汽车

行业深度报告

华源证券 HUAYUAN SECURITIES

2025 年 07 月 07 日

当自动驾驶与机器人共振：详解 VLA 与世界模型

投资评级： 看好（首次）

证券分析师

李泽

SAC：S1350525030001

lize@huayuanstock.com

联系人

板块表现：

投资要点：

- **VLA 模型具备成为具身智能基础模型潜力。**视觉-语言-动作模型（VLA）代表一类旨在处理多模态输入与输出的模型，通用性是 VLA 模型的核心特点之一，体现在其以多模态大语言模型为底座，具备“理解万物”的能力，VLA 模型的理解能力和多任务泛化能力让模型在不同的应用场景中具备出色表现，展现出超越自动驾驶乃至机器人等单独垂域应用的潜力，有望成为广义具身智能基础模型范式。

- **VLA 模型是自动驾驶向知识驱动、体验优先升级的技术基础。**我们认为汽车领域智能化的最终形式是实现驾驶领域的通用人工智能，而非简单的汽车电子软件智能化，这使得汽车从第三人称智能化向第一人称智能化、由数据驱动向知识驱动进化成为自动驾驶进化的未来趋势，而 VLA 模型特别是其中语言类大模型的成功引入则奠定了范式转变的技术基础。底层技术逻辑升维也带动了车端应用焦点由基本功能实现向人车交互性、极端场景通过性等体验提升转变，中期维度看，不同车企自动驾驶的模型性能领先、功能领先将转化为体验领先并重塑汽车行业产品生态，知识驱动范式的智能化模型也将重新定义自动驾驶，行业将真正开启电动智能化下半场角逐。

- **工程化部署能力是当前车端 VLA 核心矛盾。**现有车端 VLA 技术路线尚未收敛且工程部署较少，但由于汽车面临的结构场景和任务单一、硬件结构较统一、数据和汽车保有量高、车端算力相对充足等因素，车端 VLA 范式落地前景已经较为明朗。我们认为：**1）在汽车 VLA 模型的数据闭环构建过程中，**获取良好 3D 中间表征、强化模型长时序记忆能力与端侧计算效率、优化模型架构、构建高保真的云端模拟环境等环节是 VLA 端侧工程部署的关键。**2）在模型训练方法上，**行为克隆（初步训练端到端模型）+逆强化学习（从专家数据中初始化奖励函数）+强化学习（通过与环境交互改善模型和奖励函数）方案或将成为未来自驾模型训练主流方案。

- **具身智能本质是自动驾驶的升维问题，构建数据闭环是关键。**具身智能的场景、任务复杂度更高、本体自由度更高、感知方式更灵活、硬件构型更差异化等因素决定了具身 VLA 模型的数据闭环构建难度远超自动驾驶；但同时智能汽车实质上是物理智能体的具体形式之一，自动驾驶与具身智能在核心的智能化能力即模型构建方法论以及具体的硬件零部件领域有诸多可迁移之处，因此具身智能本质是自动驾驶的升维问题。**我们认为，当前以人形机器人为代表的具身智能关键问题在于还无法进行有效的数据采集进而实现数据闭环，而无法 Scaling 的具身模型就无从实现智能化。在当前阶段，标准化与模型性能优化是破局的关键点：**1）首先需要标准化，尤其是底层硬件、通信协议的标准化，底层零部件和软件基础标准化是机器人统一构型以进行规模化数据收集破局数据-模型能力的鸡生蛋问题以及壮大行业供应链的关键；2）模型优化核心在于闭环模型构建以及模型物理交互能力建设，其中物理交互能力的引入或是未来人形机器人向真正具身智能体转化的关键。

- **投资建议：建议关注本轮智能化进展较快的整车企业理想汽车、小鹏汽车**
- **风险提示：1）新技术迭代风险、2）市场竞争加剧风险、3）宏观经济环境波动风险**

内容目录

图表目录

1. 自动驾驶模型范式由数据驱动向知识驱动升维

1.1. 自动驾驶两大趋势：模型数据驱动转向知识驱动、模型功能至上迈向驾乘体验优先

随着自动驾驶从单一感知任务向感知-决策-执行的综合任务转化，自动驾驶不仅对于所收集数据的模态多样性与丰富度要求提升，对于模型本身的思考、理解能力要求也愈发提高。仅仅依靠大量收集自动驾驶数据训练的模型（数据驱动）只能是第三人称智能，即从旁观者角度学习、模仿人类行为却无法具备自我思考能力。**我们认为汽车领域智能化的最终目标是实现车端的通用人工智能，而非简单的汽车电子软件智能化，这要求汽车具备第一人称智能，即依靠自身思考能力探索环境、获取一般知识，**而不是执行预先定义的人类规则或从收集的数据中描绘抽象特征，这使得从数据驱动范式向知识驱动范式的转变成为自动驾驶进化的未来趋势，而 LLM 等语言类大模型的成功引入则奠定了范式转变的技术基础。

数据驱动转化为知识驱动是自动驾驶由功能实现迈向体验升级的底层技术逻辑。知识驱动范式并非完全跳脱数据驱动方法，而是在原有基础上增加了知识框架设计，知识驱动本身也需要不断从数据中进行总结提炼以获得涌现能力，**数据驱动向知识驱动转化的过程即是焦点从自驾基本功能实现向人车交互性、极端场景通过性等体验提升转化的过程。**知识驱动的方法更为关注模型类人性、泛化性与通识能力的实现，使汽车不再是单纯的驾驶工具而是成为一个能够与用户进行沟通，能理解用户意图甚至提供情绪价值的物理智能体。

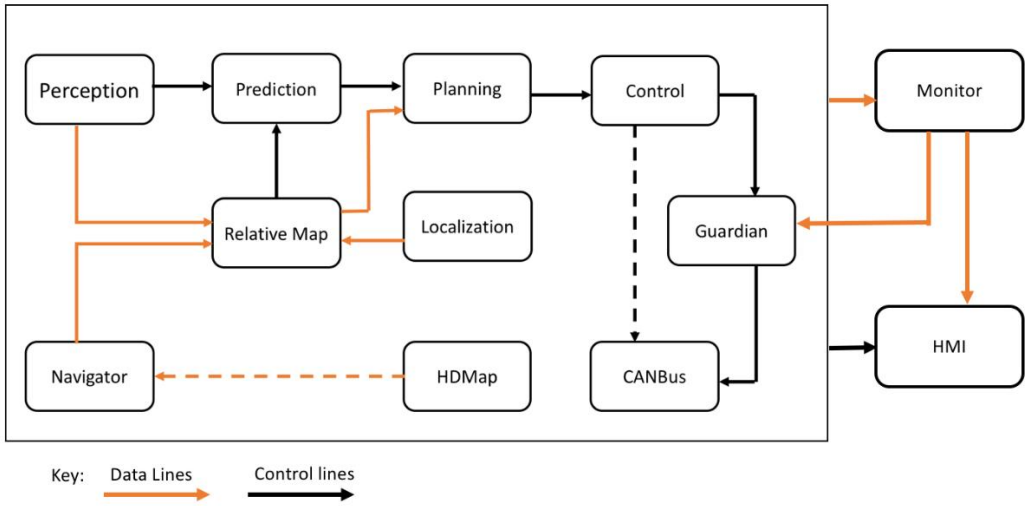
我们认为，在不久的将来，不同企业自动驾驶的模型性能领先、功能领先将转化为体验领先并重塑汽车行业产品生态，知识驱动范式的智能化模型也将重新定义自动驾驶，行业预计将迎来智能化“iphone 4 时刻”并真正开启电动智能化下半场角逐。

1.2. 传统基于规则的模块化算法核心缺陷在于无法 Scaling

传统规则驱动的模块化算法多衍生自机器人算法，该模式下通常将自动驾驶功能拆分为独立模块，这些模块通常包括地图构建、环境感知、目标检测、定位、决策规划、车辆控制等，每个模块有自己独立的算法和处理流程，不同模块间采用人为定义的接口进行连接，各个模块之间进行相对独立的开发和测试，最后将它们集成实现自动驾驶功能。模块化算法很大程度上依赖人工定义的规则和先验知识，其核心优势在于模型的可解释性，在出现系统问题或部署失败时容易调试。例如丰田 TSS、百度 Apollo 3 等早期模型都是模块化算法代表。

传统的规则式模块化算法存在固有问题，核心缺陷在于无法 Scaling：1）模块之间独立研发与人为定义接口导致信息传递损失，无法达到全局最优，且最终结果无法反向传播以优化模型性能；2）基于人为定义的规则驱动，陌生环境鲁棒性差，长尾问题难以解决；3）模块间的累积误差会影响最终结果；4）成本问题，实现一个较为稳定的传统规则式自驾系统约需要数万条各类人工输入规则，而一个无限接近人类司机的自动驾驶系统等效于数亿条规则，在实际工程落地中几乎是不可能事件。

图表 1：百度 Apollo 算法架构具备典型感知、预测、规划、控制模块化特征



资料来源：Github，华源证券研究所

数据驱动方法开始在部分子模块应用，但整体仍未跳脱规则式范畴。由于传统规则算法存在诸多问题，2019 年以后特斯拉逐步在子模块中引入神经网络算法，逐步实现感知、规控模块模型化并引起诸多自驾公司效仿，形成了“两段式”、“多段式”等多种“伪端到端”模型。例如此时典型的两段式模型中感知模块采用多传感器融合的 BEV 技术实现模块级端到端，并应用 transformer 等方法提升感知精度；规划模块则被集成在另一个神经网络中。**该阶段处于规则驱动到端到端模型数据驱动的中间态，**一方面该阶段感知、决策等子模块都由基于数据驱动的方法实现；另一方面，从接口定义和联合优化角度，此时两个模块间的接口仍表现为人为定义的显式形式，同时各模块的优化仍然局限在模块内部，可以分别做到局部最优，但难以实现全局最优，因此从严格定义看，该阶段仍属于基于规则的模块化算法。

图表 2：不同阶段的自动驾驶算法演进

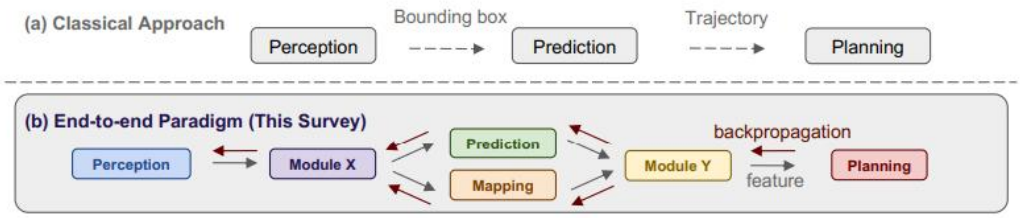
资料来源：九章 AI 产业管理咨询，华源证券研究所绘制

1.3. 端到端模型：自动驾驶从数据驱动向知识驱动演进

1.3.1. 数据驱动的端到端模型面临数据瓶颈

端到端是一种模型的组织框架而非具体的技术范式，其相对于模块化的模型组织形式如同“连续”相较于“离散”的区别，或者汽车分布式架构向域集中再到中央集中架构升级的过程。具体而言，端到端模型可以被定义为一种基于学习的、完全可微分的算法思路，它将原始传感器数据作为输入，并生成轨迹规划或低级控制动作作为输出，其中不包含任何人为设计的模块或接口。需要指出的是：**1）自动驾驶端到端模型可以依赖不同的具体技术方法实现，**即可以利用传统的神经网络架构（数据驱动式的概率输出）、也可以利用 VLA 方案（知识驱动式的理解能力输出）、或者将二者组成双系统（高维思考+低维执行）、或利用世界模型方案。不同端到端实现方案会给模型构建与性能发挥、芯片等硬件要求带来不同影响。**2）端到端模型并不一定是黑盒模式，**它可以像规则算法一样进行模块化设计并加入大语言模型以增强模型可解释性和分部优化，其核心在于不同模块间传播的是隐式特征而非具体输出结果，不同模块间可以联合优化以实现全局最优性能（信息损失最小化与联合优化）。

图表 3：端到端方案与传统模块化方案的对比



资料来源：《End-to-end Autonomous Driving: Challenges and Frontiers》_Li Chen 等，华源证券研究所

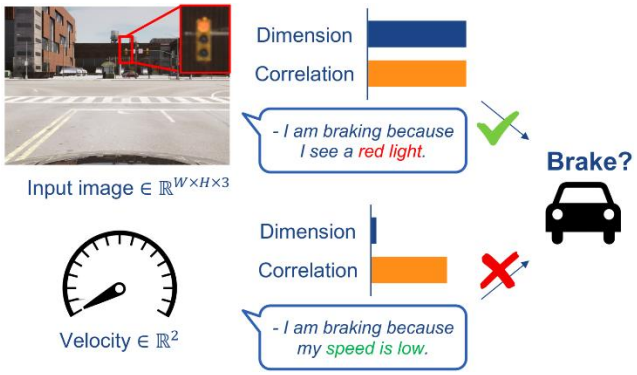
注：（b）中灰色箭头代表正向隐式特征传递，红色箭头代表输出结果的反向传播以优化模型性能

数据驱动的本质是模型进行模仿学习，其“穷举+归纳”的方法使模型发展面临数据瓶颈。数据驱动范式即是从大量具体数据中抽象出统计规律进行学习和决策，模型通过对大量标注数据进行监督训练获得相对的泛化能力，强调“数据即知识”，但模型本身基本不具备推理能力，想要获得更好的模型能力，即需要穷举更多的场景以获得数据养料，数据驱动范式的具体弊端包括：

1）数据量与数据质量要求较高，且难以穷举所有长尾场景。数据数量方面，以特斯拉为例，其 FSD 训练依赖于上千万个视频片段，累计时长达到几万小时，但起初在中国落地过程中由于本地数据量不足模型性能发挥仍然受到了限制，可能原因之一即是数据驱动方法无法穷举所有长尾场景导致模型零样本泛化能力较差；**数据质量方面，**自驾模型的质量很大程度上取决于所使用的训练数据的类型、多样性和高质量，但符合“老司机”标准的可模仿视频片段和极端场景片段并不易得。

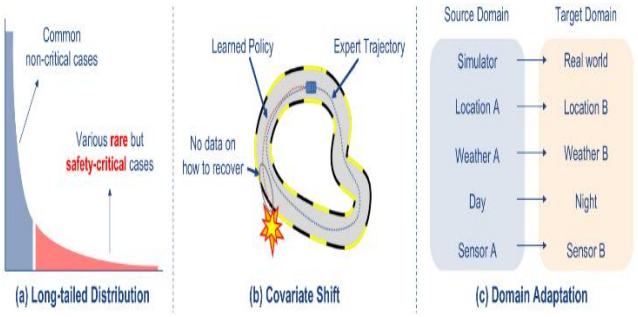
2）模仿学习存在因果混淆、数据分布偏移、专家依赖性等问题。因果混淆是指模型学习到数据间的关联性而非确定正确的因果关系，例如在十字路口停车，不能确定是因为红绿灯停车还是因为旁边的车减速而停车，模型无法自主理清逻辑；**数据分布偏移**是指训练数据与实际环境之间的概率分布不同时，模型性能可能急剧下降，即极端场景泛化能力有限，容易造成模型下限极低；**3）专家依赖性。**由于当前模仿学习主要采取行为克隆模式，这类学习方法下专家数据质量直接决定模型性能上限，模型通过模仿学习无法超越专家水平。

图表 4：模仿学习导致因果混淆



资料来源：《End-to-end Autonomous Driving: Challenges and Frontiers》_Li Chen 等，华源证券研究所

图表 5：模型学习的鲁棒性挑战（长尾问题、数据分布偏移、数据迁移问题）

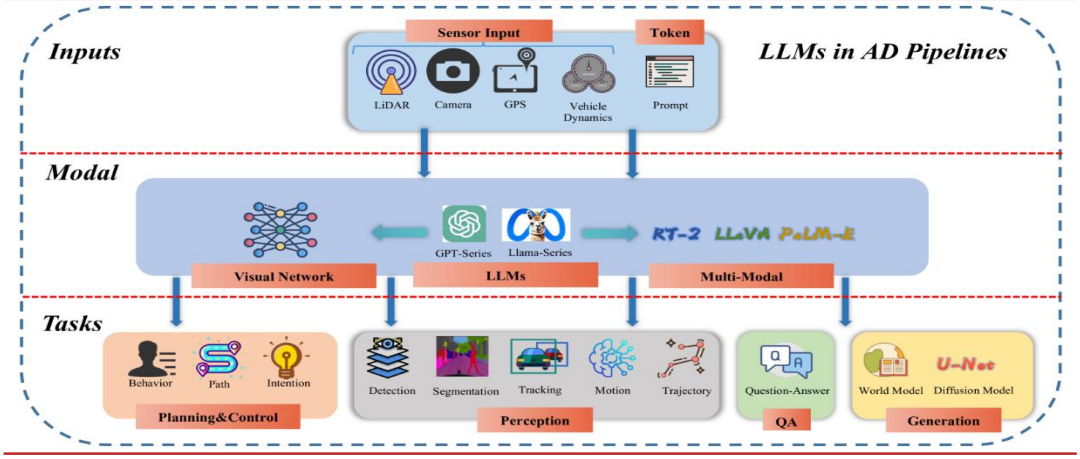


资料来源：《End-to-end Autonomous Driving: Challenges and Frontiers》_Li Chen 等，华源证券研究所

1.3.2. 多模态大语言模型引入是端到端模型实现知识驱动的关键

大语言模型引入奠定了知识驱动技术基础。近年来，如 GPT-4 等大语言模型（LLMs）在语义理解、答案生成和处理复杂任务方面展现出卓越的能力，其与多种编码器集成后形成的多模态模型实现了文本、图像、视频、点云等信息的统一特征空间映射，显著增强了模型的泛化能力，使其能够以零样本或少样本的方式快速适应新场景。而将多模态大语言模型与传统端到端模型进行有机结合形成 VLA 模型，能够凭借大语言模型丰富的知识库、强大情景理解能力更轻松地学习复杂的驾驶行为，**强调“理解即知识”，使得解决自动驾驶的长尾问题、规划决策以及为决策提供直观的解释成为可能，进而推动端到端模型由数据驱动范式向知识驱动范式的升级。**

图表 6：LLM 模型在自动驾驶流程中的应用

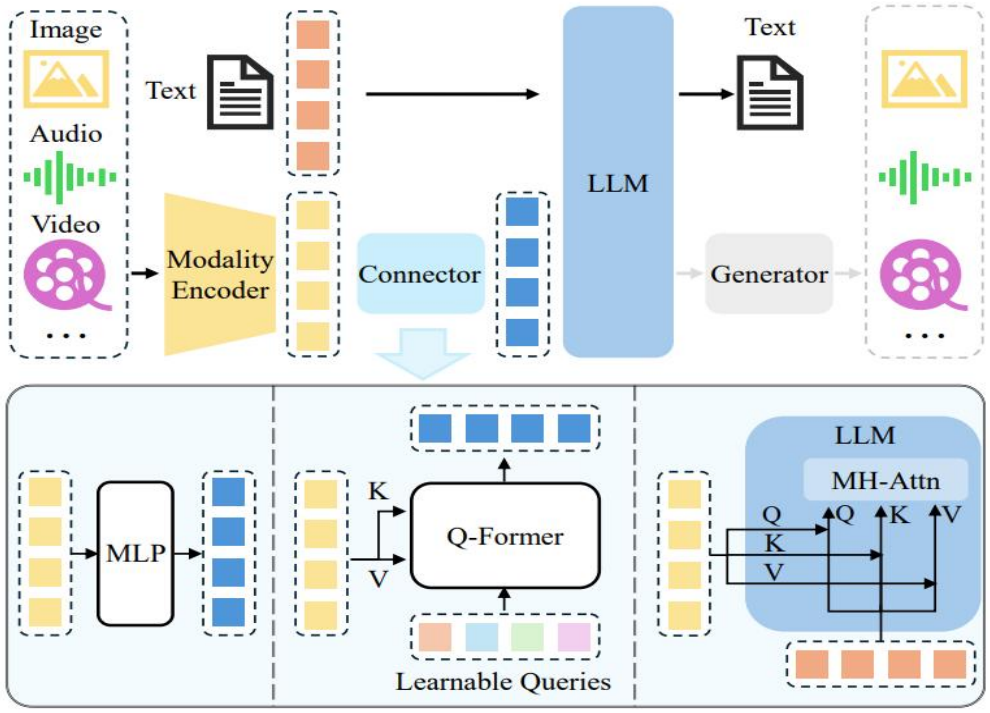


资料来源：《LLM4Drive: A Survey of Large Language Models for Autonomous Driving》_Zhenjie Yang 等，华源证券研究所

多模态性是当前自动驾驶 VLA 模型的典型特征。大语言模型（LLM）在大多数自然语言处理（NLP）任务上具有极佳推理能力，但在视觉处理上（例如理解 3D 空间）表现较差；同时例如大视觉模型（LVMs）可以轻易地处理图像与视频信息，但推理能力落后。因此将 LLM 与其他模态信息（如图像、视频、音频等）结合形成多模态大语言模型（MLLMS，例如将 LLM 与视觉编码器集成形成 VLM）**使得自动驾驶系统能像人类一样理解多模态信息，甚至在具身领域还能依靠机器人触觉、嗅觉等模态信息进一步理解物理世界，同时还能依赖 LLM 的能力进行深度思考与推理并做出相应决策。**自动驾驶领域 VLA 模型即是一类以大语言模型为基础的多模态模型，它主要关注自动驾驶中的视觉、语言、动作三种模态信息及其之间的语义连接，其多模态属性天然与自动驾驶的输入输出、人类驾驶行为的多模态性吻合，使之成为当前自动驾驶主流基座模型。

常见 MLLMs 的架构组成与训练：常见的 MLLMs 架构一般可分为 4 个模块，即预训练的多模态编码器、对齐模块、预训练的 LLM、解码器（生成器）。以理想 VLA 架构（详见图表 25）为例，其空间智能部分集成了多模态编码器和对齐模块，使用一个 3D Enconder 编码图像和激光雷达信息并输出 3D 特征，一个普通 Encoder 编码位置、导航信息等文字信息，一个 3D 投影仪（对齐模块）将编码器信息投射对齐至语言空间；语言智能部分为从零训练的 Mind GPT 语言大模型，用于理解场景和输出高层次决策规划；行为智能部分为一个扩散模型解码器，用于将语言模块输出的高层次指令（视作语言 prompt）精细化为低层次的具体车端执行动作，完成“抽象到具体”的映射。整体而言，**MLLMs 模型的训练过程主要包括模型预训练、垂域数据微调、对齐调优（例如自动驾驶中的人类行为对齐）、模型强化学习等环节。**

图表 7：典型的 MLLM 模型架构，包含编码器、连接器（对齐模块）、LLM、生成器



资料来源：《A Survey on Multimodal Large Language Models》_Shukang Yin 等，华源证券研究所

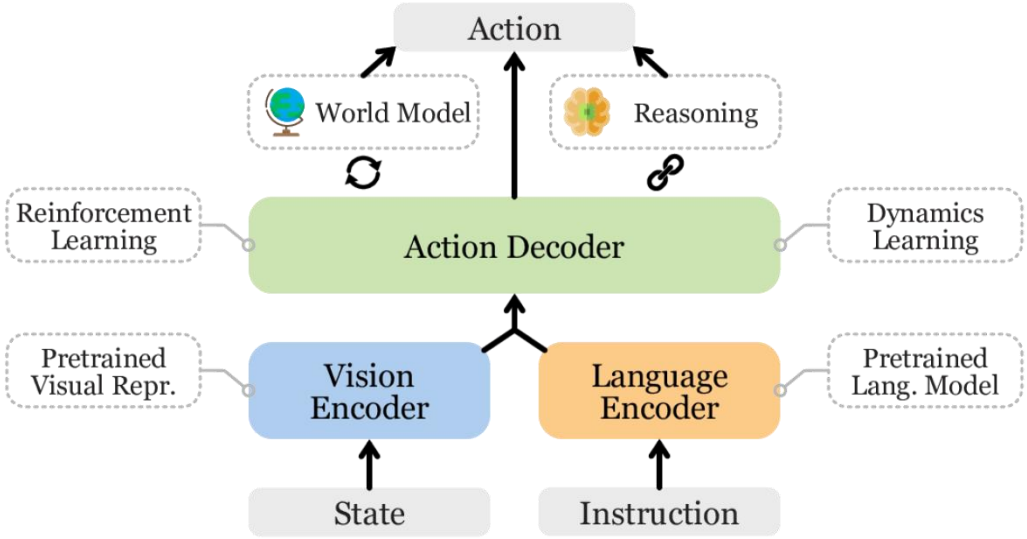
2. VLA 模型是自动驾驶端到端架构的演进方向

2.1. VLA 推动自动驾驶从功能迈向体验

视觉–语言–动作模型（VLA）是一种多模态的机器学习模型，由 VLM 模型演变而来，它结合了视觉、语言和动作三种能力，旨在实现从感知输入直接映射到控制输出的完整闭环能力，其不仅关注环境感知，也关注规划与控制问题。VLA 模型最初被开发用于解决具身智能中的指令跟随任务，其后这一理念快速应用于自动驾驶领域，**相较于“VLM+E2E”的中间态架构，VLA 深度整合了空间感知、逻辑推理、行为规划等多模态信息进行端到端训练，从根本上解决了模型信息传递损耗和不同模型联合优化训练问题，**显著提升了模型极端环境下泛化能力和决策能力，**推动自动驾驶从端到端模型“自驾功能实现”迈向 VLA 模型“交互性、类人性、泛化性体验优先”。**

一般而言，VLA 模型架构具有三个核心组成：多模态编码器（动作、文本、图像等）、大语言模型用以接收信息和进行推理、解码器用于输出轨迹和动作。但也有部分模型只含有两个模块，如 Open VLA 使用 LLM 主干直接输出 action 指令。

图表 8：VLA 模型的总体架构，包含编码器、解码器和输出动作



资料来源：《A Survey on Vision-Language-Action Models for Embodied AI》_Yueen Ma 等，华源证券研究所

VLA 模型具有强大通用能力，具备成为具身智能基础模型的潜力。VLA 通用性体现在其以多模态大语言模型为底座，具备“理解万物”的能力，针对不同的使用场景和任务，理论上只需要使用特定数据对模型进行大规模预训练并适配相应的解码模块，即能输出相应的动作指令，相当于给一个足够聪明的大脑匹配不同的躯干和感知器官以适应不同的任务需求。其高度的场景推理能力和泛化能力让模型在不同的应用场景中都能表现出色，展现出超越自动驾驶乃至机器人等单独垂域应用的潜力，有望成为广义具身智能基础模型范式。

2.2. 自动驾驶 VLA 模型构建与工程部署面临的主要问题及解决方法

自动驾驶 VLA 模型更多是一个工程化而非技术性问题。实现 VLA 模型的工程落地至少需要三个前提：即一个足够聪明的模型（大脑）在一个足够拟真的空间中（仿真环境）进行训练，并利用足够优秀的映射对齐算法实现数据、模型能力的 real2sim、sim2real 迁移。在自动驾驶领域，主要面临的是模型与环境问题，模型层面突出表现为模型的多模态性、3D 空间感知能力、计算速度与计算开销平衡、长时序记忆能力等问题；环境问题主要是如何构建优秀的仿真环境。虽然现有的车端 VLA 技术路线尚未收敛且工程部署较少，但我们认为由于汽车面临的结构化场景、任务单一、汽车自由度低且结构较为统一、数据和车队保有量高、各种数据迁移方式迭代完善、算力充足等因素，车端 VLA 技术路线已经较为明朗，其更多是一个工程化问题而非技术性问题，有望支撑汽车由 L2+走向 L3 甚至 L4 级自动驾驶，目前不同厂商都进行了模型方案在理论层面的改进，国内元戎启行、理想、小米、小鹏等已有了相关进展，其中小鹏 VLA-OL、理想 Mind VLA 工程化落地进展较快，预计年内将实现车端部署。

2.2.1.3D GS 或是车端实时获取良好 3D 中间特征的途径

多段式 VLA 模型云端训练和端侧部署都需要良好的 3D 中间特征。自动驾驶中间特征指用于连接感知层与推理决策层的抽象表示，是由原始传感器数据经过处理后的高层次特征，通常包含场景障碍物、道路语义、行人等静态信息，速度方向等动态信息，可以理解为包含自车周围所有隐式、显示信息的统一场景表达，**获取良好的 3D 中间特征，无论是在端侧服务下游如路径规划、行为预测等驾驶任务，还是云端构建良好的训练环境供模型迭代训练都有重要意义。**传统的端侧构建中间表达的方式有高精地图、BEV 鸟瞰图、占用网络、实时高精地图等方式，云端一般为 NeRF 场景重建算法+素材库+游戏引擎重建环境，但传统的方法或多或少皆有缺陷，如端侧通过稀疏查询（如实例框、地图元素）描述周围场景无法精细捕捉 3D 环境的细节导致决策过程信息不足、OCC 算法将场景表示为 3D 占用以获取更全面的细节，但稠密计算导致计算开销较大挤压了推理决策的资源，云端也存在重建速度缓慢、重建真实性、丰富度不足等缺陷。**而 3D GS 作为一种全面且稀疏的中间特征获取方式，在场景精细度和构建效率方面取得了较好的权衡效果。**

图表 9：3D GS 与其余三维重建技术的区别

技术	核心思想	优势	劣势
点云（LiDAR）	离散的三维点集合，直接记录物体表面位置	简单直观，硬件成熟	无表面连续性，需后处理（如网格化）
网格（Mesh）	由顶点和面片构成的三维表面模型	支持物理仿真和编辑	拓扑结构固定，动态场景适应性差
体素（Voxel）	将空间划分为规则立方体网格，每个体素记录属性（如密度、颜色）	规则结构适合深度学习	内存和计算开销大（分辨率限制）
NeRF	隐式神经辐射场，通过 MLP 网络建模空间点的密度和颜色	高保真重建，支持复杂光照	渲染速度慢，训练耗时长
3DGS	显式高斯分布集合，通过可微分泼溅渲染优化	实时渲染，动态场景适应性强	对初始点云质量敏感

资料来源：CSDN（Felaim），华源证券研究所

3D GS 是一种基于高斯分布的点云表示与渲染技术，有效权衡了场景重建真实性与重建效率的矛盾。3D GS 的重建过程可理解为：1）将多视角图像或点云数据（如 LiDAR）经过运动结构恢复（Sfm）处理生成稀疏点云；2）将点云转化为 3D 高斯点，并添加位置、颜色、形状分布、不透明度等信息形成场景的初步表示；3）通过可微分渲染技术和自适应密度控制进行优化，最小化渲染图像和真实场景差异；4）最后利用 GPU 生成最终图像并做到实时渲染。3D GS 的优异性能使其能够应用于自动驾驶仿真环境重建、实时渲染建图、动态障碍物检测跟踪等任务。**与传统场景重建技术 NeRF 相比，3D GS 具有计算效率较高、自监督、渲染实时性等优势，为端侧应用提供可能。1）渲染实时性高，**3D GS 能通过 GPU 并行化实现实时渲染（＞30 FPS），而 NeRF 渲染一帧需数秒至数分钟，相较之下 NeRF 更像一位精细的画家，注重写实，而 3D GS 则是一位泼墨艺术家，注重写意，泼洒的速度会显著快于精细绘画；**2）数据需求较少，**仅需少量多视角图像即可生成高保真模型，存储空间需求比 NeRF 减少 50%以上；**3）动态适应性，**3D GS 可通过调整高斯分布的位置直接建模动态物体（如移动车辆），而 NeRF 需重新训练或引入额外动态建模模块，效率较低；**4）自监督学习，**3D GS 可利用原图 RGB 信息进行自监督学习，使重建模型利用海量数据进行自我训练成为可能。

图表 10：3D GS 场景重建效果更优

资料来源：特斯拉 AIday 2022，GTC 2024，华源证券研究所

2.2.2. 强化长时序记忆能力将提升 VLA 模型长程任务规划与解决能力

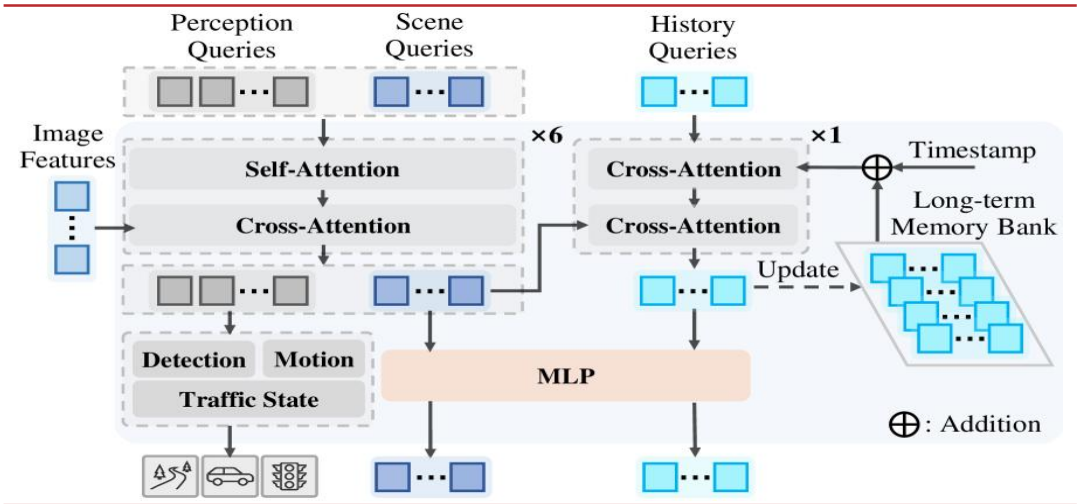
缺乏长时序记忆机制导致模型性能下降。长时序问题是指大语言模型的输入窗口能够保留的信息有限，难以关联长时间范围内的内容，因此 VLA 模型由于主干语言模块缺乏长时序记忆机制，导致模型语义跟随性较差，难以处理需多步规划的任务和行车过程中的长时序遮挡问题，在长流程任务中易出现步骤遗漏或逻辑混乱，导致驾驶行为停滞或无法正确识别目标的现象。

LLM 模型实现长时序记忆的技术难点在于：1）Transformer 架构固有缺陷，标准 Self-Attention 的计算复杂度为 O（N^2），其中 N 为序列长度，导致实际模型能够同时处理的信息有限，造成历史信息丢失；2）即使在同一文本窗口内，也会面临记忆稀释问题，即在长文本输入中，早期的信息可能被赋予较低的注意力分数而被“遗忘”；3）长时记忆不仅要存储过去的信息，还需要动态地更新和清理“过时”或“无关”的内容，对模型的架构设计和训练提出了更高的要求；4）单纯增加输入窗口长度需要更大的显存和更高的计算成本，对于车端模型而言并不经济。

针对上述问题，业界提出了多样化的解决办法，诸如拓展文本窗口、缓存与检索机制、生成段落摘要、动态记忆模块、稀疏注意力等。**我们认为“稀疏注意力机制+动态记忆模块”组合或是较好的车端方案，使模型能在“记忆能力”和“大脑容量”上获得提升：1）稀疏注意力机制**通过选择性关注输入序列中的关键部分来降低计算复杂度和内存消耗，尤其适用于处理长序列数据（如文本、图像、音频），其核心原理是通过引入稀疏连接规则，限制每个查询（query）仅与部分键（key）交互，而非全局计算，从而将 Transformer 计算复杂度从 $O(N^2)$降低到接近线性。例如谷歌 Big Bird 模型通过引入稀疏注意力机制展现了较好的性能，使模型能够处理的序列长度较传统模型提升至约 8 倍，同时显著减少了 GPU/TPU 的内存占用，

提高了模型计算效率，国内理想汽车 Mind VLA 架构中也引入了相似的处理方法。**2）动态记忆模块**通过显示存储、动态更新与历史信息检索改善传统模型的记忆能力，记忆模块相当于给模型外挂一个存储 U 盘，同时通过学习的方式，模块还能自主识别重要信息以进行选择性存储，并根据输入动态地调整存储的记忆数据，小米 QT-Former、理想早期双系统架构中的记忆模块都是该方法的代表。

图表 11：小米 QT-Former 模型架构



资料来源：《ORION: A Holistic End-to-End Autonomous Driving Framework by Vision-Language Instructed Action Generation》_Haoyu Fu 等，华源证券研究所

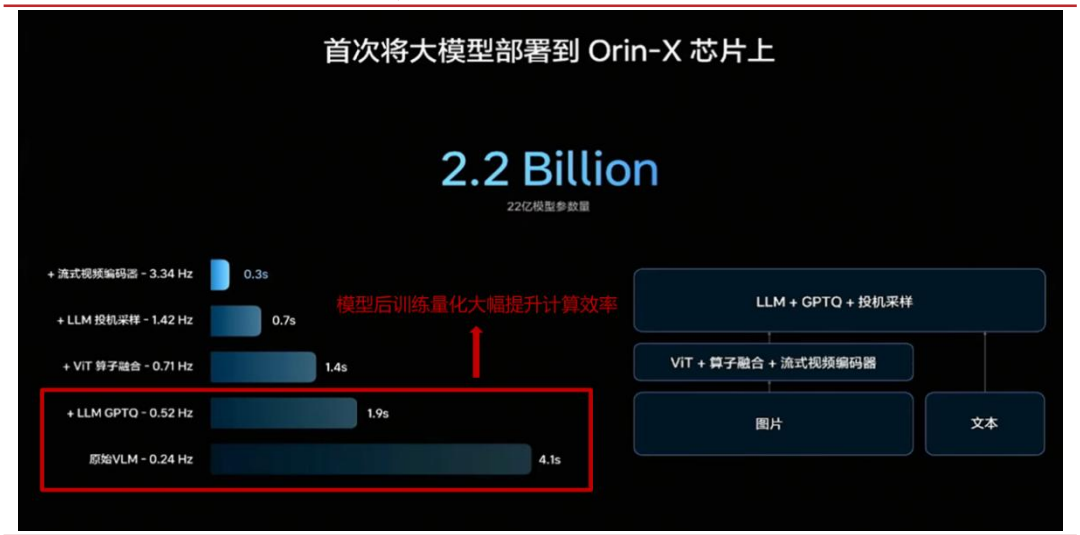
注：Long-term Memory Bank 为动态记忆模块，其通过交叉注意力机制接收感知信息和查询信息作为输入，动态更新信息并输出历史记忆；MLP 模块将更新后的历史记忆与当前的场景特征转化为 LLM 推理空间中的历史标记和场景标记。

2.2.3. 优化模型架构与推理机制可以提高端侧计算效率

端侧模型需在较小参数规模前提下尽可能提高计算效率，“稀疏化”是模型设计的关键。一般而言，模型参数与模型性能正相关，但由于汽车端侧芯片算力不足、带宽较低等因素以及端侧运行实时性需求，云端大模型在端侧部署时需要缩小参数规模和尽可能提升计算效率。除去常见的模型蒸馏、裁剪等缩小参数规模的方式外，模型量化的压缩方式，模型架构优化、推理机制改善等效率提升方式对于端侧部署也同样重要，本段以理想双系统和 Mind VLA 架构的相关技术为例探讨该过程。

模型量化可以降低模型内存空间占用并提升推理速度。模型量化核心思想是降低运算精度，即将模型中的浮点数（通常是 FP32）表示的权重和激活值转换为低精度整数（如 INT8、INT4）或半精度浮点数（FP16），从而实现模型压缩和加速的技术，其主要具有两大优势：1）**降低模型内存空间占用，**如将 FP32 模型量化为 INT8 模型，理论上模型的存储空间需求可以减少为原来的四分之一；2）**加速推理，**低精度计算通常具有更高的计算吞吐量，目前许多硬件平台（如 CPU、GPU）对低精度整数运算有专门优化，可以实现比高精度浮点运算更高的并行度和更低计算时延。量化后更小的模型规模和低精度计算使模型端侧部署算力消耗更小，例如理想 LLM 模型 GPTQ 技术（后训练量化）大幅提升了模型计算效率，使模型时延从 4.1 秒大幅降低至 1.9 秒，输出频率从 0.24Hz 上升至 0.52Hz。

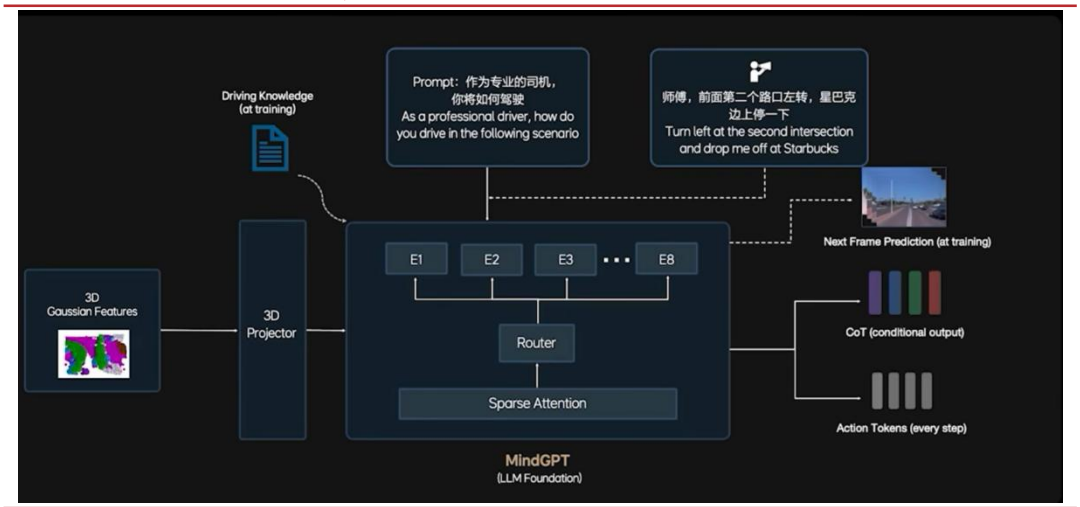
图表 12：模型量化使模型计算效率翻倍



资料来源：云见 Insight 公众号，华源证券研究所

MOE 架构在保持算力消耗相对稳定的同时实现模型扩容，进而提高模型性能。混合专家模型（MOE）是一种“分而治之”的模型策略，核心思想在于将一个大的任务分拆交由对应专家（子模型）处理。例如在 Transformer 架构中将前馈网络层（FNN）替换为一个 MOE 层，MOE 层通常由多个专家模型和一个门控网络（一般是 router）构成，当模型需要解决任务时由学习后的门控网络将任务输送给对应专家，从而实现在整体模型扩容的同时（更多的专家网络加入）其整体的计算消耗与传统稠密计算相当（同时间仅有部分专家被激活，相当于一种稀疏化机制）。例如理想的 Mind GPT 模型中引入了 8 位“专家”做相关计算，每个专家单独训练可以负责其擅长的部分，如图像分割、处理输入的语音指令、动作规划等。

图表 13：理想 MindGPT 模型中 MOE 网络有 E1-E8 8 个专家



资料来源：GTC 2025，华源证券研究所

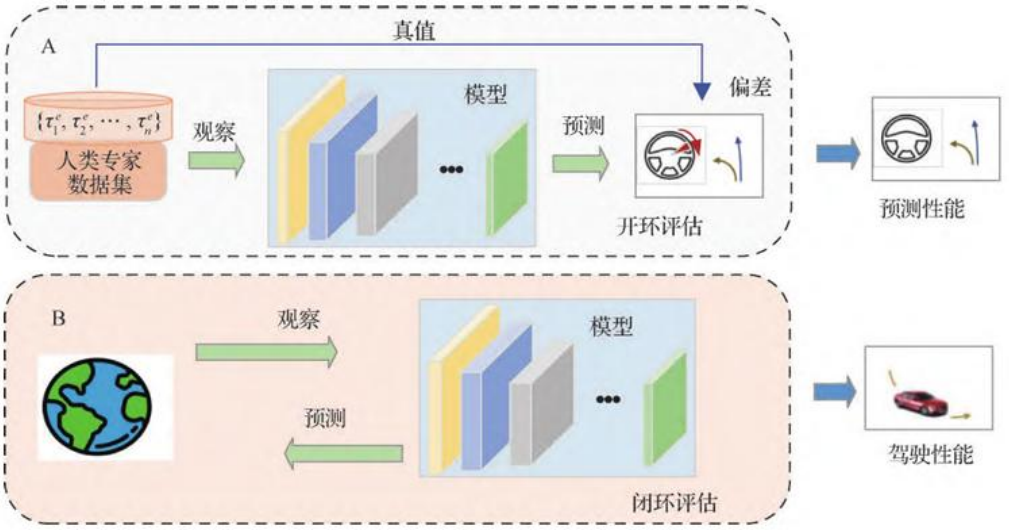
推理效率提升包括稀疏注意力机制（见前文）、投机推理+并行解码等方法。由于 LLM 模型的并行计算能力，可以近似理解其处理一个 token 和一批 token 的效率一致，在此前提下，投机推理机制通过引入一个或多个预训练的、参数较小的模型（draft model）预测生成多个候选 token，然后再利用标准模型对候选词进行批量验证，从而避免了标准模型的重复调用以提升推理效率，该方式的难点在于如何提高小模型采样准确性以避免标准模型验证次数较多；并行解码主要指在 transformer 中加入两种推理模型，如规划决策实时性要求较高的 action token 采用双向注意力机制，通过单次计算即可输出所有信息；对于时效要求较低的语言 token（如对自车行为的解释）则采用因果注意力机制逐字输出，投机推理+并行解码的方法对模型输入和输出两端计算效率提升都起到一定作用。

2.2.4. 利用世界模型构建云端仿真环境是模型闭环测试、强化学习的关键

构建高保真的仿真环境有利于 VLA 模型实现闭环测试验证。自动驾驶模型测评分为开环评估和闭环评估两类，二者核心区别在于模型输出是否有反馈与循环。目前大多数模型采用的公开数据集开环测试为一个单项流程，即传感器信息输入→算法处理→输出结果，最终结果不会产生后续反馈，一般基于预录制的数据对模型某些基础功能（如感知功能）进行测试，适用于初步验证；闭环测试则是一个循环流程，即传感器信息输入→算法处理→输出结果→执行动作和车辆反馈→将反馈作为下一时刻新的信息输入，闭环测试涉及自车与整个外部环境的交互验证、实时的数据处理和决策，更能体现模型在整个行驶过程中的规划决策性能。初步的开环测评与模型实际落地需要的交互性验证、真实测试环境等要求并不匹配，**而直接进行大规模实车闭环验证与强化学习的测评成本和安全性要求难以满足，因此构建逼真的仿真测试环境成为 VLA 模型闭环的关键。**

优秀的仿真模拟环境可以使车端 VLA 模型进行强化学习以达到甚至超越人类驾驶水平。传统的模仿学习中，行为克隆会学习从驾驶环境状态映射到人类专家采取的驾驶动作，核心目标是让模型通过监督学习的方式复制专家行为并逐步改善模型性能，但模仿学习问题在于模型上限较低（严重依赖专家数据）和泛化能力差（corner case 难以处理）。强化学习旨在让智能体与环境不断交互，通过尝试不同的行动来最大化累积的奖励，在自动驾驶领域通过强化学习可以使车辆感知、规控能力进一步优化以达到甚至超越人类专家水平。强化学习方案主要包含智能体、交互环境、奖励函数、动作策略等内容，出于与闭环验证同样的原因，优秀的仿真模拟器对于实现模型强化学习至关重要（提供“真实”交互环境）。**我们认为，未来在具有一个优秀仿真模拟器的基础上，行为克隆（初步训练端到端模型）+逆强化学习（从专家数据中初始化奖励函数）+强化学习（通过与环境交互改善模型和奖励函数）方案或将成为自动驾驶模型训练主流方案。**

图表 14：自动驾驶模型开环评估与闭环评估的结构对比



资料来源：《端到端自动驾驶系统研究综述》_陈妍妍等，华源证券研究所

注：开环评估通过将输出值与专家数据对比能实现单一功能测试，闭环评测更能检验全局驾驶性能

仿真环境构建方法多样，世界模型是未来潜力方向。目前学界对于世界模型没有明确的定义，我们认为通用的世界模型应具有几个特点：1）能够认识物理世界的表象并理解背后的运行规律（因果、物理规律等），并能够基于对物理世界的认识来预测世界的演化；2）能够进行反事实推理，即对于数据中没见过的决策也能推测出结果，具备泛化到样本数据以外的能力；3）具备基于长时记忆进行自我演进的能力。**自动驾驶领域的世界模型即利用历史场景观测信息加上预设条件预测未来智能驾驶场景变化（静态场景、动态交互的变化）和自车响应的模型，其核心任务有三大类：**1）生成未来的物理世界（场景理解、运动预测、场景仿真）；2）生成智能体的场景决策与动作规划（决策规划）；3）将二者合二为一并增加虚拟场景中智能体数量，让智能体之间产生交互从而从单一的物理模拟环境变为交互性的交通场景物理世界（端到端驾驶）。**需要指出的是，由于当前世界模型生成方案尚不成熟，我们认为当前自动驾驶仿真环境的构建中，基于部分真实数据重建+世界模型生成或是工程部署可行性较大的方案。**

目前世界模型在业界的应用主要是场景生成，即作为数据生成器进行仿真环境构建，其可以看作 VLM 模型的逆向工程，构建方案本质上是 Prompt 控制+视频场景生成。

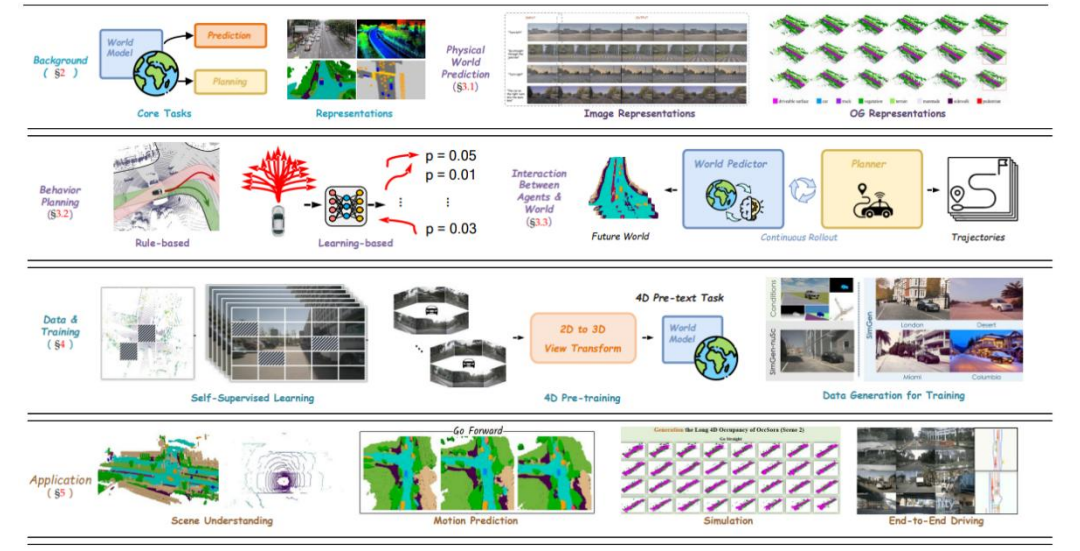
1）**从视频生成的具体步骤，**世界模型场景构建的步骤主要可分为：1）场景初始化，即收集真实的多模态数据并对数据进行标注以形成真实数据的结构化表示；2）控制条件经编码器输入并生成带噪潜在空间表示；3）扩散模型训练与结果输出；4）对生成场景进行优化和后处理。**生成式世界模型方案中控制条件的获取是其中的关键，因为生成式世界模型依赖控制条件（初始帧、相机轨迹、动静态结构化信息）来保证生成场景的环境合理性、资产可控性、物理一致性以及提升渲染效率，这些条件本质上是将人类先验知识注入生成过程，弥补纯数据驱动方法的不足。**

2）**从视频生成模型选择，**主要有对抗式生成（GAN）、Transformer 回归生成、扩散模型等不同技术路线，其中**扩散模型**相较于其余几种模型具备生成质量高且细节丰富、训练稳定性较强、生成结果多样、生成过程可控等诸多优点，成为当前生成式方案的主流。

3）**从视频生成方向角度，**当前的视频生成方向主要有三个：1）更多视角、更高分辨率，如华为 Magic Drive DiT 方案；2）更长时间，如商汤 Infinity Drive 模型能够生成超过 2 分钟的片段；3）高保真、时空一致的 3D 渲染，例如理想《Drive Dreamer 4D》和《Recon Dreamer》，**未来融合以上三种能力是世界模型视频生成发展方向。**

4）**从视频生成优势角度，相较于通常的重建或生成方案，世界模型方案至少具备三项优势：**1）摆脱对于特定的、分布受限的数据来源的依赖，能够渲染复杂操作并保持图像的时空连贯性；2）基于世界模型集成的物理引擎对物理规律的认知，生成的仿真环境除了解决 Vision Gap，还具备物理交互属性，为方案增广至广义具身智能（如机器人领域）提供可能；3）生成方案实现的场景灵活多样，且生成成本较低。

图表 15：自动驾驶中的世界模型综述



资料来源：《A Survey of World Models for Autonomous Driving》_Tuo Feng 等，华源证券研究所

注：
第一、二行为世界模型的背景和关键组件，即未来物理世界的生成、智能体的行为规划，以及它们之间的交互
第三行为自动驾驶中训练世界模型的各类方法，即自监督学习范式、预训练策略，以及数据生成的创新方法
第四行为时间模型在自动驾驶中的四个应用，即场景理解、运动预测、环境仿真和端到端驾驶

我们以理想汽车世界模型相关论文和方案为例，探讨世界模型在自动驾驶领域场景生成方向的落地进展。从整体思路看，理想汽车云端场景构建遵循重建+世界模型生成的思路并发表了 9 篇相关工作论文，其中 2 篇分别介绍 3D GS 的重建及其改进方法，其余 7 篇为生成式世界模型相关工作，整体而言其生成方案大致有以下趋势：

1）初始场景不断完善。从最初 Dive、DriveDreamer4D 方案的图像信息到后续 DrivingSphere、GeoDrive 方案中占用网络、点云信息引入，方案从单纯 RGB 信息到 3D 点云结构渲染的 RGB 图像、从静态场景到动态目标、从主要目标到树枝、房屋等细节再到潜在扩散模型对细节的补足，对于初始场景的刻画更加丰富饱满。初始场景（布局、光照、几何结构等）是后续生成渲然的基础，愈加完善的初始场景可以避免生成完全随机，确保场景生成符合基本物理逻辑也为后续的场景交互编辑提供了更好的基础；

2）生成控制条件升维。方案的控制条件从最初二维道路结构、相机位姿、车辆轨迹等静态信息向场景 3D 点云、占用网络等立体结构再向车辆可控运动等动态信息，最后升级到利用视频输入作为模型生成的指导，利用 3D 渲染、动态信息替代数值控制信号。愈加丰富的控制信息一方面能够显著提升生成场景的真实性，另一方面也为精确的场景控制提供入口，更便于闭环测验中的场景编辑。

3）更为重视闭环反馈机制构建，实现场景实时编辑与模型训练协同规划。在《Driving Sphere》论文中通过智能体协调模块引入，首次实现了自驾模型与环境交互的闭环反馈机制；GeoDrive 模型更进一步，首次在驾驶世界模型中实现场景实时编辑与 VLA 协同规划。

图表 16：理想世界模型相关论文方案总结

论文名	发表时间	基座模块或方法	主要组件	主要解决问题及优化	简要评价
《Street Gaussians: Modeling Dynamic Urban Scenes with Gaussian Splatting》	2024.1.2	3D GS	3D GS重建+4D 球面调和模型+动态物体跟踪位姿优化		
《DiVE: DiT-based Video Generation with Enhanced Control》	2024.9.3	Open Sora	1.跨视图生成模块、跨试图交叉注意力机制、ControlNet-Transformer → 实现对初始BEV布局的精确控制 2.可变分辨率和帧长度、前K帧掩码、光分类器引导输入等 → 更好的训练生成模型，确保生成结果的高质量和连贯性	首个图生视频世界模型，初步解决生成场景多视角条件下的时空一致性问题	初步提出思路：利用prompt+视频生成进行虚拟场景构建
《DriveDreamer4D: World Models Are Effective Data Machines for 4D Driving Scene Representation》	2024.10.17	Drivedreamer-2	1.NTGM（轨迹生成模块）→ 生成多样的轨迹为结构化信息提取提供新视角，并利用世界模型生成新视角视频 2.CDTS（数据对齐模块）→ 将合成数据与真实数据在时间上对齐，减轻4D GS重建中的数据偏差	1.首个利用世界模型推进自动驾驶中4D场景重建的框架，解决了以往场景重建中过分依赖前向数据分布，无法进行复杂建模的问题 2.通过提取新轨迹下道路结构、3D边界框等信息维持了交通元素的时空一致性 3.此时的结构信息更多是2D，场景更多是静态性，场景资产的可编辑性也较差，模型大范围复杂操作能力较差	首次利用世界模型进行自动驾驶4D场景重建
《DrivingSphere: Building a High-fidelity 4D World for Closed-loop Simulation》	2024.11.18	Open Sora（ST-DiT）	1.OccDreamer+Actor Bank → 构建动、静态环境组合，初始化4D环境 2.VideoDreamer → 双路径条件编码提取环境的几何、时空、位置、深度、遮挡等信息进行视频合成 3.智能体编码 → 将动态目标的身份、外观与场景位置绑定，增强时空一致性	1.利用占用网络生成的形式，从BEV视角和文本出发，进一步丰富了场景初始帧和结构化信息，理解为场景初步生成 2.在以往的结构化信息以外，引入智能体库与编码以管理动态目标，进一步实现跨帧与视图的时空一致性和保真度 3.首次支持闭环反馈机制，实现模型与环境的动态交互	利用占用网络技术进行场景初始化首次尝试动态目标管理首次支持闭环反馈与模型-环境交互
《ReconDreamer: Crafting World Models for Driving Scene Reconstruction via Online Restoration》	2024.11.29	Drive Restorer（基于Drivedreamer-2微调）	1.Drive Restorer → 通过微调的世界模型，在线修复减轻生成的重影伪影 2.PDUS → 渐进式数据库更新，实现大范围复杂操作生成	1.基本属于DriveDreamer4D的升级，对生成式世界模型用重建模型生成与真实数据形成的渲染数据集进行微调，提高了模型能力 2.采用渐进式渲染集数据更新策略，提高了大范围复杂条件下的生成能力	升级Drive Dreamer4D
《StreetCrafter: Street View Synthesis with Controllable Video Diffusion Models》	2024.12.17	StreetCrafter（基于Vista微调）	1.StreetCrafter → 基于LiDAR点云数据渲染像素级条件，为扩散模型提供几何引导 2.动态3D GS蒸馏 → 将训练好的生成模型转化为实时渲染的3D场景表示	1.首次将LiDAR数据引入初始环境构建中，不但为后续生成提供精确几何信息，还提供了整个场景的可编辑性 2.首次蒸馏生成模型，生成新视角做为额外监督信号引导3D GS优化，做到生成与重建实时进行，提到了渲染效率	引入激光雷达数据进行初始化 尝试蒸馏模型做到实时渲染
《OLiDM: Object-aware LiDAR Diffusion Models for Autonomous Driving》	2024.12.23	OLiDM框架	1.OPG → 从给定的3D文本、边界框中从目标到场景生成雷达点云数据，且重点区分前景目标和背景环境，解决目标-场景深度问题 2.OSA → 将生成空间划分为不同子空间，进行生成目标-语义对齐，解决空间错位问题	1.首次用世界模型根据文本和3D边界框直接生成点云数据，而非引入LiDAR数据，降低数据收集成本	生成激光雷达数据的数据机器
《Balanced 3DGS: Gaussian-wise Parallelism Rendering with Fine-Grained Tiling》	2024.12.23			不涉及生成，主要是提供了提升3D GS重建效率的方法	
《GeoDrive: 3D Geometry-Informed Driving World Model with Precise Action Control》	2025.5.28	GeoDrive（DiT）	1.MonSTR → 通过给定的图片，预测3D点云几何信息和相机位姿，同时区分动态目标与静态场景 2.动态编辑模块 → 渲染一个动态视频，以生成具有静态背景和移动车辆的渲染结果作为生成过程的视觉指导 3.潜在视频扩散模型 → 生成初始场景中的缺失区域，提高视觉保真度	1.将控制信号由静态的结构化信息升级为带点云几何信息、动态目标可控的视频输入，实现了控制信息的升维 2.首次在驾驶世界模型中实现实时场景编辑与VLA协同规划	控制信号升维首次实现场景编辑与模型训练实时协同

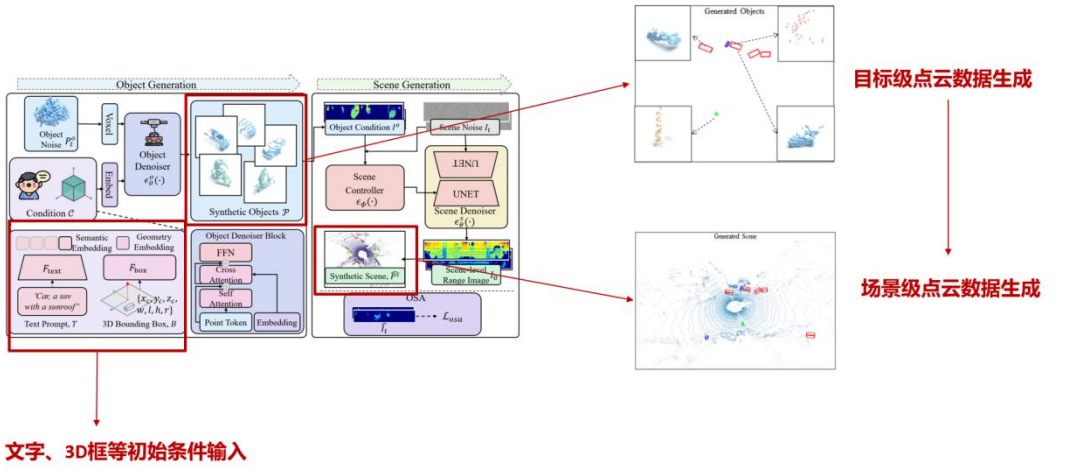
资料来源：华源证券研究所绘制

注：图中所提首次指理想汽车世界模型构建过程中的首次实现

除以上模型构建的整体趋势外，理想的生成式世界模型还可按实际应用方向归类，我们总结如下：**一是作为数据机器用以生成简单的环境数据以弥补真实重建数据的不足，是较基础的数据生成模型。**例如 Dive 模型关注视频数据生成，其利用原始 BEV 构图的 Road Sketch（道路结构）和 Layout Entries（布局条目）进行多视图视频生成；OLiDM 模型则主要解决

激光雷达数据缺乏问题，根据文本描述和 3D 边界框进行从前景目标到环境背景的渐进式生成，并且对生成的前景目标进行初步标注并利用 OSA 模块进行目标–空间语义对齐（例如解决 1 毫米像素空间对应 50m 现实距离的不合理问题），解决了自动驾驶中 LiDAR 数据规模小、标注难、场景多样性不足等问题。

图表 17：理想 OLiDM 模型 LiDAR 数据生成流程



资料来源：《OLiDM: Object-aware LiDAR Diffusion Models for Autonomous Driving》_Tianyi Yan 等，华源证券研究所绘制

二是在视频生成的基础上，进行大范围、多视角、高保真的场景渲染。理想在《DriveDreamer4D》和《Recon Dreamer》两篇论文中提出了 NTGM+CDTS 和 DriveRestorer+PDUS 两个技术集，**差异核心在于 DriveDreamer4D 方案使用公开的世界模型，且在复杂渲染中表现还不尽人意；而 Recon Dreamer 方案中的 Drive Restorer 实际上是一个经过微调的世界模型，并利用 PDUS 方法使得模型在复杂渲染（如多车道变换）中的性能更强。**

DriveDreamer4D 模型主要利用世界模型解决 NeRF 和 3D GS 等重建方案的训练数据依赖性问题，即利用先验世界模型作为数据机器来合成新颖的轨迹视频、利用结构化条件来控制要素的时空一致性以增强 4D 驾驶场景表示。具体而言，DriveDreamer4D 使用轨迹生成模块（NTGM）调整原始轨迹动作（如转向角度和速度）以生成新的轨迹；新轨迹生成以后即可获取新轨迹视角下的道路结构、3D 边界框等结构性信息；最后将结构化信息、新轨迹初始帧、文本控制信息等输入到世界模型以生成跟随新轨迹的视频；**除数据生成外，DriveDreamer4D 也关注生成数据与真实数据的对齐问题，**即利用 CDTS 在每个时间步上利用提取的结构化信息作为约束，将真实数据与生成数据进行对齐以减轻 4D GS 训练中的数据差异，具体表现为消除最终生成视频中的“鬼影”、“重影”现象。

图表 18：理想 DriveDreamer4D 模型生成效果与传统方式的比较

资料来源：《DriveDreamer4D: World Models Are Effective Data Machines for 4D Driving Scene Representation》_Guosheng Zhao 等，华源证券研究所

注：1、3 行方案中，左上角生成视图有明显“鬼影”；2、4 行方案生成效果更优

Recon Dreamer 框架通过引入 Drive Restorer 模型和 PDUS 策略来解决大范围机动下的“鬼影”问题，本质是经过自动驾驶数据微调后世界模型方案能力的进一步提升。Drive Restorer 实质上是一个扩散生成模型，理想利用未充分训练的重建模型沿自车原始轨迹渲染低质量视频并与真实视频对比形成渲染恢复数据集，以真实视频数据为监督训练 Drive Restorer 恢复渲染视频中的鬼影，并且为了增强模型能力，还对天空、图像边界等重点区域进行了掩码操作。PDUS 是一种渐进式数据更新策略，其作用类似于自驾模型的动态记忆模块，即在新轨迹生成过程中对于初始的渲染恢复数据集进行动态、分部的更新，再由 Drive Restorer 处理得到新轨迹视频，以此迭代直到模型收敛并最终提升模型在大范围机动复杂条件下的场景生成能力（即将长距离生成分解为逐步更新生成问题）。

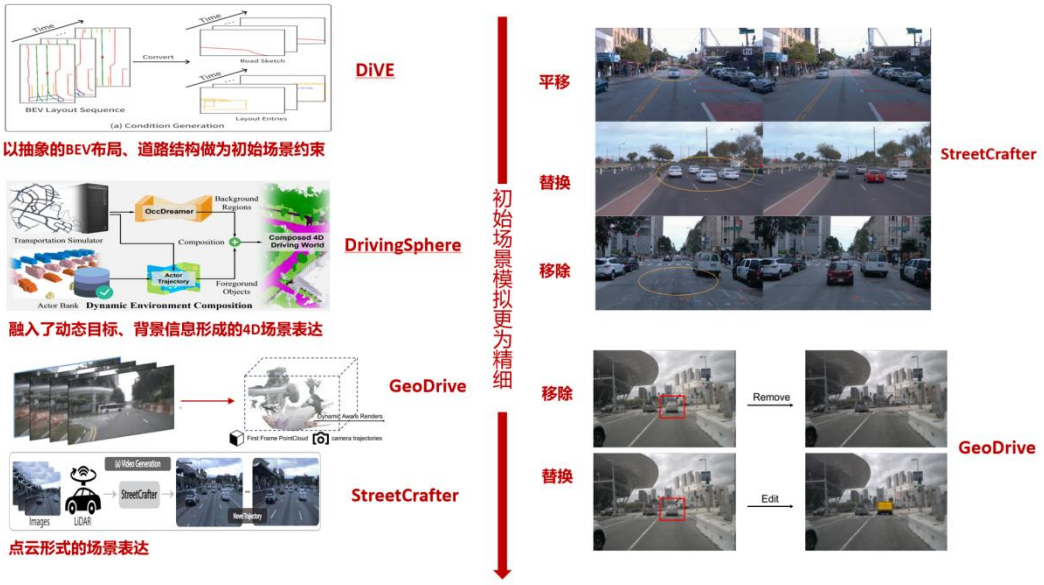
图表 19：理想 Recon Dreamer 模型长距离街景生成效果与传统方法的比较

资料来源：《ReconDreamer: Crafting World Models for Driving Scene Reconstruction via Online Restoration》_Chaojun Ni 等，华源证券研究所

注：第一行为场景真值，最后一行为 ReconDreamer 生成效果图，在 3 米级车道变化中，其生成效果更优；在 6 米级大范围变化中，其生成效果显著更优

三是进一步提升对初始化场景的精细刻画与场景编辑能力。如前所述，初始化场景的精细刻画对于后续视频生成具有基础性作用，《DrivingSphere》中通过 BEV 条件扩散模型 OccDreamer 生成城市级静态场景，结合动态交通参与者的时空位置管理，能够构建包含静态背景和动态对象的精细化 4D 世界表示；GeoDrive 模型则以单帧 RGB 图像为输入，借助 MonST3R 网络精准估计点云和相机位姿，并结合用户提供的轨迹信息构建具有三维一致性的条件序列，确保场景结构连贯真实。同时得益于场景刻画中点云信息的引入，使得部分生成模型可以在多帧点云聚合期间调整物体边界框的属性，以提供经修改的 LiDAR 条件用于视频扩散模型，而无需对每个物体分别建模且逐场景优化，即实现场景动态编辑功能，**动态场景编辑的实现奠定了模型高效训练闭环反馈的基础，例如 GeoDrive 模型首次在驾驶世界模型中实现实时场景编辑与 VLA 协同规划。**

图表 20：理想相关生成模型场景刻画与场景实时编辑



资料来源：《DiVE: DiT-based Video Generation with Enhanced Control》_ Junpeng Jiang 等，《DrivingSphere: Building a High-fidelity 4D World for Closed-loop Simulation》_ Tianyi Yan 等，《GeoDrive: 3D Geometry-Informed Driving World Model with Precise Action Control》_ Anthony Chen 等，《StreetCrafter: Street View Synthesis with Controllable Video Diffusion Models》_ Yunzhi Yan 等，华源证券研究所绘制

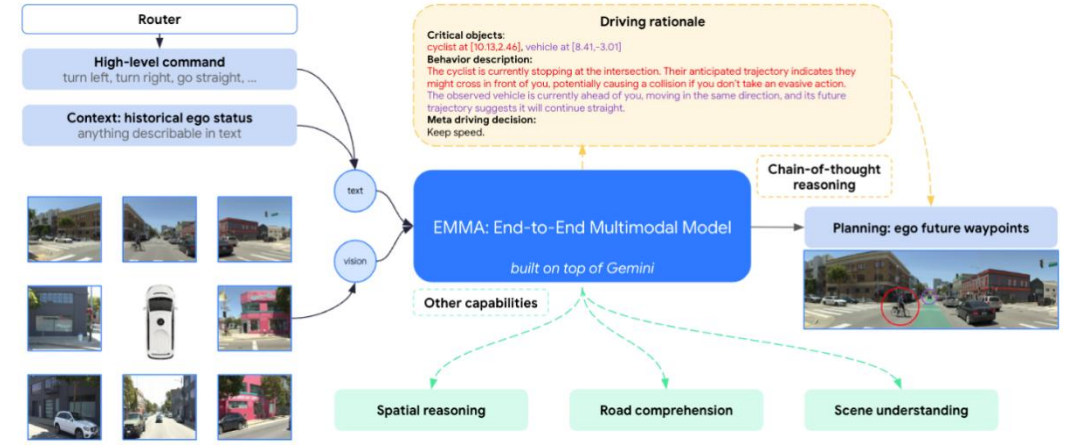
2.3. 自动驾驶典型 VLA 架构

2.3.1. Waymo EMMA：开创性的端到端多模态自动驾驶模型

作为早期开创性模型，EMMA 模型架构较为简单，主要由编码器+大语言模型构成。感知部分，EMMA 模型开创性的同时处理文本（导航指令、自车历史状态）、图像（摄像头视频感知）等多模态输入，并利用视觉-语言框架将所有的输入和输出表示为普通文本，将驾驶任务转化为视觉问答（VQA）问题，充分利用其 Gemini 大语言模型储备的大量知识，更好的理解驾驶任务中的动态变化；推理输出方面，为了增强模型的推理能力使之更符合自驾需求，EMMA 对原有大语言模型进行了微调，通过混合训练来实现更多自驾能力，具体而言，其将感知任务拆分为空间推理、道路图估计、场景理解等多个子任务，使微调的 LLM 模型能够更好的生成各种运动规划和驾驶控制信号。

EMMA 框架具有三大特性。1）EMMA 将所有的输入和输出表示为自然语言文本形式，所有任务共享统一文本表示空间，可以最大限度的调用语言模型的知识储备并提供了将其余驾驶任务继续融入系统的拓展性；2）引入 CoT 增强模型的可解释性，EMMA 将 CoT 融入到轨迹生成中，要求模型在预测时阐明相关理由，例如将推理过程结构化为场景描述、关键物体描述、关键物体行为描述、驾驶决策输出四个子任务，数据集测试结果显示，引入 CoT 的模型相较于基准模型整体性能提升了 6.7%，在驾驶决策和关键物体识别的能力上分别提升 3.0%和 1.5%；3）自监督模型，模型训练唯一需要监督数据的是自车未来位置，其余数据不需要专门人工标签，提高了数据来源的可拓展性。

图表 21：EMMA 模型架构



资料来源：《EMMA: End-to-End Multimodal Model for Autonomous Driving》_ Jyh-Jing Hwang 等，华源证券研究所

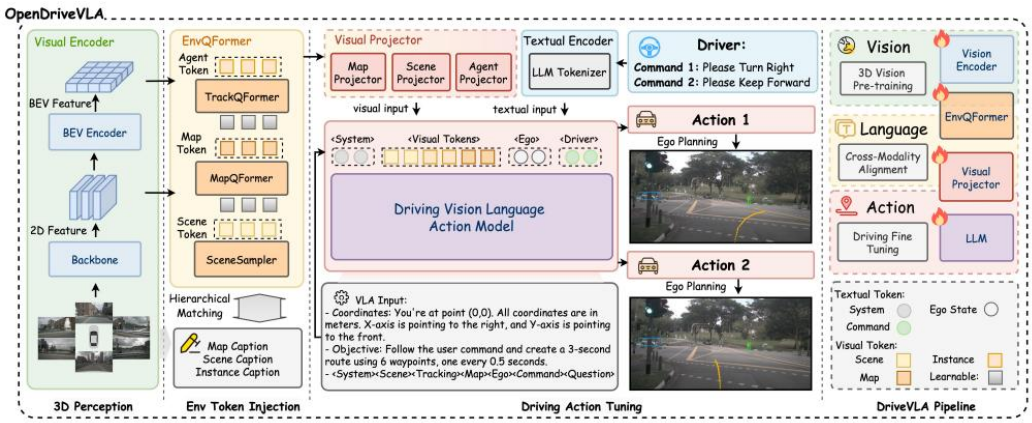
EMMA 在公开数据集的开环测试取得了较好效果。EMMA 采用最小尺寸基座模型 Gemini 1.0 Nano-1 分别在 WOMD 和 nuscens 数据集上进行了端到端轨迹预测的测试。在 WOMD 数据集中，经过内部预训练的 EMMA+模型在短时间窗口上的 ADE（平均位移误差）性能超越了基准模型，但在较长时间窗口表现较差，主要是 EMMA 只有摄像头输入，基准模型结合了激光雷达，深度感知能力更好；在 nuscens 数据集中，自监督的 EMMA+取得了 SOTA 效果，比参与测评的监督基准模型平均性能提高 6.4%，比自监督的基准模型性能提高 17.1%。

作为自动驾驶 VLA 的初步尝试，EMMA 距离工程部署尚有距离：1）模型仅能处理有限帧数，难以捕获驾驶任务所需的长时间依赖关系，自动驾驶性能较差；2）依赖预训练的多模态模型，但该模型未集成与点云相关的编码器，3D 空间感知和多模态能力受到限制；3）当前测评基于公共数据集上的开环测试，模型闭环性能不清晰，距离工程部署尚有距离；4）参数规模庞大的语言模型在车端部署对端侧芯片算力、带宽带来挑战，车端推理实时性不足，需要在模型大小、推理质量、推理效率之间实现平衡。

2.3.2. Open Drive VLA 框架的贡献在于模型 3D 环境感知和交互

Open Drive VLA 是专为自动驾驶设计的端到端 VLA 模型，主要包含一个预训练的视觉编码器和一个开源 VLM 模型。模型首先利用预训练的编码器从多视图图像中提取中间特征；然后分层视觉语言特征对齐模块将图像 token 对齐到文本域；其次在 VLM 推理空间中进行车辆-环境-自车交互推理和输出高层次的驾驶指令，最后根据高层次指令给出自车的未来轨迹。**架构的创新在于 1）引入以视觉为中心的查询模块和分层视觉-语言特征对齐模块，提升模型 3D 感知能力；2）引入条件车辆运动预测任务，提升自车复杂环境下交互能力。**

图表 22：Open Drive VLA 模型架构

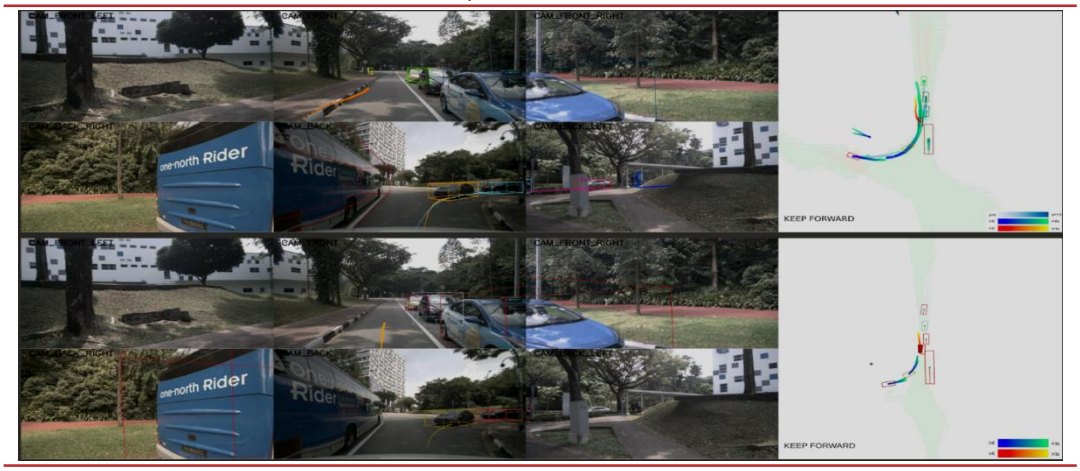


资料来源：《Open Drive VLA: Towards End-to-end Autonomous Driving with Large Vision Language Action Model》_ Xingcheng Zhou 等，华源证券研究所

3D 环境感知与对齐：传统的 VLM 模型通常依赖于 2D 视觉编码器，视觉 token 的选择和注意力权重通过语言监督间接引导，模型缺乏足够的 3D 空间感知能力会造成严重的多模态输出幻觉（即语言模型的反应与图像输入内容不一致）。针对此问题，Open Drive VLA 在感知环节采用了以视觉为中心的查询模块，使模型重点关注与驾驶相关的物体和 3D 地图信息，具体而言在模型获得 BEV 特征表示后，会用三个视觉查询模块（Track、Map、Scence）以空间定位的方式捕捉动态车辆行为和静态地图结构，以获得 3D 中间特征表示。并利用分层视觉–语言对齐机制弥补不同空间的模态差距，即指针对三个特定的查询模块引入三个特定的可训练投影机制进行视觉嵌入，使得不同模块的视觉信息都有详细的语言描述与之对应，达到对齐视觉和语言模态的效果，例如对于 Map 信息，以真实标注数据训练的文本转化机制可以将车道分隔线、人行横道和道路边界等地图元素都转化为描述性文本。

轨迹生成与环境交互：Open Drive VLA 引入了条件车辆运动预测任务，作为 3D 车辆–环境–自车交互建模的代理任务，使模型能够学习不同物体在空间中的运动模式，即模型能够在给定场景描述、地图结构以及自车状态后，在推理空间中直接预测每个实体相对于自车的未来位移，并给出自车在此条件下的未来运动轨迹预测。该任务的引入增强了模型轨迹生成能力，并改善了复杂交通场景中的决策能力。从开环评测效果角度，如下图所示，相较于 UniAD 模型，Open Drive VLA 对环境的感知能力更强，没有对周遭车辆的过度反应，生成的轨迹更为平滑。

图表 23：引入条件车辆运动预测任务后，预测通过时延更低



资料来源：《Open Drive VLA: Towards End-to-end Autonomous Driving with Large Vision Language Action Model》_Xingcheng Zhou 等，华源证券研究所

注：下图为 OpenDrive VLA 模型效果图，上图 UniAD 对周遭环境有过度反应（颜色更多，预测阻碍时长越长），而 OpenDrive VLA 有效保持了轨迹平滑性和环境感知能力，展现出其在处理复杂驾驶场景时更强的能力

OpenDrive VLA 仍面临诸多问题。1）为了平衡模型推理速度和计算开销，LLM 模型采用隐式推理，缺乏明确的 CoT 过程，导致模型在复杂场景中的推理能力和模型的可解释性较差；2）目前的测评仍是开环评测，后续的闭环测试和仿真场景搭建预计仍然存在困难；3）模型的自回归特性阻碍了高速场景中的实时推理。

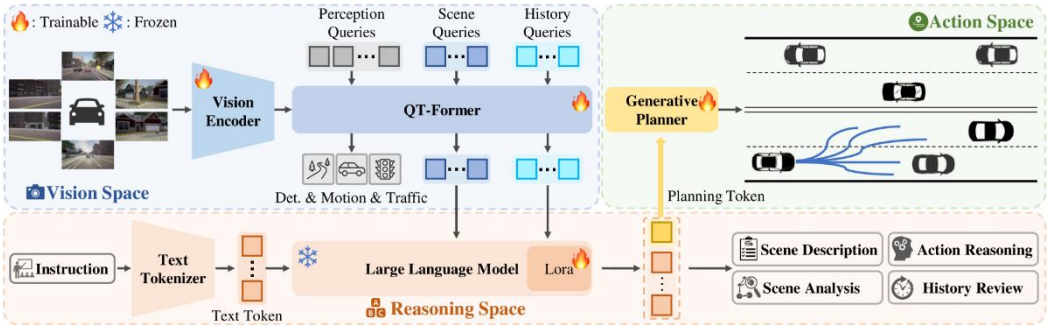
2.3.3. 小米 ORION 框架引入 QT-Former 模块实现了长时序记忆

小米 ORION 架构是典型三段式 VLA 架构，主要由三个关键组件构成：QT-Former、LLM 及生成式规划器。首先通过视觉编码器对图像编码；其次 QT-Former 实现长期上下文提取并连接视觉空间与 LLM 模型的推理空间；LLM 将场景特征、历史视觉信息、用户指令等多模态信息结合执行推理任务并预测一个规划标记；最后生成式规划器生成由规划标记条件约束的多模态轨迹。该框架利用 QT-Former 和生成式规划器分别连接了视觉-推理、推理-动作空间，实现了从图像感知到视觉问答再到动作规划的统一端到端优化，**模型创新之处在于 QT-Former 动态记忆模块的引进一定程度解决了长时序记忆问题以及 VAE 模块优化了轨迹生成。**

QT-Former 模块实现图像压缩和长时序建模。通常 VLM 模型要求输入的都是高分辨率图像，但高分辨率图像 token 化后计算量较高，不能保证端侧模型输出的实时性，因此小米引入了 QT-Former，其类似一个信息筛选机制，负责提取对语言文本生成最有用的图像特征并压缩转化为 LLM 可以理解的 token。**长时序建模层面，**传统 VLM 模型一般利用拼接多帧图像进行时序记忆，这种方法受制于 Token 长度，QT-Former 引入了动态记忆模块和历史查询机制一定程度上解决了长时序记忆的问题。其运作机理为初始化感知 Query 与场景 Query，首先原始感知 Query 与场景 Query 先通过自注意力模块交换信息；然后与带有 3D 位置编码的图像特征执行交叉注意力并分别获得感知结果及新的场景 Query，其中感知结果被输入至

任务头用于各项任务，新的场景 Query 与 long-term Memory Bank(记忆库)中的历史 Query 再执行交叉注意力以不断地更新历史 Query 并按照先进先出的替换原则再存储到记忆库中。**其创新之处在于，不同于以往记忆模块只简单存储压缩后信息而不关注提取当前场景信息的机械机制，小米通过初始化少量历史 Query，能够进一步提取与历史信息最密切相关的当前场景特征，增强了模型的长期记忆能力。**

图表 24：小米 ORION 模型架构



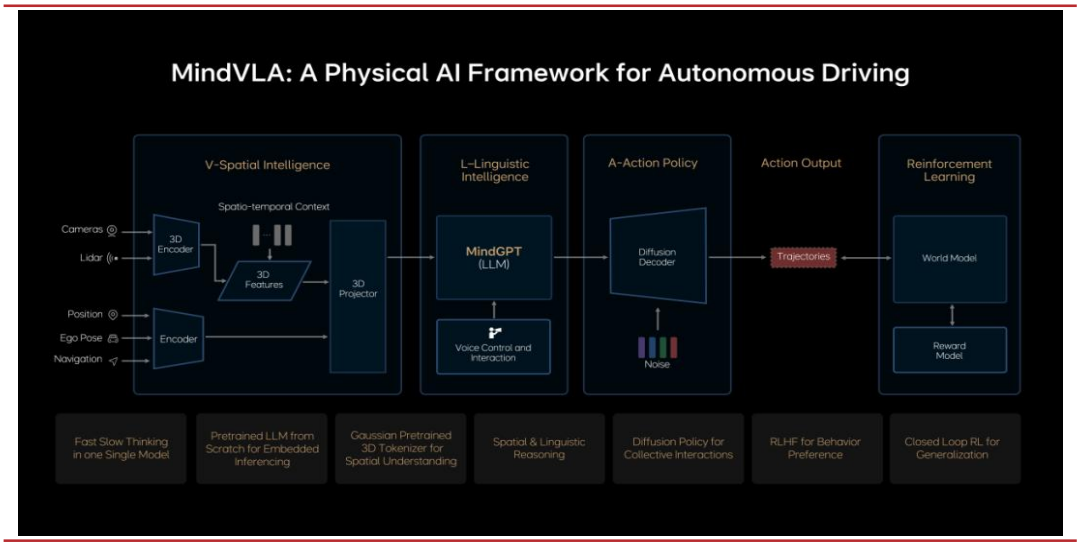
资料来源：《ORION: A Holistic End-to-End Autonomous Driving Framework by Vision-Language Instructed Action Generation》_Haoyu Fu 等，华源证券研究所

小米 ORION 架构的工程化部署面临挑战。根据小米公开数据，ORION 架构在 Bench2Drive 数据集上获得了较好的闭环测试性能，获得了 77.74 的驾驶分数和 54.62%的成功率，相较于 SOTA 方法分别增长 14.28 分和 19.61pct 的成功率。但我们认为其距离商用落地仍有较多工作：1）基座模型使用开源模型 Vicuna v1.5，没有针对自动驾驶做微调，不同模块间的配合及针对自驾任务的性能可能不足；2）LLM 模型参数规模庞大，端侧推理实时性难以保证。目前可行的路径之一是将 QT-Former 与 VAE 模型连接，将 LLM 模型用作辅助推理，形成实质的双系统模式，在端侧芯片能力足够和模型运算效率优化以后再部署全局端到端；3）图像编码器或仍沿用传统的 2D 网络，模型的 3D 空间理解能力较弱影响模型性能。

2.3.4. 理想 Mind VLA：深度融合空间、语言及行为智能

Mind VLA 六大关键技术，构建自驾模型新范式。24 年 10 月理想汽车双系统架构正式推送，但彼时的模型架构尚存在诸多问题，例如双系统联合优化困难、基于开源的 VLM 模型在 3D 空间理解能力上仍然不足、模型的多模态性处理不足(输出方式为 Transformer 回归建模，难以处理驾驶行为多模态性)、人类价值观对齐不足等。基于双系统实践和对前沿技术的吸收，理想汽车推出了自研 Mind VLA 模型，提出了 6 大关键技术：3D 空间理解能力构建、基础语言模型构建、语言模型推理效率优化、Diffuison 轨迹生成、RLHF、云端 world model 强化学习。**其模型方案可以理解为：利用 3D 空间编码器编码环境特征输入至语言空间，语言空间利用逻辑推理能力和空间理解能力将输入信息处理后给出合理的高层级 action token，然后通过 diffusion 模型进一步优化出最佳的驾驶轨迹，实现空间智能、语言智能、行为智能的统一。**

图表 25：理想 Mind VLA 模型架构

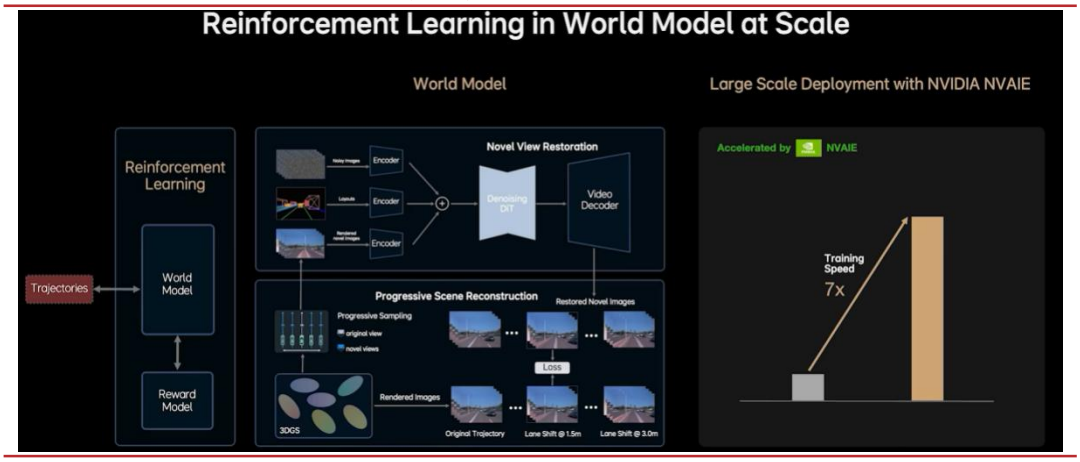


资料来源：GTC 2025，华源证券研究所

基础语言模型重构、RLHF、云端 world model 强化学习主要解决模型计算效率与类人性问题。除了从算法优化的角度提升模型计算效率，从模型自身构建角度是更为根本的解决方法，当前 VLM 一般是由开源 LLM+预训练 Vision encoder 构建，这类方法简便易行，但基于互联网数据训练却限制了模型 VL 部分的实现效果，一方面是开源 LLM 虽经过微调与后训练，但模型构成中仍有大量与自驾无关的参数占用硬件资源；另一方面是基于开源数据训练的 Vision encoder 无法充分利用自驾领域成熟的感知网络。**理想重新配比了 3D 数据、自动驾驶相关图文数据与文史类数据的比例，从零训练语言模型并自定义自驾专用 LLM input tokenizer，根本上选择了更为困难但上限更高的路径以解决模型效率问题。**

RLHF 与云端世界模型强化训练是模型后训练环节，主要解决模型类人性问题。RLHF 方法是通过筛选大量 NOA 接管数据（不符合人类预期的表现）以建立人类偏好数据集，使模型从特定的偏好数据中学习对齐人类行为，提升模型的安全下限。同时 MindVLA 基于自研的重建+生成云端统一世界模型，深度融合重建模型的三维场景还原能力与生成模型的新视角补全，以及未见视角预测能力，构建接近真实世界的仿真环境实现了基于仿真环境的大规模闭环强化学习，并利用工程化能力将 3D GS 的训练速度提升了 7 倍以上。通过创新性的预训练与后训练方式，Mind VLA 实现了优秀的模型表现与泛化能力，预计将成为部署与量产最快的车端 VLA 模型。

图表 26：理想 Mind VLA 后训练环节世界模型框架



资料来源：GTC 2025，华源证券研究所

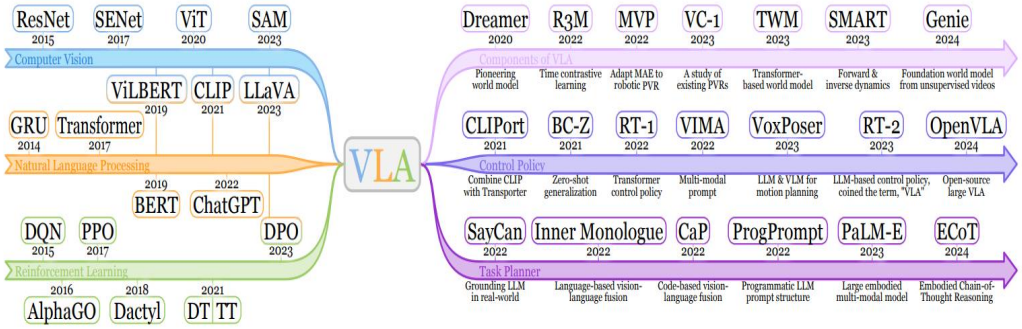
3. 具身智能本质是自动驾驶的升维问题，构建数据闭环是关键

3.1. 机器人 VLA 架构的发展历程

机器人智能化的尝试由来已久，VLA 方案或是未来通解。1966-1972 年，斯坦福研究院开发出“Shakey”，它可以理解简单的英文指令并自主避障，被视为机器人“有思考能力”的开端；2013 年，DeepMind 的 DQN 算法让学术界第一次意识到深度学习可以将机器人视觉与动作用同一张神经网络实现，并由此催生了一批“专用”机器人，但彼时机器人泛化性能极差，常因为光照条件变化或物体的稍微形态变化就宕机；2021 年 Open AI 的 CLIP、Google 的 ALIGN 将海量图片与文字对齐至同一嵌入空间，具身智能进入 VLM 时代，但仍缺乏对物理动作的直接控制能力；2022 年，Google、CMU 相继推出“SayCan”、“Instruct2Act”等工作，使模型既能看图、又能理解指令并生成动作轨迹成为可能；2023 年，DeepMind 在此前基础之上，正式推出 RT-2 模型，使机器人可以从给定的语言指令和视觉信号，直接生成特定的动作，一举奠定了 VLA 的模型范式，使机器人从“看得见”、“听得懂”正式走向“动得准”的第三阶段。

近年来对于机器人 VLA 的探索主要集中于如何进行高效数据采集与模型架构优化两个方向（本文中机器人更多代指人形机器人方案）。架构层面，从谷歌无预训练的 RT 系列到 Open VLA 到 Helix、ViLLA 等，模型发展依托于 VLM 进步，随着开源 VLM 架构优化与训练方法革新，相关成果拓展至 VLA，同时 Flow Matching、Diffusion 等技术提升了机器人动作生成能力，**模型整体架构呈现出多模态能力更强、动作生成更精细化、泛化能力更优秀、一段式向双系统演变等发展趋势。数据层面，**从谷歌私有数据集，到 Open X-Embodiment、AgiBot World 开源真机数据，再到仿真数据和互联网人类视频数据的引入，国内外机构积极探索，通过新采集、历史数据处理、合成数据等方式丰富数据源。**数据质量决定模型上限，数据是 VLA 模型发展最根本、持久的驱动力。**

图表 27：具身智能 VLA 模型发展历程



资料来源：《A Survey on Vision-Language-Action Models for Embodied AI》_Yueen Ma 等，华源证券研究所

注：左侧由上至下分别为计算机视觉、自然语言处理、强化学习；右侧由上至下分别为 VLA 主干模型、控制策略、任务规划器

图表 28：机器人智能化模型数据金字塔

资料来源：《GR00T N1: An Open Foundation Model for Generalist Humanoid Robots》_NVIDIA，华源证券研究所

3.2. 应用场景与任务的差异决定了车端 VLA 与机器人 VLA 的核心差异

我们通过对机器人 VLA 与汽车 VLA 进行对比分析以探究机器人 VLA 数据闭环构建中面临的突出问题，**本质上汽车 VLA 与机器人 VLA 所面临的应用场景与任务的不同决定了二者拥有不同的商品属性和标准化程度。**汽车可以被视为一种低自由度的特殊机器人，所面临的场景和任务是结构化场景下的单一任务，根本上决定了汽车是感知方式、输出控制、自由度、本体结构都相对标准化的耐用消费品，底层硬件结构的统一性和场景、任务单一性决定了其数据采集方式、所采集的数据、模型设计的标准化；而广义的机器人由于所面临的场景与任

务需求多样化，更多偏向于非标的可选消费电子类产品，更需要在模型设计、本体结构标准化、数据采集效率、任务覆盖范围等多个条件中进行权衡。二者场景、任务的不同，决定了二者商品属性和标准化程度不同进而决定了二者上层结构的差异。

我们尝试从数据、仿真环境、端侧算力等具体角度对比汽车 VLA 与机器人 VLA 的差异性。整体来看，机器人的数据闭环构建难度远超车端，其工程部署还需要解决标准化、本体交互能力、模型闭环等重要问题，机器人 VLA 所面临的各种问题汇集在一起突出表现为当前还无法进行有效的数据收集进而构建完整数据闭环，而无法 Scaling 的具身就无从谈起智能化，**因此我们认为机器人 VLA 模型或智能化当前还处于前期探索阶段，相较于汽车 VLA 专注于工程化，机器人 VLA 更是一个科研问题。**

3.2.1. 机器人 VLA 训练所需的数据规模或远超车端

机器人 VLA 模型所需的数据更为复杂多样，核心原因在于机器人所面临的场景和任务更为多样化，泛化能力要求远高于汽车。1）应用场景与任务不同，自动驾驶的汽车可以被理解为一个特殊的机器人，其所面临的结构化道路场景和执行的驾驶任务都较为单一；机器人如人形机器人所面临的场景和任务非常丰富，部署场景从家庭、工厂等封闭式场合到公共服务场合等开放式场景，任务从家政服务到工厂务工，几乎能囊括人类日常生活的各方面，因此从训练数据的多样性角度，其数据需求远超汽车；**2）模型能力要求不同，**从应用场景和任务出发，车端 VLA 重点在于提升端侧推理效率以增强模型的动态博弈能力，更加注重感知数据输入和 2D 轨迹规划输出以及模型实时决策，其数据维度较低（可以理解为一个动作专用模型利用同构低维数据，强化学习反复迭代提升性能）；机器人 VLA 则更为注重模型的泛化性，其所需求的数据除感知数据以外，更为关注本体与真实世界的物理交互数据（如力反馈、摩擦力数据等），且更多输出 3D 空间动作规划，所需数据的维度、复杂度更高。因此在假设完美完成一类任务所需数据规模差异不大的前提下，机器人所需的数据从多样性到复杂度相较于车端都有较大提升。

图表 29：人形机器人与汽车所面临的场景、任务丰富度不同

资料来源：Figure AI 官网、央广网、创新南山公众号、张江发布公众号，华源证券研究所绘制

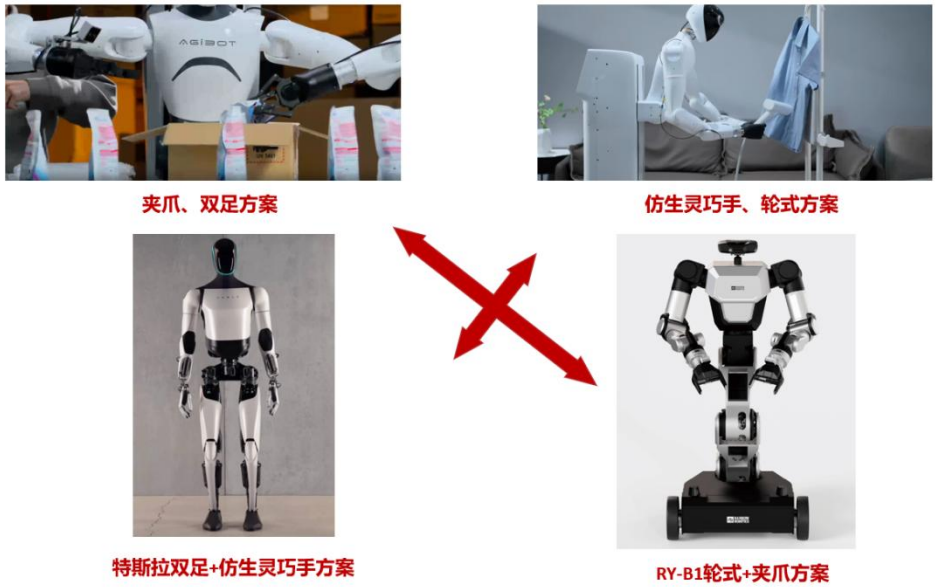
3.2.2. 硬件方案未收敛与本体高自由度限制了真实数据收集

现有公开机器人数据集相较于车端数据非常匮乏。以特斯拉为例，其在多个场合提到FSD训练依赖于上千万个视频片段，累计时长达到几万小时，相较而言，目前人形机器人领域较大的公开数据集如 X-Embodiment、AgiBot World 等，视频片段多在百万级规模。机器人硬件方案未收敛和本体高自由度导致了真实数据采集困难。

1）硬件方案未收敛导致数据孤岛。目前广义的人形机器人硬件结构尚存在不确定性，例如本体存在轮式方案与双足方案，手部结构如夹爪、灵巧手、三爪等机械结构尚未确定；传感器方案中视触觉、磁变传感器的方案选择与具体排布位置也未有定论。且当前人形机器人数据采集方法多数集中在关节层数据，如各自由度的角度、速度、力矩等，硬件不统一导致不同机器人关节结构差异较大，采集的数据具有极强的平台依赖性，可复用性低，数据孤岛问题极大提高了数据采集成本，使得产业端难以通过规模化降低数据成本。

2）高自由度导致数据采集效率低下。人形机器人与自动驾驶在自由度上差异最为显著，汽车仅有前后、速度等两三个自由度，而人形机器人灵巧手+单臂自由度可达 20+，若双臂操作+全身控制，其自由度很容易达到四五十个。高自由度带来诸多问题：1）**本体高自由度的累计传递误差带来的本体精度问题以及机器人本体-人体自由度不匹配导致映射算法构建困难**，进而导致真实数据采集精度不够，有效数据比例较低；2）较高的自由度对于当前主流的遥操、动捕等真实数据采集方式的人员及设备提出了很高的要求，提高了数据采集的成本。

图表 30：不同人形机器人本体构型尚未确定



资料来源：智元机器人、中关村在线、Rainbow Robotics 等，华源证券研究所绘制

3.2.3. 算力解放是技术进步的前提

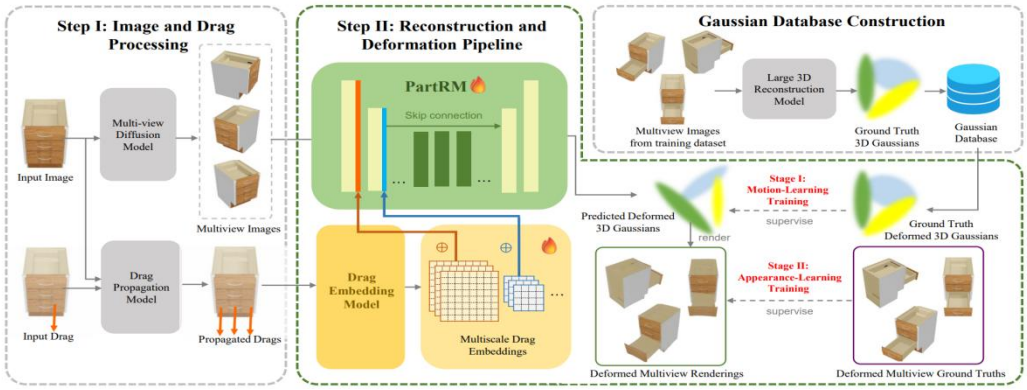
1）模型发展需要与底层算力匹配，算力解放是技术进步并应用的基础。车端以理想汽车 VLA 为例，经过权衡端侧算力与计算效率，理想率先推出了双系统架构，其后 Orin-X 向 Thor 的迭代为双系统向 VLA 的迭代创造了算力基础，人形机器人目前较为流行的双系统架构也是行业权衡端侧算力与推理效率之后的选择。**2）机器人 VLA 模型算力需求更多但端侧部署环境更为严苛。**一方面机器人面临的任务和场景更多，云端模型训练面临的数据处理需求和所需的算力需求更多；另一方面，机器人端侧相较于汽车，芯片运行工况更为恶劣，需同时处理的交互任务、自由度计算更多，散热、体积、功耗等要求更为严格。**目前行业多采取高芯低频、部分计算核睿频的方式进行端侧计算，我们认为背后核心原因是行业不成熟导致供应链不成熟所致，即缺乏一款专为机器人设计的芯片以适应机器人的工况和算力需求，目前机器人芯片更多是其余行业芯片改制而来（如英伟达 Jetson 平台），未来尖端制程下放和专用芯片研发成功或为机器人技术的快速迭代提供算力基础。**

3.2.4. 构建可供机器人使用的仿真环境需要注重可交互性建设

机器人 VLA 对于仿真环境的要求比车端更高，需要更注重交互能力建设。构建优秀的仿真环境对于机器人智能化非常重要，一方面，仿真模拟器可以提供仿真数据供模型进行基础预训练，一定程度弥补真实数据不足；另一方面，优秀的仿真模拟器也能提供相对多样化、低成本、可拓展的强化学习环境。**构建优秀的机器人仿真环境需要解决三个问题：视觉差异、资产可微分、物理动力学，**对于汽车环境的构建，视觉差异是主要矛盾点（较少的涉及物理交互，一旦交互可大致判定为出现 corner case，策略失败），**而机器人因为天然的交互属性，需要更为注重交互能力建设。**

目前在机器人仿真领域已有英伟达 Issac、清华 DISCOVERSE 等不同的仿真平台，但存在仿真数据简单、底层物理规律与物体材质仿真困难等问题，**如何解决物理规律问题是仿真领域即将面临的核心矛盾，我们认为从模型和数据角度，未来世界模型+多模态数据或是实现可交互仿真环境建设进而实现机器人物理可交互性的两条路径。**世界模型仿真路线尚未有确切的实现路径和定义，4D GS（3D GS+时间轨迹）+生成式方法或是潜在方案，例如理想 Recon Dreamer、GeoDrive 模型、清华与北大合作的 PartRM 模型，以 PartRM 模型为例，其将物体施加力到完全移动形变的过程视作一个 4D 数据，并广泛收集 4D 数据集去训练 3D GS 的形变过程以获得可微分的资产，其将动力过程视为隐式内容学习使构建的环境具备物理属性，虽然隐式学习的方法从实现效果上可能略逊于显示约束的动力学模型，但是利用视频数据作为输入和自监督的方式从 learning base 角度，这类方法或更具备 scaling 潜力。

图表 31：PartRM 模型框架，通过观察预测形变与真实形变的差值进行隐式学习



资料来源：《PartRM: Modeling Part-Level Dynamics with Large Cross-State Reconstruction Model》_Mingju Gao 等，华源证券研究所绘制

从数据角度，引入触觉信息实现多模态感知进而构建运动反馈也是构建交互能力的重要方法，当前主流传感器中，图像和深度传感器（激光雷达等）实际都是视传感一环，而触觉传感器是真实的交互式传感器，在 real2sim 过程中，当数据稀缺或视觉信息不足时，融合触觉信息能很好的提升 3D 重建效果。**目前触觉传感器的技术路线尚未收敛，我们认为从成本、传感器精度、部署灵活性、检测范围等角度考虑，霍尔磁电式、视触觉传感器或是较好方案。**

图表 32：不同机器人传感器的优缺点比较

原理	优点	缺点
压阻式	较高的灵敏度、过载承受能力强	压敏电阻漏电流稳定性差、体积大，不易实现微型化、功耗高、易受噪声影响、接触表面易碎
光电式	较高的空间分辨率、电磁干扰影响较小	多力共同作用时，线性度较低、数据实时性差、标定困难
电容式	测量量程大、线性度好、制造成本低、实时性高	物理尺寸大，不易集成化、易受噪声影响，稳定性差
电感式	制造成本低、测量量程范围大	磁场分布难以控制，分辨率低、不同接触点的一致性差
压电式	动态范围宽、有较好的耐用性	易受热响应效应影响

资料来源：焉知人形机器人公众号，华源证券研究所

3.2.5. 关于机器人 VLA 落地可能面临问题的总结

综上所述，我们认为机器人 VLA 部署仍是一个中长期事件，如何破局数据-模型-本体构型的鸡生蛋问题、如何寻求合适的落地场景，使机器人或严格定义为人形机器人从一个遥控玩具和科研展品走向真正的商业、工业落地，成为一个具备独立思考能力能够准确理解世界、具备泛化性、能够使用运动技能与现实世界交互的物理 AI，**从前置条件或发展方向看，我们认为在标准化与模型优化两个方向值得重视：**

首先需要标准化，尤其是底层硬件、通信协议的标准化，然后在此基础上实现本体部分标准化，例如在 C 端场景定义通用的人形机器人构型，在 B 端场景定义少数几种适合大部分场景、任务的类人性或其余形态机器人构型。**1）从数据收集、模型与本体适配角度，**如同大

多数人惯于使用右手且很难在短时间内变成左撇子，机器人的数据、模型都需要与确定的本体构型相匹配才能最大化模型效果，“一脑多型”方案在当前阶段对于模型能力、训练算力、映射算法的要求过于理想化，我们认为先确定基本构型然后在此基础上进行数据收集和模型构建，实现规模化数据收集与复用以解决数据孤岛问题和破局数据–模型能力的鸡生蛋问题是较为可行的方案；**2）从行业供应链角度，**没有行业基本零部件或构型的标准化就构建不了顺畅的供应链，当前从其余行业进行零部件改制供应的现象就会一直存在，导致行业整体上下游的匹配摩擦成本、制造成本无法下降，产品的最大化效能也无法发挥。

我们认为实现标准化最需要的是时间，即需要等待机器人行业走过野蛮生长的时段，待产品和应用场景进一步清晰以后等待行业的“iphone 4 时刻”，一款革命性的产品自然会引来行业跟随与供应链统一。**在此之前较为可行的现实路径是在业内建立部分标准化共识（底层逻辑为机器人可能是大部分标品+部分非标品的现实），例如实现通信协议、自由度、感知方式的逐步标准化进而尝试构建初步的数据闭环系统。**

其次是上层系统的优化，主要是机器人模型闭环与交互能力构建。1）从模型设计角度，目前业内的大小脑系统多是割裂的单独训练和线性无环结构（如缺乏图表 3 所示的反向传播），**从模型架构出发，**机器人 VLA 模型的落地需要借鉴语言模型和汽车端到端经验，**首先**需要实现真正的端到端可训，即设计一种类似人类的高低频自适应系统，在算力足够的基础上实现大小脑不同频率下的联合训练；**其次**是实现模型闭环训练，即将输出 action 反馈回模型输入并进行 test time compute，利用强化学习方案彻底提升模型行为能力的泛化性、准确性和鲁棒性等。**2）从交互能力角度，**如前所述，机器人相较于汽车更为强调物理交互能力，我们认为物理仿真能力（世界模型）和多模态数据（触觉、嗅觉等）引入或是未来机器人向真正拥有物理世界理解和交互能力的具身智能体转化的关键。

3.3. 人形机器人典型 VLA 架构

3.3.1. Open VLA：首个开源且具备商业潜力的机器人 VLA 模型

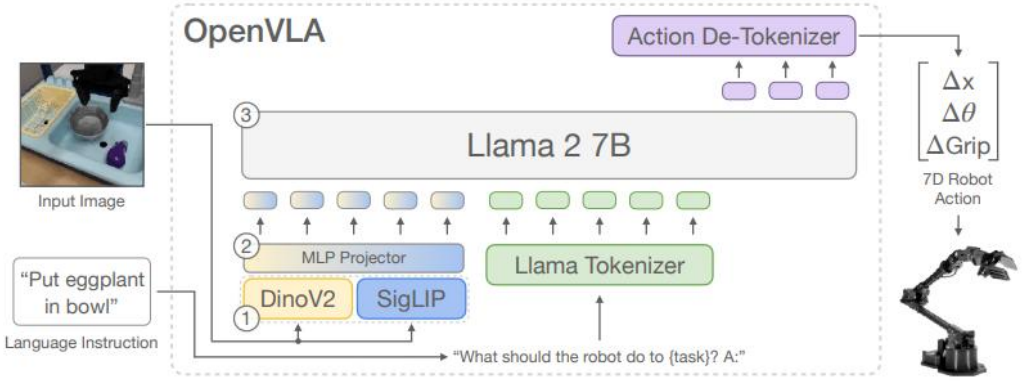
机器人 VLA 模型发展的早期阶段，模型的封闭性、庞大的参数规模限制了其大范围推广。由于机器人数据集规模远不及互联网数据集，直接复制大语言模型级别的预训练在机器人领域是几乎不可能完成的挑战，因此利用现有 VLM 方案作为基础模型并加入动作模块，再利用机器人领域真实数据做微调形成可直接生成机器人控制动作的 VLA 模型成为主要的解决思路。但在 Open VLA 之前两个问题阻碍了 VLA 方案的广泛应用：**1）模型封闭：**模型大多闭源，缺乏关于模型架构、训练过程和数据集的透明度，模型很难复现；**2）缺乏部署能力，**即以前的 VLA 模型未能提出如何在新的机器人本体、环境、任务中部署 VLA，特别是没有对 VLA 范式在消费级 GPU 上的端侧部署能力进行探讨，对于 VLA 范式的研究多停留在实验室阶段，例如 Google RT–2 模型，其闭源属性和庞大参数规模限制了模型的大范围应用。

Open VLA 为一个 7B 规模的 VLA 模型，能够在接收信息后预测 7 维的机器人控制动作，其模型架构主要由输入模块、LLM 主干两部分构成。

输入模块由两个视觉编码器、小型两层 MLP 投影器（用于映射）、Llama 语言编码器构成，其中视觉编码器部分包含预训练的 SigLIP 和 DinoV2 模型，输入图像 patch 分别通过两个编码器得到结果特征向量，相较于 CLIP 或仅 SigLIP 编码器，添加 DinoV2 编码器有助于提升模型空间推理能力，更能理解物体间的准确位置关系。

LLM 主干部分为一个 7B 规模的 Llama 2 大语言模型并在 Open X-Embodiment 机器人数据集上进行了微调，接收图像、文本等多模态信息并进行推理输出 action token。模型利用栅格法，将机器人连续的动作离散化分散到 256 个栅格中，每一个栅格可以被视作一个 token。当机器人动作被处理为 token 序列后，VLA 模型即能够以标准的自回归方式预测下一个 token 为目标进行训练。

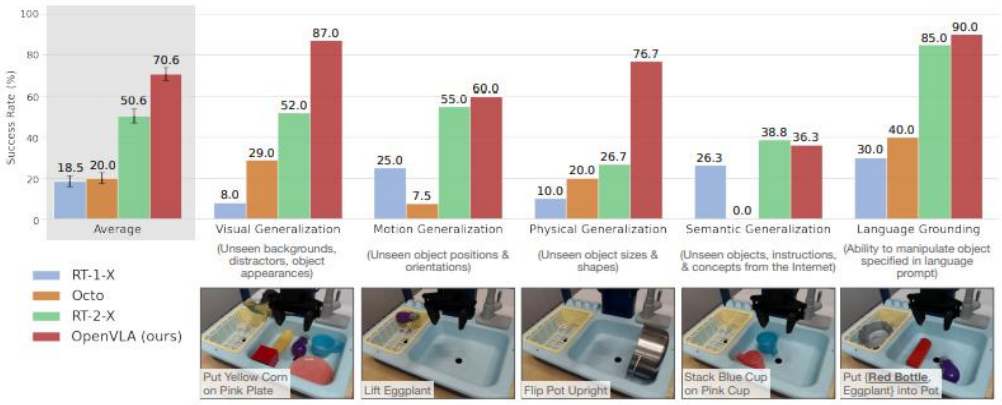
图表 33：Open VLA 模型架构



资料来源：《OpenVLA: An Open-Source Vision-Language-Action Model》_Moo Jin Kim 等，华源证券研究所

Open VLA 论证了机器人 VLA 范式的商业部署潜力。1）Open VLA 相较于前代模型如 RT-2，在整体规模缩小的情况下实现了更优越的性能，Open VLA 在通才操作方面表现出色，在 WidowX、Google Robot 两个平台 29 项任务和多种机器人实例测试中，其绝对任务成功率对比闭源模型提高 16.5%，同时模型参数规模减少为原模型的 1/7；2）利用 LoRA 和模型量化提高计算效率并压缩了模型规模，LoRA 即低秩自适应方法，是一种利用矩阵降秩计算原理实现模型微调的方法，能在模型参数不变的情况下提高模型训练速度和降低显存占用，同时结合模型量化为模型在消费级 GPU 上部署提供可能；3）完全开源，Open VLA 完全开源代码库，支持从单个 GPU 微调到多节点 GPU 集群、十亿级参数规模的 VLA 训练，并支持自动混合精度、分片数据并行等大型 transformer 架构训练技术。**整体看来，Open VLA 旨在为机器人操作提供通用策略，其较好的模型性能、更小的参数规模和完全开源特性论证了 VLA 范式的商业部署、技术生态建立的潜力。**

图表 34：Open VLA 模型在多项任务测评中相较于前代模型取得了更好的效果



资料来源：《OpenVLA: An Open-Source Vision-Language-Action Model》_Moo Jin Kim 等，华源证券研究所

3.3.2. Helix：首个人形机器人上半身高速连续控制的开源模型

Helix 是人形机器人领域首个“双系统”分层模型，用于高速灵巧地控制整个机器人上半身。如同车端“VLM+E2E”双系统架构前所面临的问题，Helix 之前的 VLA 模型也面临根本性权衡：VLM 有强大的泛化能力，但无法满足实时性要求；传统机器人运动控制策略速度快，但泛化能力不足。因此 Helix 通过设计两个互补系统来解决矛盾，系统一（S1）规模 80M，能够以 200Hz 的频率将 S2 生成的高级语义转化为精确的连续机器人动作；系统二（S2）是一个端侧部署、经过互联网数据预训练的 7B 规模 VLM 模型，以 7–9Hz 的频率进行场景和语言理解，二者之间加入了时间偏移模块以最大限度地减少异步模型的运行时间差异。同时 Helix 不同于传统的双系统模式传递语言指令等显式命令，而是依靠隐向量进行信息传递实现了模型的端到端可训。**Helix 使用了单一神经网络在一个约 500 小时、高质量的多机器人数据集上进行监督训练，使模型获得了多项特性：**

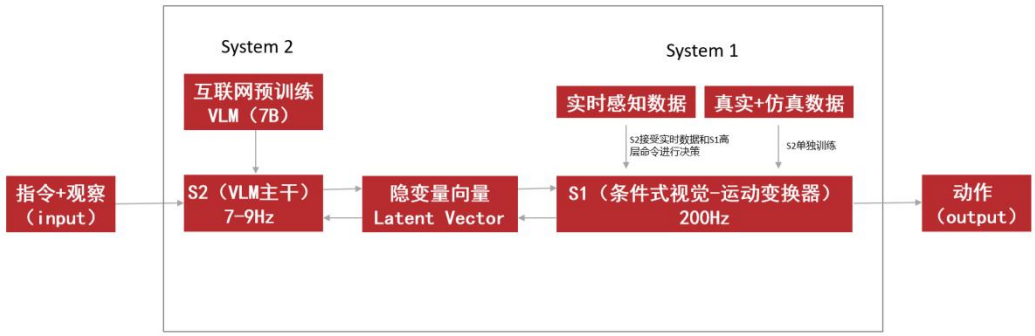
1）**训练速度与泛化能力：**模型在双系统架构下使用单一神经网络训练，双系统架构的优化使模型训练速度能够与针对单一任务进行优化的行为克隆策略相匹敌；单一神经网络使模型无需对每个新物体、任务单独训练微调，展现出其出色的通用性。

2）**可拓展性：**Helix 能够直接输出高维动作空间的连续控制，而不需要像过去的 VLA 方法使用复杂的离散动作标记方案。例如 Open VLA 的离散化方法在低维度的控制任务中很管用（通过将简单的机器人动作与特定 token 对应），但在某些需要更复杂任务步骤、更长的任务时间、更精细化连续化的机器人控制的高维任务中，传统的离散低维方法难以奏效。

3）**模型解耦：**将 S1 与 S2 解耦，Helix 可以分别优化每个系统，不受寻找统一观测空间或动作表示的约束。

4）**商业落地能力：**Helix 是首个完全在嵌入式低功耗 GPU 上运行的人形机器人 VLA 模型，可以立即投入商业部署。

图表 35：Helix 模型架构



资料来源：Figure AI 官网，华源证券研究所绘制

Helix 实现了人形机器人历史上多个“首次”出现的强大功能：

完整的、细颗粒度的上半身控制能力。Helix 是首个能够对整个人形机器人上半身以 200Hz 的频率协调 35 自由度动作空间并进行高频率连续控制的 VLA 模型，较好的解决了人形机器人上半身完整控制的两个难点：**1）超高自由度，**一般而言传统机器人最多控制个位数自由度，35 自由度意味着机器人每秒要实现 35 个部位的动作方向、角度和力度等维度的控制并进行整合优化，每个自由度的增长会使模型计算量呈指数级增长，Helix 高频连续动作控制展示了模型极强的计算能力；**2）实时协调能力，**机器人的头部或躯干移动时会改变机器人的可触及和可视范围从而形成一个新的反馈回路，这需要机器人实时的做出新的判断和动作，使系统的不稳定性增加，Helix 通过先确定躯干最优位置，再以头部视角追踪手部移动的方式实现了在高维动作空间中的协调性。

零样本多机器人协调。Figure 展示的视频中显示两个机器人通过自然语言指令（如将盒子递给你左边的机器人）进行协调，能实现对训练中从未见过的全新物品的协同操作，这项能力主要通过两种方式实现：1）充分利用系统 2（VLM 模型）的互联网先验知识实现了自然语言指令输入与全新物品识别；2）利用单一神经网络+多机器人、多任务场景数据集微调实现多机器人在同一模型控制下的协同操作，其中单一神经网络使模型可以学习所有机器人行为和跨机器人互动，“多机器人、多任务、多场景数据集微调”的能够更真实地模拟复杂环境、任务和团队协作能力，为多机器人间的交互协作奠定数据基础。

图表 36：搭载 Helix 模型的机器人实现上半身连续控制与双机器人任务协作

资料来源：Figure AI 官网，华源证券研究所

涌现“拿起一切”的能力。Figure 视频中显示，只需一个“拿起”指令，搭载 Helix 的机器人即可在没有任何先前演示或自定义编程的情况下拾取几乎所有的小型物品，同时 Helix 还可以建立互联网规模的语言理解和精确的机器人控制之间的联系，例如当被提示“拿起沙漠物品”时，Helix 不仅能确定玩具仙人掌与这个抽象概念相匹配，还能选择最近的手并能通过精确运动控制进行抓取。这种通用的“从语言到行动”的能力为人形机器人在家庭、工厂、餐厅等非结构化环境中的部署提供了可能性，凸显了 Helix 的商业部署能力。

图表 37：Helix 模型的泛化性能与抽象概念理解能力

“拿起任何物体”能力

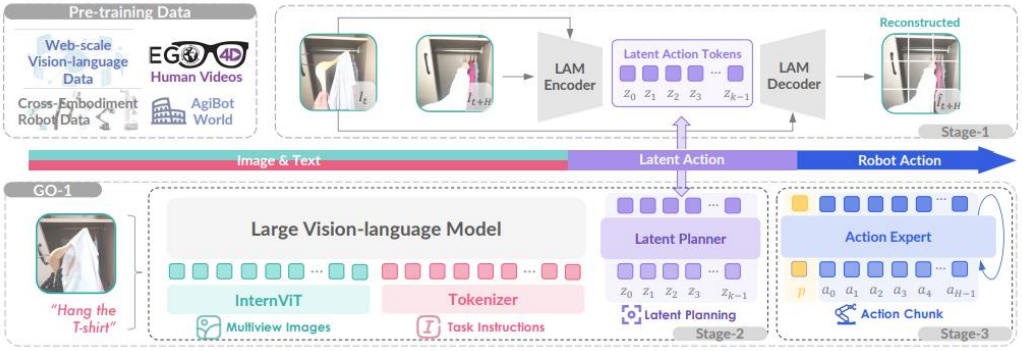
抽象概念理解+动作控制输出

资料来源：Figure AI 官网，华源证券研究所

3.3.3. 智元 ViLLA：实现大规模互联网异构视频数据高效利用

ViLLA 架构的亮点在于对大规模互联网异构视频数据的高效利用。ViLLA 架构由 VLM + MOE 组成，主要由 VLM、Latent Planner、Action Expert 三大模块协同工作，相较于传统的 VLA，ViLLA 通过预测隐式动作标记，弥合图像-文本输入与机器人执行动作之间的差距，在真实世界的灵巧操作和长时任务方面表现较好，超过了彼时已有的开源 SOTA 模型；同时 LAM 的引入，使模型不仅能够从互联网数据中增强语义理解和场景理解能力，还能够直接从人类视频演示中进行轨迹动作知识的迁移，实现对不同互联网数据的高效利用。

图表 38：智元 GO-1 机器人 ViLLA 模型架构



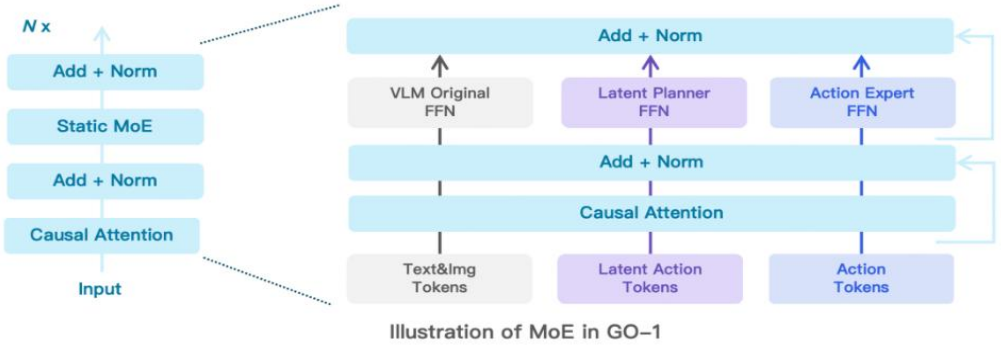
资料来源：《AgiBot World Colosseo: A Large-scale Manipulation Platform for Scalable and Intelligent Embodied Systems》_AgiBot-World-Contributors 等，华源证券研究所

VLM（多模态大模型）：采用 InternVL-2B，接收多视角图像、力信号、语言输入等，进行通用场景感知和指令理解。

Latent Planner（隐式规划器）：LP 是 MoE 中的一组专家，其基于 VLM 的输出结果预测隐式动作标记以进行高层级动作规划。**为了大规模利用互联网视频数据和训练 Latent Planner，ViLLA 架构在云端引入了 LAM 模型来建模互联网人类动作视频当前帧和历史帧之间的隐式变化，然后使 Latent Planner 预测这些变化以进行学习，从而将异构数据源中真实世界的动作知识迁移到模型能力中。**

Action Expert（动作专家）：AE 是 MoE 中的另一组专家，通过采用扩散模型，在去噪过程中实现低层级的精细动作序列生成。Action Expert 结构上与 Latent Planner 类似，与 VLM 主干网络共享 Transformer 结构但使用独立 FFN 和投影矩阵（如下图）。

图表 39：GO-1 模型中的 MOE 层，包含 VLM、Latent Planner、Action Expert 三个核心组件



资料来源：42 号电波公众号，华源证券研究所

建立在 ViLLA 架构上的 Go-1 模型充分利用了现有数据，在能力构建和模型效果上取得了较好的成果，展现了模型的通用潜力。Go-1 具备多种模型能力：1）人类视频学习，可以结合互联网视频和真实人类示范数据进行学习，并利用 AgiBot World 真机数据进行微调，一定程度解决了机器人 VLA 模型的数据难题；2）小样本快速泛化，VLM 模型赋予了模型强大泛化能力，使模型能够在极少数据甚至零样本下泛化到新场景、新任务；3）一脑多形，Go-1 作为通用机器人策略模型，能在不同机器人形态之间迁移，快速适配不同本体。在模型效果方面，在五种不同复杂度任务的测试中相较于对比模型，Go-1 平均成功率提高了 32%。**GO-1 的推出是机器人 VLA 模型通用化发展的实例，展现了机器人从完成特定任务的工具，向具备通用智能的具身智能体发展的可能性。**

4. 受益公司梳理

4.1. 理想汽车：从汽车到 AI，VLA 范式引领汽车智能化升级

营收体量增长，盈利能力逐渐修复。2024 年之前，公司营收规模高速增长并于 2023 年实现盈利，24Q1，受 MEGA 上市不及预期和竞品影响，公司销量下滑导致净利润同比下降。之后公司迅速调整战略，并在 24Q2 推出走量 SUV 车型 L6，在 L6 高销量带动下，公司 Q2-Q4 营收、净利环比逐渐改善。25Q1 公司营收达到 259.3 亿元，同/环比分别为 1.14%/-41.44%；毛利率 20.51%，同比-0.1pct，毛利率保持相对稳定。整体看来，公司增程基本盘顾虑已消除，AI+纯电产品周期将至，i 系列纯电车型有望抢占 20-40 万家庭豪华纯电 SUV 市场高溢价份额，驱动整体销量增长。

公司转型人工智能企业已初见成效，VLA 范式有望引领汽车智能化升级浪潮。2023 年，理想汽车正式提出 AI 化战略，明确提出“AI 定义汽车”战略方向，持续通过 OTA 升级车端智能化能力，23 年 12 月，全量推送城市 NOA 并实现 Mind GPT 上车；24 年分别于 7、10、11 月全量推送无图 NOA、端到端+VIM 架构、车位到车位智驾能力；2024 年 12 月，正式宣布从“智能电动车企业”转型为“人工智能企业”，并提出其愿景“连接物理世界和数字世界，成为全球领先的人工智能企业”。

从 AI 发展的数据、算力角度，理想已具备整车企业第一梯队能力。数据方面，公司月销位居在新势力品牌前列，截至 2024 年 12 月 31 日，理想智能驾驶累计里程达 29.3 亿公里，累计时长超 3382 万小时，智驾用户数达到 110.9 万人，销量、里程数据有望形成“销量优势-数据积累-模型优化-体验提升-销量增长”的闭环链路；算力方面，理想与火山引擎合作布局云端算力达 8.1EFLOPS，高云端算力支撑模型快速训练迭代，端侧 Orin 芯片即将升级为 Thor-U 芯片，为模型在端侧部署提供更高算力基础。

2025 年 3 月，理想汽车推出下一代自动驾驶架构——Mind VLA，展示了公司由整车企业向 AI 企业转型升级最新成果。**我们认为，VLA 的正式推出与落地展示了公司成为全球领先人工智能企业的决心，短期来看**随着 VLA 的落地部署与优化，公司 2025 年实现 L3 有条件的自动驾驶，2025 年、2026 年逐步将综合 MPI（城市+高速综合接管里程）提升至 500 公里、1000 公里以上的智驾目标有望逐步实现，理想有望凭借 VLA 模型的工程部署先发优势引领

汽车智能化升级浪潮。中长期来看，理想 Mind VLA 将成为公司具身智能基座模型，持续支撑理想空间、穿戴机器人业务乃至后续人形机器人等业务边界拓展，助力公司实现 2030 年成为全球领先的人工智能企业愿景。

4.2. 小鹏汽车：底层自研、全链自主打造“智驾端到端四部曲”

交付量增长，2025 年利润水平有望改善。2024 年公司交付新车 19.0 万辆，同比+34.2%，呈现稳健增长态势。随着公司降本增效措施逐渐兑现，叠加与大众合作带来的经营改善，25Q1 公司营收达到 158.1 亿元，同比增长达 141.5%，毛利率 15.6%，同比增长 2.7pct。预计 25Q2 以来改款/新款车型上市有望带动产品结构改善，叠加规模效应、技术+供应链降本的持续兑现，看好小鹏汽车后续利润水平改善趋势。

智能化是小鹏汽车核心战略，坚持“底层自研、全链自主”实现智能化从跟随到局部领先的跨越。小鹏汽车自成立伊始便将智能化作为核心战略，从传感器硬件到车端算力、从大模型算法到云端数据平台，小鹏坚持“底层自研、全链自主”，通过持续性研发投入与闭环式研发迭代体系，小鹏在智能驾驶领域目前已实现了从跟随到并跑、再到局部领先的跨越，智能化能力为其多品类车型的智能驾驶落地提供了坚实技术保障，也持续赋能小鹏 Iron 人形机器人等新业务拓展。

算法研发方面，小鹏自动驾驶系统经历了从规则驱动到端到端大模型的多轮迭代。最初的 Xpilot 以规则驱动实现 ACC 与 LCC 等基础巡航与泊车功能；2021 年小鹏汽车推出高速 NGP；2023 年推出城市 NGP 无图化版本，在全国范围内实现无高精地图覆盖；2024 年 XNGP+集成 BEV+Transformer 大模型，打通感知–预测–规划三网，实现端到端一体化决策，并在 OTA 中持续迭代 XBrain 架构，支持环岛、掉头及其他复杂场景，性能与鲁棒性不断提升；2025 年，小鹏智驾能力继续提升，720 亿参数规模世界基座模型初步验证成功，预计后续将全面赋能小鹏 AI 体系全图谱，成为小鹏 AI 汽车、机器人、飞行汽车等所有物理 AI 终端通用模型，此外小鹏 G7 首发智驾“大脑+小脑”VLA–OL 模型，实现全端侧运行，标志着小鹏正式从“软件开发汽车”进入到“AI 开发汽车时代”。**面向未来，小鹏已规划了自动驾驶技术的“端到端四部曲”，**即在 2025 年下半年推动云端大模型参数量相较于 2024 年版本提升约 5 倍，实现类 L3 级别（百公里接管＜1 次）的高品质智能驾驶体验；最终在 Ultra 平台推出 Robotaxi，通过 AI Eagle Eye、XBrain、图灵芯片与云端模型工厂的协同，实现部分低速场景下的真正意义无人驾驶商业化运营。

算力建设进展顺利，以“图灵芯片”探路自驾 L3 时代。云端算力方面，小鹏已建成国内汽车行业首个万卡智算集群，“云端模型工厂”拥有 10EFLOPS 算力，集群运行效率常年保持在 90%以上，从云到端的全链路迭代周期平均可达 5 天一次。云端模型工厂不仅支撑了 XNGP+快速上线，也为未来“一键升级”更大规模的生成式端到端大模型提供了算力保障。**端侧算力方面，**自研并成功流片“图灵”智驾芯片，该芯片采用 7nm 工艺，单颗芯片算力可达约 700TOPS，堪比三颗主流 Orin–X 芯片之和，自研端侧芯片一方面将显著优化整车成本结构，另一方面也将为未来软件算法迭代提供更好硬件适配灵活性。

图表 40：理想汽车、小鹏汽车盈利预测

	营业收入（百万元）				归母净利润（百万元）			
股票名称	2024A	2025E	2026E	2027E	2024A	2025E	2026E	2027E
理想汽车-W	144,460.0	166,928.7	219,398.3	264,635.5	8,032.4	9,861.3	13,549.5	17,427.9
小鹏汽车-W	40,866.3	91,757.5	148,413.5	197,066.0	-5,790.3	-823.0	3,079.7	7,621.0

资料来源：Wind，华源证券研究所

注：盈利预测均来自 Wind 一致预期，截至 2025 年 7 月 6 日

5. 风险提示

1）新技术迭代风险。当前无论是智能化软件技术还是汽车、人形机器人硬件技术都处于迭代升级过程中，部分技术路径尚未收敛，当前领先的企业仍面临在未来未能及时跟进技术进化节奏，面临技术领先优势被颠覆的风险。

2）市场竞争加剧风险。当前中国汽车市场竞争愈发激烈、市场增幅趋于平缓，价格战愈演愈烈，相关企业面临强劲的市场竞争和销量不及预期风险。

3）宏观经济增速不及预期风险：汽车属于耐用消费品，短期需求弹性大，若经济增速不及预期，或导致汽车市场整体增速不及预期；我们认为人形机器人更偏向于可选的消费电子，或受经济周期波动影响更大。

证券分析师声明

本报告署名分析师在此声明，本人具有中国证券业协会授予的证券投资咨询执业资格并注册为证券分析师，本报告表述的所有观点均准确反映了本人对标的证券和发行人的个人看法。本人以勤勉的职业态度，专业审慎的研究方法，使用合法合规的信息，独立、客观的出具此报告，本人所得报酬的任何部分不曾与、不与，也不将会与本报告中的具体投资意见或观点有直接或间接联系。

一般声明

华源证券股份有限公司（以下简称“本公司”）具有中国证监会许可的证券投资咨询业务资格。
本报告是机密文件，仅供本公司的客户使用。本公司不会因接收人收到本报告而视其为本公司客户。本报告是基于本公司认为可靠的已公开信息撰写，但本公司不保证该等信息的准确性或完整性。本报告所载的资料、工具、意见及推测等只提供给客户作参考之用，并非作为或被视为出售或购买证券或其他投资标的的邀请或向人作出邀请。该等信息、意见并未考虑到获取本报告人员的具体投资目的、财务状况以及特定需求，在任何时候均不构成对任何人的个人推荐。客户应对本报告中的信息和意见进行独立评估，并应同时考量各自的投资目的、财务状况和特殊需求，必要时就法律、商业、财务、税收等方面咨询专家的意见。对依据或使用本报告所造成的一切后果，本公司及/或其关联人员均不承担任何法律责任。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。
本报告所载的意见、评估及推测仅反映本公司于发布本报告当日的观点和判断，在不同时期，本公司可发出与本报告所载意见、评估及推测不一致的报告。本报告所指的证券或投资标的的价格、价值及投资收入可能会波动。除非另行说明，本报告中所引用的关于业绩的数据代表过往表现，过往的业绩表现不应作为日后回报的预示。本公司不承诺也不保证任何预示的回报会得以实现，分析中所做的预测可能是基于相应的假设，任何假设的变化可能会显著影响所预测的回报。本公司不保证本报告所含信息保持在最新状态。本公司对本报告所含信息可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。

本公司销售人员、交易人员以及其他专业人员可能会依据不同的假设和标准，采用不同的分析方法而口头或书面发表与本报告意见及建议不一致的市场评论或交易观点，本公司没有就此意见及建议向报告所有接收者进行更新的义务。本公司的资产管理部门、自营部门以及其他投资业务部门可能独立做出与本报告中的意见或建议不一致的投资决策。

信息披露声明

在法律许可的情况下，本公司可能会持有本报告中提及公司所发行的证券并进行交易，也可能为这些公司提供或争取提供投资银行、财务顾问和金融产品等各种金融服务。本公司将会在知晓范围内依法合规的履行信息披露义务。因此，投资者应当考虑到本公司及/或其相关人员可能存在影响本报告观点客观性的潜在利益冲突，投资者请勿将本报告视为投资或其他决定的唯一参考依据。

投资评级说明

证券的投资评级：以报告日后的 6 个月内，证券相对于同期市场基准指数的涨跌幅为标准，定义如下：

买入：相对同期市场基准指数涨跌幅在 20%以上；
增持：相对同期市场基准指数涨跌幅在 5%～20%之间；
中性：相对同期市场基准指数涨跌幅在－5%～＋5%之间；
减持：相对同期市场基准指数涨跌幅低于–5%及以下。
无：由于我们无法获取必要的资料，或者公司面临无法预见结果的重大不确定性事件，或者其他原因，致使我们无法给出明确的投资评级。

行业的投资评级：以报告日后的 6 个月内，行业股票指数相对于同期市场基准指数的涨跌幅为标准，定义如下：

看好：行业股票指数超越同期市场基准指数；
中性：行业股票指数与同期市场基准指数基本持平；
看淡：行业股票指数弱于同期市场基准指数。

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取决于个人的实际情况，比如当前的持仓结构以及其他需要考虑的因素。投资者应阅读整篇报告，以获取比较完整的观点与信息，不应仅仅依靠投资评级来推断结论。

本报告采用的基准指数：A 股市场基准为沪深 300 指数，香港市场基准为恒生中国企业指数（HSCEI），美国市场基准为标普 500 指数或者纳斯达克指数。