人形机器人“大脑”：神机妙算，加速进化

2025 年 04 月 02 日

➢ **具身智能有望开启万亿级蓝海市场。**在当前时点复盘机器人与人工智能的发展历程，机器人已经进入具身智能时代，与此同时，人工智能也将走向“物理 AI”发展阶段。人形机器人是两者汇聚的交点，也是具身智能时代的临界点，有望成为新一代智能终端，并开启万亿级蓝海市场。

➢ **具身智能大模型为机器人“大脑”的核心。**具身智能需要本体、智能体、数据、学习和进化架构四大核心要素，通用机器人本体又可以分为“大脑”、“小脑”和“肢体”三部分，其中，人形机器人“大脑”的核心为人工智能大模型技术，通过多模态模型建模、强化学习、地图创建和数据训练，能够管理和协调机器人的各种功能。大模型目前较为擅长需求理解、任务分解等高层级控制任务，规划级以下的控制规划属于传统机器人控制规划的范畴，更适合传统机器人更成熟的高频控制方法。**多模态大模型为机器人高层级控制带来技术突破。**多模态大模型具有理解图像、场景文本、图表、文档以及多语言、多模态理解的强大能力，可以直接用于具身智能对环境的理解，并通过提示词使之输出结构化内容如控制代码、任务分解等指令语言、图片、视频等。

➢ **国内外科技巨头与研究团队入局，具身大模型成果涌现。**谷歌、特斯拉、微软、英伟达、李飞飞团队、特斯拉、字节跳动等国内外科技巨头和科研机构争相入局，具身大模型成果不断涌现：谷歌推出 RT-1、PaLM-E、RT-2、RT-X 等多个具身大模型；特斯拉坚持端到端算法路线，实现感知决策一体化并迁移至人形机器人；英伟达推出物理 AI 开发平台 Nvidia Cosmo 及一系列世界基础模型；国内大厂字节 GR-2 在动作预测和泛化能力上表现出色。

➢ **具身大模型目前在泛化性、实时性、数据采集等方面存在挑战。**当前的具身大模型通常存在泛化能力弱的问题，已经在特定场景达成较高成功率的模型在切换至不同场景时成功率大幅降低。实时性较差则体现在输出运动频率较低，使得机器人反射弧较长，低于人类和许多实际应用场景的需求。数据采集方面的挑战则体现在真实数据收集效率偏低、收集难度和成本偏高，合成数据的使用中则需要避免生成数据与真实数据差距过大或者样式单一。

➢ **云计算与边缘计算作为“大脑”的外延，保障机器人“大脑”高效运转。**云计算是为机器人等终端设备提供算力的核心方式，云计算能够为 AI、大模型与机器人的结合提供强大的计算能力和数据存储空间，以及能够随时随地获得所需资源和算法支持的灵活性、可拓展性；此外，边缘计算为云计算的数据传输成本、时延、安全性等方面的局限性提供了补充，为具身智能人形机器人落地保驾护航。

➢ **投资建议：**2025 年人形机器人行业进入小批量量产阶段，全球将有数千台人形机器人进入工厂场景训练，加速人形机器人“大脑”的发展。我们认为，目前人形机器人硬件端技术路线趋向收敛，软件端“大脑”智能水平的提升有望成为人形机器人自主性与泛化性提升的核心推动力。建议关注：1）“大脑”领域，布局大模型与机器人业务相结合的公司，如科大讯飞、中科创达、萤石网络、柏楚电子、华依科技、芯动联科、汉王科技等；2）AI+机器人领域，具备高壁垒的公司，如 3D 视觉领域奥比中光、大脑域控制芯片天准科技、新型传感器峰岹科技等；3）同步受益的机器人本体公司，如总成方案三花智控、拓普集团等。

➢ **风险提示：**机器人算法迭代进步速度不及预期；人形机器人落地场景实际需求不及预期；市场竞争加剧。

分析师 汪海洋
执业证书：S0100522100003
邮箱：wanghaiyang@mszq.com

分析师 吕伟
执业证书：S0100521110003
邮箱：lvwei_yj@mszq.com

分析师 李哲
执业证书：S0100521110006
邮箱：lizhe_yj@mszq.com

相关研究

1.人形机器人产业周报：海外更新催化不断，深圳近期将发布人形机器人专项政策-2025/02/25

2.人形机器人产业周报：宇树科技 G1 灵动升级，软通动力发布首款人形机器人-2025/01/19

3.人形机器人产业周报：特斯拉更新机器人量产目标，OpenAI 重启机器人项目-2025/01/13

4.人形机器人产业周报：广汽发布人形机器人 GoMate，星动纪元更新大模型进展-2024/12/29

5. 人形机器人产业 2025 年度投资策略：量产元年，明日在途-2024/12/19

目录

1 具身智能打开万亿蓝海市场

复盘机器人发展历程，具身智能时代已经到来。传统的工业机器人、协作机器人等需要按照提前设定好的程序步骤进行固定的工作，或者依靠传感器部件调整自身行为。通过搭载人工智能模型，具身智能机器人则有着智能化程度高、工作场景限制小、能够自主规划复杂工作的特点。

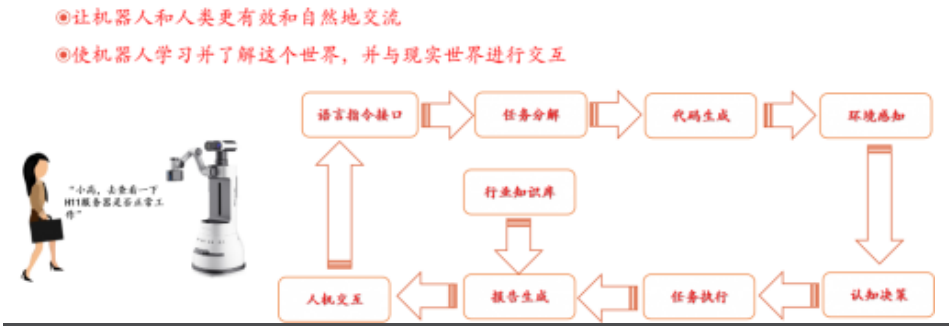
表1：智能机器人发展历程

时间	发展阶段	智能化程度	工作场景	工作任务	成熟度	代表产品
2008 年以前	工业机器人	产线自动化	固定	简单重复工作	成熟期	机械臂、轨道机器人
2008-2015 年	协作机器人	机器智能	可移动	人机协作完成复杂工作	成长期	物流机器人等
2015-2023 年	智能机器人	机器智能	可移动	自主完成简单工作	成长期	手术、陪护机器人等
2023 年及以后	具身智能机器人	人工智能	可移动	自主规划复杂工作	培育期	通用人形机器人

资料来源：奥比中光官网，甲子光年，民生证券研究院整理

具身智能机器人已经成为由“本体”和“智能体”耦合而成且能够在复杂环境中执行任务的智能系统。据高新兴机器人，具身智能机器人能够听懂人类语言，然后分解任务，规划子任务，在移动中识别物体，与环境交互，最终完成相应任务。当前，已有不少研究者尝试将多模态的大语言模型与机器人结合起来，通过将图像、文字、具身数据联合训练，并引入多模态输入，增强模型对现实中对象的理解，帮助机器人处理具身推理任务。

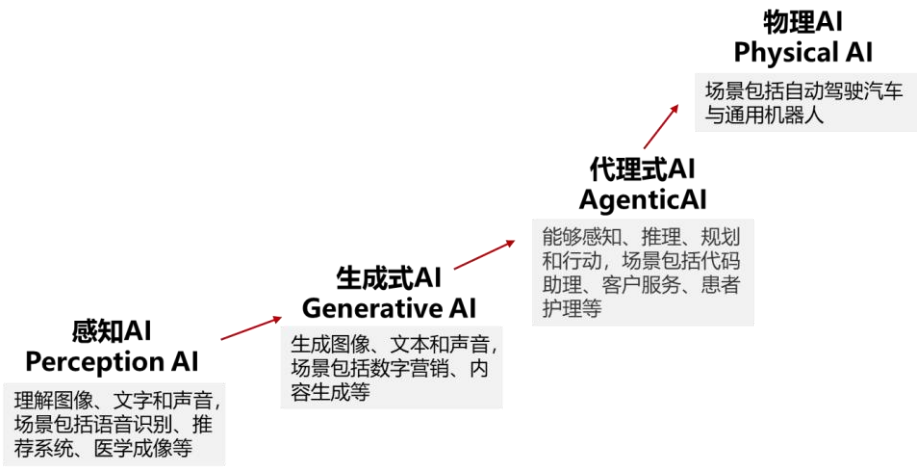
图1：具身智能机器人是一个智能系统



资料来源：高新兴机器人，民生证券研究院

复盘人工智能发展历程，下一发展阶段将是物理 AI。在 2025 CES 的演讲上，黄仁勋表示，AI 的发展有四个阶段，物理 AI 将是 AI 发展的下一个阶段，而通用机器人将是物理 AI 的核心载体。通用机器人给予人工智能身体，让人工智能有了直接改变物理世界的能力。AI 对机器人的赋能主要集中在感知与决策层，使机器人能够与环境交互感知，自主规划决策行动。

图2：英伟达定义的人工智能发展四阶段



资料来源：2025CES 黄仁勋演讲，民生证券研究院

站在具身智能时代的临界点，人形机器人有望成为新一代智能终端，并开启万亿级蓝海市场。人形机器人兼具仿人外形与人工智能，具备操作人类生产生活工具的可能性，有望成为继个人计算机、手机和智能汽车之后的新一代智能终端。马斯克于 2023 年特斯拉股东会议上预测，未来全球的人形机器人数量有望达到 100 亿到 200 亿台，在人类生活和工业制造场景中得到应用，人形机器人将开启万亿级别蓝海市场。

2 机器人“大脑”的时代机遇：具身智能大模型

2.1 多模态大模型为机器人高层级控制带来技术突破

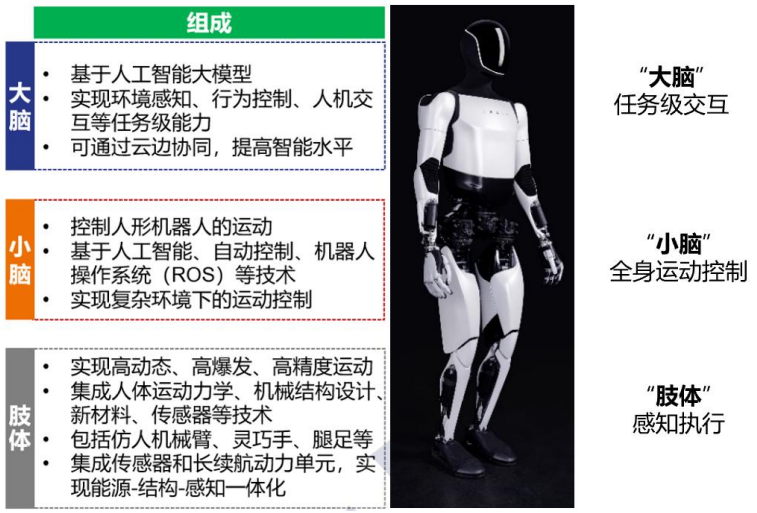
具身智能指的是机器人通过在物理世界和数字世界的学习和进化，达到理解世界、互动交互并完成任务的目标。据稚晖君，具身智能需要本体、智能体、数据、学习和进化架构四大核心要素。

图3：具身智能需要本体、智能体、数据、学习和进化架构四大核心要素

资料来源：智元机器人稚晖君演讲，民生证券研究院

一般来讲，我们可以将一台通用人形机器人本体分为“大脑”、“小脑”和“肢体”三部分，分别对应决策交互模块、运动控制模块和执行模块。其中，人形机器人“大脑”的核心为人工智能大模型技术，通过多模态模型建模、强化学习、地图创建和数据训练，能够管理和协调机器人的各种功能。“大脑”是机器人智能与高级决策的核心，也是具身智能时代机器人区别于程序控制机器人（传统工业机器人、协作机器人等）的关键环节。

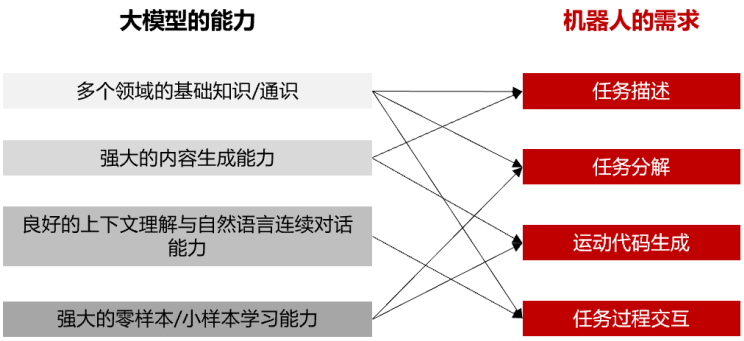
图4：“大脑”、“小脑”及“肢体”三大部分组成人形机器人



资料来源：《人形机器人产业发展研究报告（2024 年）》中国信通院，民生证券研究院

让机器人“大脑”实现突破最核心的推动力是大模型实现涌现、成为真正的生产力。大模型的能力与机器人的需求十分契合，只需要告诉机器人它要做的任务是什么，机器人就会理解需要做的事情，拆分任务动作，生成应用层控制指令，并根据任务过程反馈修正动作，最终完成人类交给的任务，整个过程基本不需要或者仅需少量人类的介入和确认，基本实现了机器人自主化运行，无需掌握机器人专业操作知识的机器人应用工程师介入。

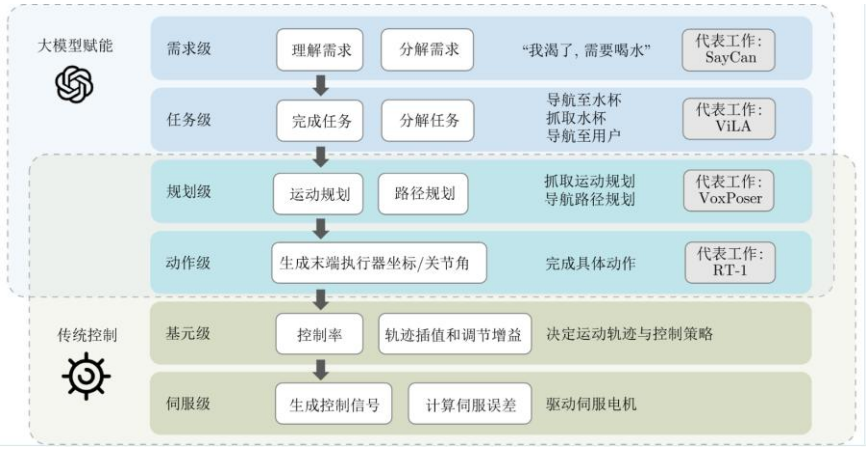
图5：大模型推动机器人产业进入具身智能时代



资料来源：机器人大讲堂微信公众号，民生证券研究院

大模型目前较为擅长需求理解、任务分解等高层级控制任务。根据《基于大模型的具身智能系统综述》，传统机器人的分层控制可以分为规划级、动作级、基元级、伺服级四个层次，具身智能机器人的控制一般可以粗略地分为高层和低层，其中高层负责全局、长期的目标，包括需求级、任务级、规划级和动作级；低层负责具体操作与及时反馈，包括基元级与伺服级。与传统机器人相比，具身智能机器人增加了需求级与任务级的控制。虽然大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，所以目前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。规划级以下的控制规划属于传统机器人控制规划的范畴，更适合传统机器人更成熟的高频控制方法。

图6： 具身智能系统的控制层级



资料来源：王文晟等《基于大模型的具身智能系统综述》，民生证券研究院

多模态大模型突破单一模态大模型的局限性，强化了机器人多模态信息整合、

复杂任务处理等泛化能力，是人形机器人大模型的技术支撑。语言、图片、视频等单一模态大模型以大语言模型（LLM）为基础，将强大的 LLM 作为“大脑”来执行多模态任务。但 LLM 只能理解离散文本，在处理多模态信息时不具有通用性。另一方面，大型视觉基础模型在感知方面进展迅速，但推理方面发展缓慢。

由于两者的优缺点可以形成巧妙的互补，单模态 LLM 和视觉模型同时朝着彼此运行，结合上部分的图像、视频和音频等等模态，最终带来了多模态大语言模型（MLLM）的新领域。形式上，它指的是基于 LLM 的模型，该模型能够接收多模态信息并对其进行推理。从发展人工通用智能的角度来看，MLLM 可能比 LLM 向前迈出一步。MLLM 更加符合人类感知世界的方式，提供了更用户友好的界面（可以多模态输入），是一个更全面的任务解决者，不仅仅局限于 NLP 任务。

图7：MLLM 的模型结构

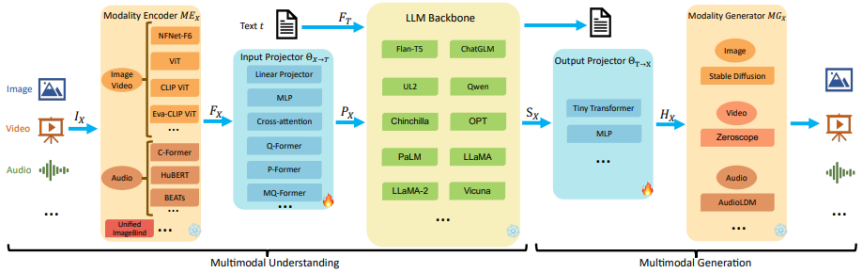


Figure 2: The general model architecture of MM-LLMs and the implementation choices for each component.

资料来源：Duzhen Zhang《MM-LLMs: Recent Advances in MultiModal Large Language Models》，民生证券研究院

2.2 国内外科技巨头与机构入局，具身大模型成果涌现

具身智能机器人操作系统有望推动人机交互的革命和人形机器人商业化落地进程，成为国内外科技巨头和科研机构的兵家必争之地：

1）**微软：**发表论文《ChatGPT for Robotics》等系列论文探究使用 GPT 控制机器人，微软建立高级机器人 API 或函数库（技能库），用户使用自然语言描述需求后，GPT 灵活选用已有 API 或自行编程完成任务；

2）**谷歌：**连续发布 SayCan、Palm-E、RoboCat、RT-1、RT-2、RT-X 等多个具身智能大模型，探究不同具身智能机器人操作系统的技术路线，包括使用真实数据训练的 VLA 路线以及通过合成数据训练的路线等；

3）**英伟达：**在 2025CES 上提出用于加速物理 AI 开发的平台 Nvidia Cosmo 及一系列世界基础模型，世界基础模型可以预测和生成虚拟环境未来状态的物理感知视频的神经网络，以帮助开发者构建新一代机器人；

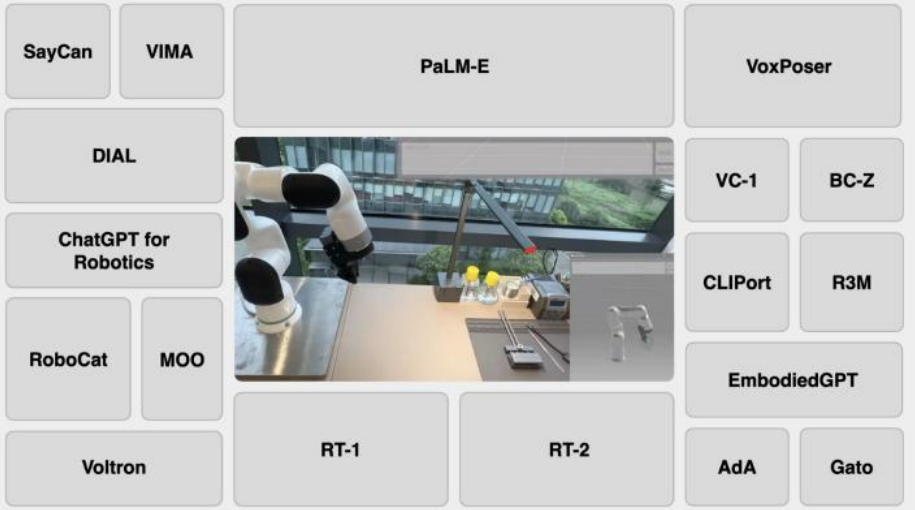
4）**李飞飞团队：**发布 VoxPoser 系统，通过 3D Value Map+LLM+VLM 相

结合的方式，根据用户自然语言直接输出运动轨迹操控机器完成任务；

5）**特斯拉：**Tesla Optimus 能够完成分拣物品、做瑜伽等操作，其神经网络训练是完全端到端的，即直接从视频输入中获取信息，并输出控制指令；

6）**国内团队：**智元机器人、字节跳动、科大讯飞等众多国内厂商已经推出具身智能系统或机器人产品。

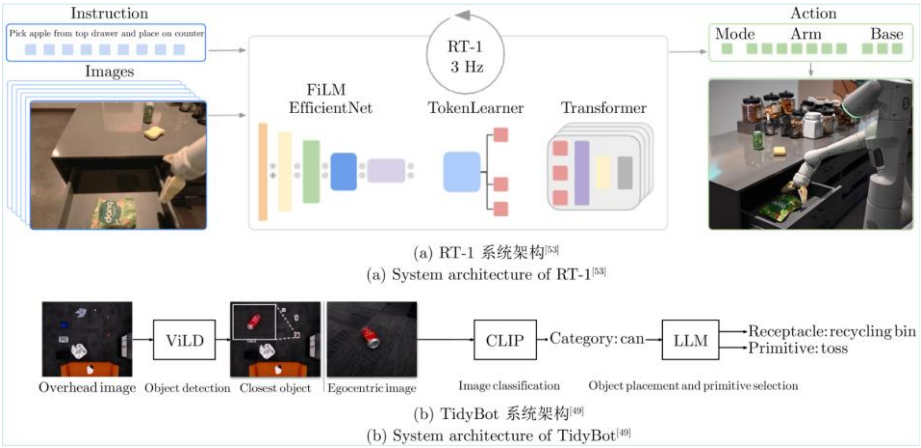
图8：全球前沿具身智能大模型或机器人操作系统



资料来源：甲子光年微信公众号，民生证券研究院

目前的具身智能架构分为端到端模型与冻结参数的大模型结合基础模型。端到端的架构可以直接从输入数据到目标结果，不需要进行提示词工程，较为简洁高效，往往在规划级、动作级中使用；冻结参数的大模型结合基础模型使用的大模型通常是在广泛的数据上预训练好的，在利用大模型的强大能力的同时保留了对特定任务进行微调的灵活性,在需求级、任务级中使用较多。使用预训练模型可以显著减少训练时间和所需的数据量，普遍适用于数据较为稀缺的任务。

图9：具身智能的不同架构举例



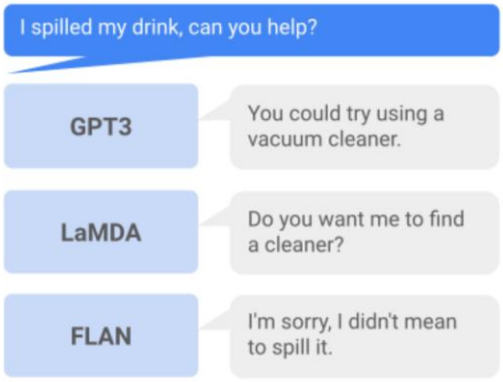
资料来源：王文晟等《基于大模型的具身智能系统综述》，民生证券研究院

2.2.1 谷歌：SayCan、RT-1、PaLM-E、RT-2 到 RT-X

1） SayCan：定位 High-Level，Do As I Can, Not As I Say

2022 年 4 月发布，SayCan 模型的核心出发点是为机器人提供既有用又可行的行动指引。PaLM-E 虽然可以将任务拆分为符合语义逻辑的子任务，但是无法判断其所设定的子任务是否能在现实世界中执行。究其原因在于，大语言模型缺少对真实物理世界的客观原理的深刻理解与经验参考，其生成的子任务虽合逻辑，但是机器人在执行过程中可能会遇到无法顺利操作的困难。以“我把饮料洒了，你能帮忙吗？”为例，现有的大语言模型可能会回答“你可以试试用吸尘器”、“对不起，我不是故意洒的”， 虽然这些回应听起来很合理，但当前环境中的机器人并不具备使用吸尘器的能力，亦或者当前环境中根本没有吸尘器。

图10：大语言模型对“我把饮料洒了，你能帮忙吗？”的回复



资料来源：Michael Ahn 等《Do As I Can, Not As I Say: Grounding Language in Robotic Affordances》，民生证券研究院

SayCan 尝试将大模型 LLM 与物理任务联系起来并解决上述问题。其中，Say 代表大模型 LLM，用于输出可用的高层级运动指令，Can 代表机器人在当前环境下能做的事情，二者通过值函数（Value Function）的方式结合起来，共同决定选择哪条指令用于实际执行。

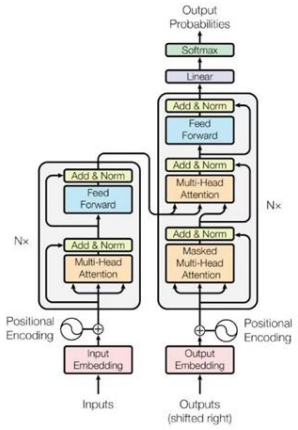
图11：SayCan 对于“我把饮料洒了，你能帮忙吗？”的决策流程

资料来源：Michael Ahn 等《Do As I Can, Not As I Say: Grounding Language in Robotic Affordances》，民生证券研究院

2) RT-1：开启 Transformer 与机器人的结合

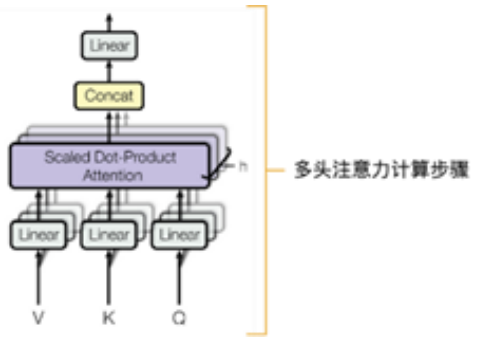
2022 年，Google 提出 Transformer 架构。Transformer 最初是为了解决翻译问题，仅仅依赖于注意力机制就可处理序列数据。这个新的深度学习模型的训练耗时短，并且对大数据或者有限数据集均有良好表现。由于 Transformer 引入了注意力机制和残差链接，也就是所谓“Attention Is All You Need”，因此其计算效率更高，能够加速训练和推理速度。

图12：Transformer 核心架构



资料来源：Ashish Vaswani, Noam Shazeer《Attention Is All You Need》，民生证券研究院

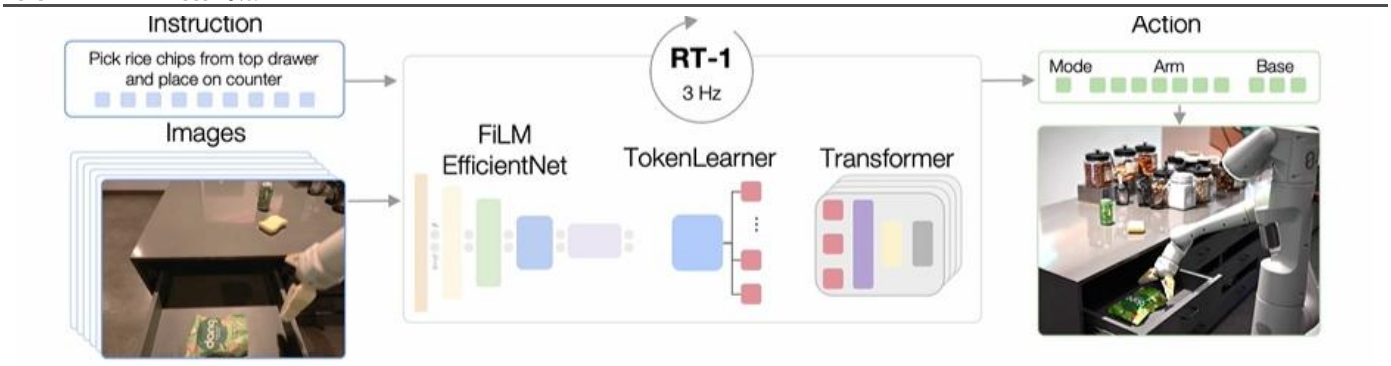
图13：自注意力机制示意图



资料来源：Ashish Vaswani, Noam Shazeer《Attention Is All You Need》，民生证券研究院

2022 年 12 月，Google 在 RT-1 上首先开启了 Transformer 和机器人的结合。RT-1 的主体是预训练的视觉模型加上用解释器处理过的语言指令，两部分再一起通过 transformer 架构输出机器人的动作指令，学习范式是模仿学习。训练数据是在 google 实验室中的两个厨房环境记录的操控移动机械臂完成抓取与放置动作时的记录，数据包括文字指令、过程中的机器人视觉图像、每一帧图像对应的机器人的动作指令（底盘速度，机械臂末端速度）等。

图14：RT-1 结构概览



资料来源：Anthony Brohan《RT-1: Robotics Transformer for real-world control at scale》，民生证券研究院

RT-1 的核心模型架构是将指令和图像 token 化，再做 token 的压缩并输出动作。RT-1 将机器人动作的每个维度进行均匀离散化，并将动作词元化，然后使用监督学习的损失进行训练。为了使视觉-语言模型能够控制机器人，还差对动作

控制这一步。该研究采用了非常简单的方法：他们将机器人动作表示为另一种语言，即文本 token，并与 Web 规模的视觉-语言数据集一起进行训练。

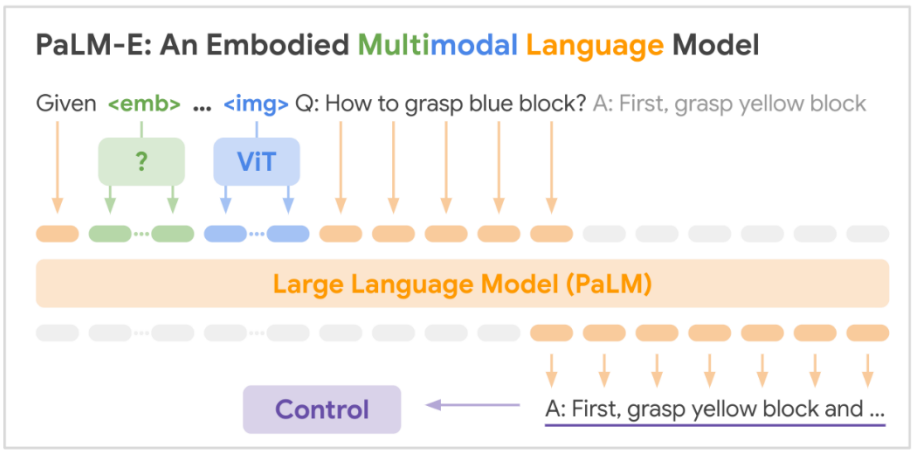
图15：机器人动作数字 token 化

资料来源：Yevgen Chebotar, Tianhe Yu《RT-2: New model translates vision and language into action》，民生证券研究院

3) PaLM-E：多模态理解能力的飞跃

2023 年 3 月发布，PaLM-E 展示了将图像和语言大模型的知识迁移到机器人领域的路径之一。PaLM-E 融合了 Google 当时最新的大型语言模型 PaLM 和最先进的视觉模型 ViT-22B，在纯文本的基础上将输入数据扩充至其他多模态数据（主要来自于机器人的传感器，比如图像、机器人状态、场景环境信息等），并输出以文本形式表示的机器人运动指令，进行端到端的训练。

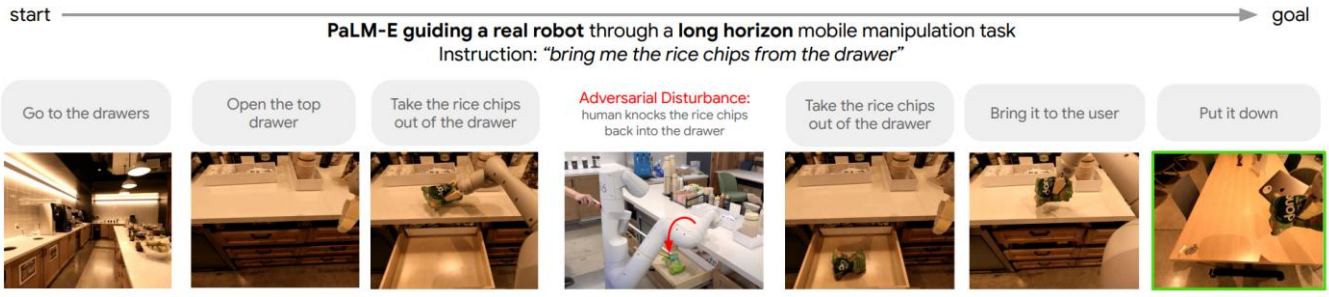
图16：PaLM-E 模型架构总览



资料来源：Danny Driess, Pete Florence《PaLM-E: An embodied multimodal language model》，民生证券研究院

PaLM-E 可以把高层级的任务拆分成若干个在语义上符合逻辑的子任务，再根据已采取步骤的历史记录和当前对场景的图像观察来生成计划的下一步。以“把抽屉里的薯片拿来给我”为例，PaLM-E 模型将输出以下机器人的运动指令：1、移动到抽屉旁边；2、打开抽屉；3、把薯片从抽屉里拿出来；4、把薯片带到用户旁边；5、放下薯片；6、任务结束。

图17：在 PaLM-E 的引导下，机器人具备拆解和执行长程任务的能力

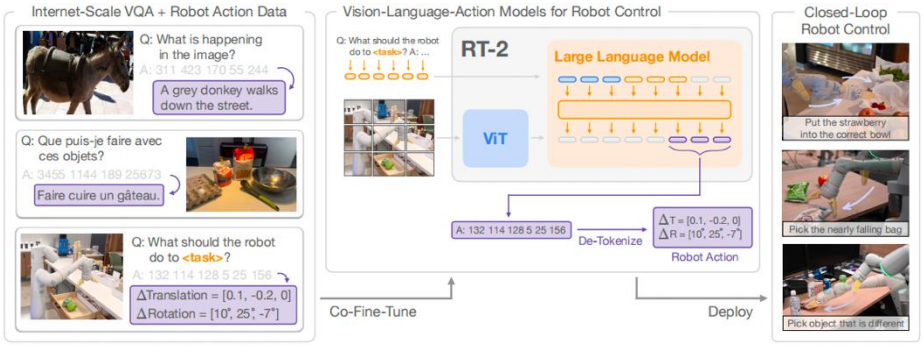


资料来源：Danny Driess, Pete Florence《PaLM-E: An embodied multimodal language model》，民生证券研究院

4) RT-2：结合 RT-1 与 PaLM-E，首个 VLA 大模型

2023 年 7 月发布，RT-2 在 RT-1 的基础上升级，可以直接理解复杂指令从而直接操控机械臂。RT-2 的目标是将 VLM 具备的数学、推理、识别等能力和 RT-1 的操作能力结合，能够用复杂文本指令直接操作机械臂，通过自然语言就可得到最终的动作。最终，Google 提出一个在机器人轨迹数据和互联网级别的视觉语言任务联合微调视觉-语言模型的学习方式。这类学习方法产生的模型被称为视觉-语言-动作（VLA）模型，具有泛化到新对象的能力、解释命令的能力以及根据用户指令思维推理的能力。

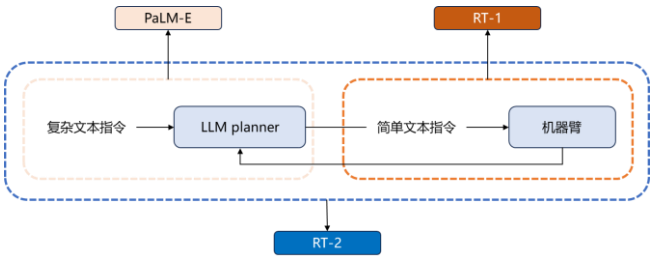
图18：RT-2 全流程概览



资料来源：Anthony Brohan《RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control》，民生证券研究院

RT-2 将输出的动作进行和 RT-1 相同的离散化操作后将词元加入视觉-语言模型原先的词表中，可以把动作词元视为另外一种语言进行处理，无需改变原有视觉-语言模型结构设计。由于 RT-2 已经在海量的视觉问答任务中进行预训练，在对图片和任务指令的理解上有更加丰富的经验，在任务集合上具有更强的泛化能力。例如在下图的拾取、移动、放置等具体任务中，智能体能够精准识别任务需求并且以过往训练经验为基础准确地完成。

图19：PaLM-E、RT-1 与 RT-2 逻辑关系



资料来源：CSDN，民生证券研究院

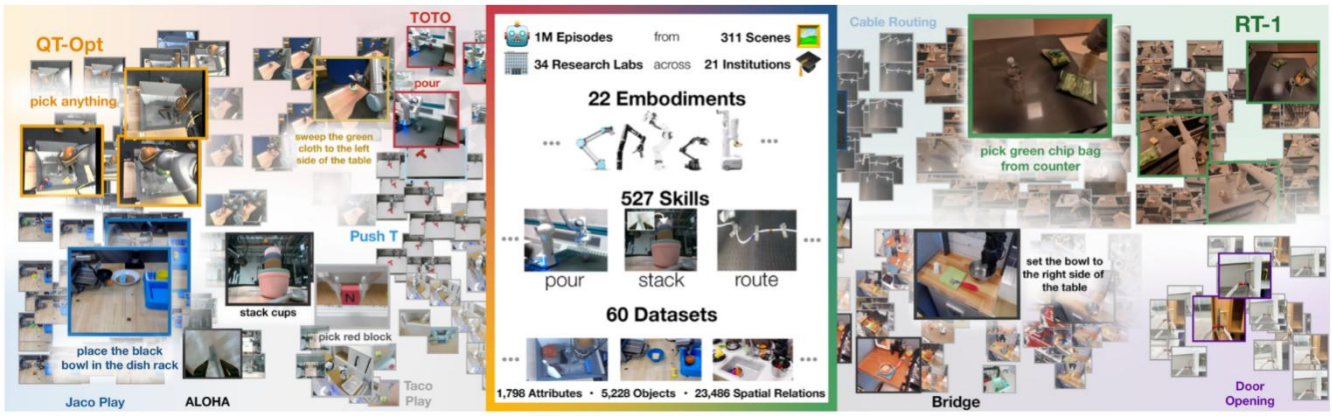
图20：RT-2 能够推广到各种需要推理、符号理解和人类识别的现实世界情况

资料来源：Anthony Brohan《RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control》，民生证券研究院

5) RT-X 系列：数据驱动泛化性及成功率跃升

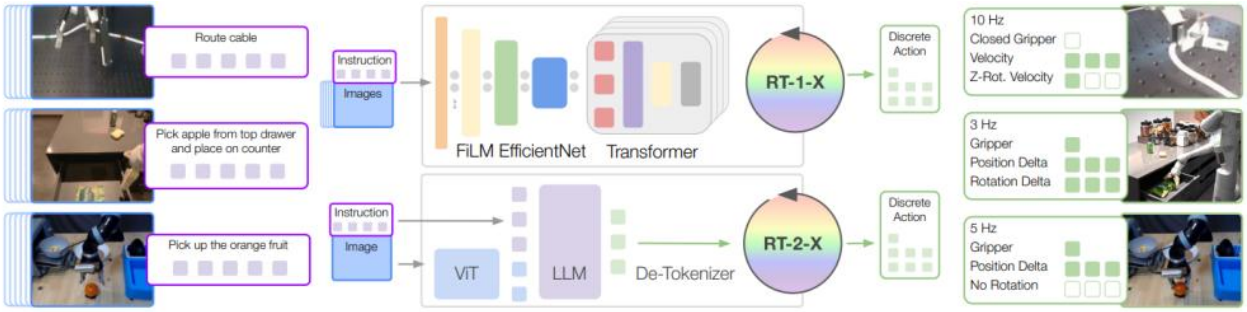
2023 年 10 月发布，RT-X 系列模型核心是让机器人学习更多机器人的“动作”，达到更强的任务泛化和更高的任务成功率。谷歌构建 Open X-Embodiment Dataset 数据库，覆盖从单机械臂到双手机器人和四足机器人等 22 个类型的机器人的 527 个机器人的“动作”。与 RT-1 相比，RT-1-X 任务完成的成功率提升 50%；与 RT-2 相比，RT-2-X 展现出更好的任务泛化能力， RT-2-X 的成功率是其之前的最佳模型 RT-2 的三倍，这也说明了，与其他平台的数据进行联合训练可以为 RT-2-X 赋予原始数据集中不存在的额外技能，使其能够执行新任务。

图21：RT-X Open X-Embodiment Dataset 数据集



资料来源：Open X-Embodiment Collaboration《Open X-Embodiment: Robotic Learning Datasets and RT-X Models》，民生证券研究院

图22：RT-X 大模型工作原理

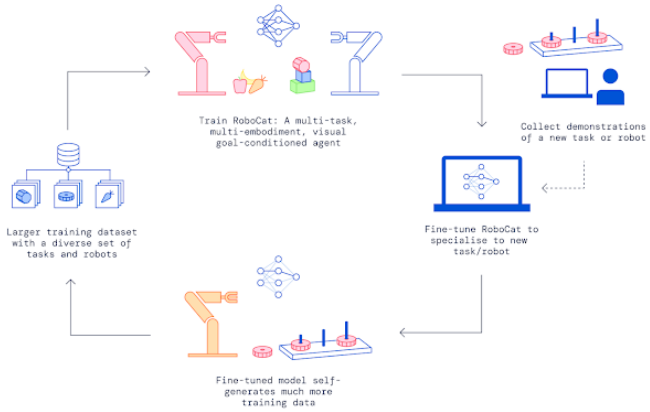


资料来源：Open X-Embodiment Collaboration《Open X-Embodiment: Robotic Learning Datasets and RT-X Models》，民生证券研究院

6) RoboCat：机器人的自我提升

2023 年 6 月发布，RoboCat 可以通过自己生成训练数据集的方式更快完善其能力。谷歌将 Gato 的架构与大型训练数据集相结合，该数据集包含各种机器人手臂的图像序列和动作，可解决数百个不同的任务。在第一轮培训之后，RoboCat 进入了一个“自我提升”的培训周期，其中包含一系列以前看不见的任务，每个新任务的学习遵循五个步骤：1）使用由人类控制的机械臂收集 100-1000 个新任务或机器人的演示；2）在这个新任务/分支上微调 RoboCat，创建一个专门的衍生代理；3）衍生代理在这个新任务/手臂上平均练习 10,000 次，生成更多训练数据；4）将演示数据和自生成数据整合到 RoboCat 现有的训练数据集中；5）在新的训练数据集上训练新版本的 RoboCat。谷歌提出，RoboCat 只需 100 个演示即可完成一项新任务，这种能力将有助于加速机器人研究，因为它减少了对人类监督训练的需求，是创建通用机器人的重要一步。

图23：RoboCat 工作原理

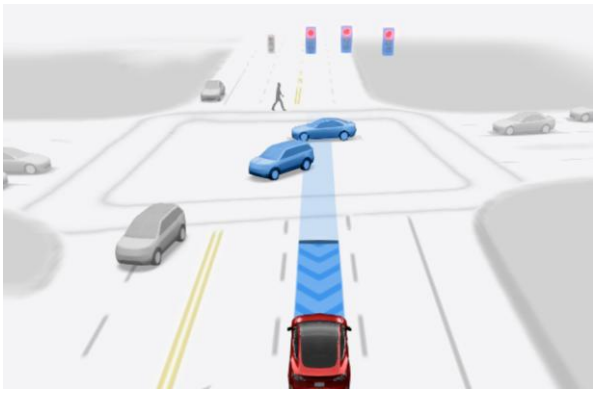


资料来源：The RoboCat team《RoboCat: A Self-Improving Generalist Agent for Robotic Manipulation》，民生证券研究院

2.2.2 特斯拉：坚持端到端算法路线，感知决策一体化

FSD 全称 Full Self-Driving（完全自动驾驶），是特斯拉研发的自动化辅助驾驶系统，目标是实现 L5 级别的自动驾驶。FSD V12（Supervised）是全新的“端到端自动驾驶”，模型架构发生了重大变化。据特斯拉 CEO 埃隆·马斯克表示，特斯拉 FSD V12(Supervised)需要人工干预的频率只有 FSD V11 的百分之一。FSD V12（Supervised）完全采用神经网络进行车辆控制，从机器视觉到驱动决策都将由神经网络进行控制。该神经网络由数百万个视频片段训练而成，取代了超过 30 万行的 C++代码。FSD V12（Supervised）减少了车机系统对代码的依赖，使其更加接近人类司机的决策过程。

图24：FSD V12（Supervised）虚拟界面显示



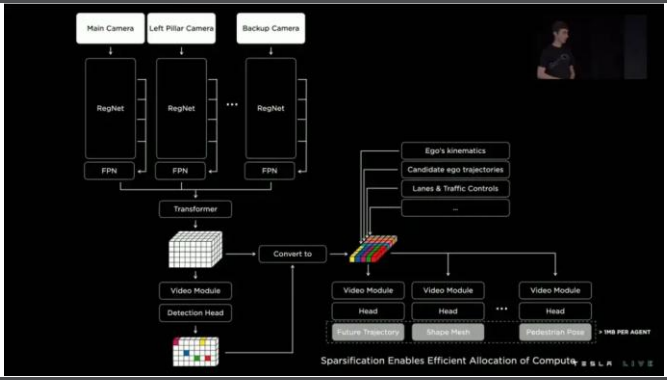
资料来源：特来讯，民生证券研究院

图25：自动驾驶的六个等级

资料来源：SAE，华为云社区，民生证券研究院

FSD V12 为首个端到端自动驾驶系统，实现感知决策一体化。特斯拉 FSD v12 采用端到端大模型，消除了自动驾驶系统的感知和定位、决策和规划、控制和执行之间的断面，将三大模块合在一起，形成了一个大的神经网络，直接从原始传感器数据到车辆操控指令，简化了信息传递过程，因而减少了延迟和误差，提高了系统的敏捷性和准确性。FSD V12 能够模拟人类驾驶决策，成为自动驾驶领域全新发展路径。FSD V12 也被称为“Baby AGI（婴儿版通用人工智能）”，旨在感知和理解现实世界的复杂性。

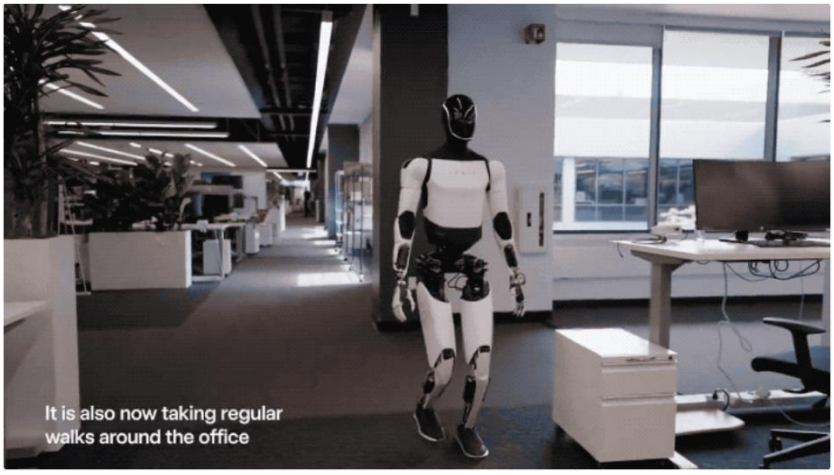
图26：Baby AGI 架构



资料来源：汽车测试网，民生证券研究院

特斯拉将车端 FSD 迁移至人形机器人。端到端算法从汽车自动驾驶迁移至人形机器人几乎不需要做太多额外工作，车本身就是一种机器人。早期的特斯拉 Optimus 机器人使用了与汽车完全相同的计算机和摄像头，通过让汽车的神经网络在机器人上运行，它在办公室里走动时仍试图识别“可驾驶空间”，而实际上它应该识别的是“可行走空间”。这种通用化能力表明了很多技术是可以迁移的，虽然需要一些微调，但大部分系统和工具都是通用的。

图27：特斯拉 Optimus 机器人避障行走



资料来源：tesla，民生证券研究院

2.2.3 字节 GR-2：高效动作预测与泛化能力

GR-2 的训练包括预训练和微调两个过程。GR-2 在 3800 万个互联网视频片段上进行生成式训练，也因此得名 GR-2（Generative Robot 2.0）。这些视频来自学术公开数据集，涵盖了人类在不同场景下（家庭、户外、办公室等）的各种日常活动，以期迅速学会人类日常生活中的各种动态和行为模式。这种预训练方式使 GR-2 具备了学习多种操作任务和在多种环境中泛化的潜能。庞大的知识储备，让 GR-2 拥有了对世界的深刻理解。

在微调阶段，GR-2 通过几项关键改进提升了其在实际任务中的表现。首先，GR-2 引入数据增强技术，通过改变训练数据中的背景和物体，使其在未见环境下更具泛化能力。此外，模型通过多视角训练，利用不同角度的视觉数据，增强了其在复杂场景中的操作灵活性和准确性。为了保证动作的流畅性，GR-2 使用了条件变分自编码器（cVAE），生成连续、平滑的动作序列，确保任务执行时的动作更加高效和精准。

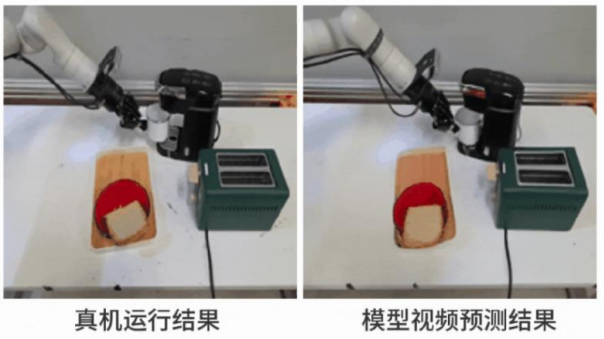
在经历大规模预训练后，通过在机器人轨迹数据上进行微调，GR-2 能够预测动作轨迹并生成视频。GR-2 的视频生成能力，让它在动作预测方面有着天然的优势，显著提高了准确率。它能够通过输入一帧图片和一句语言指令，预测未来的视

频，进而生成相应的动作轨迹。如下图所示，只需要输入一句语言指令："pick up the fork from the left of the white plate"，就可以让 GR-2 生成动作和视频。可以看到，机械臂从白盘子旁边抓起了叉子。图 29 右图中预测的视频和真机的实际运行也相差无几。

图28：GR-2 视频-语言模型与视频-语音-动作模型

资料来源：Chi-Lam Cheang 等《GR-2: A Generative Video-Language-Action Model with Web-Scale Knowledge for Robot Manipulation》，民生证券研究院

图29：真机预测结果与模拟视频预测结果对比



资料来源：Chi-Lam Cheang 等《GR-2: A Generative Video-Language-Action Model with Web-Scale Knowledge for Robot Manipulation》，民生证券研究院

GR-2 的强大之处不仅在于它能够处理已知任务，更在于其面对未知场景和物体时的泛化能力。无论是全新的环境、物体还是任务，GR-2 都能够迅速适应并找到解决问题的方法。在多任务学习测试中，GR-2 能够完成 105 项不同的桌面任务，平均成功率高达 97.7%。此外，GR-2 还能够与大语言模型相结合，完成复杂的长任务，并与人类进行互动，并可以鲁棒地处理环境中的干扰，并通过适应变化的环境成功完成任务。。

在实际应用中，GR-2 相比前一代的一个重大突破在于能够端到端地完成两个货箱之间的物体拣选。无论是透明物体、反光物体、柔软物体还是其他具有挑战性的物体，GR-2 均能准确抓取。这展现了其在工业领域和真实仓储场景的潜力。除了能够处理多达 100 余种不同的物体，如螺丝刀、橡胶玩具、羽毛球，乃至一串葡萄和一根辣椒，GR-2 在未曾见过的场景和物体上也有着出色的表现。

图30：GR-2 完成流畅端到端物体拣选示意图

资料来源：GR-2 官方项目网站，民生证券研究院

图31：GR-2 在实验中顺利完成 122 项物体拣选，其中过半物体 GR-2 未曾见过

资料来源：GR-2 官方项目网站，民生证券研究院

2.3 具身大模型的关键挑战

2.3.1 关键挑战一：泛化性弱

当前的具身大模型面临未知环境和任务时，通常存在泛化能力弱的问题。具身任务往往涉及多样的实体类型和动态的环境变化，当智能体和环境的动力学参数发行改变，目前的具身策略将很难直接适用。以 RT-2 为例，在谷歌山景城办公室的厨房测试中，RT-2 展现了极高的任务执行成功率（近 98%），然而一旦换到施工工地、嘈杂后厨等复杂场景，成功率便骤降至 30%左右。泛化性不足的原因有多个方面：一是数据不足，目前机器人操作领域的数据量远未达到互联网数据的量级；二是对错误的容忍度低，机器人操作对精度的要求远高于语言模型；三是推理频率不足，大模型在实时操作中的表现还有待提高；四是数据多样性与训练稳定性之间的平衡是一个难题。

2.3.2 关键挑战二：实时性差

当前大模型具身策略的决策存在实时性问题。机器人控制的正确性不仅依赖系统计算的逻辑结果，还依赖于产生这个结果的时间，即“迟到的结果非正确的结果”。以 Figure 机器人为例，Figure 机器人在视频中呈现的延迟时长约为 2-3 秒，因为它使用了 Pipeline、管道型路线，即自然语言发送后、机器人“大脑”可以理解并生成指令，由指令来控制。而谷歌 RT-2 只能做到 1~5Hz 的推理和控制指令生成速率，输出运动频率仅能达到 1-3Hz，使得机器人的反射弧长达 0.3 秒甚至 1 秒，远远低于人类和许多实际应用场景的需求。

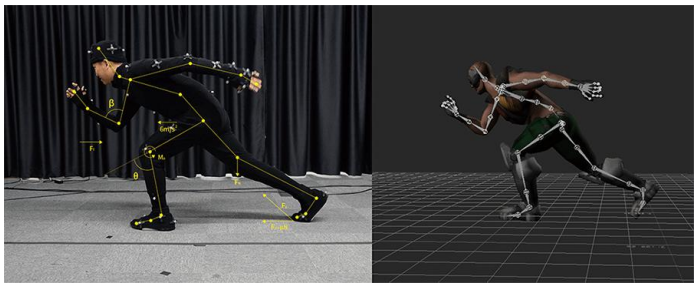
2.3.3 关键挑战三：数据收集与合成数据的使用

1）真实数据收集与标注

端到端算法需要大量连续时序的驾驶行为视频进行标注，这种数据收集、标注及闭环验证的过程在人形机器人上同样困难。人形机器人需要面对更加复杂的环境和任务，因此数据收集的难度和成本都更高。同时，由于人形机器人的操作具有更高的风险性，因此数据标注的准确性也要求更高。人形机器人需要大量实际人类真实的数据集给机器人进行训练。

动作捕捉技术和 VR 远程操作是实现人形机器人拟人化动作数据采集的有效途径。动作捕捉技术通过在人体关键部位贴上反光标记点或使用惯性传感器等方式，捕捉人体的运动姿态和动作数据。VR 远程操控技术是人类戴着 VR 眼镜和手套，通过远程操作的方式来采集机器人数据。这些数据可以被用于训练人形机器人的动作模型，使其能够模拟出类似人类的动作和行为。

图32：动作捕捉技术采集数据



资料来源：瑞立视官网，民生证券研究院

图33：VR 远程操控采集数据

资料来源：特斯拉，民生证券研究院

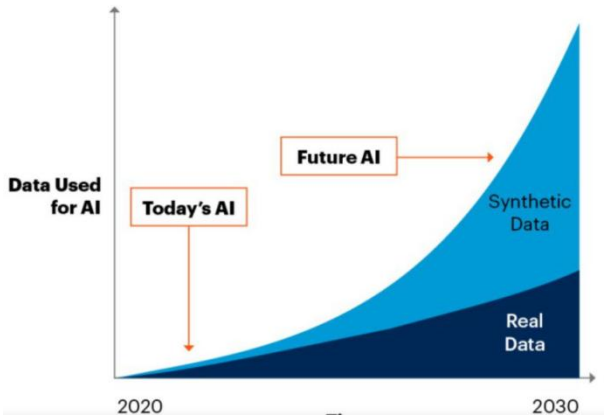
2）合成数据的生成和使用

由于扩展法则（Scaling Law）的存在，机器人的数据集大小决定了其性能的好坏，真实数据的采集消耗较大的人力物力成本，合成数据仅依赖 AI 算法实现数据生成，数据采集快并且成本低廉。

同时人形机器人面临着场景复杂性与模型泛化能力的问题，合成数据构建的世界模型就起到了很大的作用。自动驾驶场景相对结构化，主要操作在可预测和规范化的环境中。而人形机器人需要应用于多样的场景，如工厂、家庭、办公室等，对泛化能力的要求远高于自动驾驶汽车。基于世界模型生成高质量的动作视频和规划策略，在仿真环境中模拟各种复杂场景，就能够提升系统的鲁棒性。

合成数据生成的关键问题是保持数据集的熵和多样性，避免生成的数据与真实数据差距过大或者样式单一。

图34：未来合成数据的使用

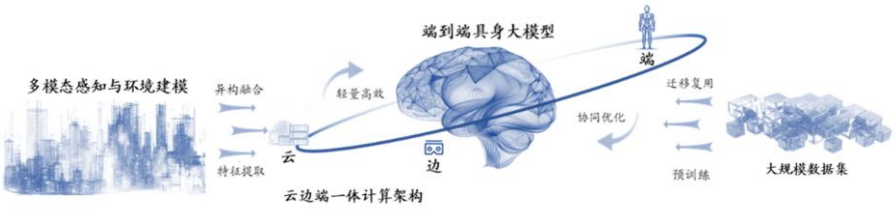


资料来源：Gartner，CSDN，民生证券研究院

3 机器人“大脑”的外延：云计算与边缘计算

“大脑”以具身大模型为核心，外延还包括模态融合、大规模数据集、云边端一体计算架构等多方面技术。具体来说，多模态融合感知技术可以将视觉、触觉等不同模态的数据直接输入到深度神经网络中，通过联合学习实现多模态信息的无缝融合，获得更全面、准确的环境表征。大规模数据集则为模型提供了广泛的先验知识，使其能够应对复杂多变的现实环境，具身大模型通过在海量多模态数据上的预训练，将多模态输入映射到一个统一的语义空间，并在此基础上进行任务理解、决策规划等高层认知。云边端一体计算架构通过软硬件协同设计，针对机器人应用的特点进行优化，可以大幅提升系统的实时性、能效比和可靠性，发挥云、边、端不同层级计算资源的优势，实现具身大模型推理、多模态感知的高效协同。

图35：围绕人形机器人“大脑”的关键技术架构



资料来源：人形机器人世界微信公众号，民生证券研究院

3.1 机器人“大脑”的运行保障：云计算

3.1.1 大模型深度赋能机器人，云计算提供算力及存储

“云计算”是分布式计算的一种，指的是通过网络“云”将巨大的数据计算处理程序分解成无数个小程序，然后，通过多部服务器组成的系统进行处理和分析这些小程序得到结果并返回给用户。它通过网络以按需、易扩展的方式获得所需资源。云计算根据服务类型划分主要可以分为三层：（1）Laas，基础设施即服务：为企业提供 IT 基础设施，如服务器、存储设备等（2）PaaS，平台即服务：提供计算、网络、开发工具等资源，用于工具和应用程序的创建（3）SaaS，软件即服务：指通过互联网按需提供软件应用程序。

AI、大模型深度赋能机器人，算力需求确定，云计算能够提供算力和存储空间。如谷歌于 23 年 3 月推出 PaLM-E 模型，融合了 ViT Vision Transformer 的 220 亿参数和 PaLM 的 5400 亿参数，集成了可控制机器人视觉和语言的能力；ChatGPT 为代表的 NLP 革命性进展未来将助力机器人的语音语义分析及交互模块的优化，强大的语义模型可以帮助泛通用机器人理解更复杂的指令和目标，从何做出更符合人类期望的决策。云计算能够为 AI、大模型与机器人的结合提供强大的计算能力和数据存储空间，以及能够随时随地获得所需资源和算法支持的灵活性、可拓展性。

3.1.2 云计算机器人及市场规模

云机器人技术指利用云计算（如云存储、云处理等）以及其他相关技术（如大数据分析、机器学习等）来提升机器人功能的所有技术。**本地机器人、网络连接和云服务器是云机器人架构的三个关键构成要素。**

（1）**本地机器人**：主要负责直接与环境进行交互，包括接收传感器输入（如视觉、触觉、声音等）和执行动作。本地机器人通常有一些基础的计算能力，能够进行一些简单的数据处理和决策。另外，本地机器人需要有网络接口，以连接到云端。

（2）**网络连接**：它是本地机器人和云服务器之间的桥梁，负责传输数据和指令。网络连接需要足够的带宽，以支持大量数据的传输，同时需要足够的可靠性和安全性，以保证数据的准确性和安全性。

（3）**云服务器**：它是云机器人架构的核心，提供大规模的数据存储和强大的计算能力。云服务器可以运行各种软件和服务，如数据分析工具、机器学习模型、数据库等。通过云服务器，机器人可以共享数据，利用强大的计算资源，进行复杂的任务。

以赛特智能推出的智塞拉为例，智赛拉是一款专为医疗领域设计的智能配送机器人，其强大的功能和独特的云机器人架构提高医疗服务效率和质量。智塞拉的**本地机器人部分**，配置有大容量箱体、高清双目摄像头、高精度定位导航系统、感知避障系统（超声波雷达、激光雷达等）和无线网络连接模块，这款自动驾驶电动车具备实时感知和自我导航的能力。通过**无线局域网/4G/5G**，智赛拉实现了与云服务器的无缝连接。它所对接的**云端**包含了医院管理系统和配送路线规划系统，使得智赛拉在实现药品追溯和人员追溯的同时，也能够自主调用电梯，并在闲时自主返回充电。

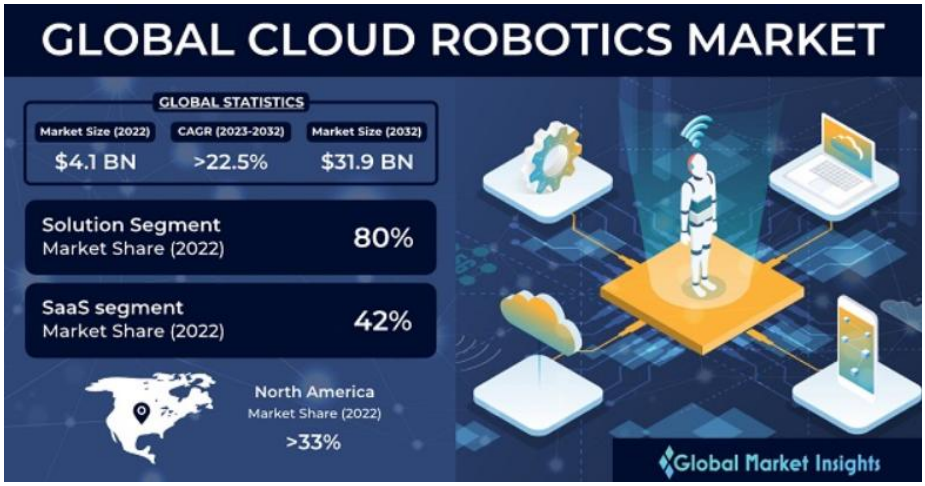
图36：智塞拉机器人

资料来源：赛特智能，民生证券研究院

根据 Global Market insights 统计，**2022 年，全球云机器人市场规模达 41 亿美元，预计 2023 年至 2032 年的复合增长率将超过 22.5%。**其中，由于 SaaS

服务为企业提供了经济高效且可扩展的访问云机器人功能的方法，且为企业节约了大量在基础设施和软件开发方面的前期投资，因此市场份额最大，超过 42%。我们预计随着云机器人市场规模的持续扩张，云计算相关基础设施将充分受益。

图37：云机器人市场规模统计



资料来源：GMI，民生证券研究院

3.2 机器人集群智能的核心：边缘计算

云计算是为机器人等终端设备提供算力的核心方式，但云计算在数据传输成本、时延、安全性等方面的局限性使得边缘计算存在发挥空间。

首先，大型数据中心存在增量算力边际递减现象，单位算力成本的增加将制约“集中式”的算力发展；第二，网络性能会限制数据中心算力的发挥，长距离数据传输还会导致较高的时延；最后，安全性也是重要因素。以工业机器人为例，部分工厂的设备管理人员并不会把机器人、传感器等设备的数据全部通过互联网上传到云端，以防止窃取信息或破坏公厂运行等极端情况发生。**边缘计算的处理能力更靠近设备或数据源，能够实现更低的时延、更好的隐私以及更优的成本，我们认为“云-边”计算结合的方式能够帮助机器人实现突破网络环境的限制，大幅缩短响应时间，提高其在复杂场景中的自适应能力和应用价值。**

目前，集成了边缘算力的模组正在成为支撑机器人边缘算力的核心形式。例如在第六届中国国际进口博览会期间，德州仪器展示了搭载 TDA4x 处理器的达明 TM5S 协作机器人，基于 TDA4x 高效稳定的数据处理能力，机械臂能够通过视觉铺桌人的动作并作出相应模仿，彰显智能制造场景下的创新应用。特斯拉针对 Optimus 机器人研发的 DOJO D1 芯片也扮演了类似的角色。

图38：TM5S 协作机器人

资料来源：达明官网，民生证券研究院

图39：Optimus 机器人

资料来源：搜狐网，民生证券研究院

国内厂商同时也在进行积极探索。例如 2023 年 5 月中科创达发布的 Rubik 魔方大模型和几有的产品、业务密切融合，提升了边缘计算在自然语言、图形图像处理、个性化推荐等领域的准确性与效率。根据公司官网，中科创达将智能音箱与机器人进行融合，通过模仿大模型的训练，实现了能够自由对话的智能销售机器人。

4 投资建议

2025 年人形机器人行业进入小批量量产阶段，全球将有数千台人形机器人进入工厂场景训练，加速人形机器人“大脑”的发展。我们认为，目前人形机器人硬件端技术路线趋向收敛，软件端“大脑”智能水平的提升有望成为人形机器人自主性与泛化性提升的核心推动力。建议关注：

1）“大脑”领域，布局大模型与机器人业务相结合的公司，如科大讯飞、中科创达、萤石网络、柏楚电子、华依科技、芯动联科、汉王科技等；

2）AI+机器人领域，具备高壁垒的公司，如 3D 视觉领域奥比中光、大脑域控制芯片天准科技、新型传感器峰岹科技等；

3）同步受益的机器人本体公司，如总成方案三花智控、拓普集团等。

5 风险提示

1） 机器人算法迭代进步速度不及预期：机器人的算法进步速度可能并非线性，在某些数据缺失的情况下，算法训练的进步速度可能下降。

2） 人形机器人落地场景实际需求不及预期：机器人的实际应用场景还需要结合 B 端/C 端客户的实际付费购买点，可能会与仿真环境中模拟的使用场景有差异。

3）市场竞争加剧：人形机器人产业处于快速发展的起点，展现出极大潜力，若其它之前未从事相关业务的公司切入市场，或导致市场竞争加剧，现有市场参与者收入、利润率水平受到影响。

插图目录

表格目录

分析师承诺

本报告署名分析师具有中国证券业协会授予的证券投资咨询执业资格并登记为注册分析师，基于认真审慎的工作态度、专业严谨的研究方法与分析逻辑得出研究结论，独立、客观地出具本报告，并对本报告的内容和观点负责。本报告清晰准确地反映了研究人员的研究观点，结论不受任何第三方的授意、影响，研究人员不曾因、不因、也将不会因本报告中的具体推荐意见或观点而直接或间接收到任何形式的补偿。

评级说明

投资建议评级标准	评级		说明
以报告发布日后的 12 个月内公司股价（或行业指数）相对同期基准指数的涨跌幅为基准。其中：A 股以沪深 300 指数为基准；新三板以三板成指或三板做市指数为基准；港股以恒生指数为基准；美股以纳斯达克综合指数或标普 500 指数为基准。	公司评级	推荐	相对基准指数涨幅 15%以上
		谨慎推荐	相对基准指数涨幅 5%～15%之间
		中性	相对基准指数涨幅-5%～5%之间
		回避	相对基准指数跌幅 5%以上
	行业评级	推荐	相对基准指数涨幅 5%以上
		中性	相对基准指数涨幅-5%～5%之间
		回避	相对基准指数跌幅 5%以上

免责声明

民生证券股份有限公司（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。

本报告仅供本公司境内客户使用。本公司不会因接收人收到本报告而视其为客户。本报告仅为参考之用，并不构成对客户的投资建议，不应被视为买卖任何证券、金融工具的要约或要约邀请。本报告所包含的观点及建议并未考虑获取本报告的机构及个人的具体投资目的、财务状况、特殊状况、目标或需要，客户应当充分考虑自身特定状况，进行独立评估，并应同时考量自身的投资目的、财务状况和特定需求，必要时就法律、商业、财务、税收等方面咨询专家的意见，不应单纯依靠本报告所载的内容而取代自身的独立判断。在任何情况下，本公司不对任何人因使用本报告中的任何内容而导致的任何可能的损失负任何责任。

本报告是基于已公开信息撰写，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、意见及预测仅反映本公司于发布本报告当日的判断，且预测方法及结果存在一定程度局限性。在不同时期，本公司可发出与本报告所刊载的意见、预测不一致的报告，但本公司没有义务和责任及时更新本报告所涉及的内容并通知客户。

在法律允许的情况下，本公司及其附属机构可能持有报告中提及的公司所发行证券的头寸并进行交易，也可能为这些公司提供或正在争取提供投资银行、财务顾问、咨询服务等相关服务，本公司的员工可能担任本报告所提及的公司的董事。客户应充分考虑可能存在的利益冲突，勿将本报告作为投资决策的唯一参考依据。

若本公司以外的金融机构发送本报告，则由该金融机构独自为此发送行为负责。该机构的客户应联系该机构以交易本报告提及的证券或要求获悉更详细的信息。本报告不构成本公司向发送本报告金融机构之客户提供的投资建议。本公司不会因任何机构或个人从其他机构获得本报告而将其视为本公司客户。

民生证券研究院：

上海：上海市浦东新区浦明路 8 号财富金融广场 1 幢 5F； 200120

北京：北京市东城区建国门内大街 28 号民生金融中心 A 座 18 层； 100005

深圳：深圳市福田区中心四路 1 号嘉里建设广场 1 座 10 层 01 室； 518048