汽车行业深度报告

AI 系列深度 09 期：从 LLM 到 VLM 再到 VLA 意味着什么？

2026 年 07 月 31 日

增持（维持）

证券分析师 黄细里
执业证书：S0600520010001
021-60199793
huangxl@dwzq.com.cn

投资要点

- **LLM、VLM、VLA 可理解为同一技术脉络上逐层扩展模态和输出能力的三类大模型。**LLM 以文本为输入输出，主要解决语言理解、生成和推理；VLM 在语言底座上加入视觉理解，使图像和文字在同一语义空间交互；VLA 进一步把动作作为输出模态，使模型能够驱动机器人或车辆在物理世界中行动。三者存在继承关系，但不能简单归结为同一种 next-token 机制。
- **从概念脉络看，**LLM 由 Transformer、GPT 等奠定，核心是以海量文本和自监督训练形成语言基础模型；VLM 由 CLIP、Flamingo 等推动，通过图文对齐、视觉编码器和模态投影，把视觉信息接入语言模型；VLA 由 Google DeepMind 在 RT-2 中明确命名，核心是把视觉、语言与动作统一到策略模型中。
- **从数据和表征看，三者难度逐级抬升。**LLM 训练依赖文本语料，VLM 增加图文对和视觉编码，VLA 则需要机器人示教、遥操作轨迹、动作反馈和真实环境数据。动作不是普通输入模态，而是会改变物理世界的控制输出；因此，VLA 既可把动作离散化为 token 自回归生成，也可用扩散策略或流匹配生成连续动作轨迹。
- **从应用边界看，**LLM 主要处理知识、代码、对话和软件任务；VLM 把能力扩展到图像理解、图文问答、视觉检索和多模态交互；VLA 则进入机器人、自动驾驶和具身执行，需要处理实时性、安全性和物理反馈。NVIDIA 的慢思考 VLM 加快思考动作模型、π0 的连续动作生成、RT-2 的动作 token 化，均体现 VLA 内部路线仍在分化。
- **产业映射上，**海外 OpenAI、Google、Meta、Anthropic 等主导 LLM/VLM，Google、Physical Intelligence、NVIDIA、Figure 等布局 VLA 和机器人；国内 DeepSeek、阿里、字节、百度、腾讯、智谱、月之暗面、MiniMax 等布局 LLM/VLM，智元、银河通用等聚焦 VLA 与机器人本体。后续竞争核心将从模型能力延伸到数据、硬件、动作控制与场景闭环。
- **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 08 期：GAN 与 Diffusion 两类生成范式有何差异？》
2026-07-31

《AI 系列深度 07 期：CNN 与 ViT 两类视觉骨干有何差异？》
2026-07-30

内容目录

图表目录

1. LLM：以文本为核心的语言基础模型

LLM、VLM 和 VLA 代表大模型能力从文本处理、视觉理解到物理行动的递进。LLM 以文本理解与生成为核心，VLM 在语言底座上接入视觉模态，VLA 进一步将动作作为输出，使模型进入机器人和自动驾驶等物理场景。

1.1. 概念界定与提出背景

大语言模型通常指在海量文本上预训练、参数规模达到数十亿乃至上千亿的语言模型，当参数规模越过一定阈值后，模型性能提升，并可能出现小模型不具备的能力。斯坦福 CRFM 则将其归入基础模型范畴。

奠基性工作包括：Transformer 架构提出完全基于注意力机制的网络结构；GPT-1 确立“生成式预训练+微调”范式；BERT 确立双向预训练。规模法则刻画了损失随模型、数据、算力的幂律关系，直接指导了大模型的设计。

1.2. 核心机制：自回归预训练与 Transformer

LLM 以自回归方式进行下一词元预测依托 Transformer 自注意力机制建模长程依赖，并依托 Transformer 自注意力机制建模长程依赖。训练流程通常包括无监督预训练、指令微调和基于人类反馈的强化学习（RLHF）。InstructGPT 证明，1.3B 参数的对齐模型在人类评测中可被偏好于 175B 参数的 GPT-3，奠定 ChatGPT 的对齐方法。GPT-3 提出的上下文学习，使模型无需更新参数即可利用少量示例完成任务。DeepSeek-R1 进一步显示，推理能力可通过强化学习激励形成。

1.3. 能力边界与应用场景

LLM 主要解决自然语言的理解与生成，并以“任务无关”的方式泛化到问答、翻译、摘要、写作、代码与复杂推理等任务。其应用已覆盖对话助手、办公与编程辅助、搜索与知识问答、智能体（Agent）等方向。

1.4. 代表性模型与公司

表1：LLM 代表性模型一览

模型	机构	时间	关键信息
GPT-3	OpenAI	2020	1750 亿参数，自回归
GPT-4 / 4o	OpenAI	2023 / 2024	多模态；参数未官方公开
PaLM	Google	2022	5400 亿参数，Pathways 训练，Google 官方，
Llama 3.1	Meta	2024	8B / 70B / 405B，开放权重，Meta 官方博客，
Claude 3 / 3.5	Anthropic	2024	旗舰多模态对话模型；参数未公开，Anthropic 官方，
DeepSeek-V3 / R1	深度求索	2024.12 / 2025.01	V3 总参 671B、激活 37B MoE；R1 纯 RL 推理 Nature 2025
通义千问 Qwen2.5 / 3	阿里巴巴	2024–2025	开源 0.5B–235B；72B 性能可比 Llama-3-405B，技术报告，
文心 ERNIE	百度	2023 起	知识增强；国内最早对标 ChatGPT 之一，百度官方，
混元 Hunyuan	腾讯	2023 起	千亿参数级，全链路自研，腾讯云官方，
GLM / Kimi / MiniMax	智谱 / 月之暗面 / MiniMax	2024–2025	GLM-4 系列；Kimi K2 总参约 1T、激活 32B，开源；MiniMax abab/M 系列

数据来源：OpenAI、Meta、Google、Anthropic、DeepSeek 等公司官网，东吴证券研究所整理

2. VLM：视觉模态如何接入语言底座

2.1. 概念界定与技术脉络

视觉语言模型，通常指以 LLM 为推理和生成底座、同时处理图像或视频与文本的模型。综述《A Survey on Multimodal Large Language Models》将其定义为以 GPT-4V 为代表、利用 LLM 执行多模态任务的模型。与仅处理文本的 LLM 相比，VLM 增加了视觉编码器与模态对齐模块。

技术脉络上，ViT 将图像切分为 patch 并进行 token 化；CLIP 利用约 4 亿图文对进行对比学习，奠定视觉-语言对齐基础；Flamingo 以门控交叉注意力实现少样本生成式 VLM；BLIP-2 提出 Q-Former 进行对齐；LLaVA 确立视觉编码器、投影层和 LLM 的主流组合范式；GPT-4V 和 Gemini 进一步将 VLM 推向商用和原生多模态路线。

2.2. 核心机制：视觉编码、模态对齐与语言推理

主流 VLM 通常采用三段式架构：1）视觉编码器，如 CLIP-ViT，将图像编码为视觉特征；2）模态对齐模块，如线性投影、MLP、Q-Former 或交叉注意力，将视觉特征对齐为视觉 token；3）LLM 主干在文本 token 和视觉 token 基础上完成推理与生成。对齐路线之外，CogVLM 提出在 LLM 内部加入视觉专家模块实现更深层融合。厂商也常

区分拼接式多模态与原生多模态两类路径。

拼接式 VLM 的核心在于把视觉特征转换为语言模型可处理的 token 空间。视觉编码器负责提取图像特征，模态对齐模块负责将这些特征映射到 LLM 的输入空间，LLM 主干再将视觉 token 与文本 token 共同建模。原生多模态路线则在训练阶段同时学习多类模态，减少后接转换模块带来的信息损耗和接口约束。

2.3. 能力边界与应用场景

VLM 使模型具备图像和视频理解能力，支持图文问答、图像描述、OCR、文档和图表理解、视觉推理及多模态智能体等任务。典型应用包括多模态助手、票据和文档解析、图形界面（GUI）操作智能体等。

2.4. 代表性模型与公司

表2：VLM 代表性模型一览

模型	机构	时间	关键信息
CLIP	OpenAI	2021	约 4 亿图文对对比学习，视觉-语言对齐基座
Flamingo	DeepMind	2022	80B，门控交叉注意力，少样本
BLIP-2	Salesforce	2023	Q-Former 连接冻结视觉编码器与 LLM
GPT-4V / 4o	OpenAI	2023 / 2024	商用多模态，图入文出 / 原生 omni，
Gemini	Google	2023 起	原生多模态
Qwen-VL / 2-VL	阿里巴巴	2023–2024	动态分辨率+M-RoPE；72B 可比 GPT-4o
InternVL	上海人工智能实验室/商汤	2023 起	InternViT-6B，开源，学术影响大
CogVLM / GLM-4V	智谱 AI	2023 起	视觉专家深度融合
Yi-VL / Step / 豆包	零一万物 / 阶跃星辰 / 字节	2024 起	开源 6B/34B；千亿级多模态；C 端规模化

数据来源：OpenAI、Google DeepMind、Salesforce、阿里巴巴等公司官网，东吴证券研究所整理

3. VLA：动作输出如何进入多模态模型

3.1. 概念界定与提出背景

视觉-语言-动作模型是具身智能和机器人领域的一类模型，通常以机器人观测、图像和自然语言指令为输入，端到端输出机器人动作。Google DeepMind 在 RT-2 中明确提出并命名该术语：论文将动作表示为文本 token，并与自然语言 token 共同纳入训练，称这类模型为视觉-语言-动作模型。其思想前身包括 RT-1 的动作 token 化、SayCan 的语言规划和技能执行，以及 Gato 的通才智能体路径。

跨本体数据基座：Open X-Embodiment / RT-X 汇集 22 种机器人本体、超过 100 万条轨迹，被业界称为“机器人学习的 ImageNet”，为后续开源 VLA 提供数据基础。

3.2. 核心机制：动作表征、连续生成与系统分工

VLA 的核心思想，是将互联网图文数据上预训练的 VLM 能力迁移至机器人控制任务。模型可在机器人轨迹和图文任务上进行协同微调，并通过三类方式输出动作：1）将动作离散化为 token；2）通过扩散策略或流匹配生成连续动作；3）采用慢思考 VLM 和快思考动作模型的双系统分工。π0 以预训练 VLM PaliGemma 为骨干，用流匹配生成最高约 50Hz 的连续动作；NVIDIA GR00T N1 采用 System 2 和 System 1 双系统架构；智元 GO-1 提出 ViLLA 视觉-语言-隐动作路线。

VLA 的动作建模可拆为三类机制。第一类是离散动作 token，将机械臂或机器人动作离散化为序列，并用自回归方式逐步预测；RT-2 和 OpenVLA 属于该路线。第二类是连续动作生成，通过扩散策略或流匹配直接生成连续动作轨迹，更适合高频、精细控制；π0 采用流匹配并实现约 50Hz 动作输出。第三类是双系统分工，由高层 VLM 负责场景理解和任务意图生成，低层动作模型负责实时控制；NVIDIA GR00T N1 和智元 GO-1 体现了相关探索。

三类机制对应两组不同问题：离散动作 token 和连续动作生成，回答的是动作如何表示和输出；双系统路线回答的是系统如何分工。前者比较动作表征形式，后者比较高层语义理解与低层控制执行之间的接口设计。

路线一，离散动作 token。该路线将连续动作切分为离散编号，并将每个编号作为动作 token 进行自回归预测。优势在于能够复用语言模型的序列建模机制，并将动作与文字纳入统一 token 空间；约束在于离散化会带来精度损失，逐步预测也可能导致延迟和动作不够平滑。代表模型包括 RT-2 和 OpenVLA。

路线二，连续动作生成。该路线不将动作离散化为 token，而是使用扩散策略或流匹配等生成模型直接输出连续轨迹。扩散策略通过逐步去噪生成动作，流匹配通过更直接的生成路径从噪声映射到动作。该路线更适合高频、精细、连续控制场景，π0 即采用流匹配生成约 50Hz 连续动作。

从表征角度看，路线一和路线二的输入侧相近，均依赖 VLM 对场景和指令的理解；差异主要在输出侧如何解码动作。路线一把动作表征为离散 token 序列，并采用自回归预测；路线二把动作表征为连续向量或轨迹，并通过生成模型整体输出。二者本质上是动作解码方式的差异。

路线三，双系统分工。该路线不直接回答动作是离散还是连续，而是将系统拆分为高层理解与低层控制：高层 VLM 负责理解场景和规划任务意图，低层动作模型负责实时生成和执行动作。快系统内部仍可能采用离散动作 token 或连续动作生成，因此双系统路线与前两类路线并非同一维度的替代关系。

双系统路线的表征创新在于引入中间意图或隐动作表征。高层模型输出任务意图或隐向量，作为低层动作生成的条件；智元 GO-1 的 ViLLA 将动作压缩为介于离散词元和

连续轨迹之间的隐动作 token，属于厂商表述中的中间层设计。该设计试图在语义规划和连续控制之间建立更稳定的接口。

双系统的关键约束在于接口稳定性和联合训练质量。高层系统决定任务意图，低层系统负责动作执行，信息通常自高层流向低层；若高层意图频繁变化或中间表征不清晰，低层动作可能出现抖动或反复调整。因此，双系统路线对中间表征设计、训练协同和实时控制工程要求较高。

双系统路线与杨立昆提出的 System 1/System 2 框架存在相似之处，但底层建模选择不同。主流 VLA 通常使用生成式或自回归 VLM 作为高层系统，在语言或 token 空间中形成任务意图；杨立昆则更强调 JEPA 类世界模型，在抽象表征空间中规划而非生成像素或 token。这一分歧也与其对生成式模型和强化学习的审慎态度相一致。

3.3. 能力边界与应用场景

VLA 旨在提升机器人和具身智能的泛化操作能力，使机器人能够理解开放式自然语言指令，并在物理世界中执行动作。相比一机一任务的传统程序，VLA 更接近通用机器人策略，应用集中于人形机器人、机械臂抓取、家庭服务和工业操作等场景。

3.4. 代表性模型与公司

表3：VLA 代表性模型一览

模型	机构	时间	关键信息
RT-1	Google	2022	动作 token 化，700+ 指令
RT-2	Google DeepMind	2023.07	首个命名 VLA，Web 知识迁移
Open X-Embodiment / RT-X	21 家机构	2023	22 本体、100 万+ 轨迹
OpenVLA	斯坦福等	2024	7B 开源，约 97 万演示，超 RT-2-X
π0	Physical Intelligence	2024	PaliGemma + 流匹配，约 50Hz 连续动作
GR00T N1	NVIDIA	2025	开源人形基础模型，双系统
Gemini Robotics	Google DeepMind	2025	基于 Gemini 2.0，动作为新输出模态
GR-2 / GO-1 / GraspVLA	字节 / 智元 / 银河通用	2024–2025	视频-语言-动作；ViLLA 隐动作；合成数据抓取基座

数据来源：Google DeepMind、Physical Intelligence、NVIDIA 等官方网站，东吴证券研究所整理

4. LLM、VLM、VLA 关系：模态扩展与约束递进

4.1. 架构关系：从文本到视觉再到动作

从架构看，三者呈模态逐层扩展的递进关系：LLM 是文本基座；VLM 在 LLM 基础上增加视觉编码器，并将视觉特征对齐到语言模型；VLA 在 VLM 基础上增加动作输

出，通常通过动作 token、动作头或连续动作生成模块实现。

4.2. 问题与数据递进：从数字语料到真实动作轨迹

从解决的问题看，三者能力边界依次递进：理解文字，理解视觉世界，在物理世界行动。NVIDIA 黄仁勋在 CES 2025 将 AI 划分为感知 AI、生成式 AI 和物理 AI（Physical AI）三阶段。从数据看，训练数据依次为文本语料、图文对和机器人轨迹；后者采集成本远高于前两者，是 VLA 的核心约束，业界因此转向人类视频和仿真合成数据补充。

表4：LLM、VLM、VLA 多维对照

维度	LLM	VLM	VLA
输入	文本	图像 + 文本	图像 + 语言指令 + 本体状态
输出	文本	文本	机器人动作
关键技术	自注意力、预训练、RLHF	视觉编码器 + 模态对齐	动作 token 化 / 流匹配 / 双系统
训练数据	文本语料	图文对，CLIP 约 4 亿，	机器人轨迹，OpenVLA 约 97 万，
代表作	GPT、DeepSeek	CLIP、GPT-4V、Qwen-VL	RT-2、π0、GR00T N1
关系定位	基座	LLM + 视觉	VLM + 动作输出

数据来源：《Attention Is All You Need》《Language Models are Few-Shot Learners》《Learning Transferable Visual Models From Natural Language Supervision》《Training Language Models to Follow Instructions with Human Feedback》等学术文献，东吴证券研究所整理

4.3. 演进时间线

表5：LLM–VLM–VLA 关键演进时间线

年份	里程碑	所属层次
2017	Transformer 架构 Google	LLM 底层架构
2018	GPT / BERT，预训练范式确立	LLM 起点
2020	GPT-3，1750 亿参数，	LLM 规模化
2021	CLIP，OpenAI，约 4 亿图文对，	VLM 起点
2022	Flamingo DeepMind/ ChatGPT 发布	VLM 成熟 / LLM 出圈
2023	GPT-4V / RT-2，首个命名 VLA，	VLM → VLA 跨越
2024	π0 / OpenVLA	VLA 扩散
2025	GR00T N1 / Gemini Robotics	VLA + 人形机器人基础模型

数据来源：OpenAI、Google DeepMind、Meta 等公司官网，东吴证券研究所整理

4.4. 中国厂商在三个层次的布局

国内格局可概括为三类：基座型，DeepSeek、阿里、字节、百度、腾讯、智谱、月之暗面、MiniMax、上海人工智能实验室/商汤，主攻 LLM/VLM；具身型，智元、银河通用等，主攻 VLA 与机器人本体；跨界型，字节、阿里、腾讯等，以基座能力下沉具身或通过投资卡位。

表6：中国主要厂商在 LLM、VLM、VLA 三个层次的布局

厂商	LLM	VLM	VLA / 具身
DeepSeek	是	部分	—
阿里巴巴，通义	是	是 Qwen-VL	生态 / 投资
字节跳动，豆包	是	是	是，GR-2，硬件双线，
百度，文心	是	是	—
腾讯，混元	是	是	—
智谱 AI	是	是 GLM-4V	—
月之暗面 Kimi	是	是	—
MiniMax	是	是	—
上海人工智能实验室 / 商汤	是，书生，	是 InternVL	—
智元机器人 / 银河通用	—	—	是 GO-1 / GraspVLA

数据来源：DeepSeek、阿里巴巴、百度、腾讯、智谱等公司官网，东吴证券研究所整理

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。

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