证券研究报告　　机械行业　　推荐(维持)

兴业证券 INDUSTRIAL SECURITIES

人形机器人系列深度报告(四)

具身大模型：人形机器人智慧内核，数据飞轮驱动迭代跃升

分析师：　丁志刚（S0190524030003）

石康（S1220517040001）

报告日期：2026年4月15日

投资要点 KEY POINTS

兴业证券 INDUSTRIAL.SECURITIES

具身大模型是人形机器人的“大脑”，主导“感知-认知-控制”交互闭环。传统大模型专注于单一或少数模态的任务处理，缺乏与物理世界直接交互的能力。具身大模型作为人形机器人的“大脑”，从“感知-认知-控制”层面赋能机器人，强调与物理世界的交互，需具备多模态感知、自主决策、实时交互执行、通用与泛化等能力。

人形机器人目前尚未实现大规模应用，主要原因或非硬件能力不足，而是大模型存在瓶颈。 从产业进程来看，当前机器人肢体层技术已较为成熟，而大模型的发展远落后于硬件。当前阶段的具身大模型已具备认知、推理与规划能力，不足之处在于难以可靠处理复杂物理世界的不确定性，同时泛化能力明显较弱。

具身大模型主流框架为分层式与端到端式，路径尚未收敛。传统决策采用分层架构，包括感知与互动、高层规划、低层执行以及反馈与增强，通过大小脑分层，人形机器人更容易落地，但分层范式存在错误累积的问题，且在跨多样任务泛化时表现不佳。端到端框架基于感知环境和机器人状态直接输出具体的机器人执行命令，将感知、语言理解、规划、动作执行和反馈优化集成到一个统一的框架中，具备高集成度与较强泛化能力，VLA模型是端到端决策的核心。

海外具身大模型：1)典型的完全端到端架构具身大模型包括谷歌DeepMindRT-2 与特斯拉FSD。RT-2 致力于通过端到端的神经网络将视觉和语言信息直接映射为机器人动作；特斯拉Optimus 可沿用汽车FSD 系统的技术栈，实现多模态输入与实时动作输出。2)典型的分层具身大模型包括Figure Al Helix、英伟达GROOT N1与Physical Intelligence πo Helix采用“系统S1 (快思考)+系统S2 (慢思考)”双系统架构；GROOT N1同样采用双系统架构，并利用流匹配技术来生成动作；π。采用“预训练VLM+ 动作专家模块”的VLA 模型。

国内具身大模型：架构持续创新，能力对标海外，典型模型包括智元机器人GO-1、 星动纪元ERA-42、 银河通用GraspVLA、 灵初智能Psi R1及字节Seed GR-3。GO-1开创性提出VLLA 架构，采用“VLM+MoE (混合专家)”；ERA-42 模型是国内首个真正意义上的端到端原生机器人大模型；GraspVLA 模型将VLM 与动作专家集成，是全球首个合成大数据驱动的基础抓取大模型：Psi R1模型采用快慢脑架构；GR-3 采用40亿参数的混合变换器架构，泛化抓取-放置能力超越πo

数据是驱动具身大模型迭代升级的关键，目前主流数据训练方案为真机、仿真与视频数据相结合。伴随具身智能转向端到端大模型，数据需求从低量单一模态数据逐步升级为海量、多模态、高精度和跨任务长程数据，其中真机数据价值最高，获取难度最大，是具身智能落地的可靠数据源。目前真实数据采集方式主要分为VR 遥操作采集、机械臂主从控制采集、数据手套遥操作等。目前主流厂家数据采集及训练方案多样，特斯拉数采方案或转向视频学习，而银河通用以物理仿真数据为主、真实数据为辅。

投资建议：1)机器人通过传感器获取外界和自身状态，为具身大模型决策提供数据支持，建议关注人形机器人传感器相关公司，如安培龙、汉威科技、福莱新材、奥比中光；2)动捕采集方案是高质量运动数据的关键来源，建议关注掌握动捕解决方案的相关公司，如凌云光。

风险提示：人形机器人量产进度不及预期：大模型技术进展不及预期；训练数据规模与质量不及预期。

目 录 CATALOGUE

01 具身大模型：人形机器人大脑，主导“感知-认知-控制”

具身大模型是人形机器人的“大脑”。人形机器人主要由“大脑”、“小脑”和“肢体”三个部分组成，“大脑”负责实现环境感知、行为控制、人机交互等任务级能力，目前主要是基于具身智能大模型技术，提高机器人的智能水平。

具身大模型从“感知一认知-控制”层面赋能机器人。具身大模型需具备以下能力：1）多模态感知；2）自主可靠决策；3）实时交互执行；4）通用性与泛化能力。

相较于传统模型，具身大模型集成多模态，强调机器人在本体与物理世界的交互中实现算法的进化。传统大模型专注于单一或少数模态的任务处理模型，缺乏与物理世界直接交互的能力，而具身大模型需要集成包括文本、视觉、音频和触觉在内的多种模态，且当模型控制机器人本体与环境交互时，可以通过反馈的数据进一步迭代算法，从而提升模型的能力，以及对于新场景和新任务的适应能力。

图：具身智能系统架构

表：具身智能大模型能力要求

能力	对大模型的要求
多模态感知	通过视觉、听觉、触觉等多种感官获取信息，大模型整合多模态感知数据，以实现对环境的全面理解。
自主可靠决策	理解任务的复杂性，将其分解为一系列可执行的子任务，这要求大模型具备强大的语言理解能力和对物理世界的深刻理解。
实时交互执行	人形机器人需要具备与人类实时的任务级交互能力，快速理解人类通过语言、手势等方式给出的指令，并有效执行。
通用性与泛化能力	大模型的泛化能力让人形机器人摆脱“一机一用”的局限，朝向通用化发展。

表：VLA 模型与传统LLM 模型对比

维度	具身智能VLA（视觉-语言一动作）模型	传统LM（大语言）模型
核心目标	物理世界交互(动作生成、闭环控制)	文本生成与推理(对话、写作)
输入输出	多模态输入(视觉+语言)，输出为连续动作	文本输入输出
训练数据	机器人轨迹数据、人类视频、仿真数据	文本语料库〔书籍、网页等)
模型架构	双系统架构等	纯语言模型
实时性需求	需低延迟实时响应	无严格实时性要求
物理约束	需考虑机器人运动学、动力学约束	无物理世界约束
应用场景	人形机器人操控、家庭服务、工业协作	文本生成、代码辅助、知识问答

资料来源：中国信息通信研究院，汪群博《基于大模型的多模态具身智能研究与展望》，机器人大讲堂公众号，菲利信科技公众号，具身智能与空间感知公众号，机器觉醒时代公众号，兴业证券经济与金融研究院整理

01 具身大模型：人形机器人大规模应用的瓶颈

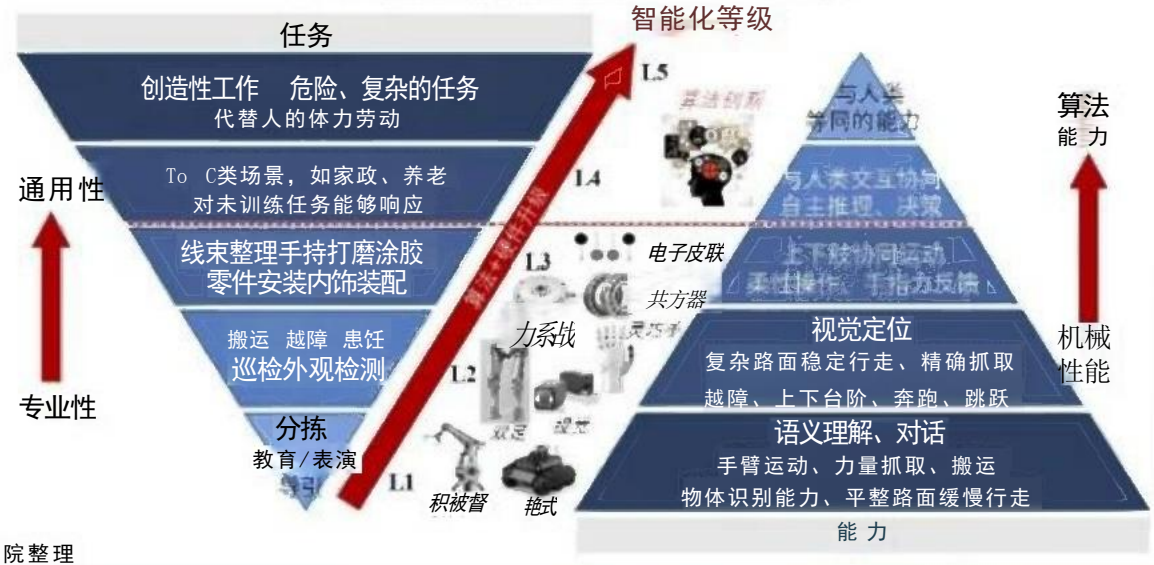
人形机器人目前尚未实现大规模应用，根本原因或非硬件能力不足，而是具身智能大模型面临瓶颈。据宇树科技创始人王兴兴，目前人形机器人硬件已经足够支持应用，机器人当下及未来最关键的挑战主要为大模型的能力还不足以支持大规模应用。具体而言：

√ **1）从产业进程来看，目前大脑层与小脑层的发展远落后于肢体层。**2025年，代表机器人物理本体的肢体层技术已处于成熟期，而代表运动控制的小脑层技术，以及代表高级认知与决策能力的大脑层技术，却基本集中在技术萌芽的导入期。

√ **2）从终局目标来看，AGI（通用人工智能）的实现，即具身智能从当前的L2向L3-L5跨越，主要依**靠大模型的迭代。灵巧手、传感器等硬件的升级使得人形机器人机械性能提升，目前已经能够实现柔性操作，但想要具备更高的通用性，并最终具备与人类等同的能力，关键在于算法创新，即大模型的迭代升级。

图：人形机器人产业进程

图：算法能力决定人形机器人智能化等级



资料来源：甲子光年公众号，勾股大数据、特斯拉官网、格隆汇，兴业证券经济与金融研究院整理

01 架构：尚未收敛，主流路线为分层式与端到端式

▶当前具身智能大模型主流框架为分层式与端到端式，路径尚未收敛。

1）分层框架：分层具身模型采用“大脑-小脑-肢体”的架构，上层大模型负责感知与决策，底层硬件层和中间响应快的小模型负责分解与执行。这类模型更适合当前的数据积累水平，且更容易融入基于学习的控制方法，因此被更多厂商采用。

2）端到端系统框架： 端到端大模型能够直接实现从人类指令到机械臂执行的过程。输入图像及文本指令后，模型输出夹爪末端的动作轨迹。这种方式简化了系统的层次结构，提高了响应速度，但由于缺乏中间逻辑推理层，对海量数据的依赖度较高，目前尚未成为主流选择。

图：具身智能大模型架构类型

端到端架构	分层架构		世界模型
	上层任务规划	下层操作策略	
输入：文本指令、图像 **输出**：动作（末端姿态）	**输入**：文本指令、环境信息 **输出**：文本序列动作	**输入**：文本指令、轨迹信息、状态信息（环境、本体） **输出**：动作（末端姿态）	处于探索期 **输入**：完整物理世界 **输出**：未来预测状态
RT-2、πo、Helix	SayCan、PaLM-E		UnifoLM-WMA-0
优势：感知-决策-控制一体化，优势误差累积小	优势：任务规划与低级控制解耦，迁移与调试灵活		优势：潜在实现通用智能，适应未知场景

资料来源：机器觉醒时代公众号，新智元公众号，深蓝具身智能公众号，上海科普网，兴业证券经济与金融研究院整理

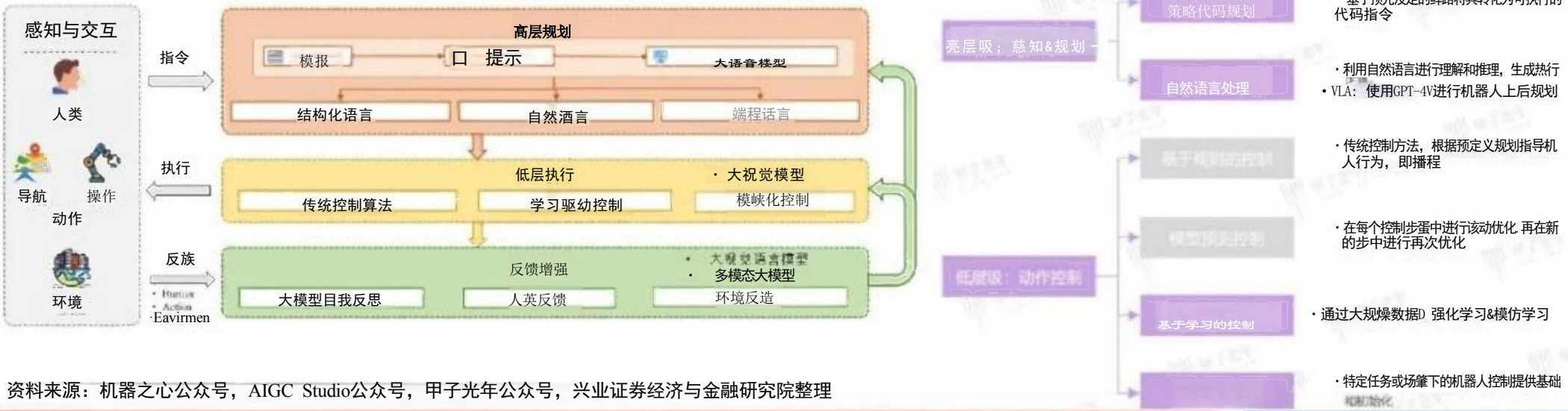
01 分层式架构：通过大小脑分层，机器人更易落地

传统决策采用分层范式，包括感知与互动、高层规划、低层执行以及反馈与增强。机器人的控制一般可以粗略地分为高层和低层。高层负责全局、长期的目标；低层负责具体操作与及时反馈。虽然基础模型具有丰富常识与较强的推理能力，但精确性、实时性较差，所以大模型往往不会直接参与机器人的低层次控制，而是通过需求理解、任务规划、动作生成等方式进行较高层级的控制。

通过大小脑分层，人形机器人更容易落地。受现实部署约束，在端侧实时跑大模型受限于端侧芯片的迭代速度，通过大小脑分层、分别部署在边缘侧和端侧的设计，机器人更容易落地。此外，分层架构更符合生物进化规律。

分层范式依赖于独立的任务规划、动作执行和反馈模块，因此存在错误累积的问题，并且在跨多样任务泛化 时表现不佳。分层架构的缺陷在于前层的微小错误会在后续环节快速放大，而且更多的人为干预往往会降低模型效果。此外，高层模型不理解物理约束，常常分配不可能完成的任务；而底层模型缺乏语义理解。

图：分层决策范式由感知和互动、高层规划、低层执行、反馈和增强组成



资料来源：机器之心公众号，AIGC Studio公众号，甲子光年公众号，兴业证券经济与金融研究院整理

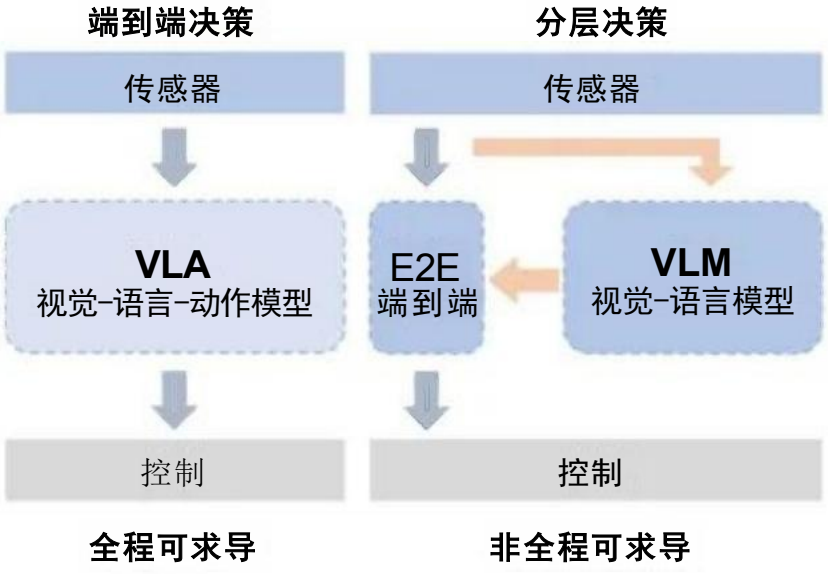
01 端到端式架构：VLA模型是端到端决策的核心

端到端架构直接将多模态输入（即视觉观测和语言指令）映射到动作，通常由VLA（视觉-语言-动作）模型实现。

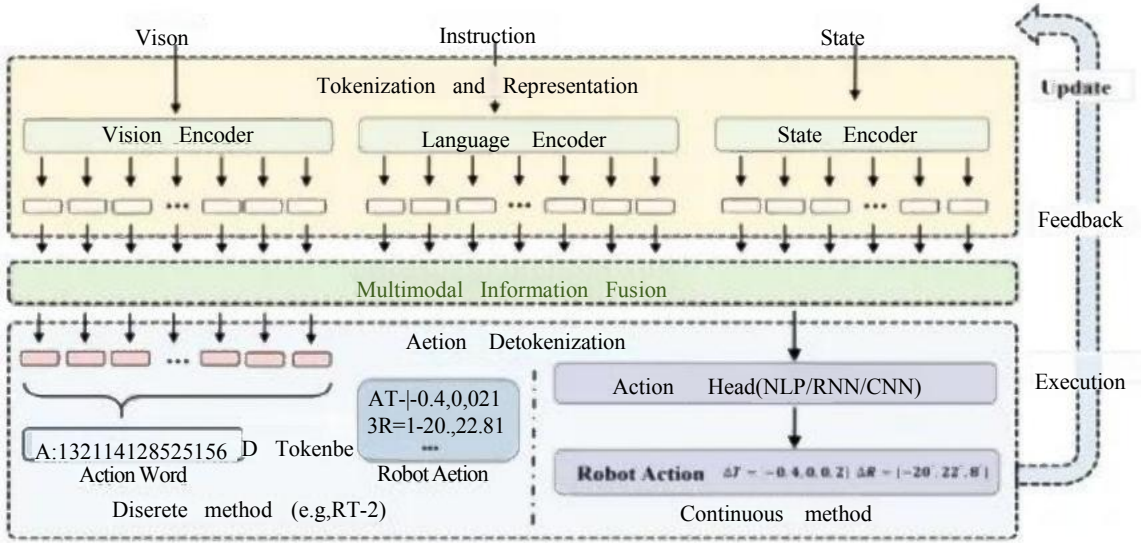
VLA模型是端到端决策的核心。将感知、语言理解、规划、动作执行和反馈优化集成到一个统一的框架中，利用大模型的丰富先验知识，能够在动态、开放的环境中实现精确且适应性强的任务执行，显著提升了具身智能代理的任务执行能力。

- **VLA模型在复杂任务中仍面临挑战，如对输入扰动的敏感性、3D空间关系理解不足、动作生成精度有限以及训练成本高昂等。**

图：端到端与分层决策对比



图：VLA 模型架构



资料来源： AIGC Studio公众号，自动驾驶之心公众号，兴业证券经济与金融研究院整理

01 世界模型：为具身智能提供大规模高质量合成数据

从LLM（大语言）模型到MLLM（多模态大语言）模型，再到视频生成模型的演进，人工智能逐渐向对世界结构、物理过程与动态演化的建模能力迈进。语言模型掌握语义与知识，但缺乏空间结构与物理规律；多模态模型具备空间感知，但无法理解动态世界；视频生成模型呈现动态一致性，但缺少可交互与可解释的内部状态。

构建能统一描述世界结构、动态变化与未来演化的世界模型(World Model)，是实现更高级别人工智能的关键步骤。世界模型旨在构建一个能够刻画物理世界结构、运行状态与未来演化的统一表示，使模型不仅能“解释过去”，还能“预测未来”，并在此基础上作出规划与决策。

世界模型技术路线分为两类：世界的隐式表示(理解当下的世界状态)与世界的未来预测(推演未来的世界演化)。世界的隐式表示侧重于对世界的抽象化建模，使智能体能够在无需外显建模的情况下完成推理、规划与决策；世界的未来预测则从显式生成的角度理解世界，核心在于让模型从数据中直接学习世界的时间逻辑与变化规律，通过直接预测未来状态来学习世界的时间结构与演化规律。

图：世界模型示意图

核心定义：给定当前的状态和一系列动作，预测未来的状态。

本质：不是简单的场景渲染，而是学习物理世界运行规律的“预测引擎”。

目标：赋予机器预测能力。

图：世界模型技术路线及发展图

资料来源：钱振兴《人工智能中世界模型的起源与研究路径》，弗若斯特沙利文《2025年中国世界模型发展洞察》，Deep into X公众号，兴业证券经济与金融研究院整理

01 世界模型：为具身智能提供大规模高质量合成数据

在具身智能领域，世界模型的重要性体现在：1)增强理解与预测能力，完成复杂任务；2)合成高质量数据；3)缩小" Sim2Real"（从模拟到现实)的差距。

参考自动驾驶的发展轨迹，具身智能中的世界模型或将经历三个阶段。1)第一阶段：数据生成器，用世界模型生成多模态训练数据，扩充数据集的多样性，训练更鲁棒的VLA 模型。自动驾驶领域已广泛应用，具身智能正在跟进。2)第二阶段：仿真2.0模拟器，替代传统仿真器，用于训练强化学习算法。3)第三阶段： WAM （世界-动作模型)： VLA模型吸收世界模型的知识，进化为WAM 模型，同时具备预测未来和生成动作的能力，部署在机器人端侧。

表：世界模型在具身智能领域的应用

能力	原理	典型模型
预测能力	将机器人动作建模为视频生成问题，使模型能够“想象”未来场景并据此规划。	UniPi(2023) VIPER(2023) GR-2(2024)
合成数据	视频世界模型能够合成高质量的机器人交互数据，有效增强策略学习。	DreamGen(2025) Roboscape(2025) EVAC(2025)
缩小Sim2Real差距	通过神经网络从真实交互数据中自动学习物理规律，使仿真环境更精准地模拟现实动态，从而有效缩小“模拟—现实”差距，使机器人能够从有限的现实交互中快速适应环境。	DayDreamer(2023) SWIM(2023)

图：世界模型在自动驾驶中的能力分析

资料来源：钱振兴《人工智能中世界模型的起源与研究路径》，弗若斯特沙利文《2025年中国世界模型发展洞察》，兴业证券经济与金融研究院整理

01 当前挑战：泛化能力、高质量数据与长时程推理

当前具身大模型在泛化能力、长时程任务与推理能力、响应速度、高质量训练数据等方面面临挑战。以VLA 模型为例：

1)泛化能力不足： 模型训练过程中容易在分布偏移下导致误差累积，制约模型的泛化能力；

2)高质量数据稀缺：模型需要包含精确动作标签和环境反馈的多模态机器人交互数据，采集此类数据需要昂贵的硬件系统、大量人力与时间成本；

3)长时程任务与推理能力不足： 当前主流的自回归动作生成方法，由于逐步预测动作，容易导致误差累积；

4)实时响应速度较慢： 采用大规模的VLM 或集成更复杂的任务规划架构，导致推理时间延长。

表：具身大模型面临挑战

核心挑战	具体情况
泛化能力不足	一方面，相比于LLM和VLM,目前机器人的数据量仍然处在非常小的量级，所训练出的机器人基础模型没有达到LLMNLM相当的泛化能力；另外一方面，机器人轨迹数据与文本数据和图像数据之间的错位，模型会在新的任务上过拟合，从而遗忘原先具备的语言理解或多模态对齐能力，引发微调后的大模型面临“虚假遗忘”问题，导致对于场景理解和泛化能力急剧下降。
高质量数据稀缺	不同于可以直接使用海量互联网图文数据训练的LLM和VLM,VLA模型需要包含精确动作标签和环境反馈的多模态机器人交互数据，采集此类数据需要昂贵的硬件系统、大量人力与时间成本。VLA系统模型能力不足限制了对高质量交互数据的采集，而数据不足又反过来限制了模型能力的提升。
长时程任务与推理能力不足	当前主流的自回归动作生成方法，由于逐步预测动作，容易导致误差累积；仅依赖自回归模型难以捕捉任务间的长期依赖与全局约束，缺乏对未来状态的预判和整体策略规划。
实时响应速度较慢	VLA模型的核心应用场景是与动态变化的物理世界进行持续的实时交互，对模型的决策实时性提出了较高的要求。当前VLA模型在性能与速度之间面临固有矛盾，为追求更强的泛化能力与长时程推理能力，研究者倾向于采用更大规模的VLM或集成更复杂的任务规划架构，但这几乎不可避免地导致了推理时间的延长。

资料来源：袁霆宇《具身智能大模型综述及展望》,兴业证券经济与金融研究院整理

目 录 CATALOGUE

02 国外人形大模型进展

国外典型的具身智能大模型有谷歌DeepMind 的RT-2、 特斯拉的FSD、Physical AI的π0、Figure AI的Helix、英伟达的GROOT N1。当前，基于基础模型的机器人模型在广泛的多模态多样本数据上进行预训练，并可以通过微调适应各种多过程复杂任务。诸如谷歌的RT-1/RT-2 和PaLM-E、Physical AI的 πdπo-FAST/π0.5、 以及Figure AI的Helix等机器人大模型已在机器人控制领域展现出实力，包括端到端控制、对象泛化性、快速高效训练、零样本能力和复杂决策和快速动作的同时实现等。

表：国外典型具身智能大模型

	RT-1	RT-2	FSD	Wo	Helix	GROOTN1
所属公司	谷歌		特斯拉	Physical AI	Figure AI	英伟达
发布时间	2022年10月	2023年7月	•	2024年10月	2025年2月	2025年3月
模型类别	VLA大模型					
模型架构	单模型架构			分层双系统架构		
	将图像和文本指令直接映射为机器人动作的、纯粹的端到端Transformer模型	以VLM模型(PaLI-X或**PaLM-E**)为主干网络	实现多层认知的统一，视觉捕捉事实、语言推理因果、动作生成结果	1)预**训练VLM:**3B参数**的PaliGemma** 2)**动作专家模块；**300M参数规模	1)**系统S1:**80M参数的Transfomer模型 2)**系统S2:**7B参数的预训练VLM模型	1)**系统S1**:基于扩散变换器(DiT)的动作模块 2)**系统S2**:预训练VLM——**Eagle-2**
训练数据类型	**真实数据：**13台机器人持续17个月采集的超13万条任务片段	**联合微调：**大规模互联网数据+机器人真机数据	**“数据-模型-验证”闭环生态：**车辆生成数据→模型学习→仿真验证→结果反馈→再训练	1)**预训练：**VLM使用大规模互联网数据；动作专家模块使用开源真机数据集和基于遥操作采集的真机数据； 2)**后训练：**高质量真机数据	1)**系统S1:**机器人真机数据 2)**系统S2:**大规模互联网数据	**预训练：**真实机器人演示数据、合成数据(Omniverse生成)以及互联网上的人类视频数据

资料来源：机器觉醒时代公众号，芝能看科技公众号，兴业证券经济与金融研究院整理

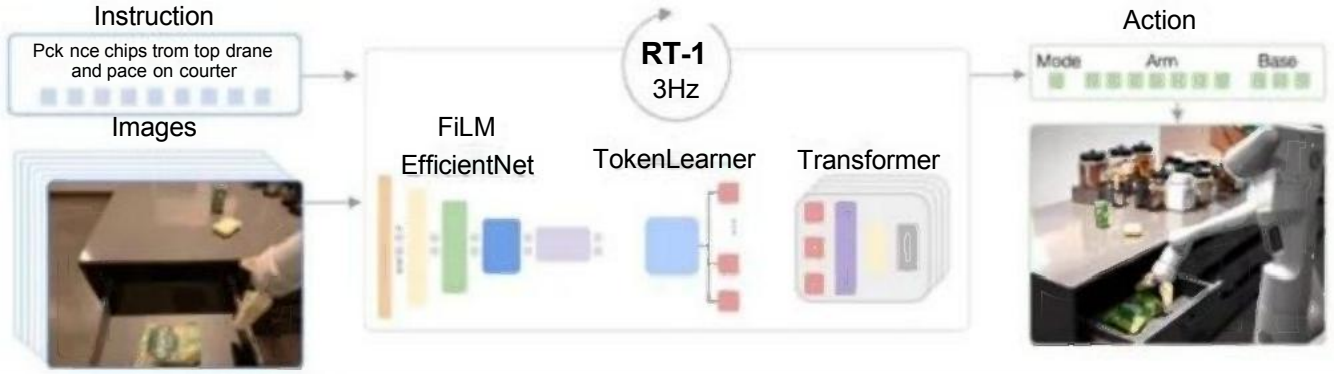
02 谷歌DeepMind RT-1：端到端技术路线代表

谷歌DeepMind 的RT 模型是端到端技术路线代表。2022年10月，谷歌DeepMind 发布RT-1 模型，其训练数据源自13台机器人持续17个月采集的超13万条任务片段。该研究开创性地将Transformer 的应用向前推进——将语言和视觉观测到机器人动作的映射视为一个序列建模问题，并利用Transformer 学习这一映射。

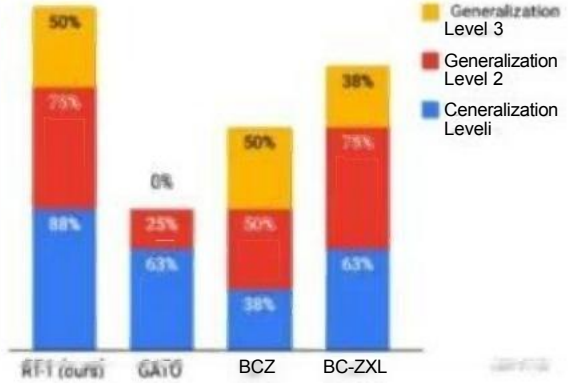
RT-1 执行闭环控制，并以3Hz 的频率持续输出动作指令，直至触发"终止"动作或达到预设时间步上限。首先通过ImageNet 预训练的卷积网络Efficient Net处理图像，该网络通过FiLM模块与指令的预训练嵌入向量进行条件调节；随后采用令牌学习器 (token Learner)生成紧凑令牌集(set of tokens)；最终由Transformer 对这些令牌执行注意力计算，输出离散化动作令牌(action token)。

RT1 限制：RT1 是纯low-level controller的任务，训练的时候不会从互联网规模的丰富语义知识中受益，机器人控制数据成本高，数据集小(130k)，模型泛化性能差模型参数量少(35M)，无法具有理解和推理能力。

图：RT-1 工作流程图



图：RT-1 泛化性较佳(2022年10月)



资料来源：机器觉醒时代公众号，机器人大讲堂公众号，CVHub, 兴业证券经济与金融研究院整理

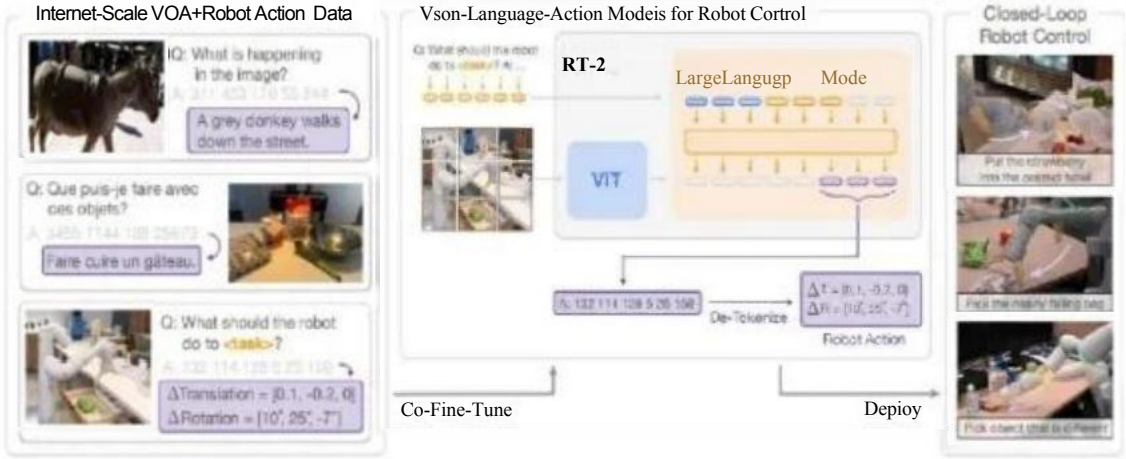
02 谷歌DeepMind RT-2：单模型端到端架构

与RT-1 关注模型的泛化能力相比，RT-2 的目标是训练一个学习机器人观测到动作的端到端模型，且能够利用大规模预训练视觉语言模型的益处。RT-1 是利用预训练模型对视觉与语言进行编码，然后再通过解码器输出动作。与之不同，RT-2 把语言、动作、图片放在一个统一的输出空间，利用VLMs 产生语言，也可以理解为“动作”为特殊的语言。总的来说，RT-2 分为两步：首先对VLMs 在大规模互联网数据进行预训练，然后在机器人任务上微调。

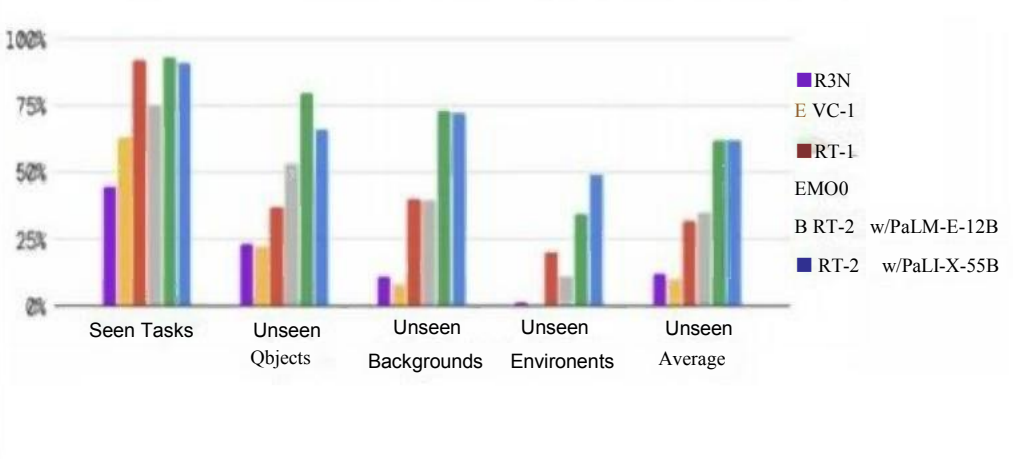
在已见任务上，RT-2 与RT-1 表现相近，而其他基线模型的成功率较低。在未见任务上，两种RT-2 模型整体泛化能力表现相近，但两者都明显优于其它基线模型，较次优的两个基线模型(RT-1 和MOO) 实现了约2倍的性能提升，较其他基线模型(VC-1 和R3M) 提升约6倍。

RT-2 限制：1) 可以执行更加复杂的指令，但是不能泛化到新的行为上，因为网络上获取的数据只能帮助模型学会更多的视觉语义信息，无法学会新的行为；2) 由于机器人数据集总体量级很少，并且无法搜集机器人没有做过的行为；3) 可以从人类行为的视频中提取数据；4) 无法实时推理：机器人控制模型需要模型能够实时推理，RT-2 参数量太大，无法实时推理。

图：RT-2全流程



图：RT-2模型与基线模型的泛化能力表现结果(2023年8月)



资料来源：机器人大讲堂公众号，机器觉醒时代公众号，Anthony Brohan《RT-2:Vision-Language-Action Models Transfer Web Knowledge to Robotic Control》，兴业证券经济与金融研究院整理

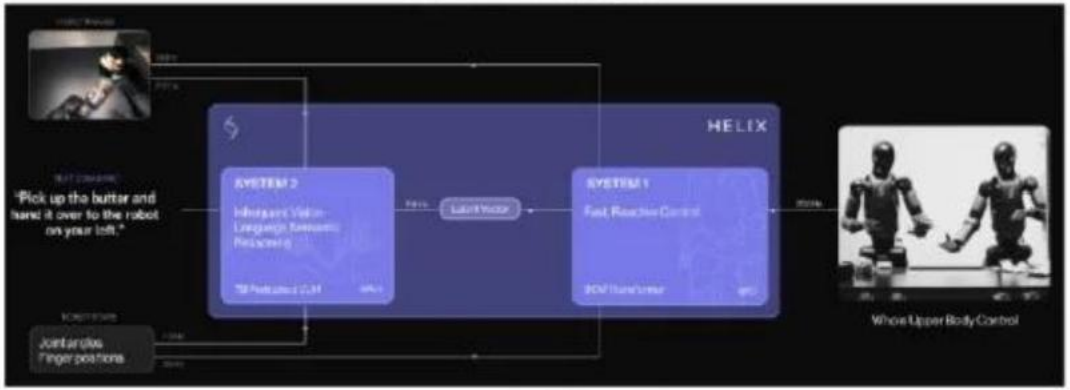
02 Figure AI Helix：首个机器人双系统VLA模型

2025年2月20日，Figure AI推出了Helix，这是第一个通过自然语言直接控制整个类人上半身的VLA 模型。

Helix是机器人领域首创“系统1+系统2”的VLA模型。VLM 主干是通用的，但速度不快，而机器人视觉运动策略是快速的，但不够通用。Helix通过两个互补的系统解决了这一权衡，允许每个系统在其最佳时间尺度上运行。S2 可以“慢慢思考”高层次目标，而S1 可以“快速思考”机器人实时执行和调整的动作。例如，在协作行为中，S1 可以快速适应伙伴机器人不断变化的动作，同时保持S2的语义目标。

训练效率较高，可以基于较少的资源实现了强大的物体泛化。Figure总共使用了约500小时的高质量监督数据来训练Helix，仅仅是之前收集的VLA 数据集的一小部分（<5%），并且不依赖多机器人具身收集或多个训练阶段。尽管数据要求相对较小，但Helix可以扩展到更具挑战性的动作空间，即完整的上身人形控制，具有高速率、高维度的输出。

图 ：Helix 双系统架构



表：Helix架构核心特征

维度	Helix模型
核心技术架构	• “系统1(S1)+系统2(S2)”双系统架构： • S1(快执行)：约8000万参数的视觉一运动Transformer，运行频率高达200Hz，负责将S2的语义意图转化为精确、连续的35个自由度全身动作； • S2(慢思考)：约70亿参数的开源视觉语言模型，运行频率7-9Hz，负责场景理解、语义解析和高层任务规划； • 两个系统通过潜空间向量进行异步通信，端到端联合训练。
关键性能	• **全上半身精细控制：**能协调控制头部、躯干、手腕及单根手指 • **零样本强泛化：**仅凭自然语言指令即可抓取和处理数千种训练时从未见过的家庭物品；**真正的多机器人协作：**单一套模型权重即可同时控制两台机器人协作完成长时序任务； • **高效训练：**仅使用约500小时的高质量监督数据就达到强大泛化能力，数据需求仅为传统VLA模型的一小部分。
商业化应用前景	• **部署优势：**模型可完全在嵌入式低功耗GPU上运行，无需依赖云端，响应速度快； • **核心场景：**主要瞄准家庭服务(整理、收纳)和物流(分拣、搬运)等复杂、非结构化的环境。

资料来源：Figure官网，机器之心公众号，兴业证券经济与金融研究院整理

02 π。:采用“双专家混合架构”的VLA模型

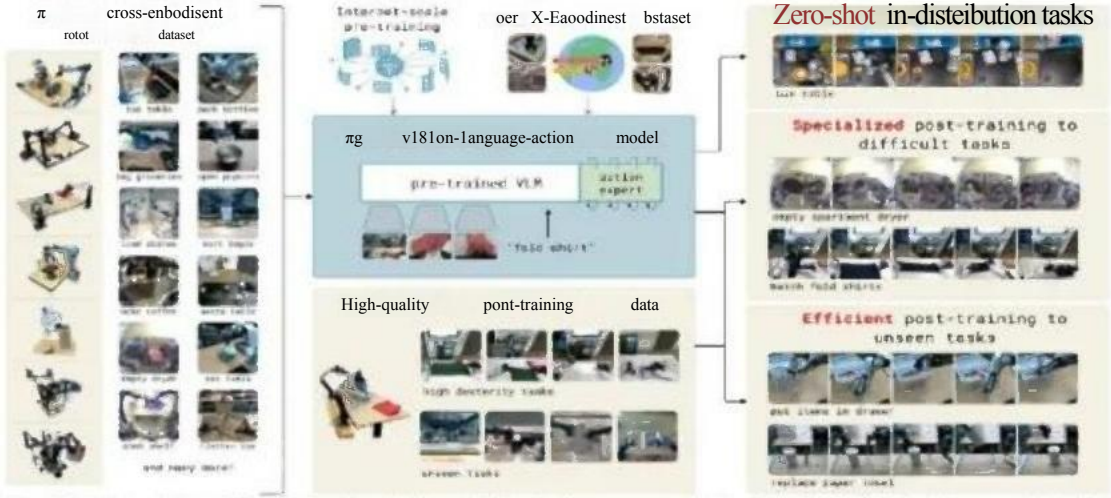
2024年10月， Physical Intelligence发布π模型。π。的创新之处包括：

1) 通过多任务大规模数据训练，使得机器人在各种复杂场景中可以展现广泛适应能力。为实现这一目标，研究团队在预训练阶段使用了OXE 数据集（涵盖22个机器人）和自采的π数据集（包括7个机器人和68个任务），这两个数据集的总时长超过10000小时，丰富了训练数据的多样性和物理交互场景，为模型提供了强大的泛化能力和物理智能。

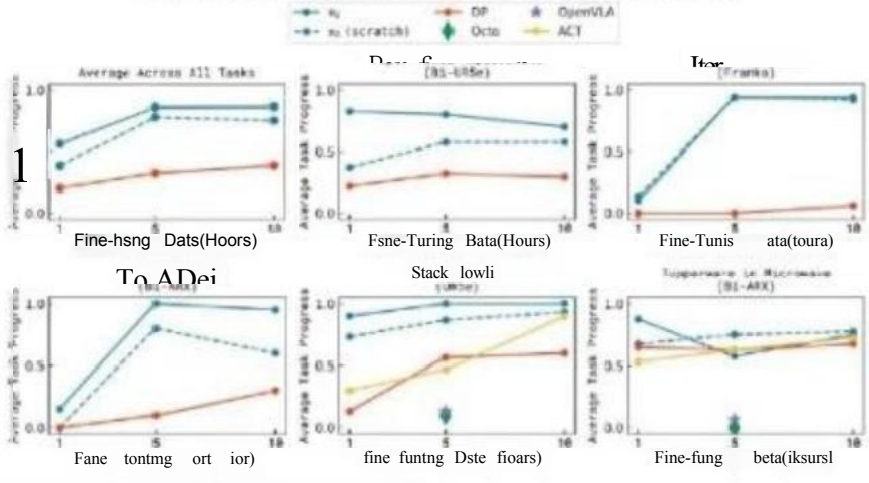
2) 采用了条件流匹配(Conditional Flow Matching)技术和动作分块算法(Action Chunk)，使得模型能够以50Hz 的高频率生成连续的动作分布，从而精准控制机器人执行复杂且高度灵巧的任务。

3) 训练策略上，采用了预训练+微调的模式。 在预训练阶段，模型通过大规模、多任务数据集学习到广泛的物理能力（预训练数据集的质量不需要非常高，以量取胜），而在微调阶段，通过少量高质量的任务数据，专注于提高模型在特定任务中的表现，如叠衣服、清理餐桌等复杂的灵巧任务。

图：π整体架构



图：微调后的π在所有任务中表现均优于其他方法



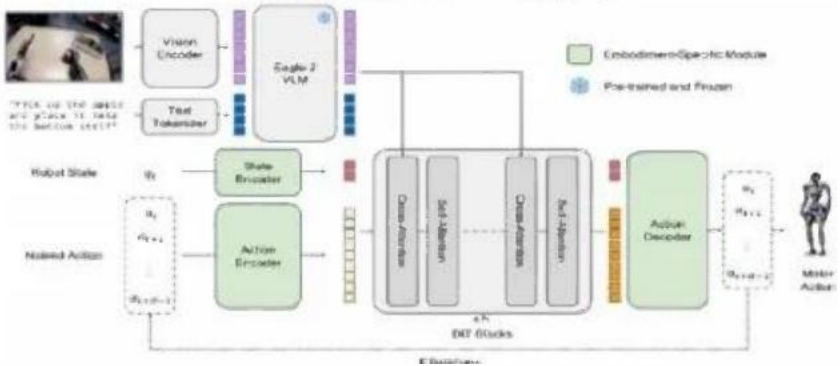
资料来源：CLEAR Lab公众号，Kevin Black《π。A VisionLanguage-Action Flow Model for General Robot Control),兴业证券经济与金融研究院整理

02 英伟达GROOT N1: 通用人形机器人开源基础模型

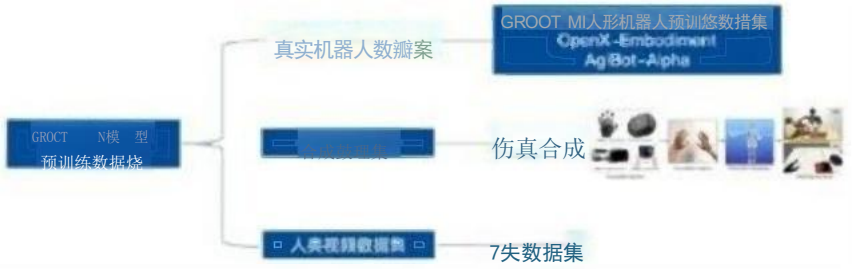
英伟达GROOT N1 模型通过创新的双系统架构的VLA 模型设计和数据金字塔策略，在复杂操作任务中实现17%的成功率提升，并支持跨硬件平台快速适配，帮助机器人高效地完成桌面操作任务。可以支持多种机器人形态，在模拟和真实环境测试中，且它泛化性强，能有效帮助机器人高效完成桌面操作任务。系统1是扩散Transformer 模块，以120Hz 的高频率生成动作；系统2是基于预训练的VLM模型，以10Hz 的频率运行，负责感知环境和任务目标。两个模块紧密耦合，通过端到端联合训练实现协同工作。

英伟达GROOT N1模型的预训练数据集分为：真实机器人数据集、合成数据集和人类视频数据集。

图：英伟达GROOT N1系统架构



图：英伟达GROOT N1预训练数据集构成



图：人类视频数据集样本

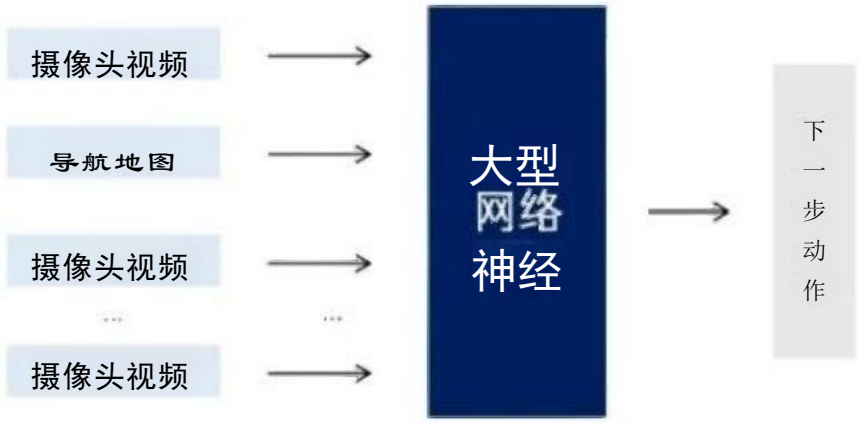
资料来源：NVIDIA《GROOT N1:An Open Foundation Model for Generalist Humanoid Robots》，深蓝具身智能公众号，兴业证券经济与金融研究院整理

02 特斯拉：FSD端到端大模型到具身智能的统一架构

特斯拉FSD 已演化为多模态大模型系统，采用VLA 形式的模型体系，将语言推理引入自动驾驶闭环中 。特斯拉的FSD 核心网络输入包括七路高分辨率摄像头视频、车辆自身运动信息、导航与音频信号。输出则包含语义分割、占用网格、3D 高斯特征、语言表达以及最终的控制动作，FSD 或已接入视觉-语言-动作 (VLA) 框架，使模型具备“解释”与“思考”的能力。现在特斯拉的FSD架构实现多层认知的统一：视觉捕捉事实，语言推理因果，动作生成结果。

Optimus 可以沿用特斯拉汽车 FSD（完全自动驾驶）系统的技术栈，采用端到端的大模型路线。汽车FSD与Optimus 大模型两者都是：多模态输入（视觉、触觉、语音、导航）；实时动作输出〔转向、步态、手臂运动〕；基于同样的世界模型推理，区别仅在于车的身体是车轮，机器人的身体是关节。特斯拉目前已经将视频生成系统直接用于Optimus 的仿真与运动规划，Optimus 可以与FSD 一样，在“虚拟世界”中学习走路、抓取、避障等操作。

图：FSD基础模型



图：不同的Optimus动作在世界模拟中被准确反映

资料来源：芝能看科技公众号，智能超参数公众号，兴业证券经济与金融研究院整理

目 录 CATALOGUE

03 国内人形大模型进展

当前，国内具身智能大模型领域发展迅速，一批高性能模型相继涌现，包括星动纪元ERA-42、 银河通用GraspVLA、 智元机器人GenieOperator-1(GO-1)、 灵初智能Psi-R1以及字节Seed 的GR-3 等。

表：国内典型具身智能大模型

	ERA-42	GraspVLA	GO-1	Psi-R1	GR-3
所属公司	星动纪元	银河通用	智元机器	灵初智能	字节跳动
发布时间	2024年12月	2025年1月	2025年3月	2025年4月	2025年7月
模型类别	VLA大模型				
模型架构	分层系统架构				
	1)高层次规划：7B参数的Instructblip视觉语言模型 **2)低层次控制：40M参数的Transformer架构模型**	**1)VLM:** IntemLM2 1.8B版本+视觉编码器+可训练投影器 **2)动作专家模块：**基于流匹配的架构	**1)VLM模型：** IntemVL-2B多模态大模型 **2)专家模块1:**Latent Planner(隐式规划器) **3)专家模块2:**Action Expert(动作专家】	**1)上层规划Planner:**基于自回归生成机制的Causal VLM架构 **2)下层控制Controller:** 采用DiT模块	采用Mixture-of-Transformers(MoT)的网络结构，把“视觉-语言模块”和“动作生成模块”结合成了一个40亿参数的端到端模型
训练数据类型	**1)预训练：**大规模互联网视频数据 **2)后训练：**少部分真机数据	**1)预训练：**大规模仿真合成数据 **2)后训练：**少部分真机数据	**1)VLM模型：**大规模的互联网图像和文本数据 **2)专家模块1:**大量人类操作和跨本体操作视频 **3)专家模块2:**机器人真机数据	—	**1)遥操作获取的机器人数据：**利用了遥操作收集机器人动作轨迹 **2)人类VR轨迹数据：**使用VR设备采集人类的轨迹数据进行学习 **3)公开可用的图文数据：**从公开可用的图片和文字中认识物体、学习抽象概念

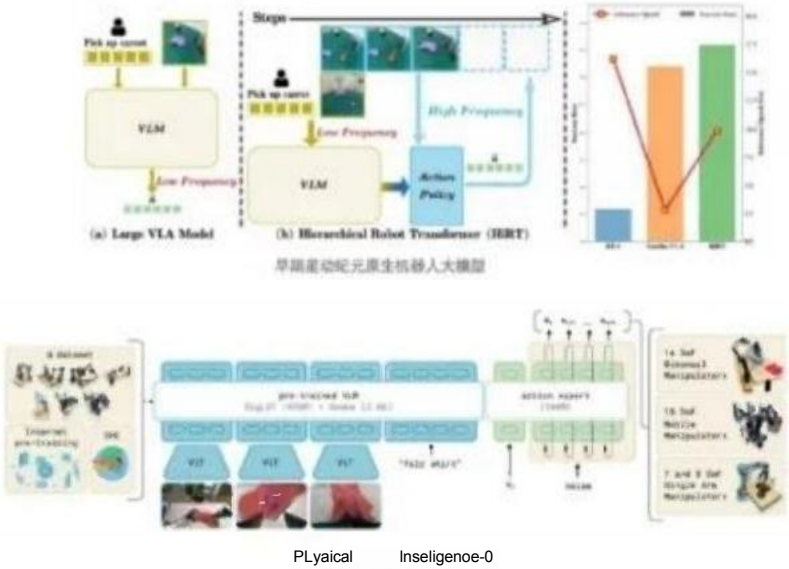
资料来源：机器觉醒时代公众号，字节跳动Seed 公众号，兴业证券经济与金融研究院整理

星动纪元ERA-42：国内首个端到端原生机器人大模型

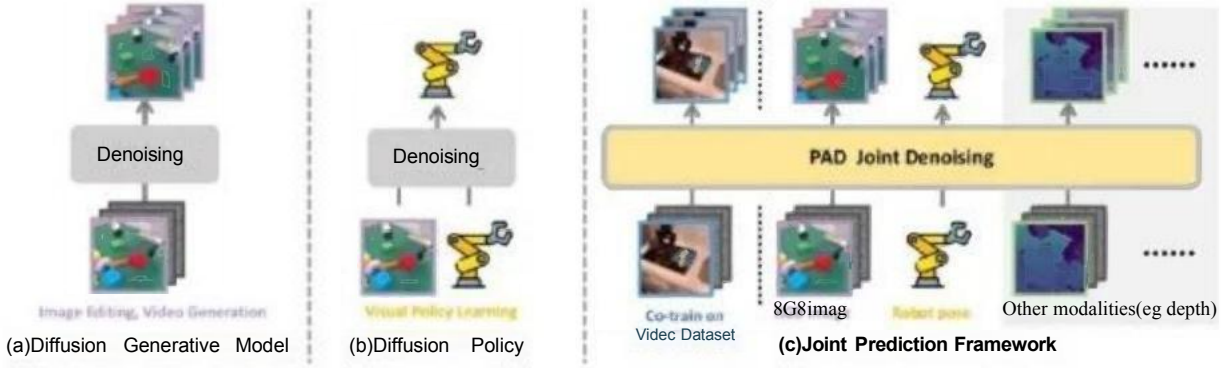
星动纪元ERA-42 是国内首个真正意义上的端到端原生机器人大模型，比肩世界领先水平。星动纪元采用端到端的算法来提升其原生机器人大模型性能。PI推出的π。模型，通过结合互联网规模的视觉-语言预训练与机器人操作数据集后训练，使得机器人能够在人类环境中自主执行多种复杂任务，而星动纪元采用大规模视频数据学习策略，涵盖无标注的视频数据、公开各类形态机器人的数据、人类活动数据以及遥操作数据等。

星动纪元原生机器人大模型ERA-42 能理解物理世界与预测未来。星动纪元已将世界模型融入原生机器人大模型中，使其模型不仅具备行动能力，还具备了对物理世界的理解能力，能够对未来行动轨迹进行预测，有效提升了机器人执行任务的高效性和准确性。

图：星动纪元早期模型与π。对比



图：星动纪元将世界模型融入原生机器人大模型



(a)Diffusion Generative Model (b)Diffusion Policy **(c)Joint Prediction Framework**

资料来源：星动纪元公众号，兴业证券经济与金融研究院整理

03 星动纪元ERA-42：国内首个端到端原生机器人大模型

软件赋能硬件：基于ERA-42 大模型，星动纪元灵巧手可实现精细化操作。相比夹爪，基于ERA-42 的能力，五指灵巧手星动XHAND1 已经能够使用包括不限于螺钉钻、锤子、取液枪等更多种多样的工具，完成更通用、灵巧性更强、复杂度更高的百种以上操作任务。

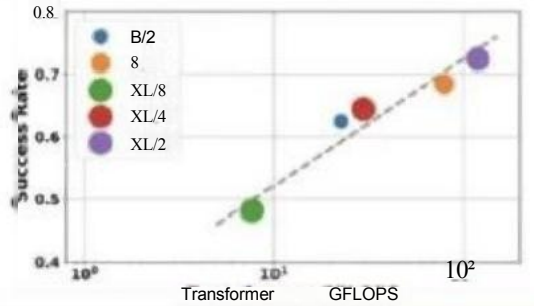
通过基于大规模视频数据的预训练，ERA-42 具备更强泛化能力，只需采集少部分数据，短时间内就能学会执行新的操作任务。以灵巧手为例，每一种操作都是通过一句自然语言文本或语音，以及摄像头的感知姿态作为输入，直接端到端输出执行操作，能够泛化到新的、未见过的环境或任务，即便面对未曾接触过的物体，灵巧手也能顺利完成操作任务。此前，星动纪元技术团队就通过这种训练方式采集简单的红黄蓝方块抓取数据，成功实现了从未见过的多样化物体(如胡萝卜、茄子等)的抓取泛化。

ERA-42 **初步体现“Scaling效应”，即模型规模与性能之间的正相关性。**研究结果表明，随着模型规模的扩大，任务成功率也明显提升，初步体现了和大语言模型训练中类似的“Scaling 效应”。

图：通过简单彩色方块的抓取数据实现多样化物体的泛化抓取操作

图 ：ERA-42 初步体现“Scaling效应”

	PAD-XL/2	PAD-XL/4	PAD-XL/8	PAD • L/2	PAD-B/2
Layers	28	28	28	24	12
Hidden size	1152	1152	1152	1024	768
Heads	16	16	16	16	12
Token length	257	65	17	257	257
Parameters	661M	661M	661M	449M	128M
Gflops	119.1	29.5	7.3	79.1	22.5
Average SR	72.5%	64.5%	48.2%	68.4%	624%



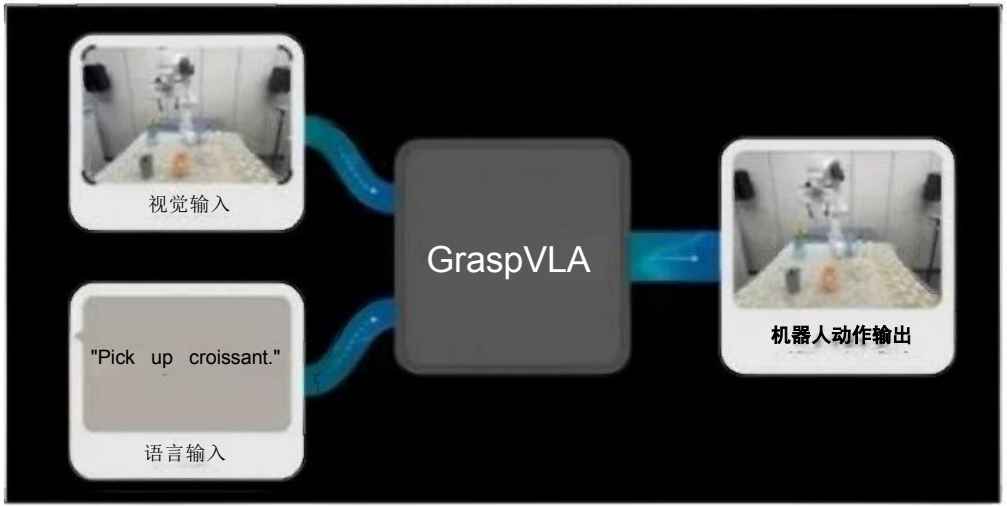
资料来源：星动纪元公众号，兴业证券经济与金融研究院整理

03 银河通用GraspVLA：合成数据预训练的抓取模型

总体架构：GraspVLA 将视觉-语言模型(VLM) 与动作专家集成，并通过渐进式动作生成(PAG) 机制连接。VLM 获取观察图像和文本指令，用于视觉-语言联合感知。用条件流匹配动作专家进行细粒度的末端执行器动作生成。进一步引入PAG，以便将从互联网基础数据集中学习的知识有效地迁移到抓取技能。

GraspVLA是全球首个合成大数据驱动的基础抓取大模型，使用合成数据预训练+少部分真机数据后训练。该模型通过十亿帧合成数据的预训练，掌握了包括光照、背景、位置、高度、动作策略、动态干扰和物体类别在内的七大泛化能力，能够在真实场景中实现零样本(Sim2Real) 抓取，无需额外训练即可应对未见过物体的复杂摆放和动态环境变化。针对特定工业需求(如抓取接线座、三角板等特殊零件)，仅需少量真实数据进行后训练即可快速迁移能力。

图：端到端具身抓取基础大模型



图：GraspVLA使用合成大数据驱动

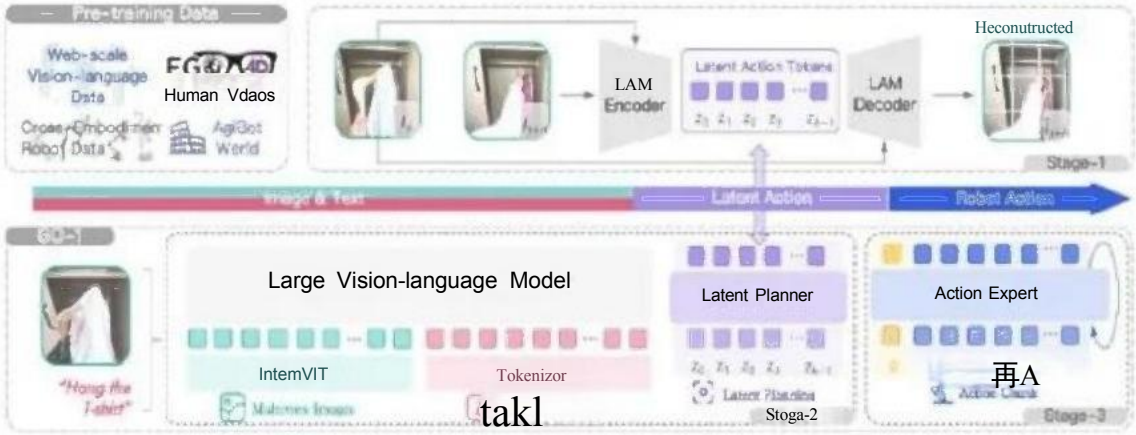
资料来源：大语言模型和具身智体及自动驾驶公众号，银河通用机器人公众号，焉知机器人公众号，兴业证券经济与金融研究院整理

03 智元机器人G0-1：基于ViLLA架构的具身智能模型

2025年3月10日，智元机器人与上海人工智能实验室联合推出通用具身基座大模型GO-1。GO -1 模型基于智元机器人2024年年底发布的AgiBot World数据集开发，并开创性提出了Vision-Language-Latent-Action(ViLLA) 架构，实现了利用人类视频进行学习和小样本快速泛化，显著降低了具身智能的技术门槛。

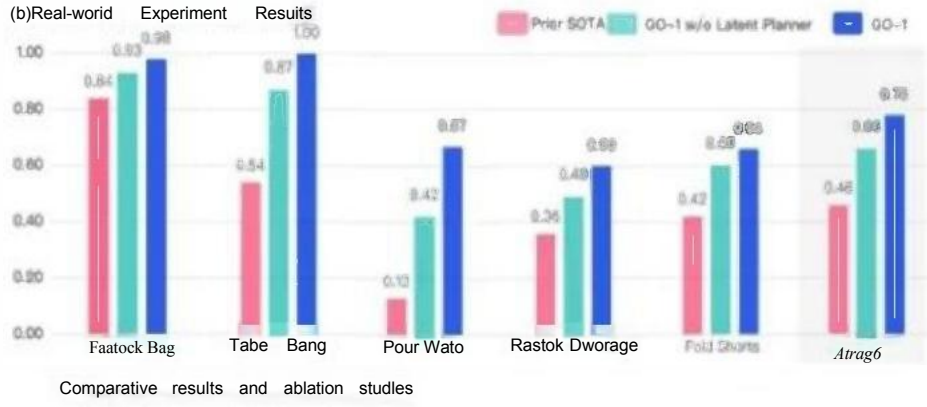
与VLA 架构相比， ViLLA 能有效利用高质量的AgiBot World数据集以及互联网大规模异构视频数据，增强策略的泛化能力。 ViLLA 架构由VLM （多模态大模型）+MoE （混合专家）组成。VLM 作为通用具身基座大模型的主干网络，借助海量互联网图文数据获得通用场景感知和语言理解能力， MoE 中 的Latent Planner（隐式规划器）借助大量跨本体和人类操作数据获得通用的动作理解能力， MoE 中 的Action Expert（ 动 作专家）借助百万真机数据获得精细的动作执行能力。在推理时， VLM、Latent Planner和Action Expert三者协同工作。

图：GO-1 ViLLA架构



The framework of GO-1

图：相比已有的最优模型，GO-1 处理复杂任务成功率提高32%（2025年3月）



Comparative results and ablation studies

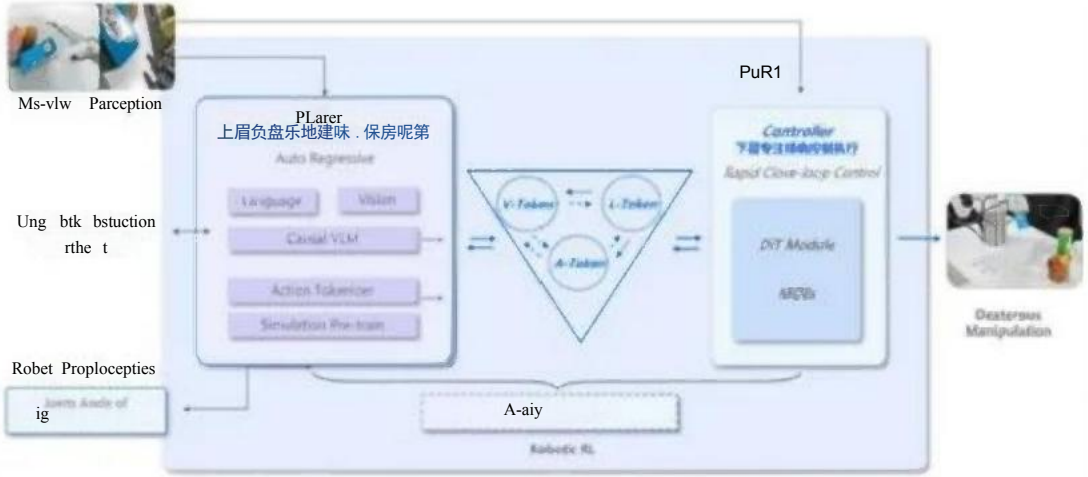
资料来源：IETCSR智能系统与机器人公众号，人形机器人联盟公众号，上海经信委公众号，兴业证券经济与金融研究院整理

03 灵初智能Psi R1：分层端到端快慢脑架构

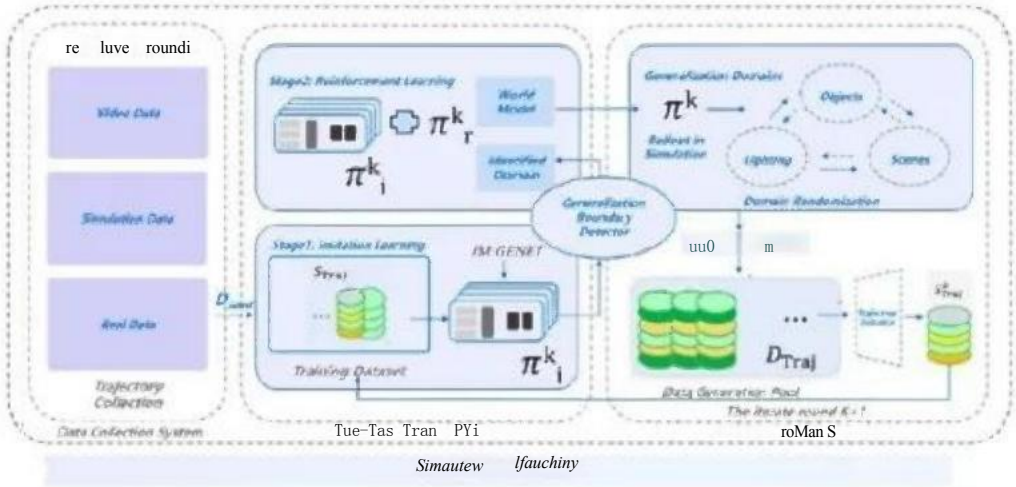
技术路线：分层端到端快慢脑架构， 灵初智能与Figure AI、Google、Nvidia技术方案一直，均采用分层端到端架构。其中S1 为快脑，负责运动控制、快速响应； S2 为慢脑，负责推理、规划、深度思考。

数据架构：基于强化学习，构建了仿真-真机-互联网一体的数据采集范式，突破模仿学习的瓶颈，较大提升数据的使用效率。灵初智 能会先通过遥操作和互联网收集等方式获得一个冷启动的数据集，先输入进模型进行模仿学习训练，接着用强化学习训练灵巧操作技能，然后结合模仿学习和强化学习，对这些场景数据里的物体、灯光、背景等要素随机化处理，提升模型的泛化能力。经过反复迭代，可以大幅提升模型的技能泛化表现，同时提高数据的采集利用效率。

图：R1模型理解和执行人类任务



图：灵初智能PsiR1数据架构



资料来源：灵初智能公众号，甲子光年公众号，兴业证券经济与金融研究院整理

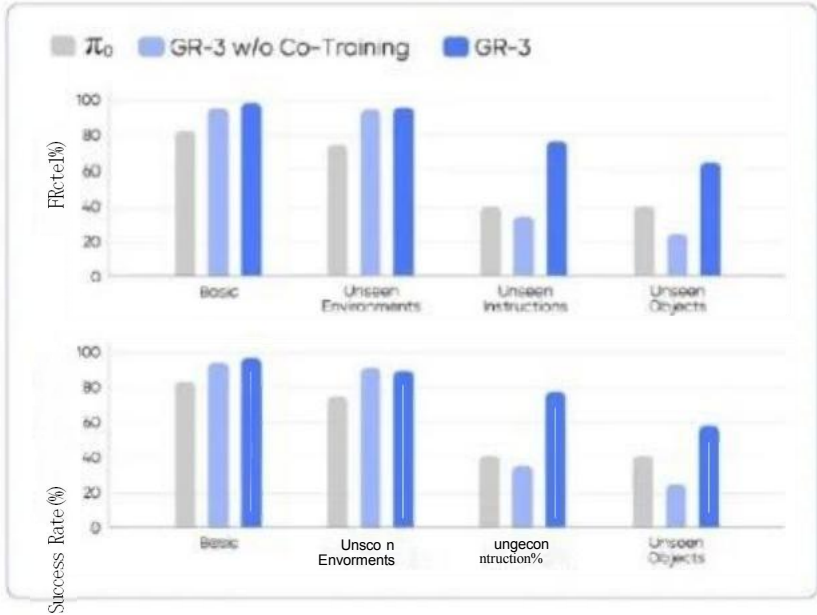
03 字节跳动GR-3:40 亿参数的大规模VLA模型

架构：VLA 模型GR-3 由字节跳动Seed 团队发布，采用40亿参数的混合变换器架构。GR-3 分别在机器人轨迹上使用流匹配目标和在视觉-语言数据上使用下一个标记预测目标进行共同训练。

训练策略： GR-3 能够从三种类型的数据中学习：视觉-语言数据、机器人轨迹数据和人类轨迹数据，通过多源数据融合训练策略，在机器人操作任务中实现了较强的泛化能力，在精细操作任务中的成功率提升高达250%，为通用机器 人的实际部署提供了可行的技术路径。

图：GR-3 数据类型及模型架构

图：GR-3 可泛化抓取-放置能力较强(2025年7月)



资料来源：字节跳动Seed公众号，深蓝具身智能公众号，兴业证券经济与金融研究院整理

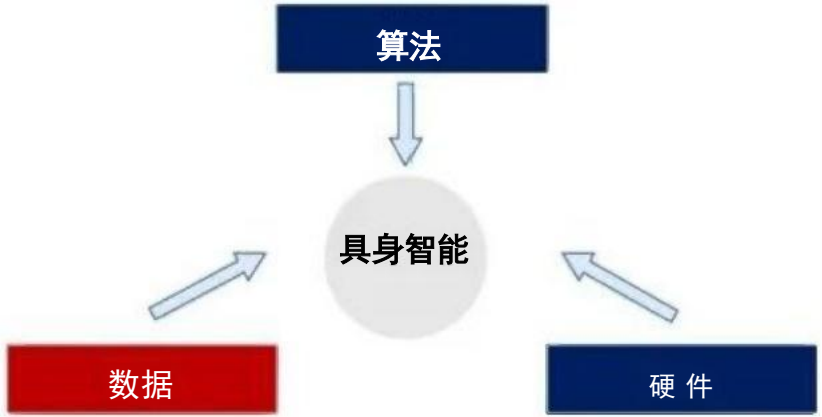
目 录 CATALOGUE

04 数据：驱动具身大模型迭代升级的关键

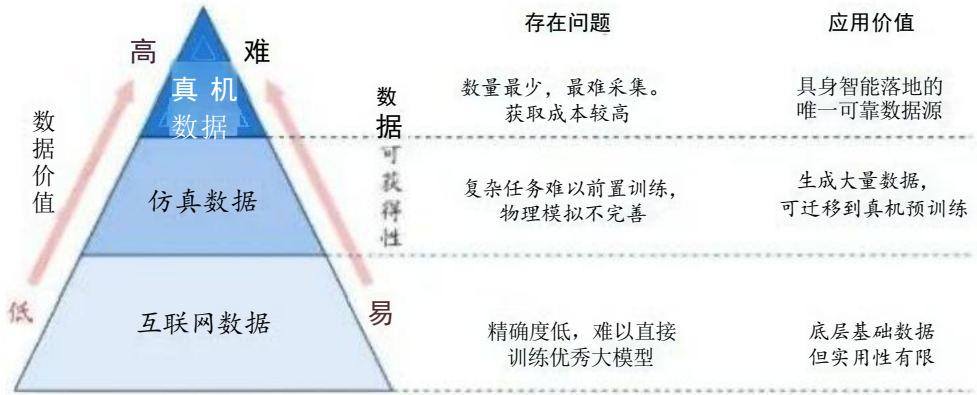
具身智能的实现需要高密度、高质量的数据。伴随具身智能技术架构由手工编程转向数据驱动的端到端大模型，数据需求从低量单一模态数据逐步升级为海量、多模态、高精度和跨任务长程数据。

真机数据价值最高，获取难度最大，是具身智能落地的可靠数据源。当前具身智能训练数据集呈金字塔结构分层，顶层为高质真机数据，是唯一能支撑复杂任务落地的数据源，但数量稀缺、采集成本较高。

图：构建通用机器人大模型的三个关键要素



图：具身智能训练数据集呈金字塔结构分层



表：具身智能的实现需要高密度、高质量的数据

阶段	数据量	数据类型	数据预量	数据处理方式	核心目标
G1	低	单一传感器数据 如视觉、触觉	低情度，实时性为主	基于预输程规刚	特定场景任务扶行
G2	中	多场果任务数据、环境数据	需标注清流	机器学习与任务编排	原子能力复用与初步泛化
G3	高	多模态数据(视觉、触觉、动作序列)	高精度对齐	端到端数据整动训练	送用训练框架开发
G4	极高	跨场景真实数据+仿真数据	物理原理级标注	大操作模里(LAM	跨技能泛化与物理理解
G5	极大量	多模态长程任务数据、开放场景数据	多样化精度	LL M+LAM融合	开放场景全面泛化

表：真机数据价值最高，获取难度最大，是具身智能落地的唯一可靠数据源

AI类别	大语言模型	自动驾驶	具身智能
代表	ChaGP135	特斯拉FSD	替元机器人
数据类型	文本	视频	真机数据一仿真数据
数据量	45TB	每月上亿英里	100万条真实轨迹+1000万条仿真数据
采集成本	低	中	高

资料来源：Yuke Zhu《Pathway to Generalist Robots:Scaling Law,Data Flywheel,and Humanlike Embodiment》，智元2024年度新品发布会，量子位，汽车知识图谱，智元机器人官网，兴业证券经济与金融研究院整理

04 人形机器人数据采集方式

机器人数据飞轮： 通过部署机器人执行真实世界任务，自动收集数据并用于优化模型，从而使机器人表现更好、适应更多场景，进而收集更大量、更多样数据，形成越用越强的正向循环。

真实数据采集方式： 主要分为VR 遥操作数据采集系统、数据手套 遥操作系统、机械臂主从控制数据采集系统及UMI(Universal Manipulation Interface, 通用操作接口)。

图：机器人数据飞轮

图：机器人学习数据采集方案

方案	图例	原理	优点	缺点
VR遥操作		通过人工穿戴VR眼镜实时检测手部(或手中拿的手柄)的位姿，进而映射到机器人的操作任务上	采集到的数据精度高	数据采集成本高昂，人员培训成本高，效率低
主从臂摇操作		通过人工操作从臂，带动主臂做相同的运动	数据精度较高	体积较大，对人手操作不友好；采集数据类型有限，难以适应复杂操作
数据手套		通过直接采集人手的位姿，映射到灵巧手上，包括IMU惯性数据采集手套、光纤数据采集手套、光学动捕手套	高精度动作捕捉，实时动态映射校准	仅捕捉手部、需要配合其他设备使用
UMI (通用操作接口)		人类通过手持夹爪来执行操作，收集人类在真实环境的演示	便携、低成本、硬件门槛低	存在精度等限制

资料来源：Yuke Zhu《Pathway to Generalist Robots:Scaling Law,Data Flywheel,and Humanlike Embodiment》，3DCV公众号，机器人产业应用公众号，AGI具身智能公众号，Xbotics具身智能实验室公众号，兴业证券经济与金融研究院整理

04 特斯拉：数采方案或转向视频学习

特斯拉第一代数采方案：动捕手套+动捕服。2023年5月，特斯拉的遥操作设备和动捕首次对外披露，该操作系统由特斯拉自制，五个摄像头安置于头盔上，并连接到采集人员背负的沉重背包上。手上是动捕手套，上衣是动捕服。较低延迟和高保真度的远程操作系统，用于收集模仿人类执行某些任务的机器人的人工智能训练数据

特斯拉第二代数采方案： VR 眼镜。2024年5月，特斯拉的遥操作设备转变为VR 眼镜，主要优势为设备轻量化程度提升，提升采集效率。

特斯拉数采方案或转向视频学习。2025年5月， Optimus 学习了更多技能，如炒菜、倒垃圾、拉窗帘、放置零件等，根据前负责人介绍，Optimus 通过学习第一视角的视频数据学习新的技能，未来的目标是通过第三视角视频来学习新的技能。特斯拉已经大大降低了真实数据的比例，提高了视频数据的比例，有助于降低数采成本。

图：特斯拉第一代数采方案：动捕手套+动捕服

图：特斯拉第二代数采方案：VR 眼镜

资料来源：具身纪元公众号，兴业证券经济与金融研究院整理

04 银河通用：物理仿真数据为主，真实数据为辅

银河通用数据范式为“虚实结合、以合成为主、真实为辅”。银河通用主要通过计算机图形学手段，将真实世界的物理完全搬进仿真环境，即构建一个大规模的物体资产库，让机器人可以与各种柔性物体进行交互，例如抓取、放置、开门、开抽屉、使用遥控器等，继而大规模生成这些动作数据。

合成数据技术细节： 在可交互的物体资产上，利用银河通用积累的合成管线来产生大量的动作，这些大量的动作可以通过强化学习的方式让机器人自主摸索生成，不管是直接合成还是强化学习试错生成的数据，都可以通过仿真器进一步检验，再用渲染器转变成视觉数据，最终直接把模型部署在真实世界。

图：银河通用大模型训练范式

图：银河通用机器人数据合成仿真+Sim2Real路线

资料来源：银河通用机器人公众号，银河通用创始人王鹤在世界机器人大会演讲，兴业证券经济与金融研究院整理

目 录 CATALOGUE

05 投资建议

传感器收集外部环境和人形机器人自身状态数据，为具身智能大模型决策提供依据，从而实现机器人与物理世界的交互，建议关注人形机器人传感器相关公司：安培龙(一/六维力传感器)、汉威科技(柔性触觉传感器)、福莱新材(柔性触觉传感器)、奥比中光(卡位3D视觉传感器)。

作为高质**量运动数据的关键来源，动捕采集方案是人形机器人学习与优化其控制策略中不可或缺的一环，建议关注具备动捕解决方**案的相**关公司：**凌云光(全资子公司元客视界推出Al动捕产品Fzmotion，服务宇树科技、优必选等客户具身智能产品的研发与训练)。

表：人形机器人传感器标的业绩预测及估值分析(2026.4.7，iFinD机构一致预期)

公司简称	股票代码	市值(亿元)	归母净利润(亿元)				PE		
			2024	2025/2025E	2026E	2027E	2025/2025E	2026E	2027E
安培龙	301413.SZ	110.8	0.83	1.08	1.41	1.84	102.6	78.7	60.2
汉威科技	300007.SZ	135.5	0.77	1.50	1.57	1.93	90.4	86.5	70.2
福莱新材	605488.SH	97.4	1.39	1.22	1.55	2.08	80.2	62.9	46.8
奥比中光	688322.SH	299.3	-0.63	1.27	3.02	4.75	235.6	99.0	63.0
凌云光	688400.SH	203.0	1.07	1.61	2.60	3.75	126.1	78.1	54.2

资料来源：iFinD，各公司公告，兴业证券经济与金融研究院整理

注：安培龙、福莱新材2025年利润采用iFinD机构一致预期；汉威科技、奥比中光、凌云光2025年利润采用业绩预告或业绩预告区间中值；兴业证券为奥比中光、凌云光做市商