形式翻版、复制和发布。经授权刊载、转发本报告或者摘要的，应当注明出处为东吴证券研究所，并注明本报告发布人和发布日期，提示使用本报告的风险，且不得对本报告进行有悖原意的引用、删节和修改。未经授权或未按要求刊载、转发本报告的，应当承担相应的法律责任。本公司将保留向其追究法律责任的权利。

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>