机器人大模型深度报告
我们距离真正的具身智能大模型还有多远?

首席证券分析师：周尔双
执业证书编号：S0600515110002
zhouersh@dwzq.com.cn

证券分析师：钱尧天
执业证书编号：S0600524120015
qianyt@dwzq.com.cn

研究助理：陶泽
执业证书编号：S0600125080004
taoz@dwzq.com.cn

2025年8月9日

投资要点：

1. 人形机器人为何需要高智能的大模型?

尽管人形机器人的形态早已实现工程可行，但其真正实现产业化落地的关键，在于摆脱传统工业机器人“控制刚、泛化弱”的局限，补足对不确定性的理解与适应能力。工业机器人主要基于确定性控制逻辑运行，缺乏感知、决策与反馈能力，导致高度依赖集成，成本高、通用性差。相比之下，人形机器人以“通用智能体”为目标，强调感知—推理—执行的完整链路，必须依托大模型支撑的多模态理解与泛化能力，才能适应复杂任务与动态环境。当前多模态大模型的兴起，为人形机器人提供了“初级大脑”，开启从0到1的智能进化，并通过数据飞轮实现模型能力与产品性能的持续提升。然而整体智能化仍处于L2初级阶段，通往泛化智能仍面临建模方法、数据规模与训练范式等多重挑战，高智能大模型将是贯通通用人形机器人路径的核心变量。

2. 从架构端和数据端看，目前机器人大模型的进展如何?

当前机器人大模型的快速演进，主要得益于架构端与数据端的协同突破。架构上，从早期的SayCan语言规划模型，到RT-1实现端到端动作输出，再到PaLM-E、RT2将多模态感知能力融合至统一模型空间，大模型已逐步具备“看图识意、理解任务、生成动作”的完整链条。2024年π0引入动作专家模型，动作输出频率达50Hz;2025 年Helix实现快慢脑并行架构，控制频率突破至200Hz，显著提升机器人操作的流畅性与响应速度。数据端，已形成互联网、仿真、真机动作三类数据协同支撑的结构化体系：前两者提供预训练量级与泛化场景，后者则直接提升模型在物理世界中的实用能力。其中，真机数据采集对高精度动捕设备依赖度高，光学动捕以精度优势适配集中式训练场，有望成为具身模型训练的核心数据来源。当前主流训练范式正由“低质预训练+高质后调优”快速迭代，模型智能的跃迁正转向“从数据堆料到结构优化”的阶段。

3. 未来大模型的发展方向是什么?

面向未来，具身大模型将在模态扩展、推理机制与数据构成三方面持续演进。当前主流模型多聚焦于视觉、语言与动作三模态，下一阶段有望引入触觉、温度等感知通道；Cosmos 等架构尝试通过状态预测赋予机器人“想象力”，实现感知—建模—决策闭环，构建更真实的“世界模型”，提升机器人环境建模与推理能力；数据端，仿真与真实数据融合训练成为主流方向，高标准、可扩展的训练场正成为通用机器人训练体系的关键支撑。

4. 投资建议

模型端建议关注【银河通用(一级公司)】【星动纪元(一级公司)】【智元机器人(一级公司)】，数据采集领域建议关注【青瞳视觉(一级公司)】【凌云光(688400.SH)】 【奥比中光(688322.SH)】，数据训练场领域建议关注【天奇股份(002009.SZ)】。

5. 风险提示

大模型技术进展不及预期，高质量数据获取受限，人形机器人需求不及预期。

1. 人形机器人为何需要高智能的大模型?

2. 从架构端和数据端看，目前机器人大模型的进展如何?

3. 未来大模型的发展方向是什么?

4. 相关标的

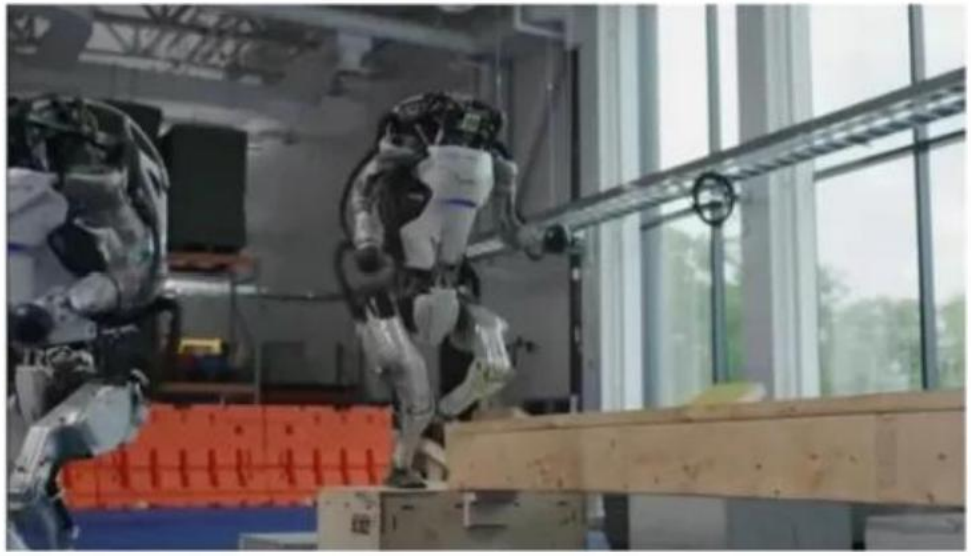
5. 投资建议与风险提示

1.1人形形态并非技术难点，核心在于通用智能的补足

- **人形形态的机器人早已实现工程落地，但长期停留在“仿形不仿智”阶段。**过去的人形机器人主要以模仿人类形态为目标，相关技术早在数十年前已初步成熟。早期典型代表如2000年本田推出的ASIMO 与2013年波士顿动力的Atlas,虽具备出色的运动能力，但执行逻辑高度依赖预设行为库。这类机器人可完成跑跳等复杂动作，体现了运动控制硬件的成熟度，但其行为均来自人工设定的指令序列，无法自主理解任务或适应环境变化。因此，本质上这类产品仍是“人形的机器”，而非“具备人类智能的机器人”。它们缺乏对环境的感知、任务的理解与泛化能力，尚不具备真正的智能交互与通用任务执行潜力。

图：日本本田ASIMO 机器人跑步图

图：美国波士顿动力Atlas机器人运动图

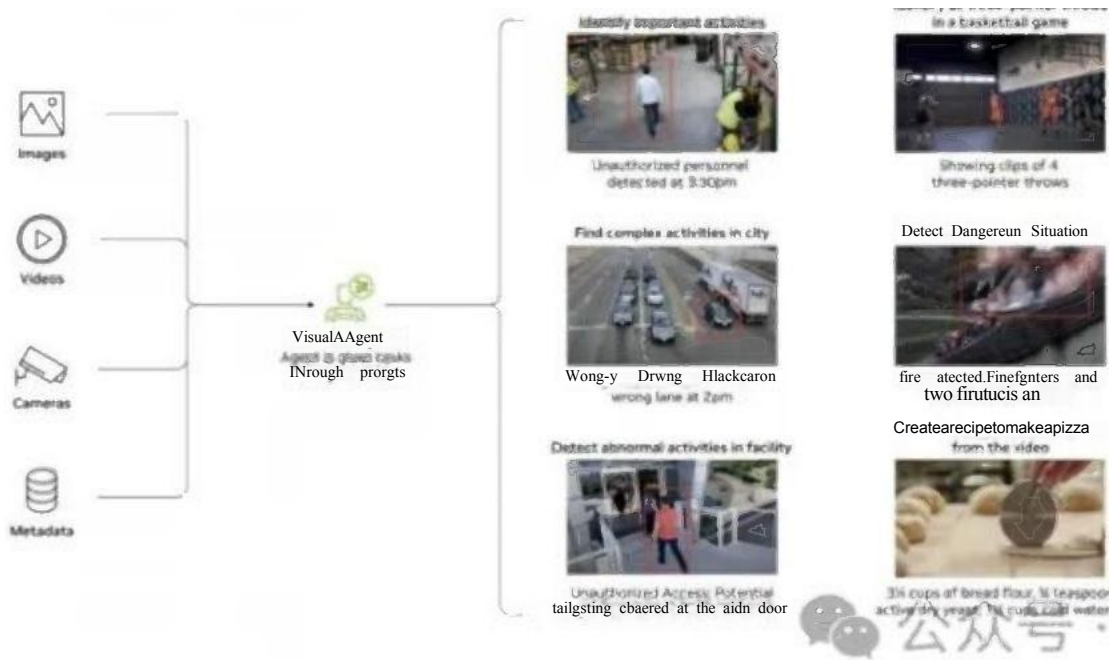


数据来源：极果公众号，环球网科技，东吴证券研究所

1.2多模态大模型的出现，为人形机器人装上“智能大脑”

- **本轮人形机器人热潮的底层驱动力，是市场对其“智能性”的高度期待。** 随着多模态大模型的突破，机器人首次具备了“感知—理解—决策”的潜力，被视为拥有“大脑”的关键起点。大语言模型 (LLM) 的成功，验证了通过大规模互联网文本训练神经网络具备推理能力的可行性；而视觉语言模型 (VLM) 进一步拓展模态边界，使模型可以“看懂图像、理解语言”。 LLM 专注于文本推理，VLM 则通过融合图像/视频与语言等模态信息，构建起跨模态的统一表征体系，从而支持模型理解现实世界的更多维度。
- **动作模态的融入，让模型端真正赋予机器人执行操作的能力。** 仅能感知、理解世界并不是机器人大脑的终极目标，机器人的最终目标是在认知的基础上实现与现实世界的动作交互。**目前机器人模型的核心迭代方向，是将动作模态融入现有的视觉语言模型。**

图：从LLM到VLM,AI 对现实世界感知不断丰富

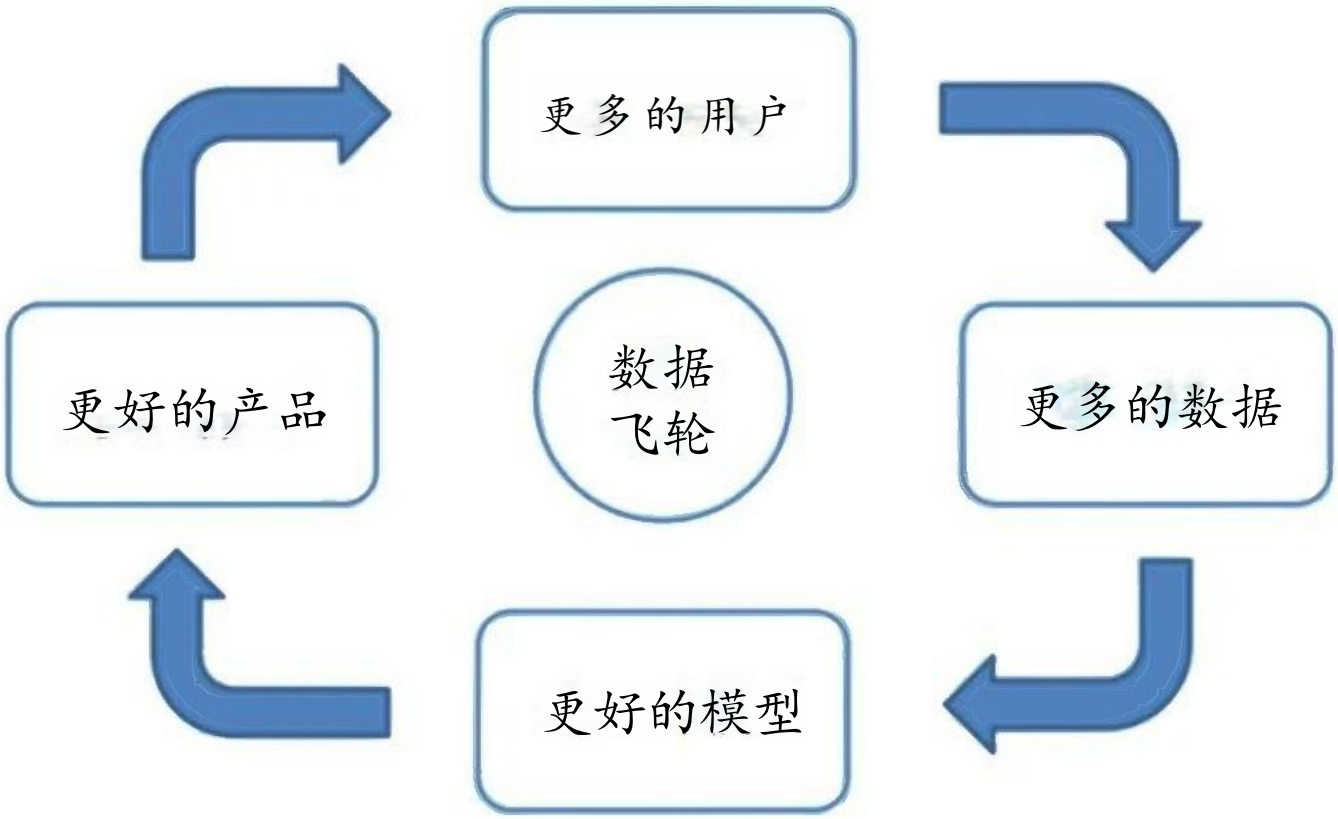


数据来源：芝能智芯公众号，东吴证券研究所

1.3初级具身智能模型撬动人形机器人产业0-1落地

- 当前大多数人形机器人仍处于展示阶段，核心瓶颈在于智能程度不足。 一旦具备初步智能化能力，机器人即可在特定场景中落地应用，并通过任务反馈不断优化模型，开启数据飞轮与产品迭代循环，从0-1迈向1-100的演化。
- **数据飞轮是驱动智能系统能力提升的核心机制。** 本质是“收集数据—改进模型—提升产品—吸引更多用户和数据—再次改进”的正向循环，有望带动人形机器人快速迈入迭代加速期。

图：数据飞轮



数据来源：东吴证券研究所整理

1.4当前模型水平有限，距离真正泛化仍有较远距离

- 现阶段人形机器人仅在智能化的初级阶段。北京市人形机器人创新中心牵头，联合上海市、浙江省人形机器人创新中心，以及优必选、宇树科技、中国信息通信研究院、工业互联网研究院等多家头部企业与科研机构，共同制定了全球首个《人形机器人智能化分级》标准，从感知、决策、执行、协作四维度划分L1-L5五级。目前主流产品智能水平普遍仅在L2左右，尚未具备自主泛化与应变能力。未来向更高智能等级进化仍需突破模型、数据与算力多重门槛。 真正实现通用智能机器人仍有较长路径要走，需在技术、数据体系和生态协同上持续积累。

表：L1-L5五级智能化能力分级体系

维度	能力等级	核心能力描述
P感知认知能力	P1-P5	单模态感知→多模态融合→场景理解→跨领域认知→ 自主知识构建
D决策学习能力	D1-D5	规则执行→简单推理→任务规划与学习→知识迁移→ 自我演进
E执行表现能力	E1-E5	基础运动→多任务协调→工具运用→复杂操作→类人灵活执行
C协作交互能力	C1-C5	单模态响应→多模态理解→情绪识别→个性化交豆→群体协同

数据来源：中关村融智特种机器人产业联盟公众号，东吴证券研究所

1. 人形机器人为何需要高智能的大模型?

2. 从架构端和数据端看，目前机器人大模型的进展如何?

3. 未来大模型的发展方向是什么?

4. 相关标的

5. 投资建议与风险提示

2.1发展历程：三条主线驱动，加速技术衍变

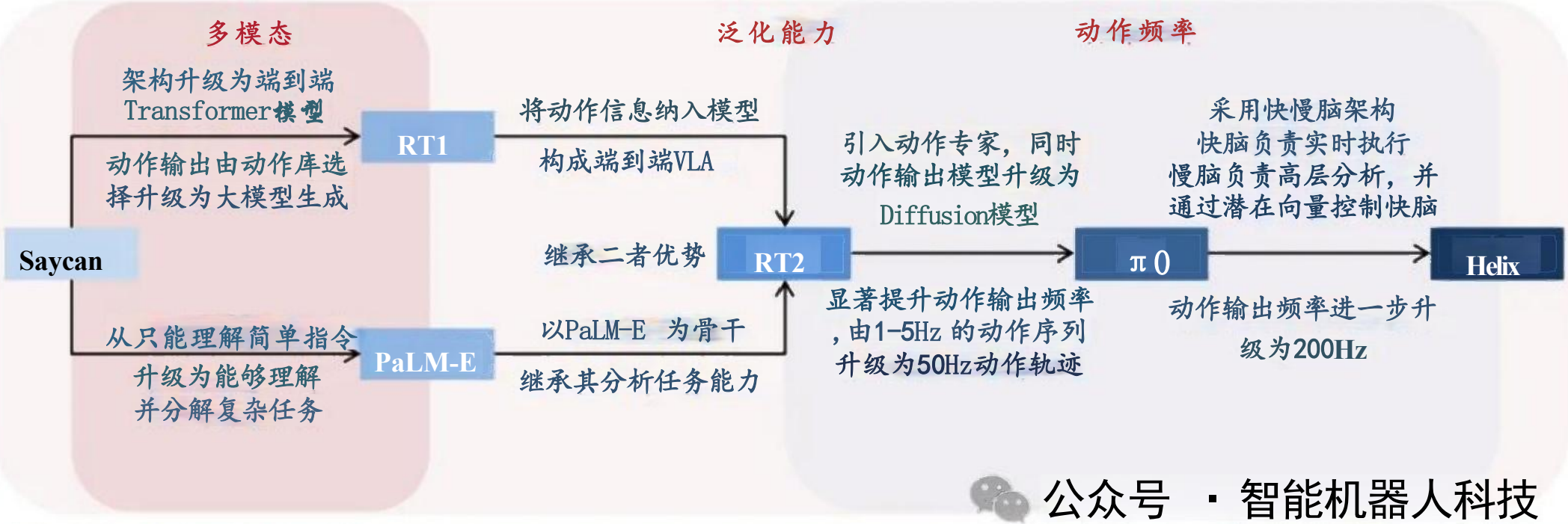
●多模态、动作频率和泛化能力三条主线驱动技术衍变。

1）多模态： 22年4月Saycan发布，能够根据任务指令在动作库中输出最优动作。22年12月RT1 发布，动作输出升级为由Transformer生成的动作Token。23年3月PaLM-E 发布，较Saycan在任务理解能力上显著升级。23年7月RT2发布，结合RT1和PaLM-E两者优势，将动作信息纳入模型输出空间。

2）动作频率： RT2 只能输出1-5Hz 的动作序列，为克服这一问题。24年10月π0发布，引入采用FlowMatch 模型的动作专家，动作输出升级为50Hz 的动作轨迹。25年2月Helix发布，采用快慢脑结构，操纵频率进一步提高，输出200Hz动作序列。

3）泛化能力： 由于现实世界极其复杂，不可能通过枚举穷尽所有场景，因此机器人必须具备“零样本泛化”能力。纵观模型发展史，各模型均强调多任务联合训练、预训练迁移能力以及跨平台迁移能力，核心目的就是提升零样本泛化表现。

图：机器人大模型发展历程



数据来源：东吴证券研究所整理

2.2.1 SayCan：语言模型与可行性评估结合的任务规划框架

- SayCan 是由Google 于2022年提出的一种将语言模型（LLM）与机器人可行性模型相结合的任务规划系统。它的核心思想是将自然语言指令拆解为一系列子任务，然后由LLM（如PaLM）生成可能的动作候选，再由一个训练好的“Can”模型(基于Q-learning 的 Affordance Model)评估每个动作在当前环境下的可行性。

- 具体流程：SayCan 首先接收一段用户指令(如“请帮我从厨房拿一瓶水”)，通过语言模型推理出多个可能的操作步骤(如“走到冰箱前”、“打开冰箱”、“拿出水瓶”)，之后使用训练得到的可行性模型为这些步骤打分，最终选取可执行性最高的动作序列传给机器人控制器执行。该方法巧妙地将LLM的语义理解能力与机器人在现实场景中的动态感知能力连接起来，实现了“说得通也做得到”的人机交互模式。

图：SayCan模型架构

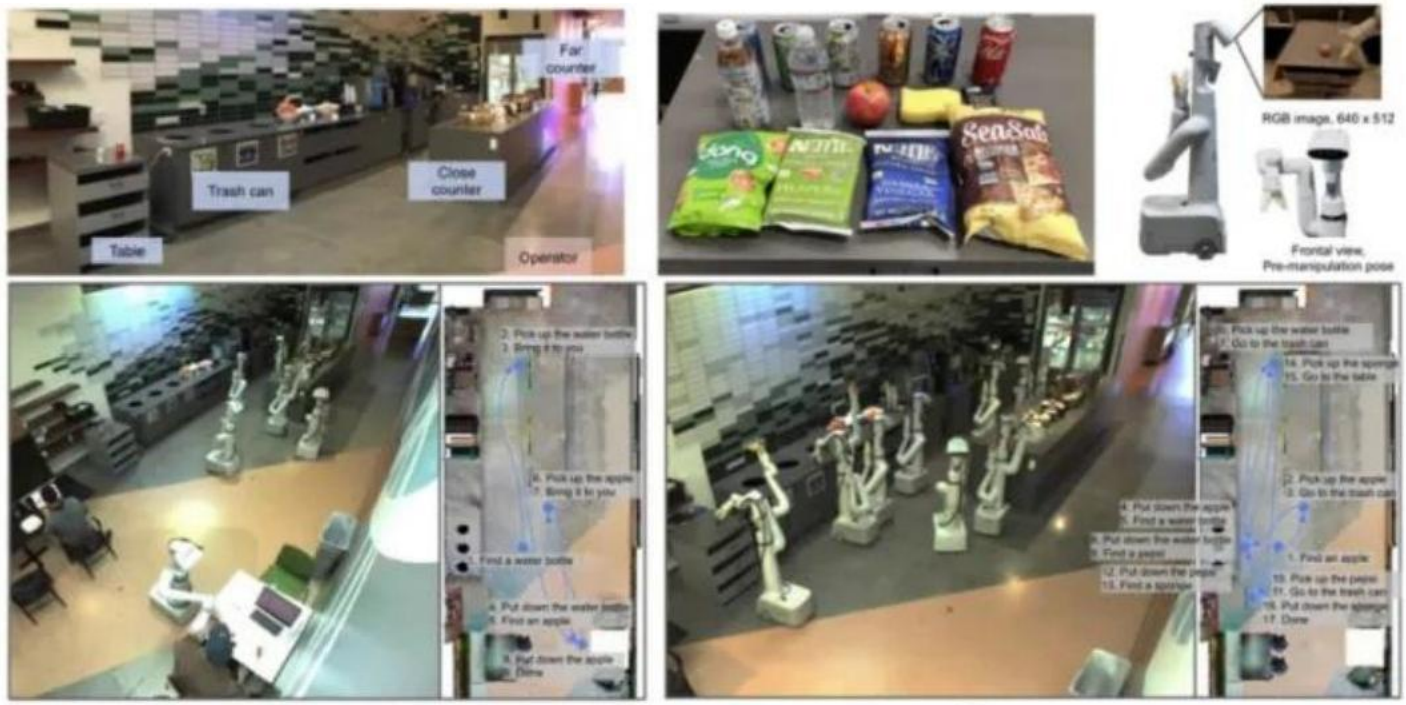
数据来源：《Do As I Can,Not As I Say:Grounding Language in Robotic Affordances》,东吴证券研究所

2.2.1 SayCan：语言模型与可行性评估结合的任务规划框架

东吴证券

- SayCan **在真实机器人平台上展示出显著优于传统方法的任务执行能力，验证了“语言+可行性打分”**架构的**有效性。** 在实验中，SayCan 在真实世界的厨房环境中执行了101项任务，规划成功率为81%，执行成功率为60%，比未经过现实世界约束的LLMs执行准确性提高了约15%
- SayCan 的两阶段结构在工程部署和模型泛化上存在一定限制，难以满足大规模通用机器人系统需求。1）语言理解模块和可行性模型是独立训练，无法实现全局端到端优化，导致部分情况下任务分解与动作打分之间存在语义脱节；2）可行性模型依赖预定义的状态特征和Q值训练，对不同场景需重新标注和学习，迁移成本高；3）系统整体对任务顺序依赖强，缺乏自主replanning 能力，难以处理开放环境中的任务中断或失败恢复。

图：SayCan 厨房任务实验场景



(a)"I just worked out, can you bring me a drink and a snack recover?"

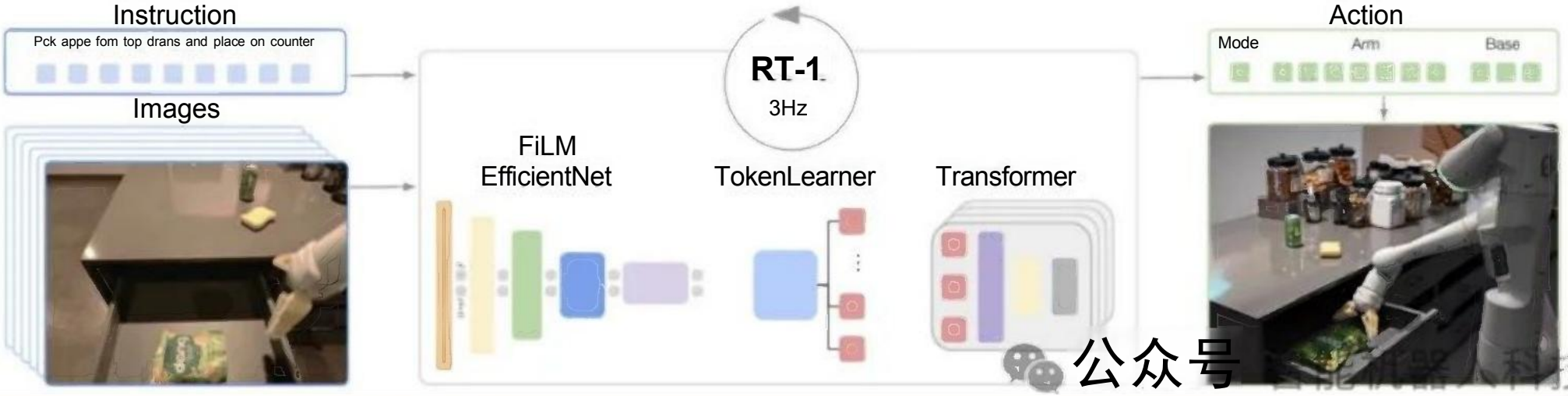
(b) "I left out a coke, apple, and water, can you throw them away and then bring me a sponge to wipe the table?"

数据来源：《Do As I Can,Not As ISay:Grounding Language in Robotic Affordances》,东吴证券研究所

2.2.2 RT-1：端到端Transformer控制模型

- **Google Robotics于2022年发布的RT-1(Robotics Transformer 1)，基于Transformer模型及**简约标记化方法，利用大规模开放式语言及视觉数据实现实时、可扩展、可泛化、适用于实际场景的机器人运动控制。RT-1使用Transformer 编码器将图像帧与语言指令联合编码，输出动作token（例如控制机械臂的末端位姿与夹爪状态）。训练数据来自于机器人远程操作演示，总共包含13万条动作轨迹，涵盖700多个任务目标。
- 具体流程：当听到用户的简单指令如“帮我拿起桌子上的一瓶水”时，EfficientNet会接收图像信息，然后FiLM层把语言指令与图像结合，输出token到Tokenlearner,并由其提取关键Token输送给Transformer,最后由Transformer 输出具体动作token序列，如关节旋转几度、电机如何运行等。
- RT-1 是业界首个大规模部署的机器人端到端Transformer 控制模型，实现了从感知输入到控制输出的全流程统一。相比SayCan 模块化架构，RT-1提供了更高效率和更强适应能力的系统训练范式，大大简化了机器人模型设计与训练流程。模型在厨房、实验室等家庭环境中实现了可扩展的多任务执行，平均任务成功率大幅超过传统控制策略。

图：RT-1模型架构



数据来源：《RT-1:ROBOTICS TRANSFORMER FOR REAL-WORLD CONTROLATSCALE》,东吴证券研究所

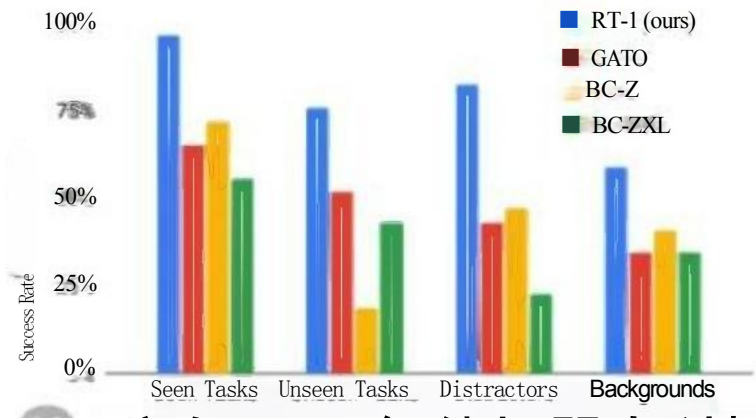
2.2.2 RT-1：端到端Transformer控制模型

- **RT-1实现了端到端的高效控制流程，在真实环境中的多任务执行中展现出极高的稳定性、泛化能**力与工程适应性。实验表明：1)RT-1可在家庭厨房场景中执行超过700项具体任务，在3000多次真实测试中平均成功率达97%,典型操作如“移动物品” “打开抽屉”等成功率超过90%;2)模型具备较强的语义泛化能力，能够理解并正确响应指令的多种表达方式，例如“请递杯子”与“帮我拿那个水杯”均能正确执行；3)具备良好的任务扩展能力，新任务可通过行为克隆(Behavior Cloning)快速适配，无需重训练整个模型，显著提升数据利用效率与部署灵活性。

- RT-1仍受限于任务平台耦合、语义理解能力弱等问题，在通用性与认知层智能上尚未突破。1)模型在特定机器人平台和场景(如厨房)上训练，迁移到其他平台需重新收集大量数据，缺乏跨平台泛化能力；2)仅使用图像和指令做输入，缺乏触觉、语音等其他模态的感知，对复杂任务(如操作失败后的反馈修正)处理力有限；3)缺乏高阶规划机制，执行策略主要依赖短期视觉反馈，难以完成逻辑顺序复杂的任务链；4)语言指令解析深度不够，面对多条件或因果逻辑类表达(如“先清理再放杯子”)的执行准确率仍不理想。

图：RT-1多维评估表现优于Gato 与 BC-Z系列模型

Model	Seen Tasks	Unseen Tasks	Distractors	Backgrounds
Gato(Reed et al.,2022)	65	52	43	35
BC-Z(Jang et al,2021)	72	19	47	41
BC-ZXL	56	43	23	35
RT-1(ours)	97	76	**83**	**59**

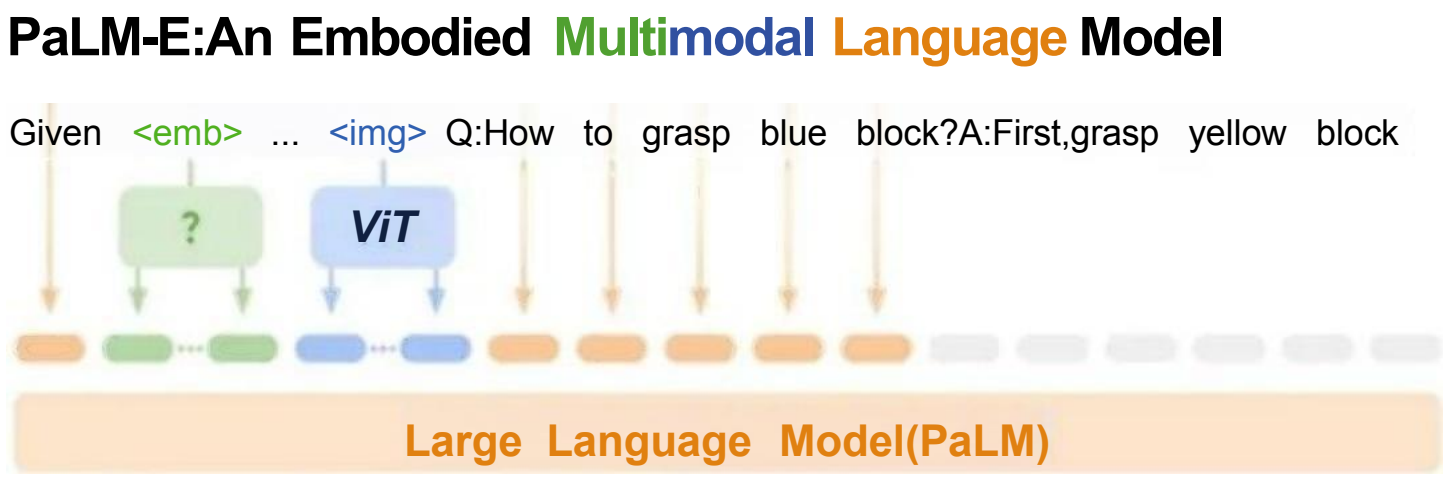


数据来源：《RT-1:ROBOTICS TRANSFORMER FOR REAL-WORLD CONTROLATSCALE》 东吴证券研究所

2.2.3 PaLM-E：多模态具身语言模型

- PaLM-E是Google 于2023年发布的融合语言、视觉和传感器状态信息的大规模多模态语言模型，能够实现具身智能任务中的高层推理与决策生成。其核心机制是在预训练的大语言模型(如PaLM) 中，注入来自视觉、状态估计等模态的连续输入，统一转化为“多模态句子”(multimodal sentence)，并作为Transformer解码器的输入。这些连续模态通过专门训练的编码器(如ViT、OSRT、MLP)映射到语言token 所在的嵌入空间，与语言token 共同组成输入序列，最终输出自然语言形式的回答或动作指令。
- 具体流程：面对“请把蓝色积木放在黄色积木上”的指令，PaLM-E 可生成一个多步文本决策：“1. 拿起蓝色积木；2. 移动到黄色积木上方；3. 放下蓝色积木”。该文本决策随后由底层控制策略(如RT-1) 执行动作。PaLM-E 在架构上实现了感知—语言—行动之间的自然耦合，具备端到端推理能力。

图：PaLM-E模型架构

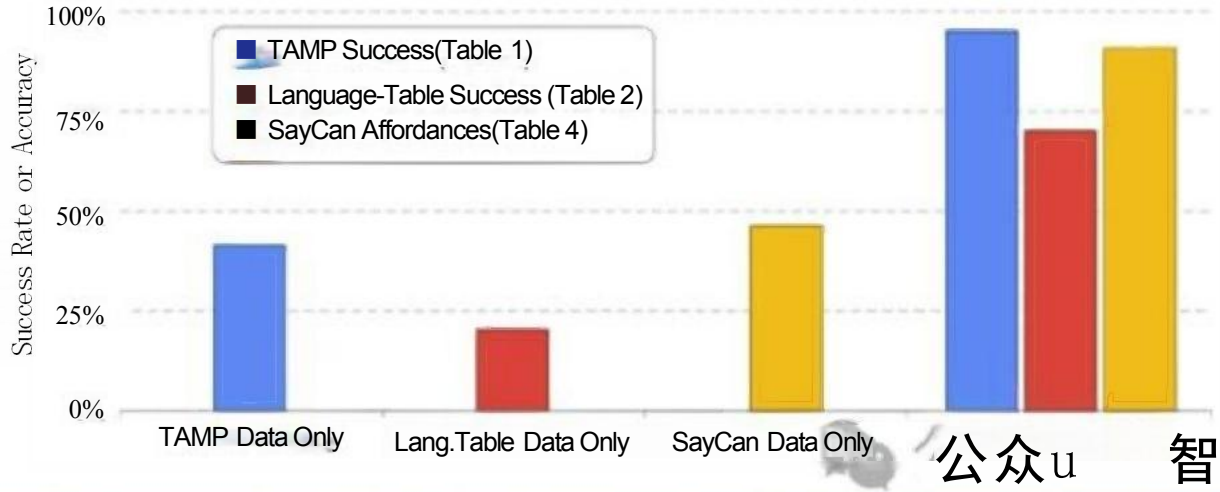


数据来源：《 PaLM-E:AnEmbodiedMultimodalLanguageModel》，东吴证券研究所

2.2.3 PaLM-E：多模态具身语言模型

- PaLM-E 在多个具身任务和视觉语言任务中展现出优秀的泛化能力和任务迁移性能。在桌面操作与移动操作环境中，PaLM-E 能生成多步语言计划并驱动真实机器人完成如“分类推积木”“从抽屉中取物”等任务，实现one-shot 和 zero-shot 泛化。此外，PaLM-E-562B 在OK-VQA 等通用视觉语言任务中取得领先成绩，并能进行多图推理、数学运算与时序感知问答等复杂推理。联合训练实验表明，通过融合多源数据， PaLM-E 在仅用少量具身数据时依然能维持高性能表现。
- **PaLM-E 在实际部署中仍面临一定挑战，主要包括模型规模、推理效率与训练门槛问题**。1) 模型体量庞大：如 PaLM-E-562B 包含540B 的语言模型与22B 的视觉编码器，推理速度与资源需求高，不适合部署在资源受限的机器人边缘设备上；2) 训练成本高：需要预训练的大模型、图像编码器与高质量具身数据，训练门槛高，数据采集效率有限；3) 低层控制依赖预设策略：高层生成的文本决策仍需靠RT-1等低层策略执行，系统整体仍未完全闭环自动学习；4) 对三维感知场景效果有限：虽然OSRT 引入了神经三维结构表示，但在高度复杂、动态交互场景中的空间理解仍有提升空间。

图：多模态混合训练显著提升PaLM-E 在多任务场景中的成功率

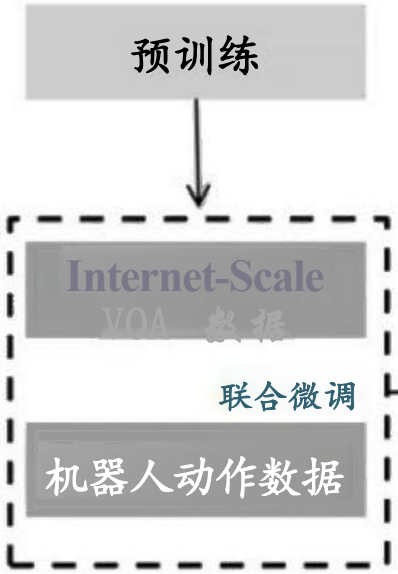


数据来源：《PaLM-E:AnEmbodiedMultimodalLanguageModel》,东吴证券研究所

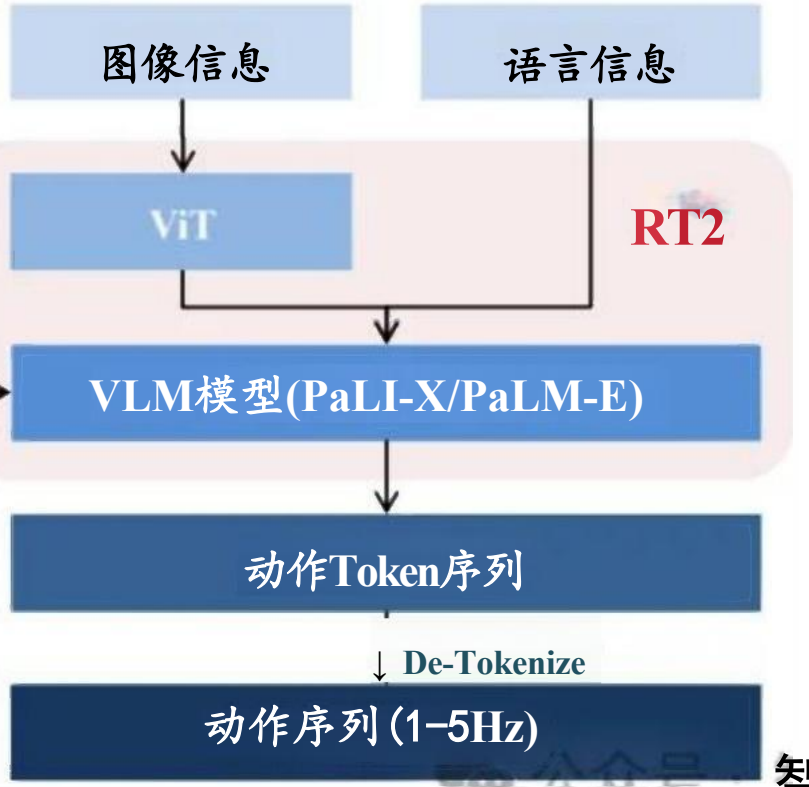
2.2.4 RT2：动作信息纳入VLM 模型，构成端到端VLA

- 架构&输出：采用经动作信息训练的VLA 模型，输出1-5Hz 的动作序列。VLM 模型以PaLM-X 或PaLM-E 为骨干，经过上述方法训练后成为端到端的VLA 模型。后者在应用中可直接分析经ViT处理的图像信息和语言信息，最后视模型大小输出1-5Hz的动作序列。
- 具体流程：当听到“帮我从冰箱里拿一瓶水的指令时”，由VLM 模型分析图像和语言信息，直接理解任务要求，并输出如手臂旋转几度、电机如何运行的动作Token序列。其相较于RT1，主要结合了PaLM-E 推理和决策的优势，增强了对任务的理解能力。

图：RT2模型训练方法



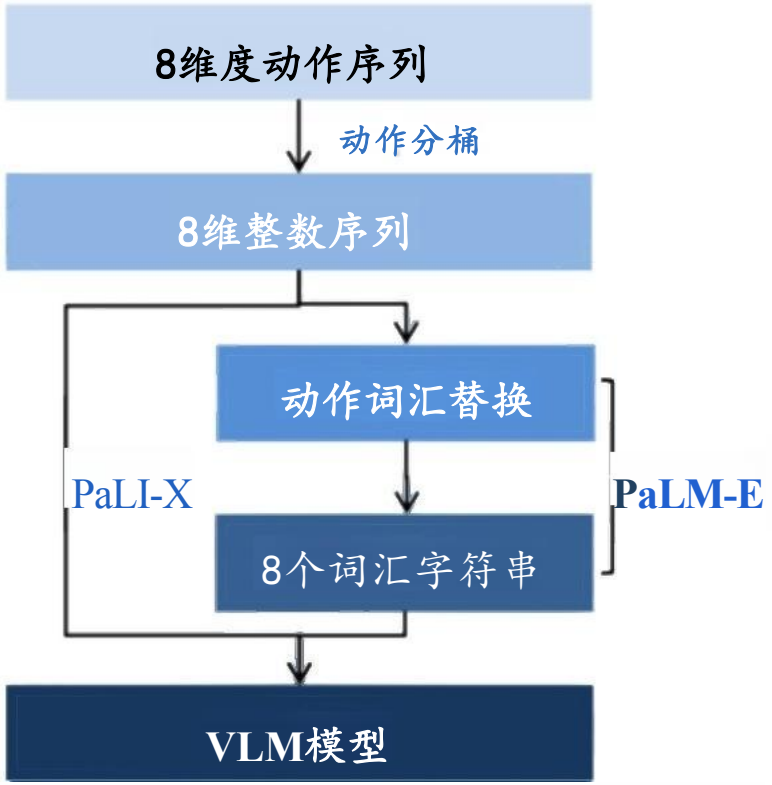
图：RT2模型架构



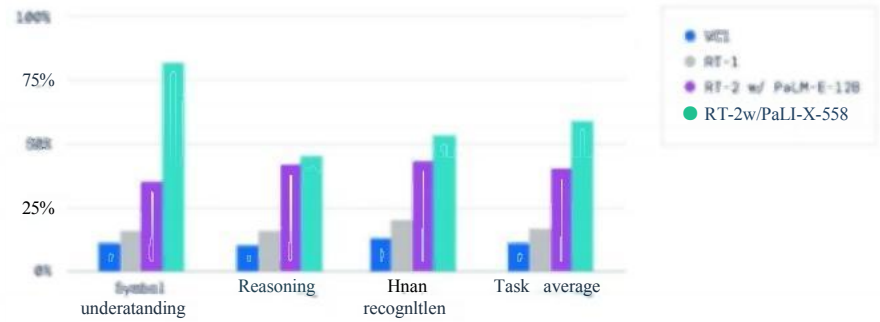
2.2.4 RT2：动作信息纳入VLM 模型，构成端到端VLA

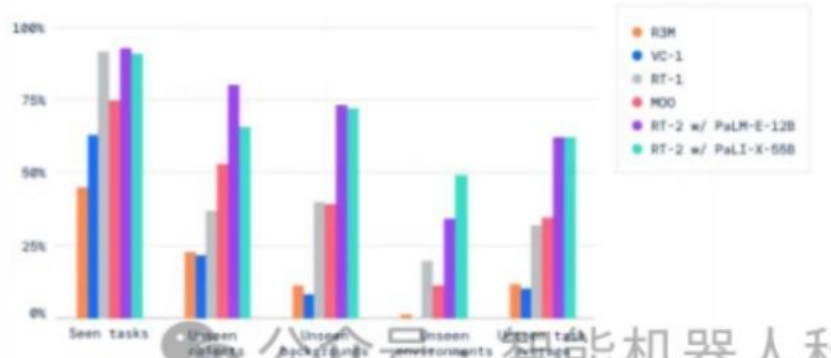
- 训练方法：利用分桶替换将动作信息转为语言信息纳入训练。RT2 使用的机器人共有8个自由度，训练过程中先将每个自由度的连续运动数据(如手臂能够偏转0-360°)划分为256个区间，每个区间对应一个二进制数。此时以PaLM-E为骨干的模型再将256个二进制数替换为256个使用频率最低的生僻字，从而将运动信息转为语言信息置入LLM 模型中训练。而以PaLI-X为骨干的模型具备1000个独立Token，二进制数序列可直接置入模型训练。
- **优势：**将动作信息纳入模型，泛化性得到增强。
- **劣势：**动作输出频率较慢，难以完成正常工作。

图：动作信息分桶替换后转为语言信息



图：较先前模型，RT2泛化性显著提升



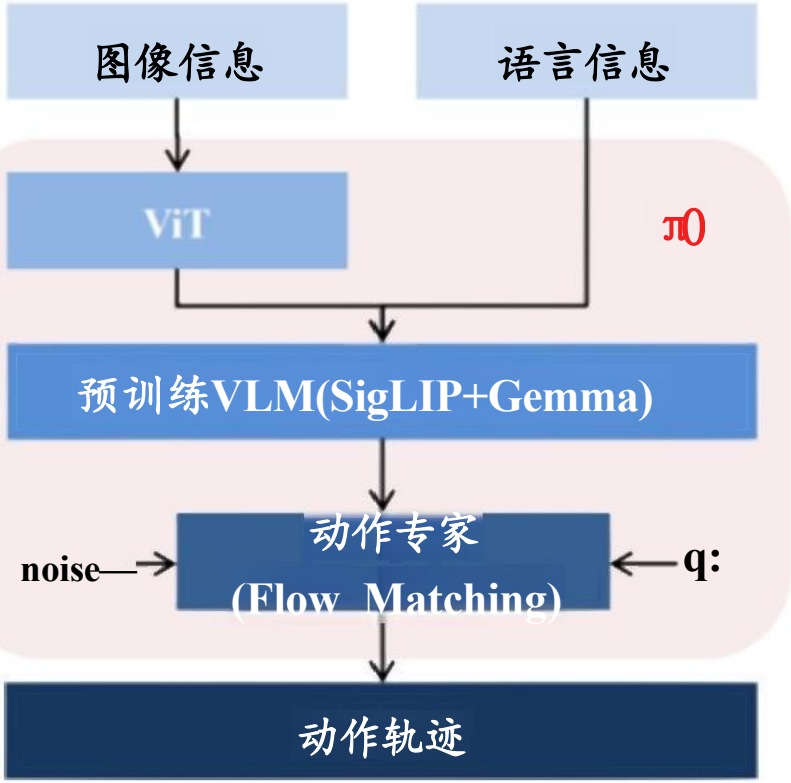


数据来源：《RT-2:Vision-Language-Action Models Transfer Web Knowledge to Robotic Control》,东吴证券研究所

2.2.5 π0/π0-Fast/π0.5：引入动作专家，输出50Hz 动作轨迹

- π0：采用VLM+ 动作专家，输出50Hz动作轨迹。 π0由预训练的VLM （视觉模型SigLIP+LLM 模型Gemma） 和使用Flowmatch 模型的动作专家组成。图像信息经ViT后和语言信息一同输入给VLM，经其处理后输入给动作专家，后者结合当前状态q， 输出50Hz连续动作轨迹。
- π0-Fast：采用Fast算法+Transformer 动作专家，训练时间缩短5倍。 Fast算法先将动作轨迹用DCT（离散余弦变换）压缩，再由BPE （字节对编码）后生成离散动作Token， 进而可将运动数据放入动作专家模型中训练，实际应用中Transformer输出的动作Token 经Fast解码后转为动作轨迹。
- π0.5：采用内置策略规划器的VLA。 类似π0-Fast，将VLM 训练为VLA， 同时内嵌任务分级模块。

图： π0模型架构



图： π0-Fast采用Transformer+Fast

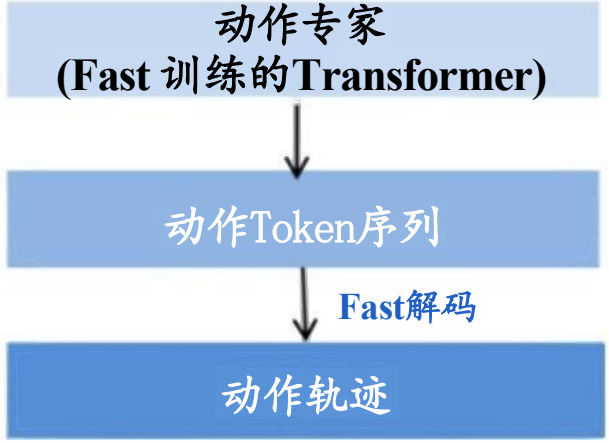
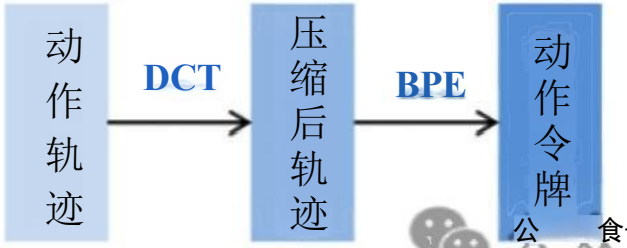


图 ： Fast 算法详情



图： π0.5采用VLA

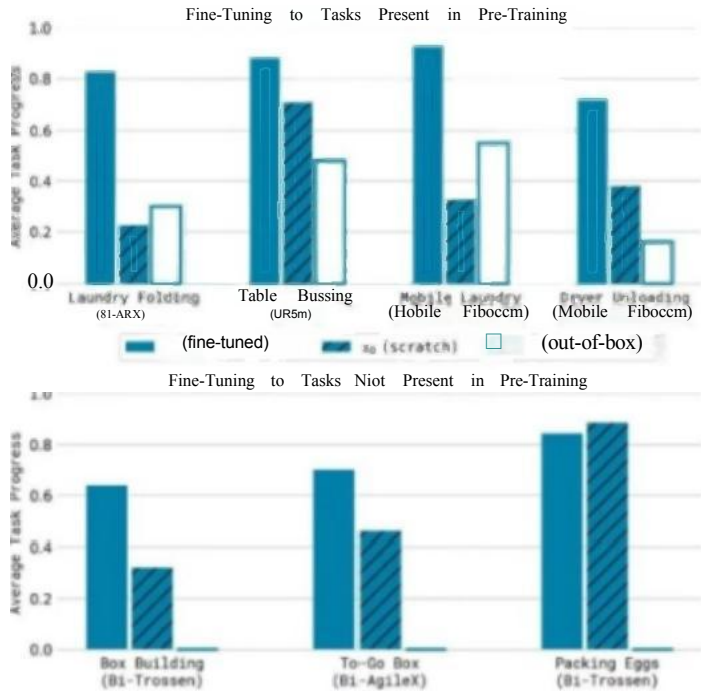
数据来源：《π0:AVision-Language-Action Flow Model for General Robot Control》,东吴证券研究所

2.2.5π0/π0-Fast/π0.5: 引入动作专家，输出50Hz动作轨迹

- 具体流程：工作流程整体与RT2 类似，不过π0引入了动作专家。当听到指令“帮我从桌上拿瓶水”时，动作专家会分析来自VLM 处理后的信息，结合当前状态，输出从A 状态到B 状态的一小段连续的运动轨迹，如手臂先旋转到几度、再旋转到几度。
- 优势：π0采取基于Difussion原理的FlowMatching架构，更加适配Action Chunking，动作输出的稳定性和成功率更高。在较慢推理速度下，输出50Hz高频运动轨迹，对于复杂环境的处理能力显著增强。
- **劣势：**但尚未达到100Hz 的输出频以满足复杂场景工作需要。

图：π0能够完成多种简单工作

图：对于复杂阶段任务，π0也有较好表现

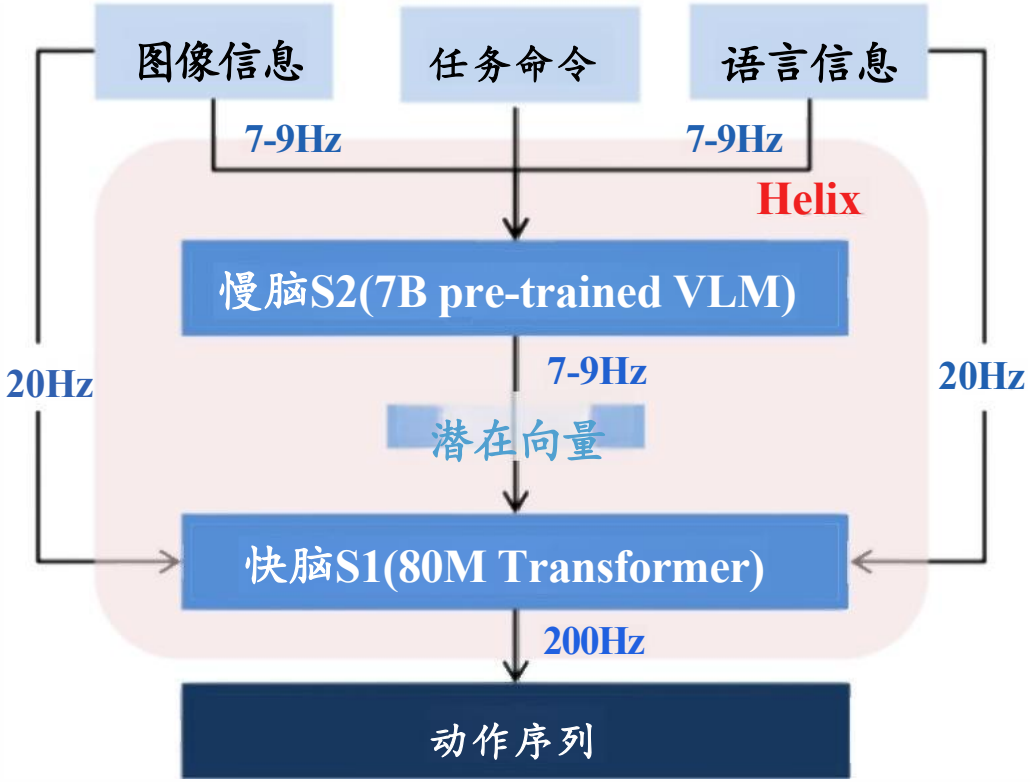


数据来源：《π0:A Vision-Language-Action Flow Model for General Robot Control》,东吴证券研究所

2.2.6 Helix：快慢脑结构，输出200Hz动作序列

- **架构&输出：采用端到端的快慢脑架构，输出200Hz 动作序列。**Helix采用一个7B 参数量的预训练VLM 作为慢脑，以及一个80M 参数量的Transformer模型作为快脑。两个模型解耦，在实际应用中以不同频率同时处理图像及语言讯息，慢脑负责思考高层目标，并以潜在向量指挥快脑，快脑负责实时执行和调整动作，并输出200Hz动作序列。同时由于潜在向量的存在，快慢脑可进行梯度回传，从而两者构成一个整体的端到端模型。
- **创新点：实现零样本多机器人协同以及拾取能力涌现。**实验中，两台Figure 02使用Helix首次实现了多机器人间的协作任务。同时，Figure发现，Helix涌现了拾取任意物品的能力。

图：Helix 模型架构



图：Helix 训练高效&使用统一的模型权重系统

训练高效：仅使用500小时的高质量监督数据

统一模型：无需根据任务微调动作输出层

图：两台Figure 02协作完成任务

数据来源：Figure官网，东吴证券研究所

2.2.6 Helix：快慢脑优势明显，未来或通过纳入统一模型&结合Diffusion模型优化

- 优势：双系统架构符合人类思考方式，同时由于模型参数增多会拉慢推理速度，因此若想在兼具较强推理和运动输出能力，思考执行分层的快慢脑架构必不可少。
- 改进方向1-优化双系统架构的融合性：智平方的FiS-VLA 为双系统架构的融合性提出了创新。现有的双系统模型存在两个系统相对独立，无法充分共享“慢思考”系统预训练知识的问题，协同效率低，“快执行”系统缺乏对“慢思考”系统语义推理结果的充分利用。FiS-VLA 提出创新架构，将VLM末端2层Transformer 模块重构为“快执行”的执行模块，嵌入“慢思考”内部，形成统一的高效推理与控制模型。这种思路既保留了双系统架构的动作输出能力，又拥有融合型模型的贯通理解能力。
- 改进方向2-优化动作输出模块能力：FiS-VLA 采用了双系统感知协同训练策略，利用扩散建模增强了“快执行”系统的动作生成能力，更好适配Action Chunking的优势，动作输出稳定性提升。

图：丹尼尔 • 卡尼曼的双重过程理论

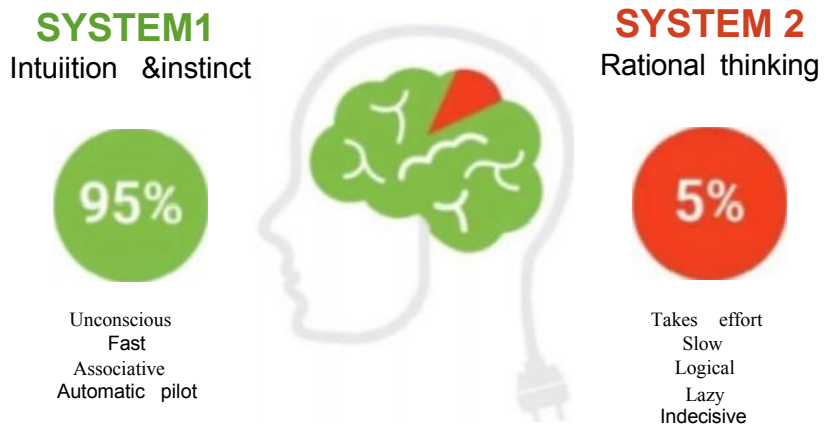
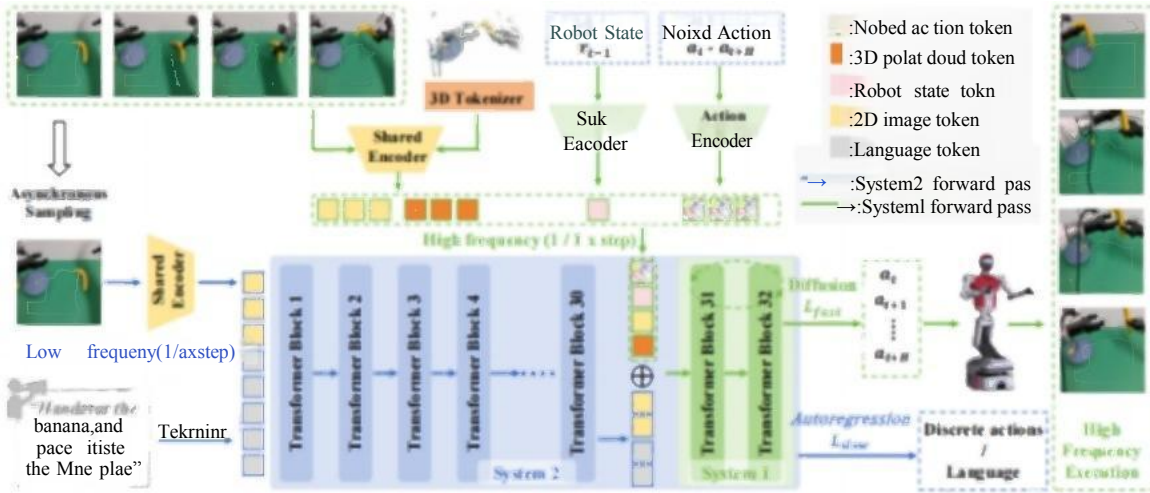


图 ： Fis-VLA 模型架构

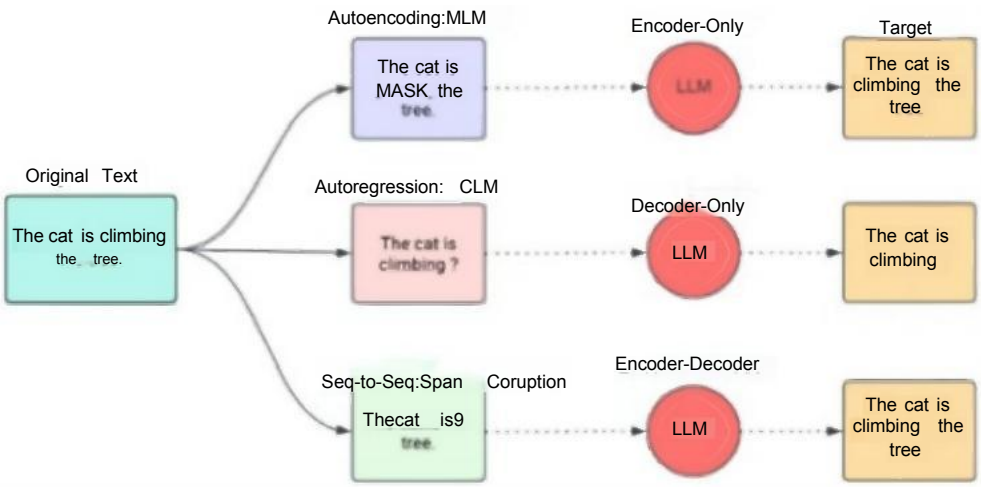


公众号 · 智能机器人科技

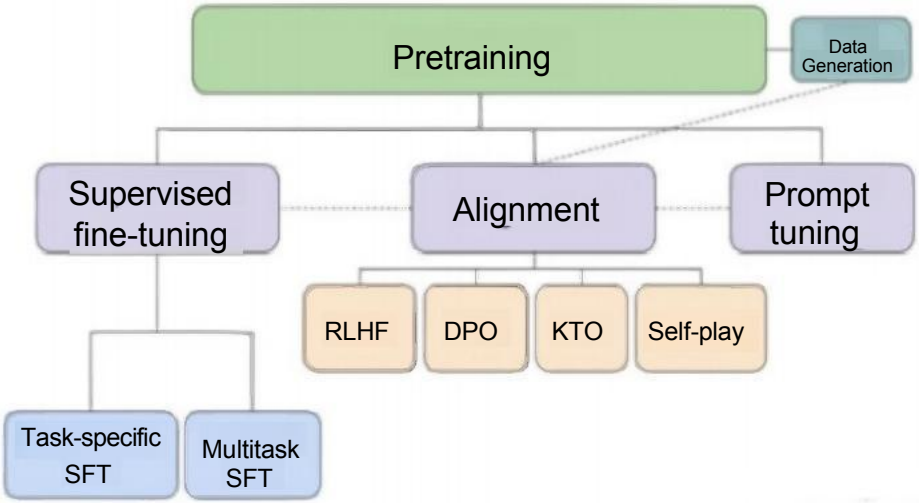
2.3训练：预训练赋予通用能力，后训练强化专项能力

- **大量低质量数据预训练，少量高质量数据后训练。**当前机器人大模型训练主要分为预训练和后训练。
- 1)预训练阶段，采取自主监督学习，即利用大量的不带标注的互联网数据和视频数据，让模型真正理解世界和场景，为后续的任务执行做准备。
- 2)后训练阶段，采取微调的思路，让机器人从理解世界到融入世界。首先进行监督学习微调，通过使用高质量的仿真或真实数据集(带有标注)进行训练，让机器人学会执行特定类型的任务。

图：预训练策略图



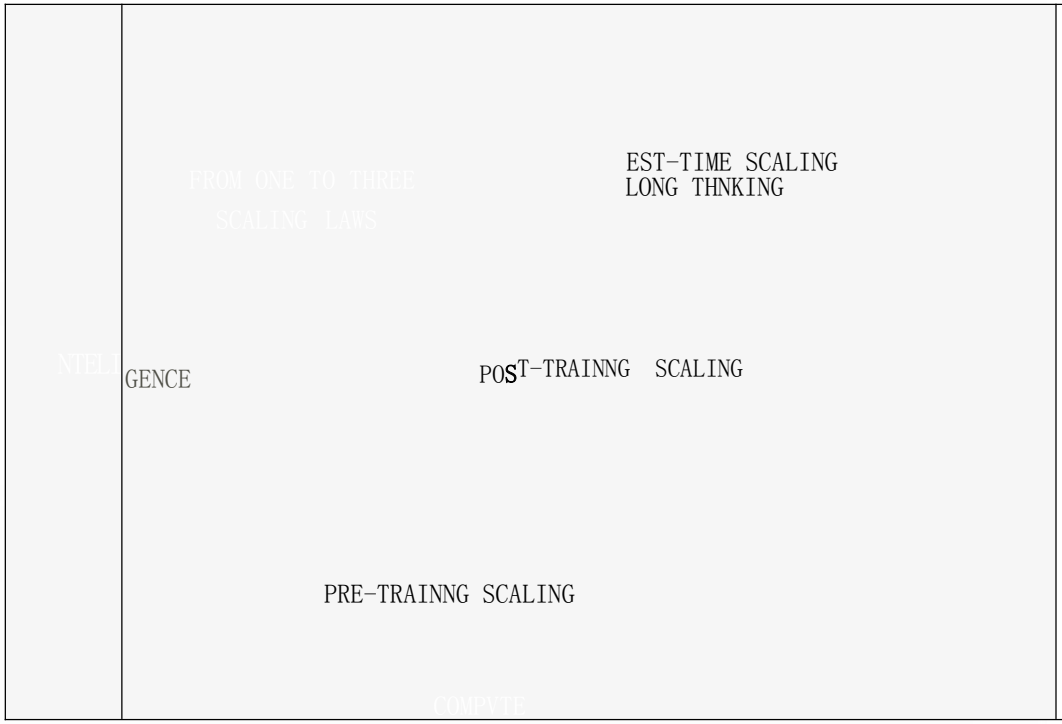
图：后训练流程图



2.4数据：行业基石，依三大法则影响模型效果

- 在具身智能大模型的发展中，数据端占据着举足轻重的地位。数据是大模型训练的基础，如同燃料对于引擎，是驱动模型性能和正确率提升的关键要素。
- 目前已确认在具身智能大模型领域，模型的性能同样遵循Scaling Law。 根据Scaling Law，当模型的参数或计算量按比例扩大时，模型性能也随之成比例提升。但只有当参数规模突破了某个阈值，大模型才会“涌现”出上下文学习、复杂推理等能力。而随着参数规模的增加，需要更多数据来训练模型，即模型参数与训练数据量之间也存在类似的比例关系。因此，在一个优秀的模型架构基础上，大量高质量的数据是迈向更高智能性的关键。
- 然而，随着AI技术的发展，已经出现了三条不同的法则来描述不同方式下计算资源的应用如何影响模型性能，分别是预训练Scaling 法则、后训练Scaling法则和推理阶段Scaling法则。

图：大模型的三大Scaling法则对模型性能的影响图



表：大模型的三大Scaling法则对模型性能的影响图

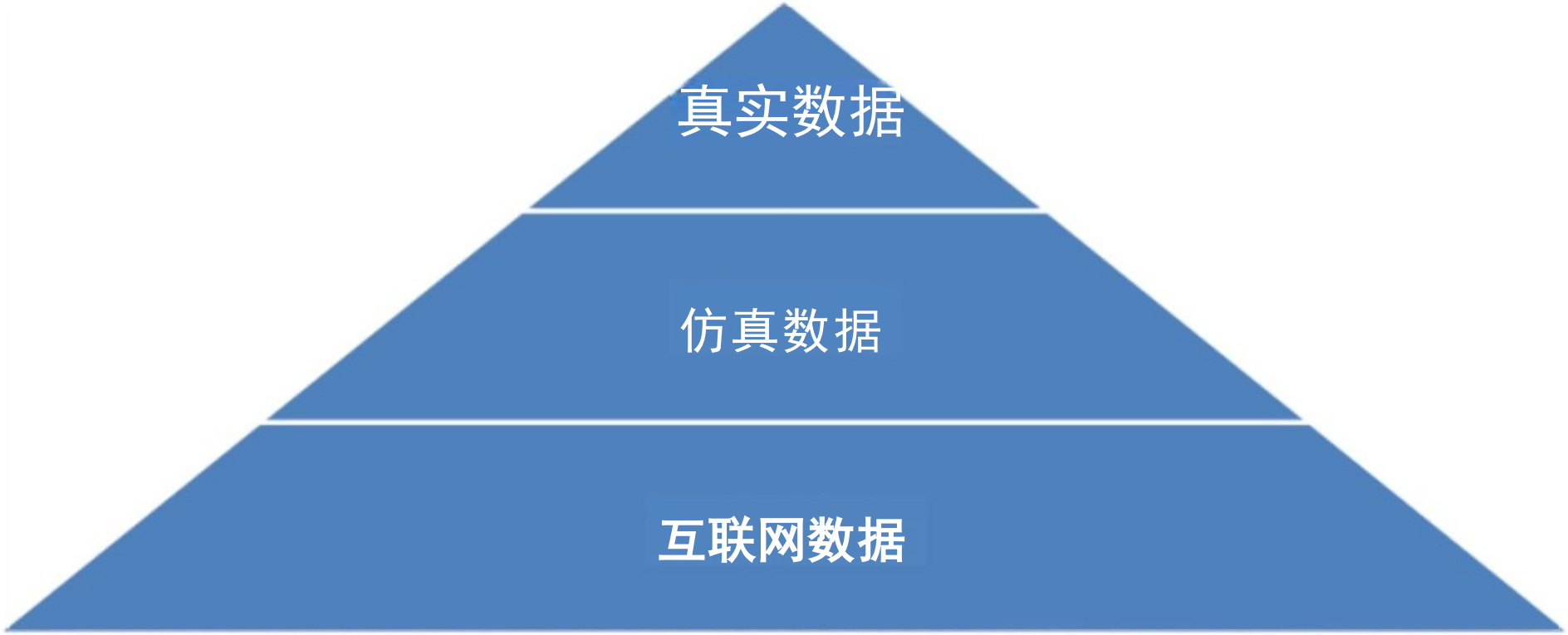
法则类型	核心定义	关键技术/方法
预训练 Scaling Law	基础法则，通过扩大训练数据规模、增加模型参数及计算资源，可提升模型性能。	-超大参数模型 -专家混合模型 -分布式训练技术
后训练 Scaling Law	基于预训练基础模型，通过调优技术进一步提升预训练模型的计算效率、准确率或领域适应性，模型针对特定应用的适应性与相关性，实现模型定制化。	-微调(领域数据适配) -蒸馏(教师-学生模型) 强化学习(RLHF、RLAIF) -Best-of-n采样、搜索方法 -合成数据增强
推理 Scaling Law	聚焦推理阶段，通过动态调整奖励机制、增加计算.投入提升复杂问题处理能力，不依赖参数或训练数据增加。	-多次采样与并行采样 自我原则批评调整(SPCT) -元奖励模型 复用后训练方法如Best-of-n采样）科技

数据来源：中国证券报《数据“燃料”催化具身智能：训练工厂涌现行业加速破局》，东吴证券研究所

2.4.1数据类型：三类数据结合，共同支撑模型训练

- **真实、仿真及互联网数据结合，共同支撑模型训练。**在数据的世界里，每一种数据类型都有其独特的优势和局限，形成了我们所知的数据金字塔。
- **数量之外，数据质量亦十分关键。**在这个金字塔的基座，是海量的来源于互联网的视频数据，它们的获取成本微乎其微，但相应的，它们的数据价值也相对较低。随着我们向上移动，仿真数据居于中间层，而顶端则是珍贵的真实数据，它们无疑拥有最高的数据价值。

图：机器人大模型数据金字塔图



数据来源：具身数据集开发者论坛-OpenLoong，东吴证券研究所

2.4.1互联网数据：数据量大且成本低，适用于预训练

- 视频数据是从互联网爬取的海量现实世界视频（如人类日常活动、物体交互等），主要用于机器人大模型的预训练阶段。通过监督学习和无监督学习的模式，让机器人大模型从大量的视频数据中寻找规律。
- **视频数据的优缺点：**

1）优点：数据量极大（容易取得，互联网可批量获取）、获取成本极低。

2）缺点：若采用监督学习模式，需要给数据打标签，成本极高；
场景泛化难，视频场景与机器人实际任务场景差异大。

图：机器人大模型基于观看未标记的在线视频的行动学习图

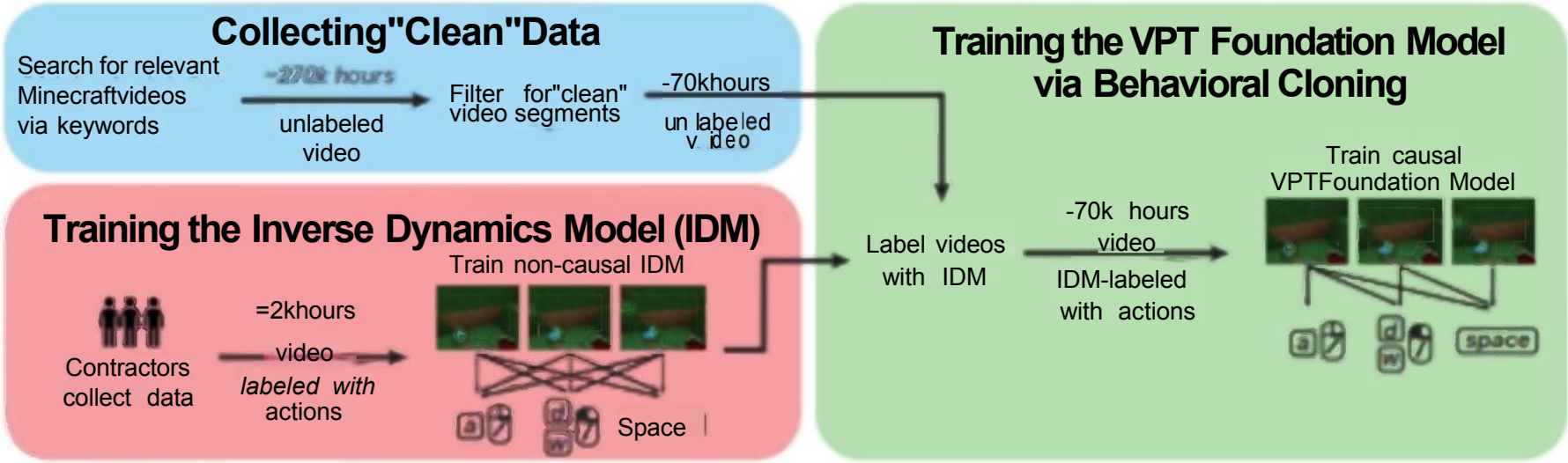


Figure 2:Video Pretraining(VPT)Method Overview.

公众号 · 智能机器人科技

数据来源：智源社区-OpenAI，东吴证券研究所

2.4.1 仿真数据：质量较高且成本较低，具备高性价比

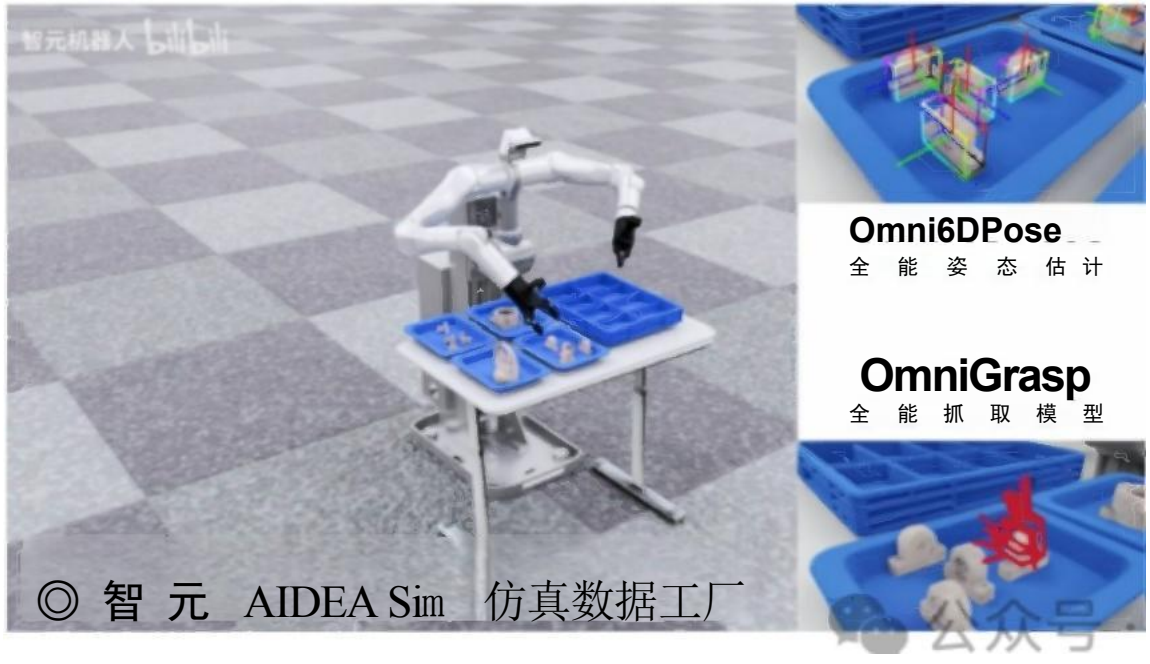
●仿真数据是在仿真平台中，通过渲染虚拟场景，生成、控制类似虚拟机器人完成任务进而收集的数据，支持强化学习。目前的仿真平台有如英伟达Isaac，其仿真能力已经非常强大，支持非常多种类的机器人构型以及不同场景的渲染。

● **仿真数据的优缺点：**

1）优点：①仿真场景无需搭建平台，无需购买机器人，成本低；
②多种渲染场景可以快速切换，数据的多样性快速提升；
③可以大批量复制复刻，同步24小时不停仿真生成，效率高。

2）缺点：①Sim2Real Gap，仿真到现实难以对齐，仿真平台中有很多现实世界的因素考虑不到，比如电机内部的摩擦、材料的柔性形变、发热、流体难以仿真；
②过拟合，若仿真场景单一，模型可能过度适配虚拟，难以应对现实中未模拟的变量。

图：智元机器人AIDEA仿真数据工厂场景图

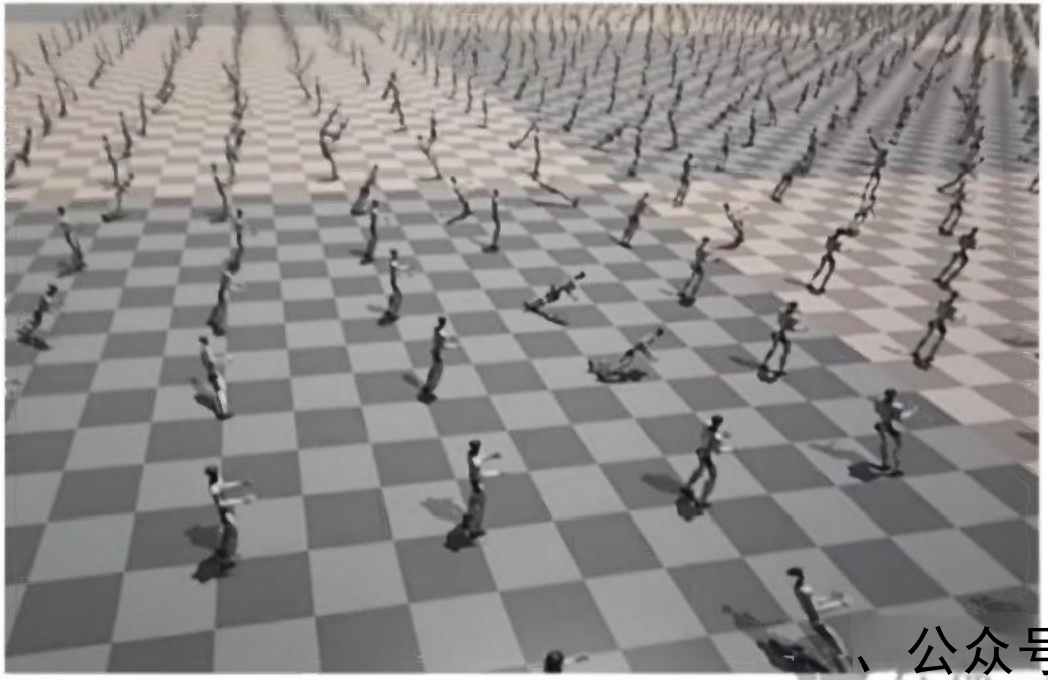


数据来源：智元机器人哔哩哔哩官方账号，东吴证券研究所

2.4.1仿真数据：赋予人形运动能力，多应用于预训练

- **人形运动能力训练方式收敛，各家多应用仿真数据。**相较于互联网数据，仿真数据所模拟的物理环境更为真实，同时能满足大量生成和低成本两项要求。同时由于所有数据都在虚拟世界中产生，机器人和物体的所有状态(如精确的三维位置、速度、关节角度)都是已知的，可以零成本地生成完美的标签，这对于预训练和强化学习训练至关重要。
- **在实际应用中，仿真数据多用于模型的冷启动。**以仿真数据启动，用真实数据校准是目前行业的主流选择：先利用仿真数据完成模型的基础能力训练，然后将其部署到真实环境中，在实践中持续收集真实数据，再利用这些真实数据反过来对模型进行微调和迭代，逐步缩小虚拟与现实的差距。**值得一提的是，利用仿真平台进行强化学习是目前机器人下肢运动能力训**练的主流范式。

图：宇树机器人在仿真环境中训练双足运动能力



数据来源：智元机器人官网，东吴证券研究所

2.4.1真实数据：质量最高，受限于采集效率数据量小

- 真实数据是通过遥操作控制实体机器人，直接捕捉其在真实世界中的动作数据。
- 另外还可以通过让人/人形机器人穿戴动捕设备运动，捕捉运动轨迹数据，捕捉关键节点的运动数据信息。
- **真实数据的优缺点：**

1）优点：数据质量高，直接反映真实物理规律和环境约束，是后训练阶段提升模型实用性的关键。

2）缺点：采集效率低，成本高昂。

- **真实数据必不可少，对后训练效果至关重要。** 在实际应用中，真实数据多用于后训练微调阶段。高质量的真实数据是微调的关键环节，主要用于提升机器人任务执行的成功率。

图：智元机器人AIDEA数采超级工厂场景图

智能机器人科技

数据来源：智元机器人哔哩哔哩官方账号，东吴证券研究所

2.4.2数据采集成本高企，数据结构差异显著

- 当前真机数据采集成本仍处高位，成为限制模型泛化能力提升的重要瓶颈。在传统采集方式下，双足机器人每小时仅能采集3-4条有效数据，单条采集成本约为19-20元。即便采用终端携带式采集设备，预计100台机器人每日数据产出也仅为8-10万条，且设备使用周期短、单机投入高。若人工参与采集，按每人每日采集500条计，年人工成本也接近30万元。
- 行业目前主要按照“仿真+真机”比例混合训练模型，不同机构侧重路径有所分化。由于真实数据获取难度大、场景覆盖受限，行业普遍按照1:9或1:10的比例配置数据结构。北京人形机器人创新中心采用7:3仿真占比，而银河通用、优选达等企业更多依赖仿真数据进行预训练，并辅以小规模真机数据进行后调。智元则实现VLA 模型100%真机数据驱动，路径差异凸显。

图：主流机器人大模型训练数据对比

公司	数据占比
北京人形机器人创新中心	仿真数据和真实数据使用占比约为7:3
(上海)国家地方共建人形机器人创新中心	真实数据和仿真数据使用占比约为3:1
智元机器人	智元所有的多模态大模型、VLA模型真机数据使用占比100%
银河通用	主要依靠合成仿真数据，预训练占比99%+， 后训练采集少量真机数据进行快速对齐
优必选	仿真数据和真实数据使用占比约为4:1

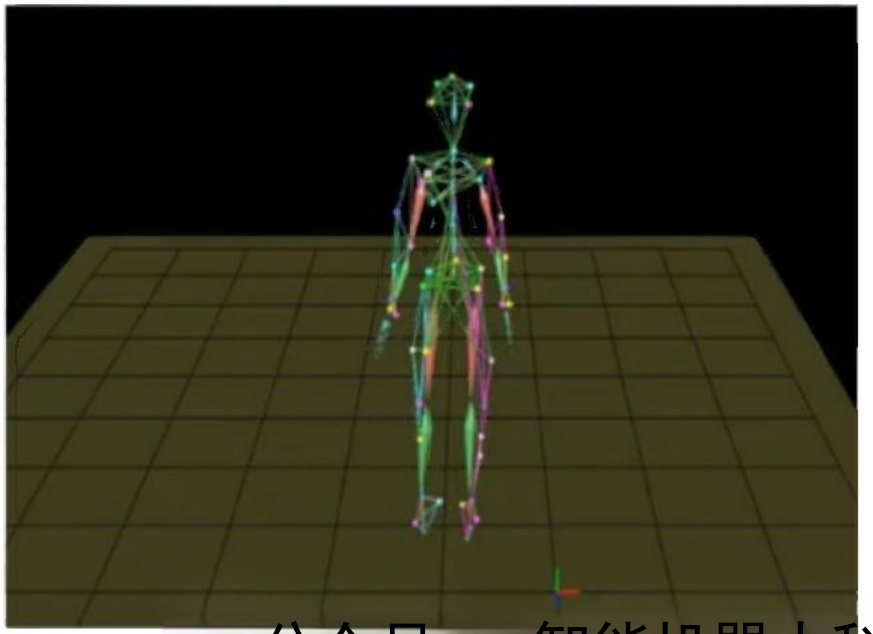
数据来源：各公司官网，东吴证券研究所

2.5高精度动捕系统是真机数据采集的关键基础设施

- 当前机器人真机数据的采集主要依赖动捕设备完成，分为光学捕捉与惯性捕捉两类。动捕设备通过将传感器安装于机器人或操控者身上，并结合遥操作方式采集运动轨迹，为后续模型训练提供高质量数据。
- 光学动捕依托高密度深度摄像头与反光标记点，能够精准重建人体或机器人的动态姿态。具体方式是搭建光学动捕棚，在空间顶部布设高帧率摄像头，通过识别人体或机器人关键部位的Mark点，实时获取位置信息，并在计算机中还原三维动作数据。
- 光学动捕系统在建设成本上差异较大，主要受精度需求与场地规模影响。 如凌云光训练场建设报价显示：8-10工位中型方案约300万元(含运动捕捉系统、服务器等)，单工位约30万元；若为100工位以上的大型配置，单位成本可降至10-20万元。相比之下，海外系统如Vicon或OptiTrack的价格普遍为国产方案的3-4倍。

图：光学动捕设备

图：动捕生成的计算机动画



数据来源：调研鹅公众号，东吴证券研究所

2.5惯性动捕：灵活便捷的数据采集方案

- **惯性捕捉系统通过IMU 传感器记录动作轨迹，具有部署灵活、便携性强的特点。** IMU 包含加速度计、陀螺仪和磁力计，可捕捉各关键关节的加速度、角速度与方位角信息。这些数据反馈至计算机后，即可重构人物或机器人的三维动态姿态，实现远程实时建模。
- **惯性动捕在成本与配置上更为弹性，适合多场景数据采集。** 价格区间从数万元至十余万元不等，具体取决于传感器精度、通道数量及配套软件能力。典型产品如诺亦腾PN3 Pro，售价约为4.58万元，适用于中小规模动作采集需求。

图：惯性动捕设备

图：动捕数据驱动人形机器人学会“像人一样运动”

数据来源：诺亦腾商城，东吴证券研究所

2.5光学vs 惯性各有侧重，适配不同采集场景

- 光学动捕具备高精度、低误差、易处理等优势，是目前精度要求最高的采集方案。 其典型优势在于亚毫米级追踪精度、无累积误差与抗干扰能力强，适用于高精度建模场景。但其缺点也较为明显，如成本高、场地限制大，仅适用于固定棚内采集。
- 惯性动捕则以轻便、低价、可移动性强著称，适合灵活部署与户外任务采集。其无需搭建动捕棚，便于随身穿戴，部署门槛低，广泛应用于中低精度场景。但受限于惯性传感误差叠加与磁场干扰，整体精度与稳定性仍逊于光学系统。

表：光学捕捉与惯性捕捉的优劣势比较

维度	光学捕捉	惯性捕捉
优势	1.精度高：亚毫米级追踪精度(被动式)支持毫米级细节捕捉； 2.无累计误差，位置数据依靠摄像头捕捉不存在惯性系因加速度产生的漂移误差； 3.不受磁场干扰，只依赖光学； 4.数据容易处理。	.不受空间限制，可以在任何地点随时采集； 2.容易上手，相比于光学棚复杂的搭建和调试，惯性捕捉可以随穿随用； 3.成本低，比光学成本低很多。
劣势	1.造价贵，需要搭动捕棚； 2.应用场地受限，只能在棚中捕捉，不能到室外或灵活变化的场地。	1.存在位置漂移(位置信息通过对加速度双重积分得到，加速度测量的微小误差会累计放大，出现漂移)，误差比光学大 ； 2.受磁场干扰，懊捕中的磁力计会受磁场干扰出现误差； 3.穿戴的束缚感比光学更高； 4.算法要求高，需要融合加速度、陀螺仪、磁力计等多个传感器数据，再融合骨骼求解器才能反算出位置信息。
代表企业	凌云光、青瞳视觉、OptiTrack(国外)、Vicon(国外)	诺亦腾、Xses(国外能 机器人科技

数据来源：研报瞭望塔公众号，东吴证券研究所

2.5训练场建设有望加速带动动捕设备加速放量

- **随着人形机器人迈入智能演化周期，对高质量真机数据的需求将显著上升。**相比图像、语音等成熟模态，动作数据获取高度依赖真实物理采集，具身大模型的训练效果高度取决于高质量、多样性、结构化的动作样本。
- **训练场建设打开动捕设备放量成长空间，可穿戴设备有望提供新思路。** 全球范围内训练场正加速建设，为光学和惯性动捕打开成长空间。同时，轻量化穿戴服动捕设备作为一种新技术逐渐受到重视。穿戴服设备能够实现工人无感穿戴，不影响工作，从而降低采集成本并扩大采集数量。当前，该技术在数据保存和质量把控存在难点，技术尚未形成闭环，但潜在空间较大，未来有望实现突破。

图：北京人形机器人数据训练中心采用FZMotion光学动捕技术

数据来源：机器人产业应用公众号，东吴证券研究所

1. 人形机器人为何需要高智能的大模型?

2. 从架构端和数据端看，目前机器人大模型的进展如何?

3. 未来大模型的发展方向是什么?

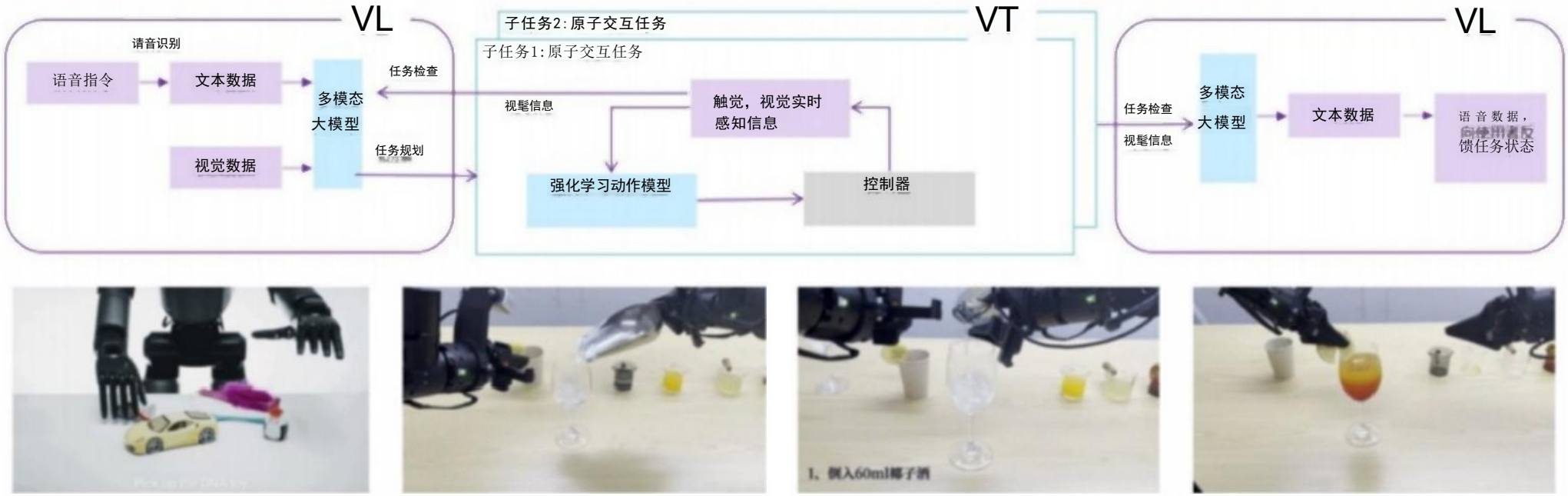
4. 相关标的

5. 投资建议与风险提示

3.1模态：融入更多模态，构筑世界模型

- **当前主流模型仅涵盖三个模态，未来扩展空间大。** 当前主流的机器人大模型多是VLA，即只包括视觉、语言和动作三个模态，未来若想加强人形与世界的交互，构筑更真实的世界模型，或将需要融入如触觉、温度等更多模态。
- **触觉或为下一个模态，VTLA 已有相关研究储备。** 触觉的引入可助力VLA 模型进一步泛化。通过引入触觉这一关键信息，VLA 模型可进一步延申为VTLA 模型。目前包括戴盟、帕西尼在内的公司已有相关VTLA 技术储备，预计未来触觉将成为下一个融入模型的模态。

图：VTLA模型的构架原理

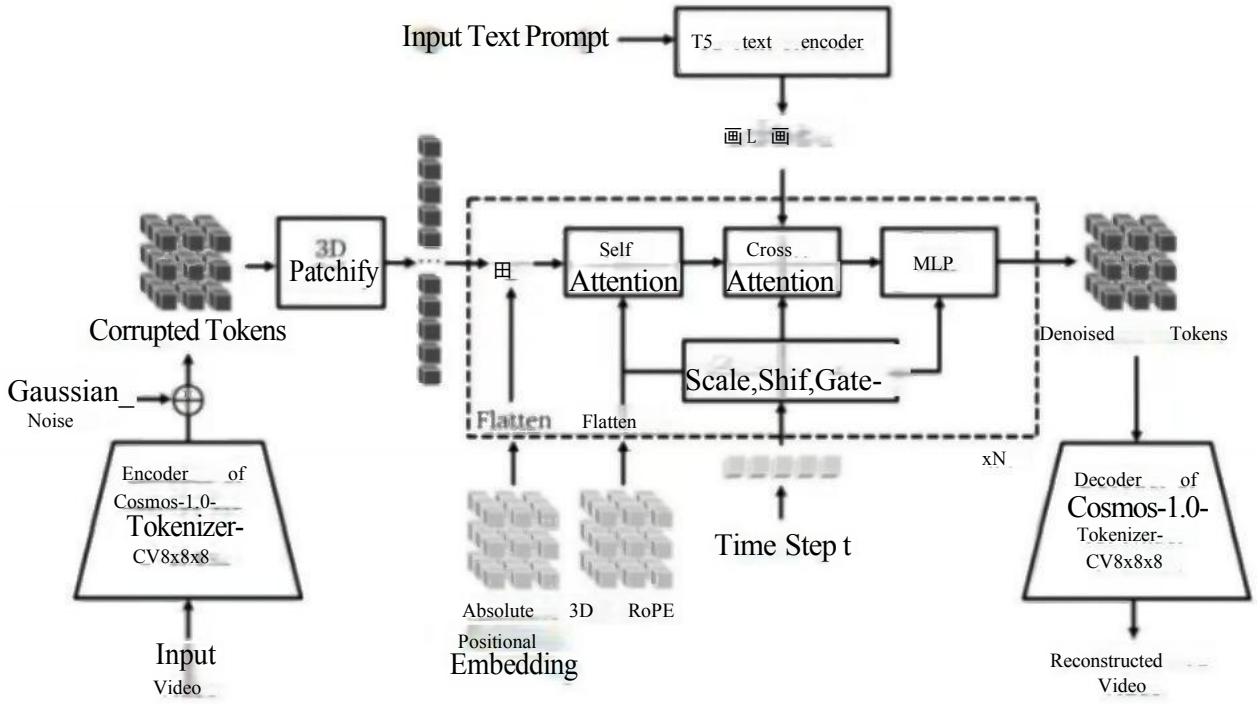


数据来源：他山科技，东吴证券研究所

3.2架构演进：引入“世界模型”作为核心推理机制

◆未来机器人通用大模型的架构演进方向之一，是将“世界模型”引入决策推理流程，作为具身智能的核心支撑模块。当前的大模型大多基于感知和语言指令直接生成动作，但缺乏对环境物理规律的建模能力，导致其泛化性与高阶推理能力受限。世界模型(World Model)本质上是一类可以模拟环境动态的神经网络，以 Cosmos 为代表的架构能够通过学习状态转移规律，基于当前状态与输入预测未来状态，实现“感知—建模—预测—决策”的闭环认知。这类机制的引入有望赋予机器人“想象力”，让其不仅能看到当前、听懂指令，更能推演未来，从而在面对复杂任务、多变环境时具备更强的适应能力与泛化能力。

图：Cosmos-1.0-Diffusion WFM的整体架构



数据来源：《Cosmos World Foundation Model Platform for Physical AI》,东吴证券研究所

3.3世界模型：英伟达发布Cosmos世界模型平台

◆ 英伟达发布世界模型平台，提供大量仿真数据。25年1月，英伟达发布Cosmos 世界模型平台，上面有一系列开源、开放权重的视频世界模型，参数量从4B 到14B 不等。这些模型的作用非常明确，就是为机器人、自动驾驶汽车等在物理世界中运行的AI 系统生成大量照片级真实、基于物理的合成数据，以解决该领域数据严重不足的问题。Cosmos 一共包括四大功能模块：扩散模型、自回归模型、视频分词器，以及视频处理与编辑流程。

图：Cosmos 提供的模型种类

Type	Models		Tokenizer	Enhancer
Diffusion	Cosmos-1.0-Diffusion-7B-Text2World	Cosmos-1.0-Diffusion-7B-Video2World	Cosmos-1.0-Tokenizer-CV8x8x8	Cosmos-1.0-PromptUpsampler-12B-Text2World
	Cosmos-1.0-Diffusion-14B-Text2World	Cosmos-1.0-Diffusion-14B-Video2World		
Autoregressive	Cosmos-1.0-Autoregressive-4B	Cosmos-1.0-Autoregressive-5B-Video2World	Cosmos-1.0-Tokenizer-DV8x16x16	Cosmos-1.0-Diffusion-7B-Decoder-DV8x16x16ToCV8x8x8
	Cosmos-1.0-Autoregressive-12B	Cosmos-1.0-Autoregressive-13B-Video2World		

数据来源：《Cosmos World Foundation Model Platform for Physical AI》,东吴证券研究所

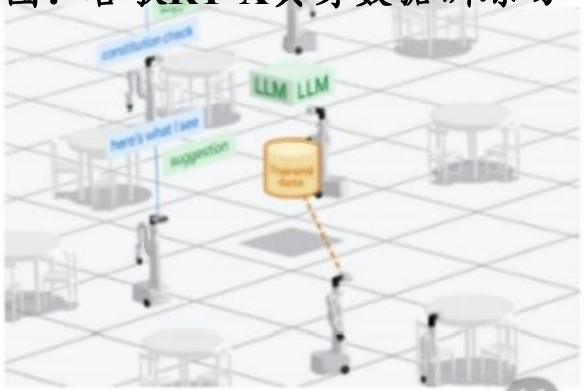
3.4数据构成：仿真数据与真实数据融合共用

◆**未来机器人大模型的训练数据将呈现“仿真+真实”共存的融合态势，这是提升模型泛化性与智能性的必然选**择。纯真实数据训练虽然更贴近实际，但采集效率低、成本高，同时由于大多为“成功范式”，模型难以从失败中学习，缺乏负样本经验。而单靠仿真数据又存在明显的Sim2Real Gap，仿真环境难以完全还原现实世界中的感知噪声、物理扰动与交互复杂性。因此，真实数据用于纠偏与对齐，仿真数据用于规模扩展和多样性覆盖，二者融合训练可有效兼顾效率与表现，是行业公认的发展方向。

◆为了支撑大模**型对海量多样化数据的需求，构建标准化、可扩展的机器人数据训练场，已成为业内普遍共识与行动方向。** 根据Scaling Law的经验推演，1亿条高质量行为轨迹数据是支撑具身智能大模型能力跃迁的关键门槛。当前，包括优必选、机器人创新中心、Tesla、华为在内的多个企业和研究机构，正加速搭建“仿真-真机融合”的数据训练场，通过并行机器人、远程操控、仿真回放等机制，高效采集覆盖不同场景、任务和交互模式的大规模数据。这些训练场不仅是数据源，更是模型评估、数据标准化和迭代反馈的基础设施，将成为未来具身智能训练体系的关键底座。

图：特斯拉具身智能遥操作训练场

图：谷歌RT-X具身数据训练场



图：斯坦福ALOHA具身作业训练场

数据来源：人形机器人世界公众号，东吴证券研究所

1. 人形机器人为何需要高智能的大模型?

2. 从架构端和数据端看，目前机器人大模型的进展如何?

3. 未来大模型的发展方向是什么?

4. 相关标的

5. 投资建议与风险提示

4.1银河通用(一级公司):全仿真数据训练路线，发布GraspVLA

◆北京银河通用机器人有限公司是市场领先的具身多模态大模型通用机器人企业。成立于2023年，银河通用致力于为全球用户提供通用机器人产品，并已率先在商业、工业、医疗等场景中广为应用。

◆银河通用成立时，公司的核心战略方向即为研发和应用“具身多模态大模型”，要实现真正意义上的通用机器人，仅仅拥有强大的硬件本体是远远不够的，更为关键的是要为机器人赋予一个能够理解复杂环境、进行自主决策并与人类自然交互的“智慧大脑”。公司研发的具身大模型旨在融合视觉、语言、动作等多种模态信息，使机器人能够像人类一样感知世界、理解指令并执行复杂的任务。

◆ 此 外，公司坚定不移地走全仿真数据路线，在应用实际场景时采集少量线下数据做算法微调。

1)预训练阶段：公司构建自研仿真数据生成管线，其训练了一个大型视觉-语言-抓取模型DexVLG，能够根据语言指令，通过单视角RGBD输入，预测灵巧手抓取位姿。从而公司能够以极低的边际成本批量生成高度多样化的合成数据，用于预训练。这部分数据约占整个训练数据的99%甚至更高，支撑起模型的泛化能力。

2)后训练阶段：公司针对特定任务需求，采集少量真机数据进行快速对齐。

图：银河通用全球首个完全依靠合成数据预训练的VLA大模型GraspVLA

数据来源：银河通用官网，东吴证券研究所

4.1银河通用(一级公司):模型数据体量大，泛化性强

◆公司首款机器人Galbot(G1) 采用独特的可折叠单腿+轮式底盘设计，能满足精细操作和泛化抓取能力的需求，且可解决Mobile,Pick and Place这类“简单”操作的泛化性问题。

◆ 公司GraspVLA 模型具备全球最大数据体量。GraspVLA 的预训练完全基于合成大数据，训练数据达到了有史以来最大的数据体量——十亿帧 r 视觉-语言-动作」对，掌握泛化闭环抓取能力、达成基础模型；预训练后，模型可直接Sim2Real 在未见过的、千变万化的真实场景和物体上零样本测试，全球首次全面展现了七大卓越的泛化能力，满足大多数产品的需求；而针对特别需求，后训练仅需小样本学习即可迁移基础能力到特定场景，维持高泛化性的同时形成符合产品需求的专业技能。该大模型一举打破了世界范围内具身通用机器人当前发展的数据瓶颈和泛化瓶颈，具有重要意义。

◆ 发布首款零售商业化大模型。另外，公司发布全球首款面向零售商业化的端到端模型GroceryVLA, 该模型解决了复杂零售环境的两大关键挑战，实现智能抓取全覆盖和场景适应高效率。

◆ 风 险 提**示：模型泛化能力有待观察，零售场景商业化节奏存在不确定性，市场竞争加剧。**

图：银河通用首款机器人Galbot(G1)

图 ：G1 在执行任务

科技

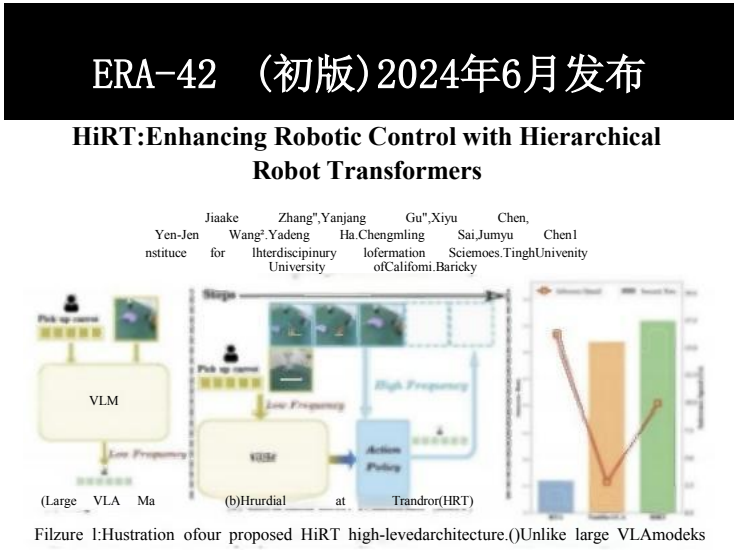
数据来源：银河通用官网，东吴证券研究所

4.2星动纪元(一级公司):双系统架构，融入世界模型训练

◆北京星动纪元科技有限公司孵化于清华大学交叉信息研究院，是国内第一个能够做到One policy for multipletasks,实现端到端原生机器人大模型落地真机的公司。作为原生通用具身智能体定义者，公司以赋予机器智能体以最通用、最原生的方式和物理世界交互的能力为使命。公司认为端到端架构是实现通用机器人的关键，只有通过端到端，机器人才能从感知到行为执行的整个链路打通，从而实现通用性，打造终极具身智能体。

◆ **公司模型先于Figure发布，最早提出双系统架构。** 星动纪元自研端到端机器人大模型ERA-42 应用了端到端具身大模型HiRT模型，由清华姚班团队在2024年6月发表，与Figure最新发布的具身大模型Helix在模型架构上高度相似。此外，公司技术团队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：星动纪元自研端到端机器人大模型ERA-42



Filzure l:Hustration ofour proposed HiRT high-levedarchitecture.()Unlike large VLAmodeks

图：Figure具身大模型Helix

数据来源：星动纪元官网，东吴证券研究所

4.2星动纪元(一级公司):双系统架构，融入世界模型训练

◆ **双系统架构，统一规划下实现快速响应。**ERA-42 支持端到端训练，能够直接从输入数据到输出动作进行学习，无需复杂的中间表示。通过端到端的训练，ERA-42可以更好地利用数据中的信息，学习到更加有效的特征表示和行为策略，提高了模型的性能和效率。 ERA-42 采用高层次规划和低层次控制的双系统架构。通过层次化的Transformer模型，ERA-42能够更好地处理复杂控制任务，并且在不同的环境和任务中表现出良好的泛化能力；能够实现对机器人动作的实时控制，确保机器人在复杂环境中快速响应；能够在一个统一的框架下处理多种机器人控制任务，无需为每个任务单独训练模型。

◆ **融入世界模型训练。**公司采取了一条不同的训练道路，将世界模型融入，使ERA-42 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。同时， ERA-42 也是全球首个融合世界模型的端到端全模态具身机器人大模型。此外，ERA-42 后期训练还采用了强化学习技术，通过奖励机制引导模型学习最优的动作策略。

表：ERA-42引领具身大模型进入通用灵巧操作时代

重要意义	优势表现
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	具备以下三个要素：统一一个模型泛化多种任务和环境、端到端、Scaling up(规模化)
ERA-42引领具身大模型进入通用灵巧操作时代	相比夹爪，基于ERA-42的能力，星动XHAND1已经可以完成100多种精细化、智能化的复杂灵巧操作任务；能理解物理世界与预测未来；具备更强泛化能力和自适应性；初步体现“Scaling效应”
ERA-42和为AI打造的全新硬件平台协同进化，共建原生具身智能体	星动纪元打造了为AI定义的全新硬件平台。以人形机器人的核心执行末端灵巧手为例，星动自研推出的五指灵巧手星动XHAND1共有12个主动自由度

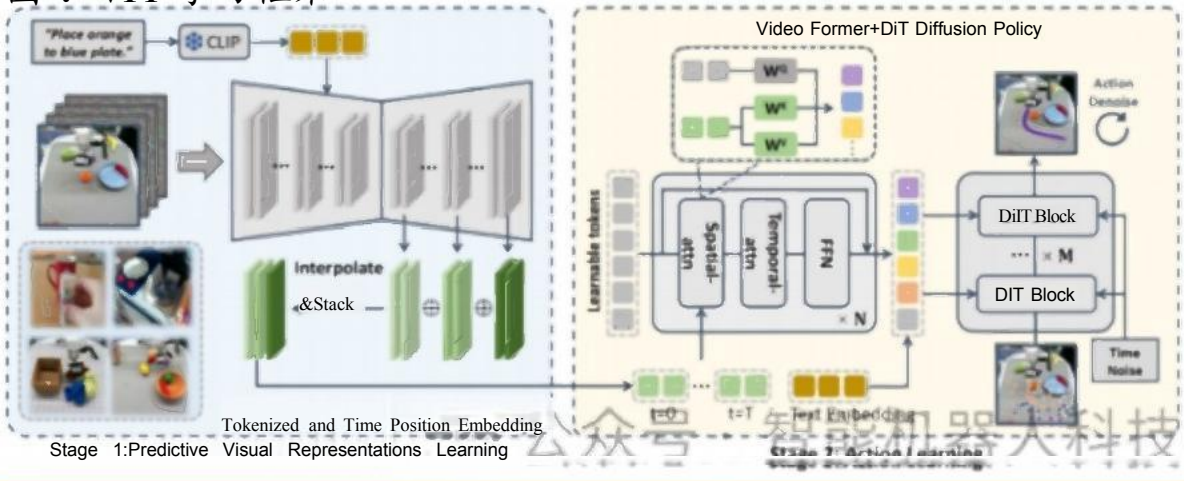
数据来源：星动纪元官网，东吴证券研究所

4.2星动纪元(一级公司):发布AIGC 生成式大模型

- **联合清华大学叉院，发布AIGC 生成式大模型。**清华大学叉院的ISRLab 和星动纪元开源ICML Spotlight高分作品AIGC 生成式机器人大模型VPP(Video Prediction Policy)。利用预训练视频生成大模型，让AIGC 从数字世界走进具身智能的物理世界，就好比“机器人界的Sora”。
- **互联网数据训练，克服Diffusion推理速度慢难题。**1) VPP 利用了大量互联网视频数据进行训练，直接学习人类动作，极大减轻了对于高质量机器人真机数据的依赖，且可在不同人形机器人本体之间自如切换，这有望大大加速人形机器人的商业化落地。2)VPP 将视频扩散模型的泛化能力转移到了通用机器人操作策略中，巧妙解决了 diffusion推理速度的问题，开创性地让机器人实时进行未来预测和动作执行，大大提升机器人策略泛化性，并且现已全部开源。3)VLM 更擅长高层级的推理，而AIGC 生成式模型更擅长细节处理。 VPP 基于AIGC 视频扩散模型而来，在底层的感知和控制有独特的优势。
- VPP分成两阶段的学习框架，最终实现基于文本指令的视频动作生成。第一阶段利用视频扩散模型学习预测性视觉表征；第二阶段通过Video Former和 DiT扩散策略进行动作学习。
- **风险提示：模型架构研发进展不及预期，世界模型推理能力尚处早期，行业竞争加剧。**

图 ：VLAvsAIGC生成式

图 ：VPP学习框架



数据来源：星动纪元官网，东吴证券研究所

4.3智元(一级公司)：ViLLA 架构，融合VLM 和动作专家

◆智元机器人成立于2023年2月，由包括”稚晖君”彭志辉在内的多位专家联合创办，公司使命为“以智能机器创造无限生产力”,是一家致力于以AI+机器人的融合创新，打造世界级领先的具身智能机器人产品及应用生态的创新企业。

◆ **ViLLA架构，融合VLM 和动作专家。**启元大模型（GO-1） 是智元机器人发布的全球首个通用具身基座大模型，采用Vision-Language-Latent-Action(ViLLA) 架构，融合多模态大模型（VLM） 和混合专家系统(MoE), 实现感知-规划-执行的闭环认知体系。其中VLM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视频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。

表：VLM、Latent Planner和Action Expert三者环环相扣

在推理时，VLM、Latent Planner和Action Expert三者协同工作
VLM采用InternVL-2B,接收多视角视觉图片、力觉信号、语言输入等多模态信息，进行通用的场景感知和指令理解
Latent Planner是MoE中的一组专家，基于VLM的中间层输出预测Latent Action Tokens作为CoP(Chain of Planning,规划链),进行通用的动作理解和规划
Action Expert是MoE中的另外一组专家，基于VLM的中间层输出以及Latent Action Tokens,生成最终的精细动作序列

图：GO-1中的混合专家系统

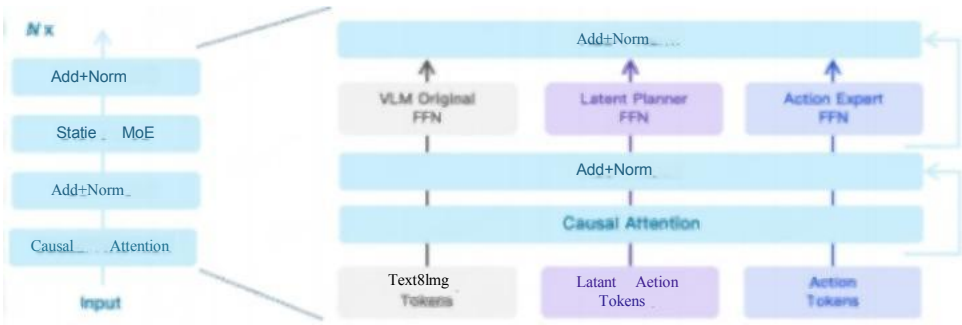


Illustration of MoE in GO-1

数据来源：智元官网，东吴证券研究所

4.3智元(一级公司):发布世界模型及评测基准

- **智元机器人重磅发布具身智能领域双重里程碑式突破:** 全球首个基于机器人动作序列驱动的具身世界模型EVAC (EnerVerse-AC), 以及具身世界模型评测基准EWMBench。这两大创新成果现已全面开源, 旨在构建“低成本模拟-标准化评测-高效迭代”的全新开发范式, 持续赋能全球具身智能研究, 加速技术落地与产业发展。
- EVAC是一个能够动态复现机器人与环境复杂交互的世界模型, 标志着从传统仿真到生成式模拟的跃迁。同时, 为了科学、系统地衡量具身世界模型的性能表现, 智元机器人推出了全球首个具身世界模型评测基准——EWMBench, 旨在填补行业空白, 构建统一、可信的评测标准。
- **EnerVerse和EVAC 形成技术优化闭环。**EnerVerse作为强大的世界模型基础架构, 为EVAC 提供可靠的基础框架与预训练能力, 而EVAC 生成的多样化高质量数据又能反哺EnerVerse 模型的持续优化, 二者形成“训练-验证”技术闭环, 不断推动模型性能突破。通过EWMBench 提供的精细化、多维度量化分析, 研发团队可以精准定位EVAC在处理如“多物体交互”“动态环境避障”等复杂场景的潜在不足, 从而进行更具针对性的优化。
- 风 险**提示:ViLLA 架构实际效果尚待规模化验证, 世界模型训练闭环构建难度高, 行业竞争加剧。**

图:全球首个机器人动作序列驱动的世界模型EVAC

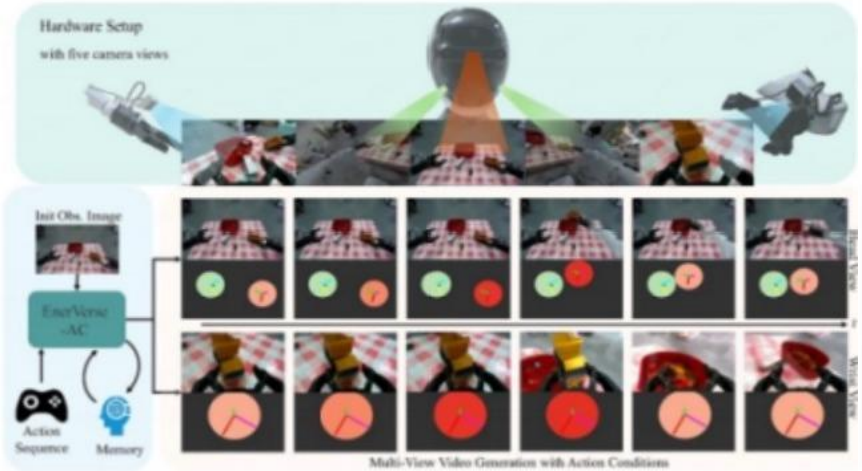
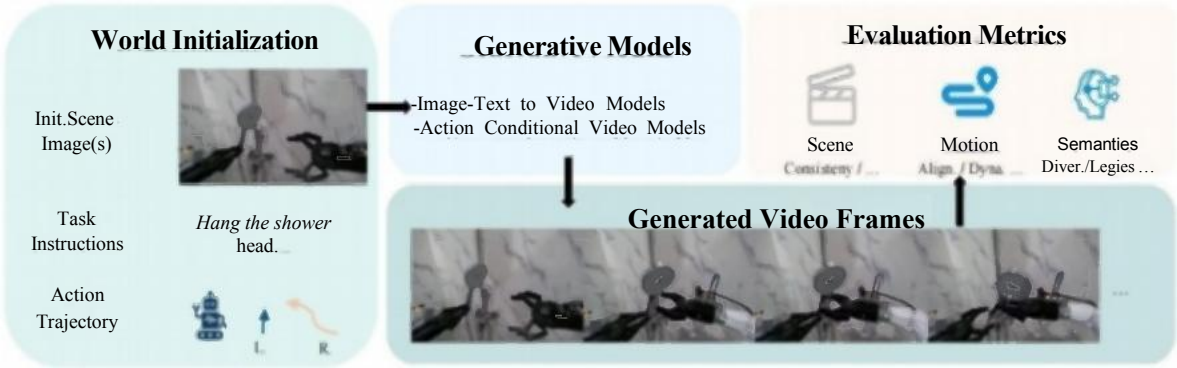


图:具身世界模型评测基准EWMBench



公众号 · 智能机器人科技

数据来源:智元官网, 东吴证券研究所

4.4青瞳视觉(一级公司):与原力深度合作，优化动捕系统

◆青瞳视觉是国内领先的光学动作捕捉企业，自主研发并生产具有国际领先水平的红外光学动作捕捉系统，并提供全栈式的解决方案及全流程的服务。可满足影视动画、游戏制作、数字人、虚拟制片、虚拟仿真、科研与工业自动化、运动分析、医疗康复等各领域及场景下的应用需求。

◆ 公司与原力深度合作。公司将原力Kinetiq AI解算器独家集成到青瞳的CMAvatar 动捕系统中，极大提升了动捕解算效率和使用者的动捕体验，减少人工修正成本，最新的CMAvatar3.0实现了对四足动物与复杂道具的动捕突破。

◆公司的高端光学动捕相机K26 拥有远超传统动捕相机分辨率与捕捉距离，搭配全新升级的CMAvatar3.0智能算法支持与强抗干扰能力的技术支撑，从数据采集到动作生成几乎无需人工干预，大幅提升动捕效率。

◆依托高精度动作捕捉系统与高保真光学手指追踪技术，青瞳视觉对机器人6DOF 运动学、触觉、视觉等多模态数据进行捕捉，并细化到手指这类精细部位，轻松应对各种复杂环境下的手指捕捉需求。

◆由青瞳视觉与索尼中国研究院联合推出的CMVolcap 三维重建系统，集成先进的毫米级精度三维重建算法，精度与效率并存；模块化相机阵列可根据实际需求灵活部署，实现全方位无死角三维重建；拥有全自研软硬件，可与青瞳视觉动捕系统深度联动。

◆**风险提示：动捕系统商业化节奏不及预期，核心软硬件性能迭代不及预期，数采中心建设不及预期。**

图：青瞳视觉光学动捕棚

图：青瞳视觉CMVolcap三维重建系统

数据来源：青瞳视觉官网，东吴证券研究所

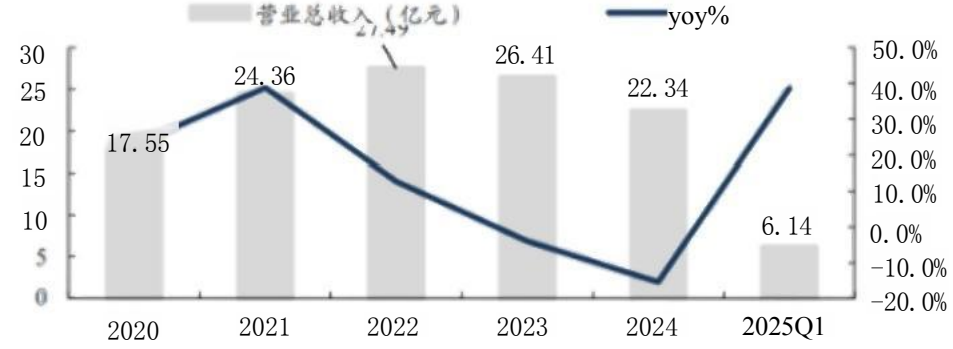
器人科技

4.5凌云光(688400.SH)：　动捕需求驱动公司收入加速上行

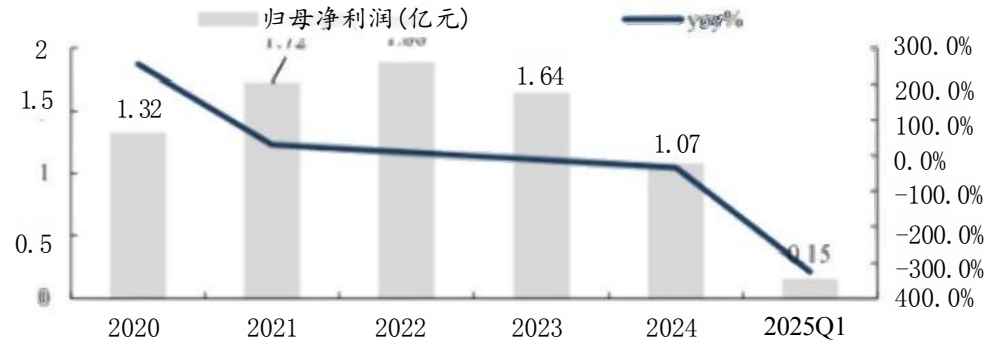
◆ 凌云光以光技术创新为基础，围绕机器视觉与光纤光学开展业务，致力于成为视觉人工智能与光电信息领域的全球领导者。

◆ 2025年一季度营收6.14亿元，同比增长38.57%，主要系“视觉+AI”产品组合的不断完善，工业智能制造赛道固定资产投资的温和复苏，以及国内具身智能产业蓬勃发展带动元客视界光学动捕产品FZMotion　收入大幅增长，同时公司收购的JAI年初顺利交割并表，相关业务快速整合协同也助力营收提升。2025年一季度净利润1498.5万元，同比扭亏为盈，主要系收入增长及JAI A/S并表。

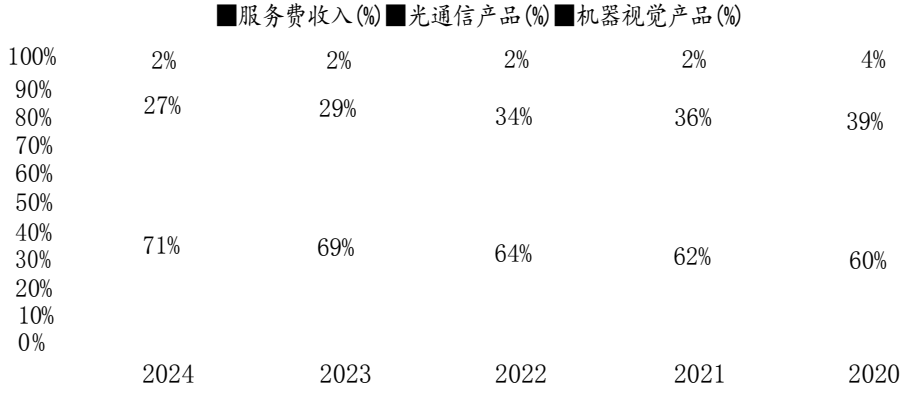
图：2020-2025Q1 年公司营业收入(亿元)



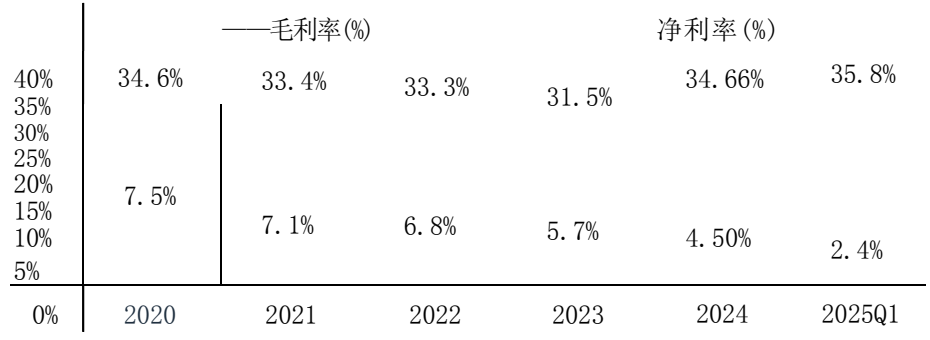
图：2020-2025Q1 年公司归母净利润(亿元)



图：2020-2024年分业务收入占比(%)



图：2020-2025Q1 年公司毛利率与归母净利率(%)



公众号 · 智能机器人科技

数据来源：凌云光官网，wind，　东吴证券研究所

4.5凌云光(688400.SH)：客户资源雄厚，自研通用视觉大模型

◆ 客户资源雄厚，相关产品市占率第一。公司通过高精度运动捕捉(亚毫米级)采集人体动作数据，映射至机器人构型，优化运动控制算法，解决人形机器人“拟人化运动”痛点，已服务宇数科技、优必选、小米等头部机器人厂商。公司光场重建与动捕系统国内市占率第一。

◆ 自研通用视觉大模型E.Brain算法平台。具备AIGC 生成、智能化标注、基于主动检测的模型迁移以及产品失效分析的能力，凭借在典型行业卓有成效的落地实践，成功跻身案例名单。 F.Brain算法平台以视觉成像为基础，以AI视觉大模型为赋能路径，以工厂产线的质量管理为抓手，创新性地挖掘质量数据新价值，将生产工艺数字化知识化，反向溯源制造环节漏洞，激发新型工业化的新动能。

◆在通用工业智能方面，凌云光推出了LuserLVM 工业领域通用视觉大模型。该模型通过分层设计，既满足了基础大模型的通用性需求，又针对不同行业和应用场景进行了优化。在缺陷生成、辅助标注和缺陷提示等方面，该模型展现出了卓越的性能，大幅提升工业质检的效率和精度，而无需依赖大量的算力，能够快速部署在3C、锂电等10多种工业视觉检测环境中。

◆风险提示：动捕系统商业化节奏不及预期，核心软硬件性能迭代不及预期，数采中心建设不及预期。

图：凌云光LuserLVM 工业领域通用视觉大模型

图：凌云光F.Brain 入选【AI赋能新型工业化创新应用优秀案例】

公众号 · 智能机器人科技

数据来源：凌云光官网，东吴证券研究所

4.6奥比中光(688322.SH): 布局五大技术路线，自研引擎芯片

◆奥比中光总部位于深圳，国内3D 视觉领域龙头企业，产品覆盖消费电子、工业检测及机器人领域。

◆ 公 司Gemini 系列双目3D 相机支持毫米级深度感知，可实时生成物体三维模型，用于机器人避障(如普渡送餐机器人)、物流分拣(如NarGo 订单拣选机器人)等场景。

◆奥比中光是全球少数全面布局五大技术路线的企业之一，能够根据不同应用场景灵活选择最适合的视觉技术方案。3D视觉感知技术是通过光学、机械、电子、计算等多学科交叉融合，实现对物体三维信息的精确获取和处理的技术。目前主流的3D视觉感知技术包括结构光、iToF（间接飞行时间）、 dToF（直接飞行时间）、双目立体视觉、激光雷达等技术路线。不同技术路线各有优缺点，适用于不同的应用场景。奥比中光构建了“全栈式技术研发能力+全领域技术路线布局”的3D视觉感知技术体系，形成了从底层到应用的完整技术链条。这种全栈式布局使其能够灵活适配不同场景需求。

◆ 自研引擎芯片，降低生产成本。公司技术体系的核心是自主研发的MX 系列深度引擎芯片，这是3D 视觉传感器的“大脑”，负责处理复杂的深度计算任务。通过芯片的自主设计，奥比中光不仅大幅降低了产品成本，还提高了系统性能和稳定性，为大规模产业化奠定了基础。

◆风险提示：视觉方案商业化节奏不及预期，自研芯片性能有待规模化验证，多技术路线协同落地复杂度高。

图：奥比中光Gemini 系列双目3D 相 机

图：奥比中光四代结构光深度引擎芯片与Astra 3D摄像头

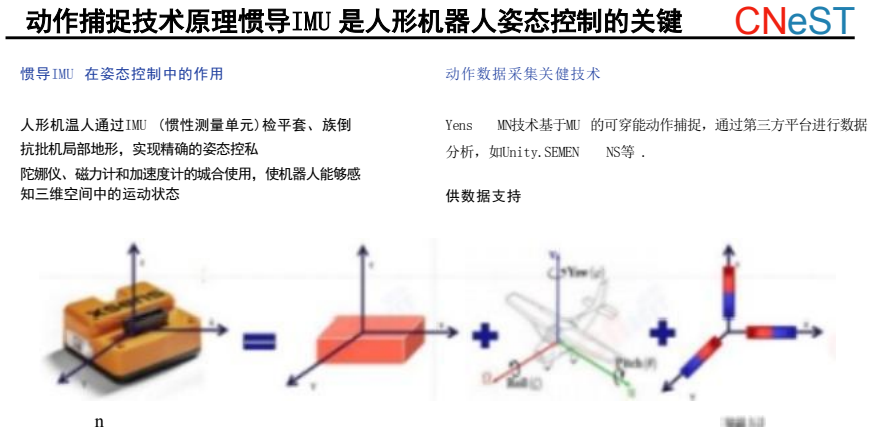
数据来源：奥比中光官网，东吴证券研究所

4.7 Xsens(一级公司):产品矩阵完善，全条件抵抗磁干扰

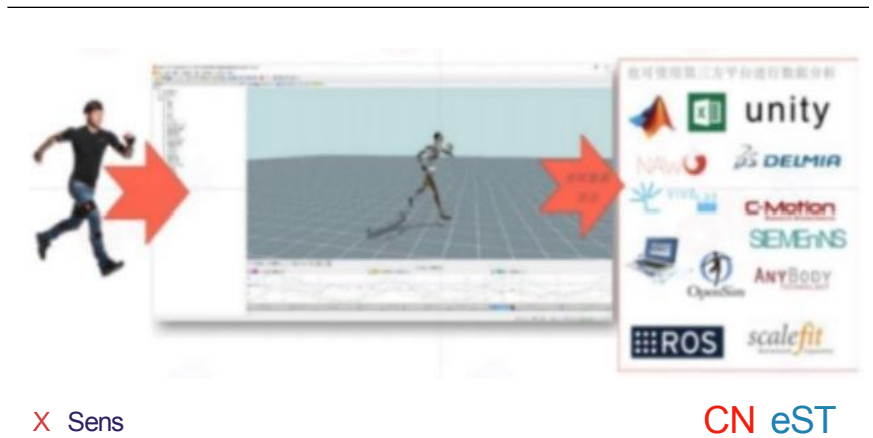
- Xsens成立于2000年，总部位于荷兰恩斯赫德，并在洛杉矶、中国香港和上海设有分支机构。公司隶属于Movella 集团，是全球3D 运动追踪与惯性动作捕捉领域的标杆企业，拥有近200项专利，产品线覆盖IMU 模组、穿戴式惯性动捕系统(MVN 系列)以及配套软件**IMU 传感器技术：**通过**穿戴式惯性测量单元**(如Xsens MVN Link) 采集关节运动数据，无需外部摄像头，适用于复杂环境(如工厂车间、户外)。
- Xsens基于微型惯性测量单元(IMU) 构建“人形机器人拟人化动作AI 训练系统”，可在无摄像头、无标记点的条件下，实时采集操作员的全身关节角、线加速度、动态平衡等高维生物力学特征。
- Xsens称行业唯一能做到全条件抵抗磁干扰，加上设备的便携性和易操作性，其系统可以实现随时随地动作捕捉。配合Motion Cloud,东部数据可在云端共享与批量处理，支持异地协同标注，降低50%后期清理时间。

◆风险提示：惯性动捕抗磁干扰能力待验证，惯性动捕位置漂移问题待进一步解决。

图：配套软件IMU 传感器技术是关键



图：穿戴式惯性测量单元



数据来源：Wind，东吴证券研究所

4.8天奇股份(002009.SZ): 智能装备企业，携手优必选切入机器人赛道

◆天奇股份创立于1984年，2004年于深交所上市，总部位于江苏无锡，是一家以“汽车全生命周期智能装备”为核心业务的国家级高新技术企业。公司长期为全球主流车企(吉利、比亚迪、宝马、大众等)提供总装、焊装、涂装、物流自动化整线交钥匙工程，累计交付汽车智能制造产线700余条，市占率国内第一。

◆公司构建基于汽车制造Know-how 的“场景语义库”和“工艺动作基元库”，为具身大模型提供百万条真机轨迹及千万级合成数据。

◆ 公司与优必选合作在极氪汽车工厂部署Walker S1机器人。Walker S1在完成料箱搬运、CTU 入库等任务上，动作执行效率较传统设备提升30%;与银河通用合资成立天奇银河机器人公司，利用汽车车间环境生成训练数据，优化机器人在装配、检测等环节的动作鲁棒性。与银河通用联合训练“Genie-Assembly”大模型，已在车身焊点检测 、SPS 零件抓取等任务上实现零样本迁移，精度≥99.2%，节拍提升15%。

◆ 风险提示：汽车客户验证周期较长，数据采集中心落地不及预期，实际任务迁移精度与效率仍需持续验证。

图：天奇股份实应用场景

现实应用场景	具体说明	特性
无锡市具身智能机器人工业数据采集与实训中心	2024年11月投运，占地6000㎡，按真实焊装、总装、电池包装配工位1:1复刻，可同时容纳50台人形机器人并行训练；中心具备“三高一低”特点。	高保真 高并发 高安全 低成本
已落地的“实景训练+数据采集”案例	极氪5G智慧工厂：2024-07起分两阶段导入Walker S1进行车身外观质检、车门卡接训练，累计采集2.3万条轨迹，模型迭代3版后误检率由3%降至0.7%;	生产线即训练场 科 技
	比亚迪长沙工业园：UQI优奇无人物流方案实现人形机器人与无人叉车、MES协同，完成料箱搬运8000余次，形成全球首个“人形机器人+无人物流车”混场作业数据集;	
	吉利晋中基地：2025-05起部署12台天奇自研双臂机器人，进行底盘螺栓拧紧，车身擦拭打蜡等精细作业单工位训练周期由30天压缩至5天。	

数据来源：天奇股份官网，东吴证券研究所

1. 人形机器人为何需要高智能的大模型?

2. 从架构端和数据端看，目前机器人大模型的进展如何?

3. 未来大模型的发展方向是什么?

4. 相关标的

5. 投资建议与风险提示

5. 投资建议与风险提示

投资建议：

模型端建议关注【银河通用(一级公司)】 【星动纪元(一级公司)】 【智元机器人(一级公司)】，数据采集领域建议关注【青瞳视觉(一级公司)】 【凌云光(688400.SH)】【 奥比中光(688322.SH)】， 数据训练场领域建议关注【天奇股份(002009. SZ)】。

风险提示：

(1) 大模型技术进展不及预期。 当前具身智能仍处于技术快速演化阶段，相关大模型在推理能力、泛化能力、动作输出频率等方面尚未完全成熟，若未来技术瓶颈迟迟难以突破，可能影响机器人智能化进程及产业化落地节奏。

(2) 高质量数据获取受限。具身大模型训练高度依赖高质量真机数据，若训练场建设进度不及预期，或采集成本与效率难以有效控制，可能导致模型性能难以持续提升，进而影响行业发展。

(3) 人形机器人需求不及预期。 人形机器人作为大模型能力的重要应用载体，其商业化路径尚处早期探索阶段，若终端需求增长缓慢或应用场景拓展不达预期，将间接影响具身大模型的产业化价值空间。

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
智公司 网址：http://www.dwzg.com.cn

东吴证券
SOOCHOW SECURITIES

东吴证券财富家园

公众号 · 智能机器人科技