证券研究报告 行业专题报告

华鑫证券 CHINA FORTUNE SECURITIES

从谷歌看机器人大模型进展

海外科技研究

投资评级： 推荐 （ 维持 ）

报告日期： 2023年11月19日

- **分析师：傅鸿浩**
- SAC编号：S1050521120004
- **分析师：臧天律**
- SAC编号：S1050522120001

研究创造价值

大模型是远期人形机器人的必备要素：

人形机器人的特点在于通用性和泛化能力，远期可以替代人类完成多项任务。而大模型具有庞大的先验知识库与强大的通识理解能力，可以满足人形机器人通用性的场景要求和技能要求，不再仅限于完成某一类特定工作，而是进一步完成多类型任务。在机器人大模型上，思维链可以帮助机器人拆分与分解一件事件如何完成，先解码出计划的步骤，再解码需要完成任务需要输出的动作。

谷歌：从Saycan到RT-X，软件领军者，步步为营，模型高速迭代

从2022 年4月谷歌推出 Say-can 模型，初次引入大模型用于做任务理解和拆分，到RT-1使用传统神经网络的方法来执行SayCan的任务，再到RT-2将VLM大模型与RT-1的机器人执行数据集一起微调训练，最后创建Open X数据集训练出模型RT-X。谷歌的模型持续高速迭代，逐步向底层运动控制方面发展。

机器人产业仍然处于较为早期阶段，数据、数据与细分场景模型搭建均有产业机会

目前大部分机器人大模型仍然以单机械臂抓取为主，且模型的框架仍然在持续变化。可以明确看到大模型现在对底层的控制仍然偏弱。我们认为未来产业机会主要有三个方面，1、算力：机器人需要快速与环境交互，同时大模型本身要计算和存储空间。二者叠加之下机器人所需的参数和算力比自动驾驶以及大语言模型都要更大，因此对于算力的需求将在后续逐步有所体现。2、数据：机器人需要通过多种传感器感知环境状态，然后执行实际动作来完成任务，一方面需要3D环境数据，另一方面需要的是主动数据，此类数据量极度稀缺。3、细分场景的模型：未来大模型在机器人的应用，或许是通过底层的通识大模型+细分场景模型微调获得，其中底层架构的通识大模型有望参考类似手机安卓的模式由头部的AI企业开源，而细分场景的模型（同时也包括所需的数据）才是未来大部分企业可以竞争的市场。在这个赛道中，数据仍然是模型的基础。

机器人下游发展不及预期

算力与算法模型更新迭代不及预期

行业竞争加剧风险

目 录

CONTENTS

01 大模型是人形机器人的必备要素

研究创造价值

大模型是人形机器人的必备要素

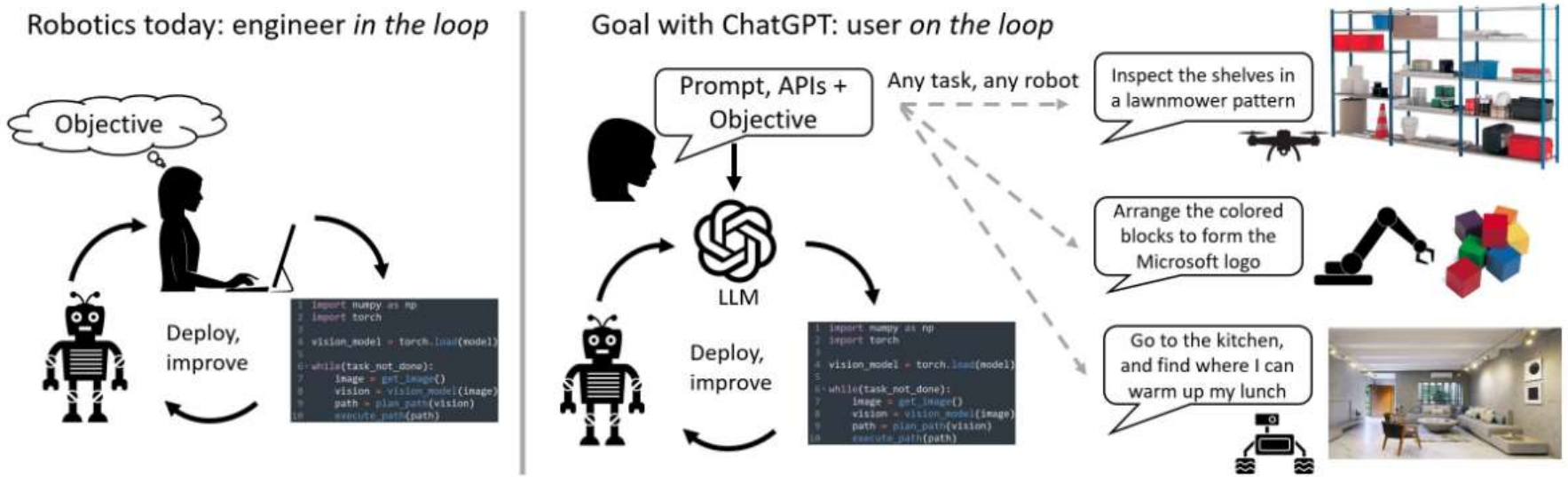
长期来看，人形机器人的最大优势在于通用性：

人形机器人的特点在于泛化能力。如果只为解决单一或少数场景的应用，则特定专用机器人足以满足要求（如酒店服务机器人，扫地机器人等），从第一性原理来说，机器人之所以拟人，其根本目的在于完成多样化的任务——能爬楼梯，能按电梯，能提重物等完成所有人类所需的各种任务。

通用性的实现依赖大模型的应用（体现在感知与识别）：

大模型具有庞大的先验知识库与强大的通识理解能力。可以满足人形机器人通用性的场景要求和技能要求。不再仅限于完成某一类特定工作，而是进一步完成多类型任务。目前机器人的应用基础是代码，机器人工程师需要编写代码和规范来控制机器人行为，这个调试过程缓慢、昂贵且低效，使用场景有限。ChatGPT带来一种新的机器人应用范式，我们可以通过LLM将自然语言快速转换为代码。这样就可以解决大量的场景以及任务需求，有望大幅度降低了算法开发的复杂度，同时可以简化合并算法模型数量，提升开发效率。而传统算法模型即使经历大量的训练，仍存在较多小概率场景（corner case）难以覆盖，泛化能力较低

图表 1：机器人有望通过大模型将自然语言快速转换成代码



资料来源：《ChatGPT for Robotics: Design Principles and Model Abilities》微软，华鑫证券研究所

大模型是人形机器人的必备要素

人形机器人大模型所需的视频数据足够充足（体现在后续的动作）：

深度学习的本质是模仿，可以用大量的人类视频来进行预训练/模仿学习，之后再通过标注用Reinforcement Learning进行微调。机器人做成人形也是为大模型在机器人上的发展铺垫。

思维链条：

思维链(Chain of Thought，CoT)是一种思维工具，通过逐步延伸和拓展一个主要想法，帮助人们进行更深层次的思考，并得出更复杂、更全面的结论。在机器人大模型上，思维链可以帮助机器人拆分与分解一件事件如何完成，增加了先解码出计划的步骤，再解码需要完成任务需要输出的动作，在需要语义推理任务上效果更好。

在谷歌7月发布展示的具身大模型中RT-2中，机器人展示了类似视觉语言模型（VLM）的思维链，如：选出与其他物品不同的物品；告诉机器人很困，让机器人拿饮料，机器人会拿红牛；让机器人完成锤钉子任务，但桌子上只有耳机线、石头、纸，使用思维链后机器人会拿石头等。

图表 2：RT-2的锤钉子实验

资料来源：《RT-2: Vision-Language-Action Models》谷歌，华鑫证券研究所

02 谷歌机器人大模型进展

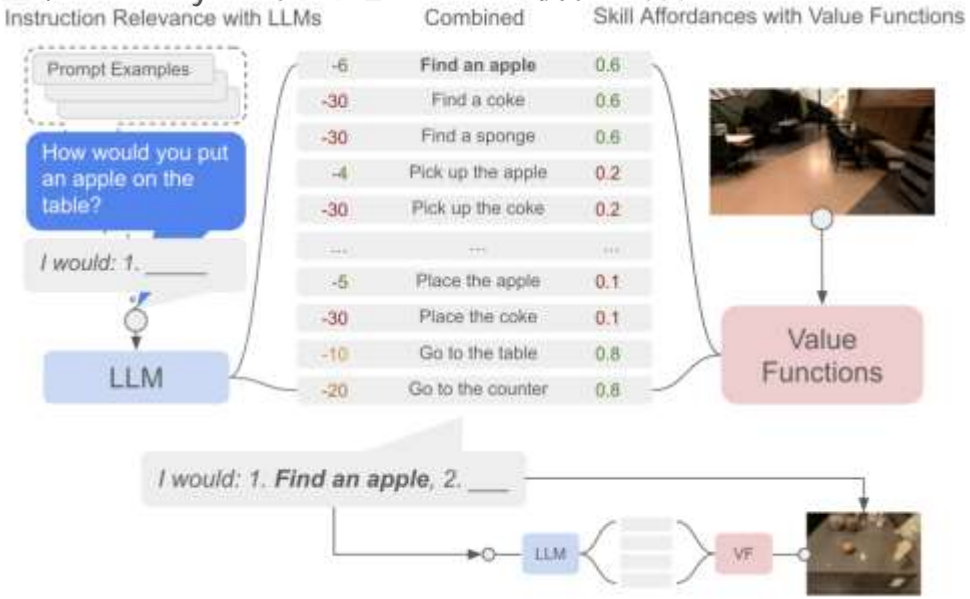
SayCan：谷歌机器人大模型的开端，连接LLM与具身智能

- 2022 年4月，谷歌推出 Say-can 模型。将任务拆分成两个部分，先是“Say”，模型通过与谷歌的大语言模型结合，把获得的任务进行分解，找到最适合当前行动；之后是“Can”，模型计算出当前机器人能够成功执行这一任务的概率。机器人通过将二者结合起来，进行动作。例子：对机器人说“我的饮料撒了，你能帮助我吗”机器人会首先通过语言模型进行任务规划，这时可能最合理的方式是找到一个清洁工、找到一个吸尘器，找一块海绵自己擦等。然后机器人会通过价值函数计算出作为机器人，找到海绵自己擦是最佳方案。之后，机器人就会选择寻找海绵的动作。
- 亮点：首次引入大语言模型帮助理解任务，选择合适的任务规划。
- 不足：机器人的动作仍然是预设好的，因此只能完成特定任务。底层技能通用性和泛用性较差。只能输出高级指令。

图表 3：传统LLM不与世界互动，而Saycan通过预训练的价值方程转换成具体指令

资料来源：《Do As I Can, Not As I Say》谷歌，华鑫证券研究所

图表 4：Saycan同时结合LLM与价值函数

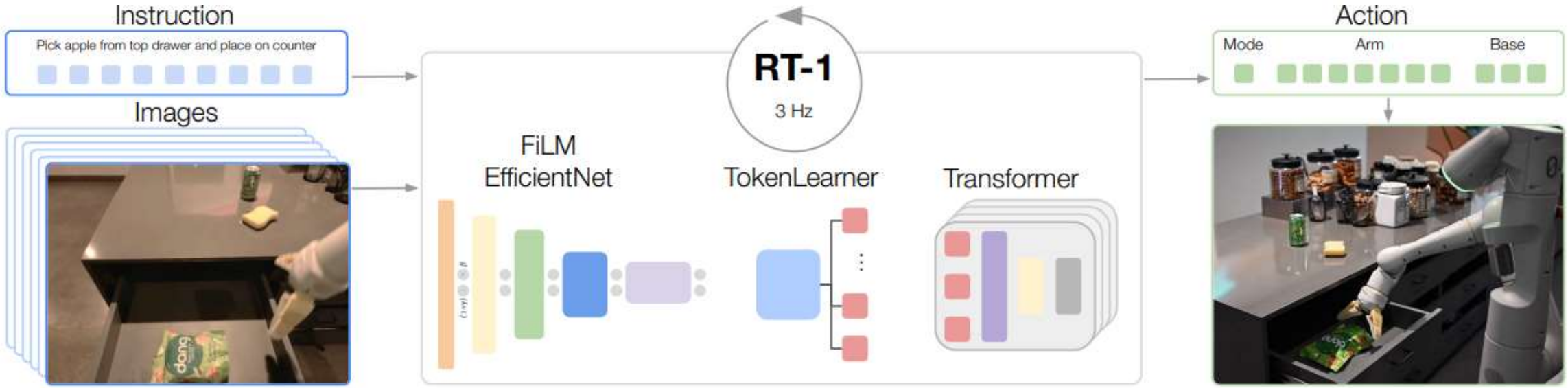


资料来源：《Do As I Can, Not As I Say》谷歌，华鑫证券研究所

RT-1：用于动作控制的端到端模型

- 原理：RT-1模型输入图片以及自然语言指令，通过基于image net（图像分类数据集）的高效卷积神经网络将其输出成为一系列与图片中任务相关的token,通过特征学习器将其转换成压缩的图像特征（image token），经过Transformer模型解码得到离散的动作指令。
- 亮点：将任务通过Saycan拆分成具体的任务，然后使用RT-1去执行。可以执行700个现实中文字指令，并且泛用到新的任务中（可以在三个未见过的厨房执行任务）。可以接受图片作为输入。训练了宝贵的数据集供使用，使用13个机器人历经17个月收集了超过13万个轨迹。端到端的控制模型。
- 不足：对新任务的泛化实际上是以前见过的案例，只能接受出现过的指令。本质上是模仿学习，无法超越数据集的遥操作。严格意义上不是“大模型”，无法从互联网规模（internet-scale）数据中受益。

图表 5：RT-1可以接受图片以及自然语言，输出运动指令

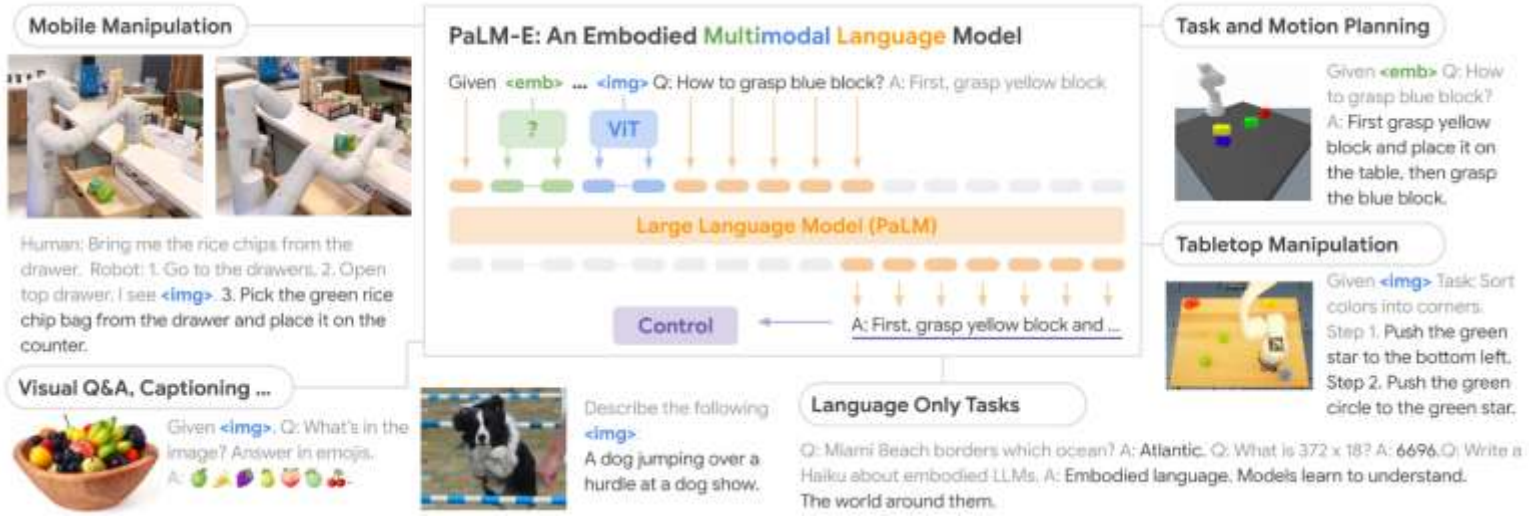


资料来源：《RT-1: Robotics Transformer for Real-World Control at Scale》谷歌，华鑫证券研究所

PaLM-E：多模态视觉语言具身大模型（VLM）

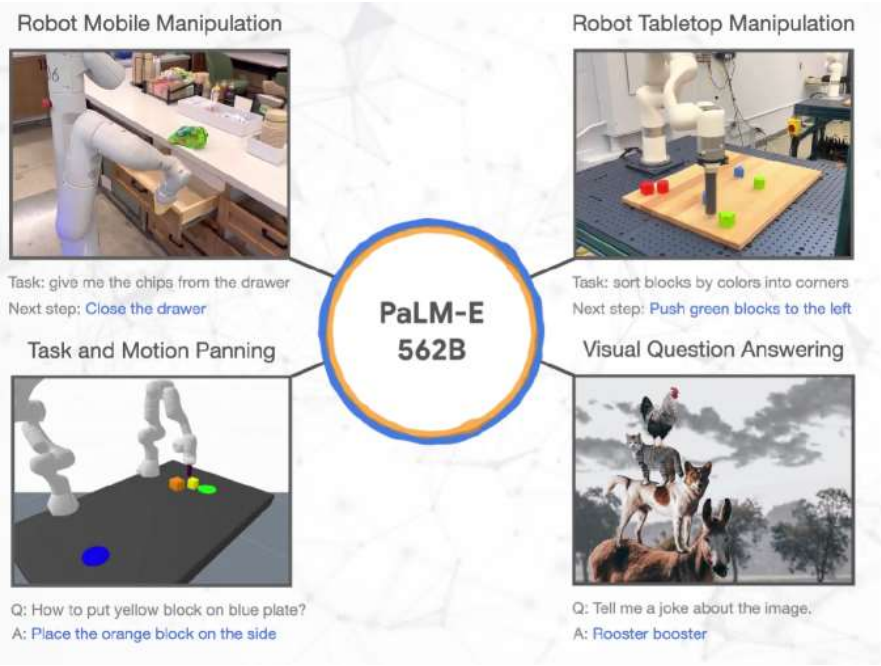
- 原理：由谷歌大语言模型PaLM与拥有220亿个参数的最大视觉模型ViT-22B结合而成，输入连续的视觉、状态、文字之后，在已经预训练的大语言模型PaLM基础上进行端到端训练，用于多个具体任务，包括顺序机器人操作规划、视觉问题解答和图像视频字幕描述。最终输出文本形式的高级任务指令（可以是问题的答案，也可以是PaLM-E以文本形式生成的一系列决策，这些决策应由机器人执行）。
- 亮点：让机器人能够接收持续的多模态的输入（包括文本，图片，状态以及其他传感器模态），连续信息以类似于语言标记的方式注入到语言模型中，并具有一定的推理能力。参数量级有明显提升，5620亿的参数模型。
- 不足：本质为大语言模型，对于动作的完成和指导较弱。只解决机器人的高级别指令，没有更基础层级的具体运动控制相关指令

图表 6：PaLM-E可以接受多模态的输入



资料来源：《PaLM-E: An Embodied Multimodal Language Model》谷歌，华鑫证券研究所

图表 7：PaLM-E参数量高达562B



资料来源：《PaLM-E: An Embodied Multimodal Language Model》谷歌，华鑫证券研究所

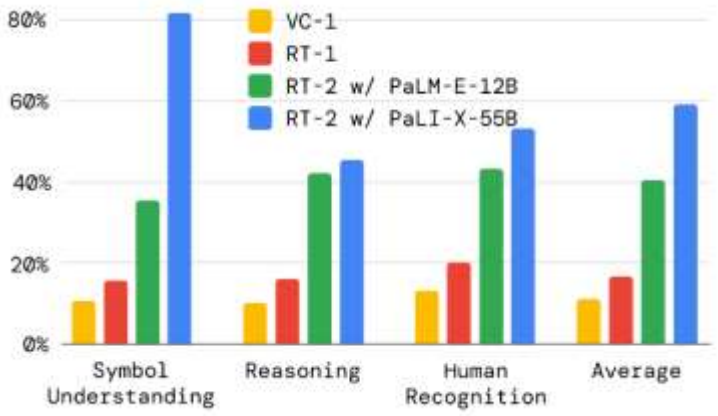
RT-2：控制机器人的视觉 - 语言 - 动作（VLA）大模型

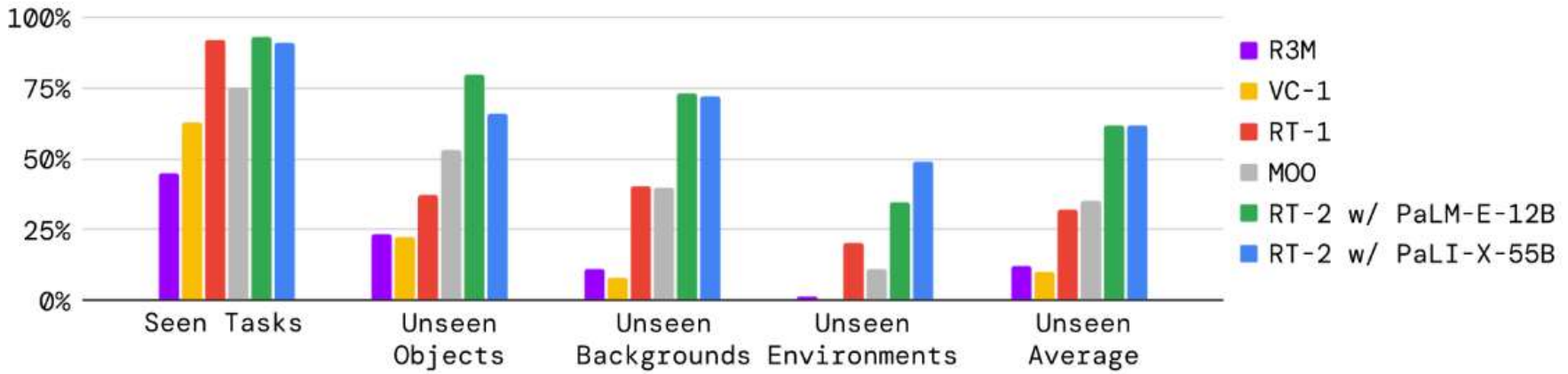
- 原理：机器人数据仍然稀缺的背景下，收集到海量机器人数据难度太大。因此谷歌RT-2抛弃了RT-1从头训练Transformer模型的方式，而是直接采用已有的视觉语言模型（VLM）作为主模型，再使用更适合机器人任务的方法对其进行微调（结合RT-1的视觉/语言/机器人动作数据集与互联网级别数据共同微调co-fine-tuning），最终输出机器人行为字符串。

- 在这种训练方式下，机器人模型拥有一个已经预训练好的VLM模型，可以理解成一个互联网数据级别的常识系统，能够识别物体、了解物体。而在后续的微调阶段，再加入机器人实际抓取物体的数据集，

- 效果：在符号理解，推理和人类识别三项考核中，RT-2的正确率是对照组（RT-1/VC-1）的三倍。而在泛化性上，没见过的物体，没见过的背景，没见过的环境等方面，RT-2相比对照组有一倍的提升。

图表 8：RT-2在见过的任务和未见过的任务表现均明显优于RT-1

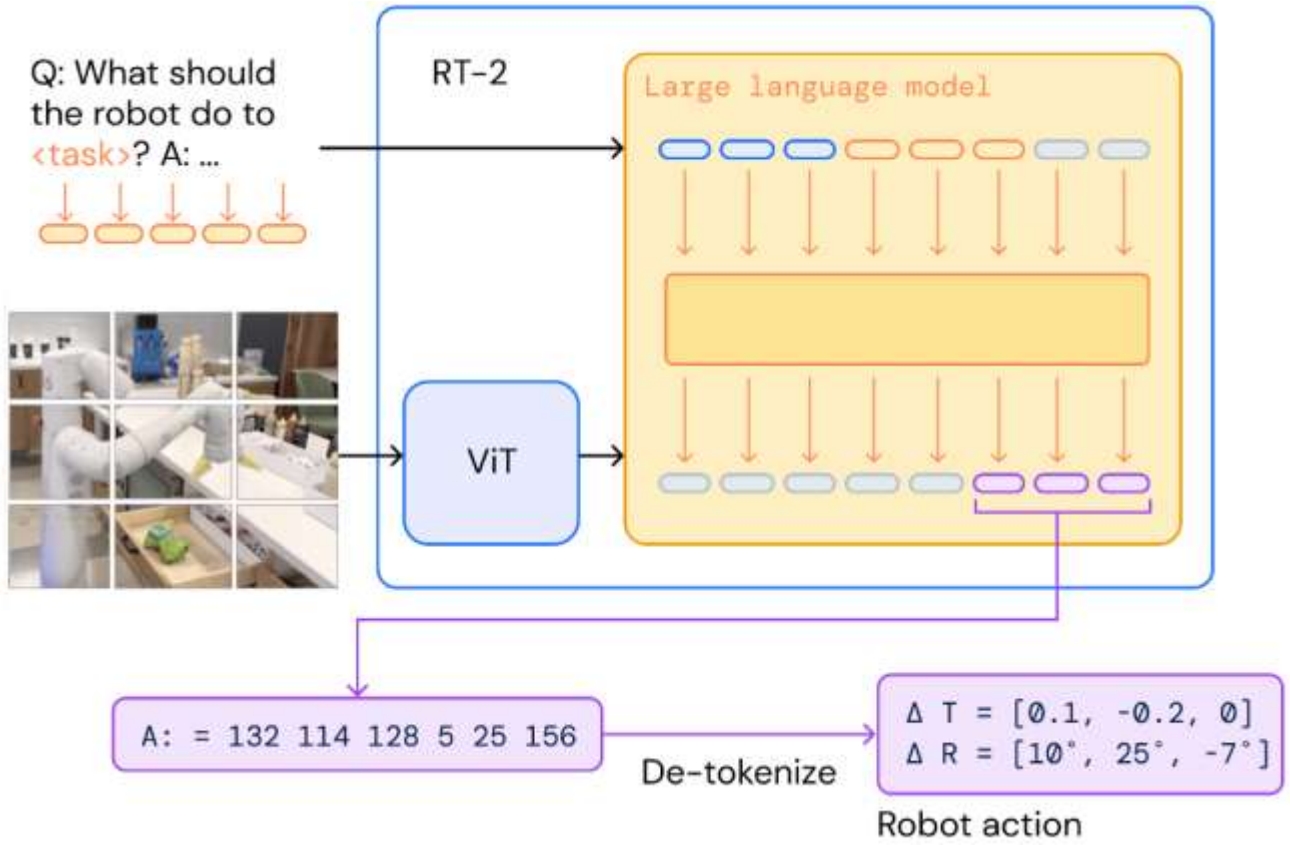




资料来源：《RT-2: Vision-Language-Action Models》谷歌，华鑫证券研究所

RT-2：控制机器人的视觉 - 语言 - 动作（VLA）大模型

图表 9：RT-2本质上是ViT与LLM的结合，输出机器人的坐标作为动作



- 亮点：包含chain of thought的第一次涌现。直接生成较为具体的运动人运动指令。既能够从互联网规模数据中学习（RT-1做不到），又能够输出机器人所需的具体的动作指令。（SayCan、PaLM-E做不到）。相较于SayCan与RT-1的分拆执行的双层模型架构，RT-2在训练模型时候就学习视觉、语言、机器人行为，直接产生动作输出。

- 不足：场景仍然局限，主要为桌面任务。虽然RT-2对于物体和位置的认知拥有了互联网级别的数据训练，可以去拓展新的任务，但是具体动作微调较为依赖RT-1的数据集，而此类数据仍然较为昂贵。本质上大模型主要还是体现在VLM相关方面，也就是在语言和视觉概念，在物理控制层面没有办法获得更强的能力。后期改善将依赖于视频学习的方法。最后，目前运行VLA模型的成本仍然太高，后续希望能够有更多新的底层架构VLM模型出炉。（目前论文主要是PaLM-E和PaLI-X）

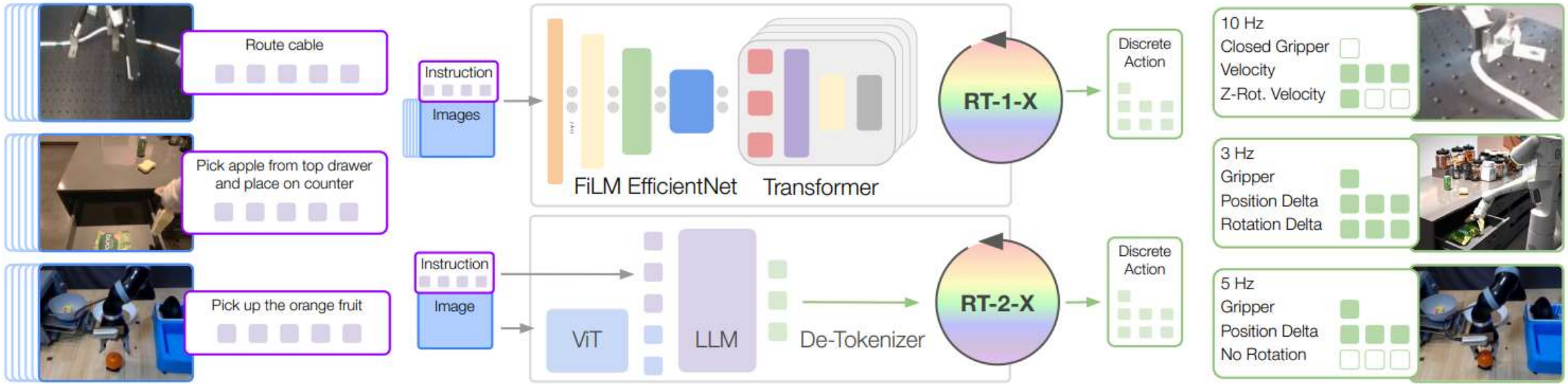
资料来源：《RT-2: Vision-Language-Action Models》谷歌，华鑫证券研究所

RT-X：具身智能大数据集Open X加持的RT-2与RT-1

- 背景：希望能够开发一个通用X-robot，可以高效地适应到新的**机器人、任务、环境**。

- 原理：创造了新能的具身智能大数据集Open X：汇集了来自21个机构的22个不同机器人的数据，包含527个技能和160266个任务。并用此数据集训练前述的机器人模型RT-1和RT-2得到新的模型RT-1-X与RT-2-X。

图表 10：RT-1与RT-2在数据集的帮助下训练成RT-1-X与RT-2-X

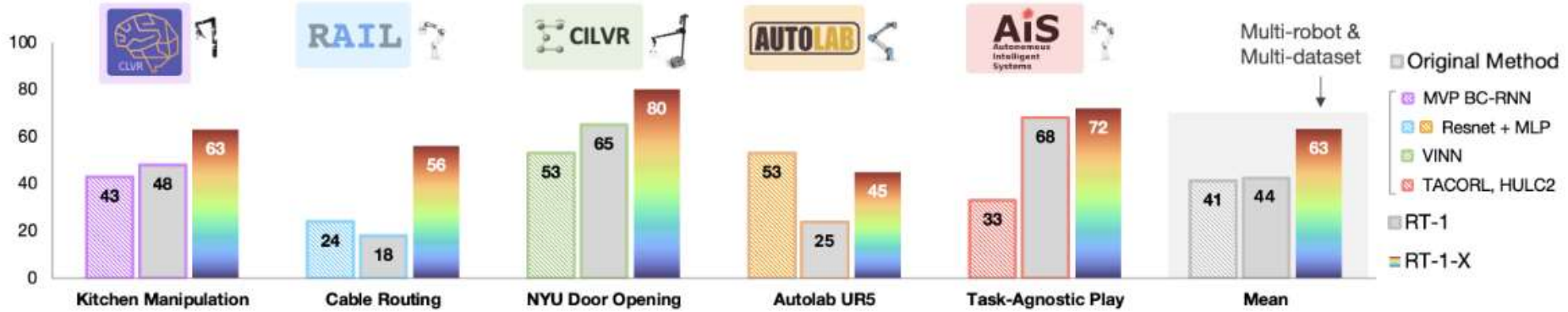


资料来源：《Open X-Embodiment: Robotic Learning Datasets and RT-X Models》谷歌，华鑫证券研究所

RT-X：具身智能大数据集Open X加持的RT-2与RT-1

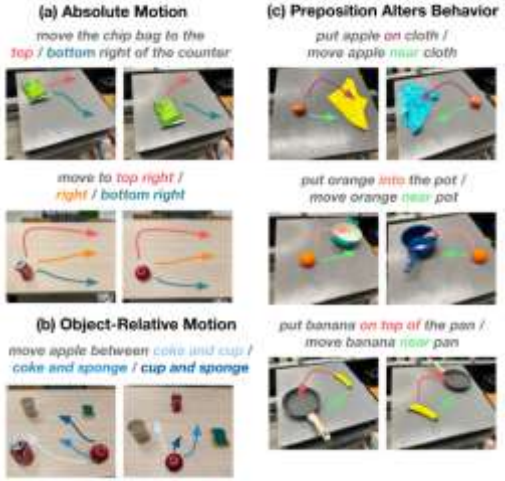
- 效果：RT-1-X模型较原有RT-1或数据集原始模型的成功率有50%的提高，值得注意的是RT-1-X与RT-1的架构是相同的，因此性能的提高完全是依靠数据训练的提升。

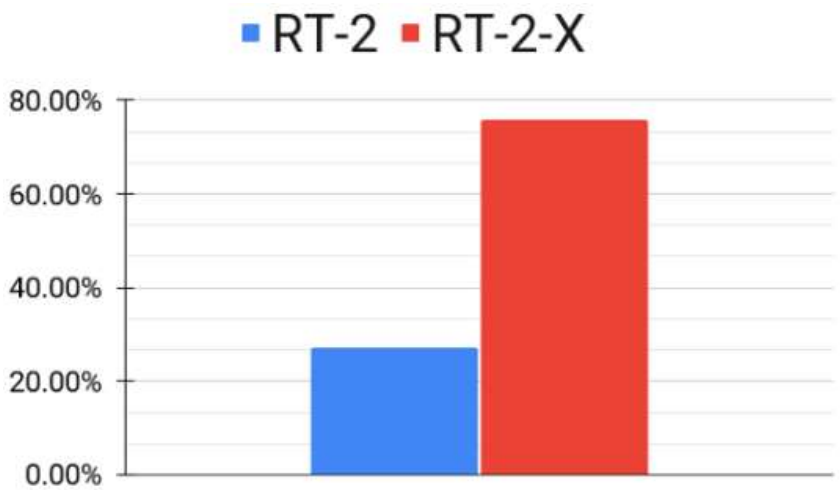
图表 11：RT-1-X在多项任务中的表现明显强于RT-1



- 效果：RT-2-X模型能够展现RT-2此前所不具备的技能，比如对相对和绝对位置的认知。在涌现能力上RT-2-X也是RT-2的三倍。

图表 12：RT-2-X模型展现出的涌现能力很强



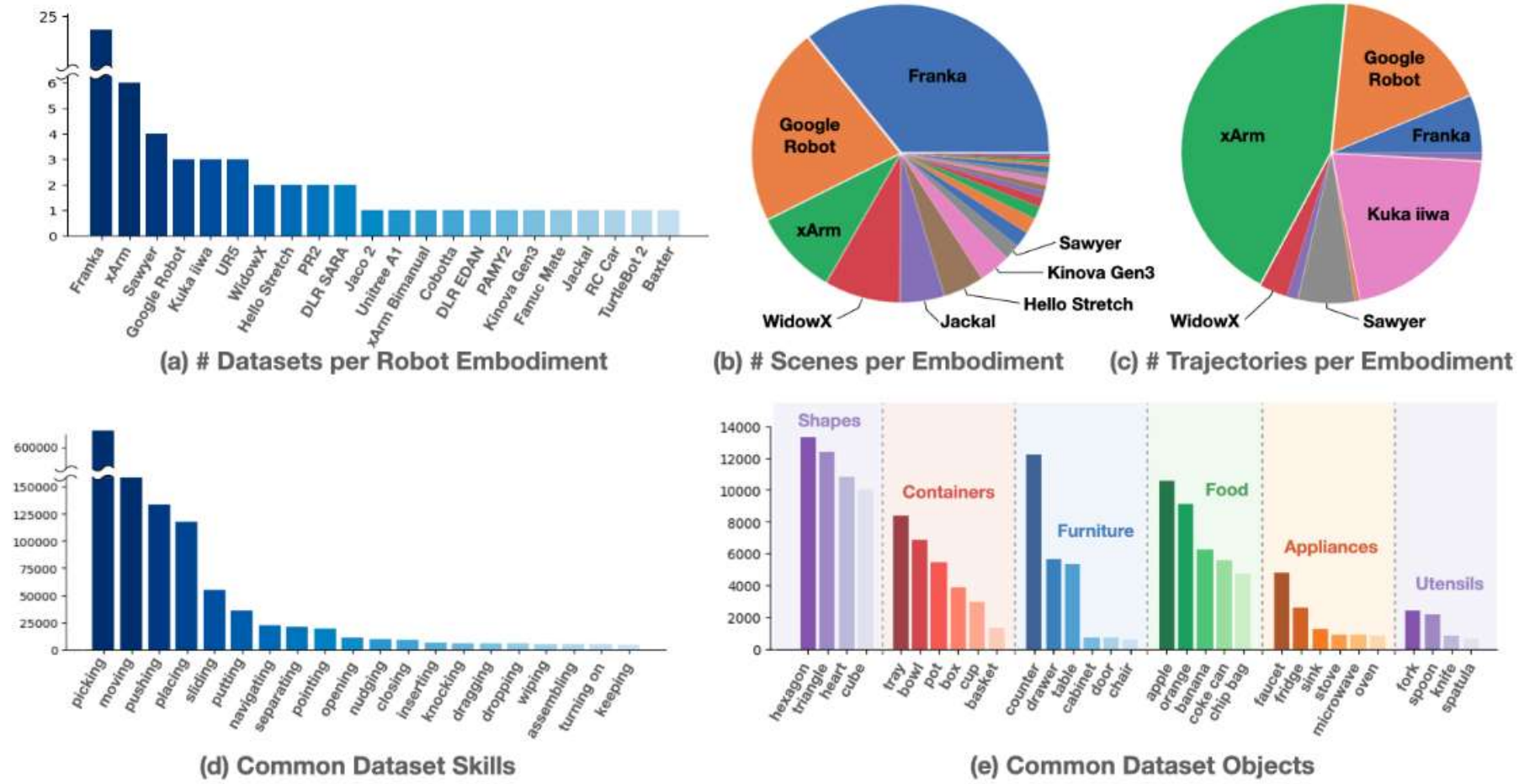


资料来源：《Open X-Embodiment: Robotic Learning Datasets and RT-X Models》谷歌，华鑫证券研究所

RT-X：具身智能大数据集Open X加持的RT-2与RT-1

- RT-X最大的意义在于创造了一个仍然在持续增长的**共享与开源**的数据集（因为数据量在持续增长，RT-1-X与RT-2-X在训练的时候并没有用到全部数据）。机器人数据不像大语言模型中的图片和文字那样容易获得，能够将多个机器人数据开源共享，覆盖足够多的场景与任务。这一点与深度学习中的ImageNet相似。

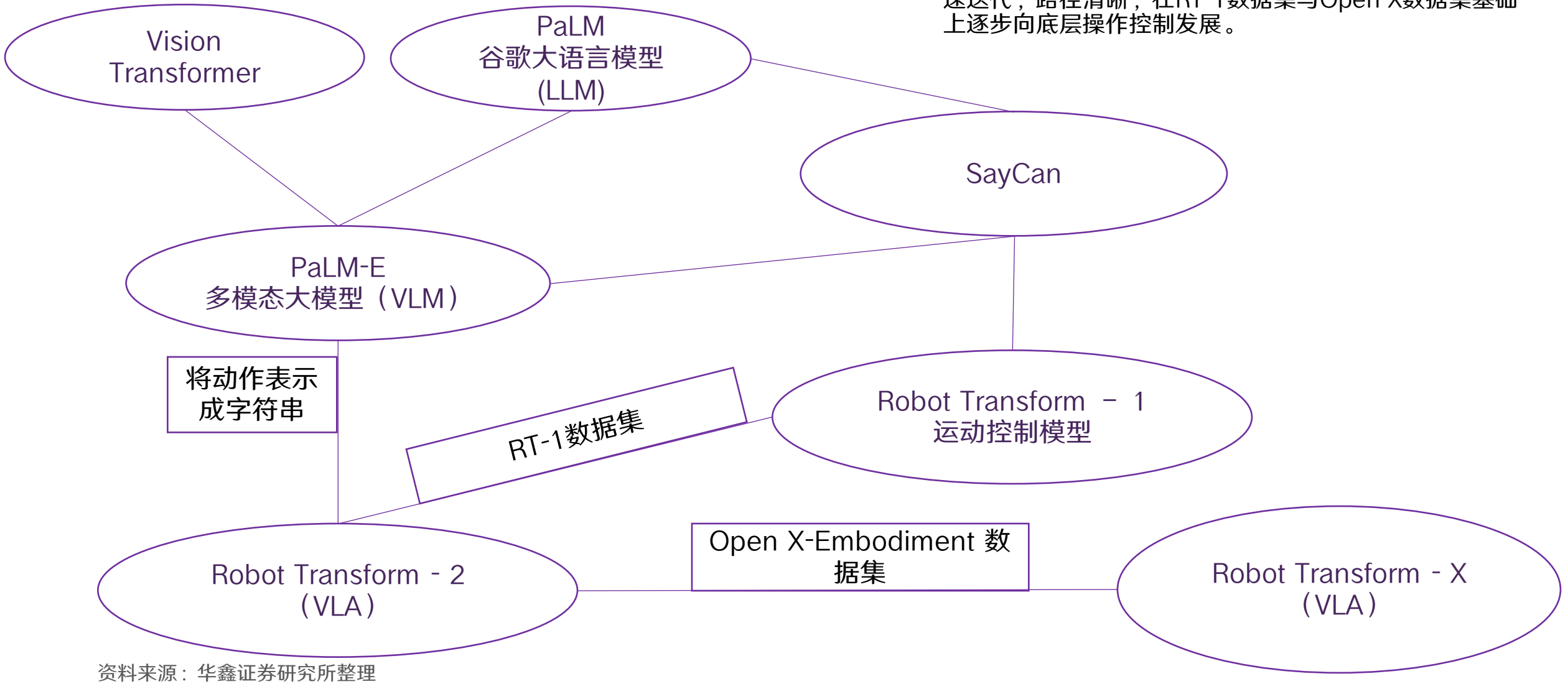
图表 13：Open X融合了目前最全的机器人场景与任务数据



资料来源：《Open X-Embodiment: Robotic Learning Datasets and RT-X Models》谷歌，华鑫证券研究所

谷歌：软件领军者，步步为营，模型高速迭代

图表 14：Open X融合了目前最全的机器人场景与任务数据



资料来源：华鑫证券研究所整理

- 谷歌的机器人大模型由大语言模型演化而来，持续高速迭代，路径清晰，在RT-1数据集与Open X数据集基础上逐步向底层操作控制发展。

03 目前机器人大模型产业化发展的问题与展望

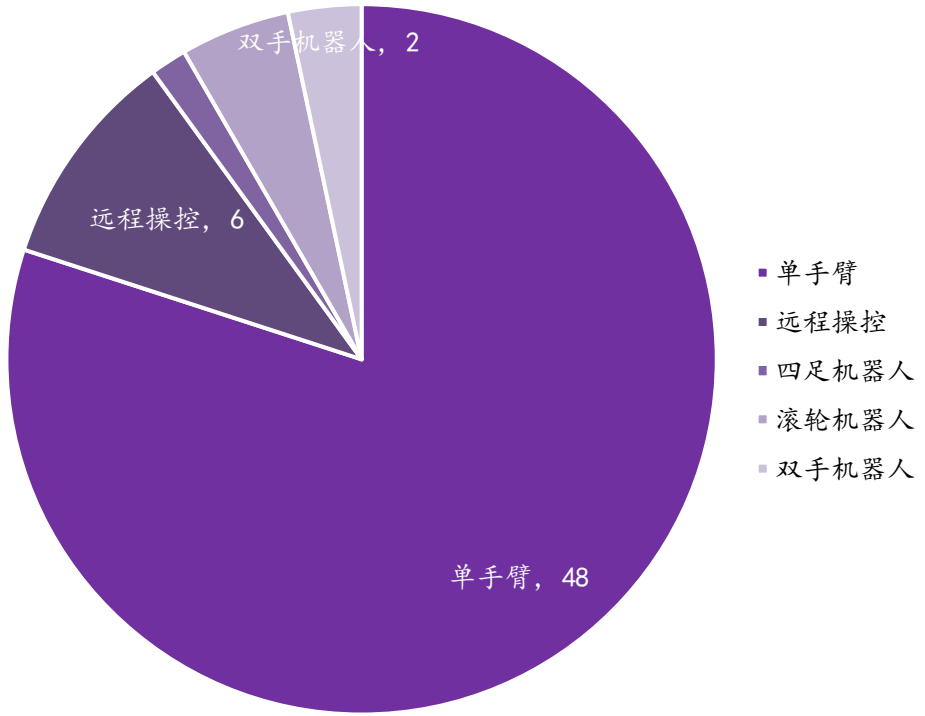
研究创造价值

全都是手臂和抓取——机器人大模型仍然处于非常早期阶段

可以看到目前大部分机器人大模型仍然以单机械臂抓取为主，从谷歌最新的Open X数据来看，单手臂的机器人形式仍然占据了绝大多数。显然光机械臂无法满足大家对于人形机器人泛用性的需求。不光是谷歌，包括斯坦福李飞飞团队的VoxPoser等也都是停留在物品抓取的阶段，距离常规理解的操作（从拧螺丝钉到组装宜家家具）还有较大差距。

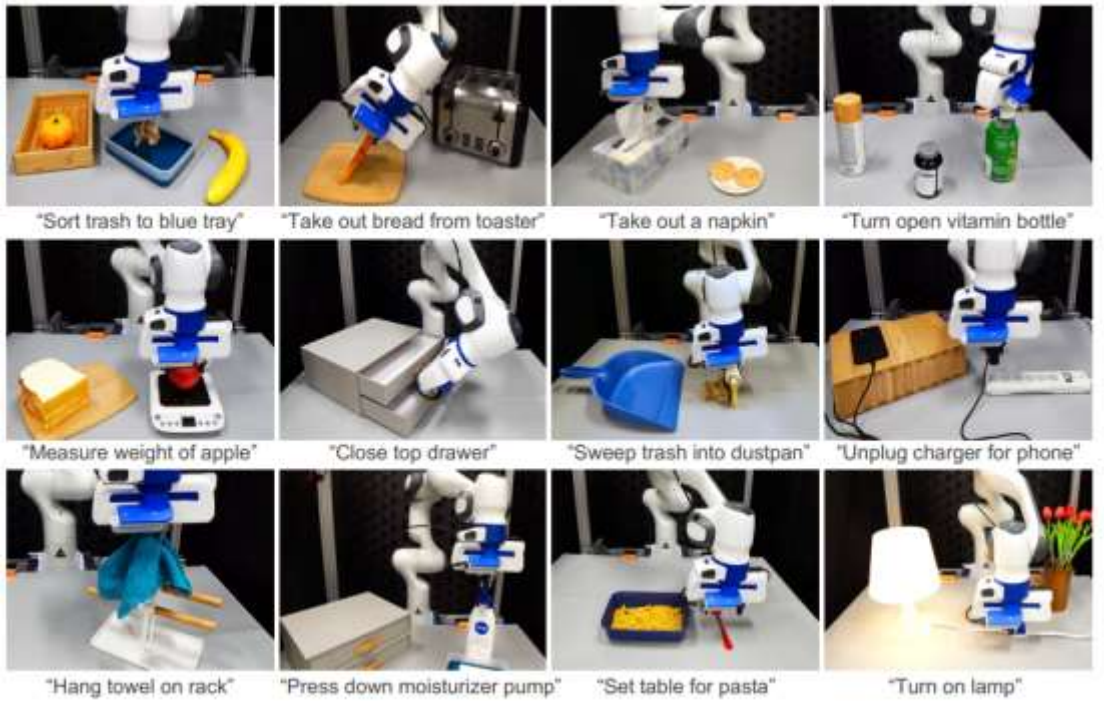
对比手机产业发展进程，如果以IPHONE4出炉代表着智能机产业化的标志，那目前机器人行业仍然处在类似功能机的阶段——可以听歌、可以发短信，但是无法成为一个互联网的载体。

图表 15：Open X的数据集中，大部分仍然是单手臂的数据



资料来源：《Open X-Embodiment: Robotic Learning Datasets and RT-X Models》谷歌，华鑫证券研究所

图表 16：李飞飞团队的模型VoxPoser同样以机械臂抓取为主



资料来源：《 VoxPoser: Composable 3D Value Maps for Robotic Manipulation with Language Models》，华鑫证券研究所

模型尚未定型、精细化操作不足——机器人大模型仍然处于非常早期阶段

- 从2022年4月谷歌推出SayCan模型以来，两年不到的时间，用于机器人的大模型已经经历过多次的模型迭代，但直至现在仍然尚未有最终的模型定型。大模型仍然层出不穷，微软谷歌斯坦福等均有论文持续出炉，彼此之间甚至连任务定义还都不一样。例如在将做任务规划的大脑与下层运动控制的方案之间的通信渠道的方案中，谷歌RT-2采取的是action tokenization, VoxPoser采取的是value map, SayCan采取的是value function, 这些技术路线还没开始收敛。我们认为目前机器人大模型的技术路线还远未开始收敛。随着后续语言类大模型的持续发展，机器人相应的底层架构同样有变化的可能。

- 其次大模型目前展现出来的精细化控制能力较弱，很难做偏底层的运动控制。这主要是因为目前大模型直接输出离散的tokenization的位置和状态，未考虑连续运动的轨迹平顺性、时间最优、功耗等额外因素。另一方面，对于精细化程度要求较高的任务，可以通过Model-based方法处理，并且可以看到在工业机器人领域，很多控制精度要求较高的行业已经有满足生产要求的工业机器人，因此大模型如果希望能够在工业中得到应用，精细化操作道路仍然较远。

图表 17：RT-2目前最终也只是输出离散的动作token，包括手臂的位置坐标以及底座的位置坐标等

资料来源：《RT-2: Vision-Language-Action Models》谷歌，华鑫证券研究所

产业启示：大模型需要到做到哪一步？

- 目前市场上公开的大模型大多以软件企业/学校为主导，无论是谷歌还是微软，都没有实际量产落地人形机器人，这种AI类型公司的模型特点主要是期望能够打造一条端到端的，包含感知-决策-规划-控制全部流程的大模型。以谷歌为例，从SayCan初次引入大模型用于做任务理解和拆分，到RT-1使用传统神经网络的方法来执行SayCan的任务，再到RT-2将VLM大模型与RT-1的机器人执行数据集一起微调训练，谷歌的意图不仅仅是用大模型做基础的认知和理解，而是将底层的动作也囊括其中。
- 而另一边，传统的机器人公司也在加入语言大模型方便理解任务，试图从底层控制入手，使用大模型进行一些常识、决策与推理，类似SayCan，在这里大模型只扮演一个前期任务拆解的角色，而不会涉及后续运动方面的控制。
- 从目前进展来看，显然后者产业化进度相对而言更快，不少现有的机器人公司已经开始接入AI大模型系统，如智元科技的远征A1，采用了语言任务模型 WorkGPT，结合了 LLM 和 VLM 等 AI 技术，能够为机器人提供自主感知环境、理解任务、编排动作的能力。我们认为短期之内使用大模型来做通识理解、抽象思维、多级推理，配合model-base的小模型是能够较快落地的商业方案。中长期看，随着人形机器人逐步落地，后续积累了足够多的机器人操作数据与3D环境数据，再使用大模型+微调（类似RT-2）进行全部流程的训练，有望能够涌现出更好的表现。

图表 18：大模型对于运动控制的把握目前还停留在动作级

	控制层级	任务级	动作级	基元级	伺服级
目前进度	传统/工业机器人		■	■	■
	AI/大模型目前进展	■	■		
未来发展	可能性1	■	■	■	■
	可能性2	■	■	■	■
	可能性3	■	■	■	■

资料来源：《Robotics modelling, planning and control》，华鑫证券研究

图表 19：智元机器人的EI Brian框架同样有类似的分类

资料来源：智元机器人，华鑫证券研究

以通识大模型+底层细分场景微调的框架为最终版本的机器人大模型来看，我们认为产业未来在算力、数据要素、软件模型三个维度具有投资机会：

1、算力：机器人需要快速与环境交互，同时大模型本身要计算和存储空间。二者叠加之下机器人所需的参数和算力比自动驾驶以及大语言模型都要更大，因此对于算力的需求将在后续逐步有所体现。看好算力相关产业链发展。

2、数据：语言类模型的数据可以较为简单地从网上落得，而机器人的数据不同，机器人需要通过多种传感器感知环境状态，然后执行实际动作来完成任务，一方面需要3D环境数据，另一方面需要的是主动数据，即需要机器人主动实际执行才能得到的数据，此类数据量极度稀缺，以谷歌的RT-1的数据集为例，为了获得13万个执行轨迹，需要使用人工遥操作演示示教的方式进行为期17个月的收集。后续虽然有可能可以通过模仿学习来进行学习，降低数据成本，但是我们预计最终的底层操控的实现必然需要通过机器人实际操作才行。（类似学习自行车的过程，虽然可以通过视频以及网络上的一些要领学习大概，但是真的掌握骑自行车必然需要实际操作过才能掌握）因此看好具有先发优势，已经发布机器人产品，掌握机器人相关数据的公司。

3、细分场景的模型：未来大模型在机器人的应用，或许是通过底层的通识大模型+细分场景模型微调获得，其中底层架构的通识大模型有望参考类似手机安卓的模式由头部的AI企业开源，而细分场景的模型（同时也包括所需的数据）才是未来大部分企业可以竞争的市场。在这个赛道中，数据仍然是模型的基础，因此看好同时具有较强机器人本体硬件与软件能力的公司。

机器人下游发展不及预期

算力与算法模型更新迭代不及预期

行业竞争加剧风险

新材料、电力设备组介绍

傅鸿浩：所长助理、碳中和组长，电力设备首席分析师，中国科学院工学硕士，央企战略与6年新能源研究经验。

杜飞：碳中和组成员，中山大学理学学士，香港中文大学理学硕士，3年大宗商品研究经验，负责有色及新材料研究工作。

臧天律：金融工程硕士，CFA、FRM持证人。上海交通大学金融本科，4年金融行业研究经验，覆盖光伏、储能领域。

证券分析师承诺

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注册为证券分析师，以勤勉的职业态度，独立、客观地出具本报告。本报告清晰准确地反映了本人的研究观点。本人不曾因，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。

免责条款

华鑫证券有限责任公司（以下简称“华鑫证券”）具有中国证监会核准的证券投资咨询业务资格。本报告由华鑫证券制作，仅供华鑫证券的客户使用。本公司不会因接收人收到本报告而视其为客户。

本报告中的信息均来源于公开资料，华鑫证券研究部门及相关研究人员力求准确可靠，但对这些信息的准确性及完整性不作任何保证。我们已力求报告内容客观、公正，但报告中的信息与所表达的观点不构成所述证券买卖的出价或询价的依据，该等信息、意见并未考虑到获取本报告人员的具体投资目的、财务状况以及特定需求，在任何时候均不构成对任何人的个人推荐。投资者应当对本报告中的信息和意见进行独立评估，并应同时结合各自的投资目的、财务状况和特定需求，必要时就财务、法律、商业、税收等方面咨询专业顾问的意见。对依据或者使用本报告所造成的一切后果，华鑫证券及/或其关联人员均不承担任何法律责任。本公司或关联机构可能会持有报告中所提到的公司所发行的证券头寸并进行交易，还可能为这些公司提供或争取提供投资银行、财务顾问或者金融产品等服务。本公司在知晓范围内依法合规地履行披露。

本报告中的资料、意见、预测均只反映报告初次发布时的判断，可能会随时调整。该等意见、评估及预测无需通知即可随时更改。在不同时期，华鑫证券可能会发出与本报告所载意见、评估及预测不一致的研究报告。华鑫证券没有将此意见及建议向报告所有接收者进行更新的义务。

证券投资评级说明

股票投资评级说明：

	投资建议	预测个股相对同期证券市场代表性指数涨幅
1	买入	>20%
2	增持	10%—20%
3	中性	-10%—10%
4	卖出	<-10%

行业投资评级说明：

	投资建议	行业指数相对同期证券市场代表性指数涨幅
1	推荐	>10%
2	中性	-10%—10%
3	回避	<-10%

以报告日后的12个月内，预测个股或行业指数相对于相关证券市场主要指数的涨跌幅为标准。

相关证券市场代表性指数说明：A股市场以沪深300指数为基准；新三板市场以三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的）为基准；香港市场以恒生指数为基准；美国市场以道琼斯指数为基准。

华鑫证券
CHINA FORTUNE SECURITIES

研究创造价值

报告编号：231119183554