证 券 研 究 报 告

从发散到收敛，具身模型和数据体系的产业演进

机器人系列深度之39

证券分析师：胡书捷 A0230524070007
王珂 A0230521120002
联系人：胡书捷 A0230524070007
2026.8.6

SWS RESEARCH

投资要点

- **机器人产业发展的四大关键要素为模型、数据、本体和场景。**其中，模型是机器人大脑，负责推理决策；数据是模型训练的燃料，没有高质量数据，模型无法进行有效训练；本体是机器人物理载体，执行能力的物理上限，所有智能指令最后依靠本体执行；场景是最终商业化落地的重点，和机器人的需求来源，也决定了机器人的任务边界，机器人必须有场景才能产生价值。我们认为，**在当前阶段，模型和数据是行业最大的瓶颈。**市场对该环节的关注和研究较少。
- **模型：主流玩家的模型持续迭代，模型架构尚未收敛。**主流玩家的模型持续迭代，从PI的π0到π0.7，谷歌的RT1到Gemini Robotics2，在跨本体泛化、操作流畅度、数据生成瓶颈方面持续突破；模型架构尚未收敛，但相似度较高，多为VLA、扩散模型、世界模型等的融合架构，大脑推理和小脑动作模型分层。参与者呈现百花齐放态势：1）科技大厂：谷歌、字节；2）具身模型&本体公司：PI、Figure、智元、星海图等；3）大语言模型厂商：OpenAI等。
- **数据：模型训练的核心燃料，关键在于数据数量、质量和scale-up。**相比大语言模型有海量的互联网数据训练，具身智能天然缺乏机器人轨迹数据，因此数据成为模型迭代的关键瓶颈。当前数据来源包含互联网数据、仿真合成数据、第一人称视角数据、遥操作数据等。当前，不同公司押注不同的数据路线，普遍采用多种数据类型、开源+内部数据集融合的方式。在模型训练阶段，数据生产商及采集设备商迎来投资机遇。
- **投资提示：**人形机器人进入具身智能驱动的新阶段。本体零部件是硬件基础，量产落地带动减速器、伺服、传感器需求释放；具身大模型作为机器人智能核心，算法与仿真平台构筑软件壁垒；训练数据是模型迭代核心约束，真实示教采集＋第一视角采集+仿真合成数据等多路线并行，数据供给端稀缺性持续提升。中长期建议沿三条主线布局：①人形机器人核心零部件厂商；②具身模型和本体厂商；③机器人数据采集硬件与数据服务相关标的。
- **风险提示：**模型技术迭代不及预期风险、数据供给和质量瓶颈风险、商业化落地风险。

机器人产业发展的四大关键要素

- **机器人产业发展的四大关键要素为模型、数据、本体和场景。**
- **1）模型：**机器人大脑，负责推理决策，决定了机器人智能化水平的上限；
- **2）数据：**模型训练的燃料，没有高质量数据，模型无法进行有效训练；
- **3）本体：**机器人物理载体，执行能力的物理上限，所有智能指令最后依靠本体执行；
- **4）场景：**商业化落地的重点，和机器人的需求来源，也决定了机器人的任务边界，机器人必须有场景才能产生价值。

图1：机器人产业发展的四大关键要素

模型

VLA模型、世界模型、强化学习

参与者：OpenAI、Google Deepmind、字节、阿里千问、PI、Field AI、Generalist AI、Skild AI等

本体

人形、轮式、四足、机械臂等

参与者：特斯拉、宇树、智元、云深处等

数据

遥操作数据、人类视角数据、仿真合成数据、互联网数据

参与者：Scale AI、XDOF、各地数采厂、光轮智能等

场景

工业、商用、酒店、家务、通用

参与者：科沃斯、石头科技、埃斯顿、云迹、极智嘉等

注：部分公司参与多个环节，我们仅列出部分代表性公司

主要内容

1.1 机器人模型迭代：从运动到认知，再到多模态和世界模型

阶段 1：经典控制模型时代（1960–2010），基于数学建模

- **过程：**1960s工业机器人兴起，建立拉格朗日机器人动力学模型；针对双足机器人的控制算法迭代，如PID（单关节独立控制）、自适应控制、鲁棒控制等。**典型代表：**传统工业机械臂、早期 ASIMO 初代。
- **局限性：**需要精确 CAD 参数、重量、摩擦系数；负载变化、柔性、间隙没有感知，无法自主理解环境；只能复现预设轨迹，四足、人形这类强耦合、非线性机器人很难精确建模。

阶段二：数据驱动模型萌芽（2010–2018），机器学习补充动力学，感知接入

- **过程：**机器学习推动数据驱动模式出现，包括：高斯过程 GP、神经网络补偿动力学；强化学习初步用于机器人；模型预测控制MPC开始用于人形和足式机器人。**典型代表：**MIT Cheetah四足机器人的简化刚体模型+MPC；DeepMind 早期机器人抓取。
- **局限性：**感知服务于运动控制，而非认知模型；感知能力不足。

阶段三：大语言模型诞生 —— 机器人认知模型启蒙（2018–2022）

- **过程：**Transformer、LLM 出现，机器人开始拥有语义理解能力，但大小脑分层：①运动模型：以动力学+MPC/RL 为主，变化不大；②认知模型：引入外部大模型做任务规划。例如，LLM 作为“规划大脑”接收自然语言指令、拆解子任务，运动控制器作为“小脑”执行动作；**典型代表：**Google PaLM-SayCan（2022）
- **局限性：**LLM只懂文本，没有空间几何感知、没有具身常识，无法理解物体位置、尺寸、可操作性；语言和机器人视觉、运动空间割裂。

1.1 机器人模型迭代：从运动到认知，再到多模态和世界模型

阶段四：多模态具身基础模型时代（2022年至今，当前主流方向），视觉+语言+动作统一模型

- **过程：**不再分开训练语言模型、视觉模型、机器人控制模型；构建统一多模态具身基础模型，输入图像、文本、动作；输出机器人动作序列。例如RT-1 / RT-2、π0、OpenAI RoboCat等，是当前最主流的模型架构。
- **局限性：**泛化能力较差，仿真到现实迁移难度大，物理幻觉、无法应对不可预测性；世界模型派别认为LLM有根本性局限，难以实现能理解物理世界的智能。无法理解空间、物理规则、因果、持续感知与预测现实场景。

阶段五：世界模型（下一代，演进中）

- **过程：**模型的核心目标为，机器人不只输出动作，还能预测执行动作之后环境会发生什么，实现闭环逻辑：感知观测 → 世界模型预测未来状态 → 规划最优动作 → 执行 → 新观测，具备自我进化和自我学习能力。现在还在早期探索阶段。

表1：机器人模型的迭代过程

时代	模型核心	建模方式	泛化能力	适用场景
经典控制 (1960-2010)	刚体动力学解析模型	第一性原理推导	极低，固定任务	传统工业机械臂
机器学习 (2010-2018)	动力学 + 浅层神经网络 / RL	原理模型 + 数据残差拟合	中等，单一场景	四足机器人、简单抓取
LLM 规划+控制器 (2018-2022)	LLM 任务规划 + 下层运动控制	模块化分离架构	中等，语言理解强，空间弱	桌面机械臂服务机器人
多模态具身基础模型 (2022至今)	视觉 - 语言 - 动作统一 Transformer	多模态联合预训练	高，跨任务、跨物体	通用操作人形机器人
世界模型驱动具身智能（下一代）	感知 + 预测世界模型 + 动作策略	交互式闭环预测建模	极高，未知环境自主探索	通用人形机器人

1.2 当前具身模型玩家持续扩容，大模型厂商投入加码

- **当前具身模型玩家持续扩容，百花齐放态势：**1）科技大厂：谷歌、字节、腾讯、阿里等；2）具身模型&本体公司：PI、Figure、智元、星海图等；3）大语言模型：openAI等。模型路线各有侧重，尚未收敛。

表2：机器人模型的主要玩家

公司		模型	发布时间	模型架构
创业型公司	星海图	G05	2026年5月	VLA策略模型
		G0 Plus	2026年1月	
		G0	2025年8月	双系统VLA模型
	智元	WITA-Omni Preview	2026年7月	Thinker–Talker–Actor端到端原生具身交互架构
		GO-2	2026年4月	Genie Operator-2 VLM+动作
		GO-1	2025年3月	Genie Operator-1VLM+动作
	银河通用	GraspVLA	2025年1月	VLA抓取基础模型
		LDA-1B	2026年4月	1B参数 世界-动作统一模型
	千寻智能	Spirit v1.5	2026年1月	通用端到端VLA基座模型，基于Qwen3-VL主干+扩散策略头
		Spirit v1.6	2026年7月	新一代融合世界模型VLA基座
	PI（Physical Intelligence）	π0	2024年10月	VLM+流匹配动作
		π0.5	2025年4月	
		π0.6	2025年11月	
		π0.7	2026年4月	VLM+流匹配动作专用模块，引入更多token
科技大厂	小米	Xiaomi-Robotics-0	2026年2月	47亿参数VLA模型，基于Qwen3-VL，MoT混合架构
		U1	2026年7月	VLM+动作
	阿里	Qwen-VLA	2026年5月	基于Qwen3-VL的VLA架构
		Qwen-Robot Suite	2026年6月	全系列机器人模型套件（VLA、导航、世界模型）
	字节	GR-3	2025年7月	VLM+动作
		GR-RL	2025年10月	基于强化学习的机器人策略模型
	腾讯	HY-Embodied-0.5-X	2026年4月	MoT架构，分为MoT-2B端侧版和MoE-32B复杂推理版
	谷歌DeepMind	RT-2	2023年7月	视觉-语言-动作统一Transformer架构
		RT-1	2022年12月	Transformer-based机器人策略模型
		Gemini Robotics 1.5	2026年5月	Gemini 2.0多模态模型+Motion Transfer层
	英伟达	Isaac GR00T N1	2025年3月	VLM+DiT动作生成架构
		GRT N.	年月	Frozen VLM+Eagle . grounding +FLARE训练目标
		GR00T N1.6	2025年12月	Cosmos-2B VLM主干+32层DiT架构
大模型公司	OpenAI	RoboCat	2024年1月	自监督VLA模型

资料来源：各公司官网，申万宏源研究

1.2.1 Physical Intelligence

- **Physical Intelligence成立于2024年，总部位于旧金山，其创始人团队汇聚全球机器人学和大模型领域的人才，包括来自UC 伯克利大学、Google DeepMind、斯坦福大学、特斯拉等机构的顶尖人才，具备“学术+工程”双重基因。**
  - PI 认为机器人的进化应当复制大语言模型的Scaling Law，致力于开发一个通用的基础模型，输入同一套代码，就可以直接注入到人形机器人、双臂机械臂、甚至带有轮子的移动底盘中。让机器人在不需要任何代码重写的情况下，零样本（Zero-shot）泛化地去折衣服、组装纸箱、甚至清理完全陌生的厨房。
  - **24年发布π0模型，25-26年推出π0.5、π0.6、π0.7模型，连续引入FAST、RECAP、MEM、RL等模块。**进化路径为：**1）架构：**从VLM+流匹配动作专用模块，逐步叠加动作分词、高低层解耦、知识绝缘、强化学习价值头、长短时记忆模块；**2）数据：**从单一机器人演示数据，逐步融入互联网多模态数据、多环境多机型数据、百万级开源轨迹、真实部署产生的强化学习经验数据、长时任务视频数据；**3）训练范式：**从纯模仿学习，走向多任务联合训练、梯度隔离防干扰、离线+在线强化学习、带记忆的推理训练。

表3：Physical Intelligence的模型迭代过程

时间	节点	核心价值
2024.10	**π0**	**首提“VLM+流匹配动作专用模块”架构，输出连续高频率动作**
2025.01	FAST	动作分词技术，训练速度提升5倍
2025.04	π0.5	开放世界泛化，陌生环境直接作业
2025.11	π0.6	开箱即用，无需微调即可落地
2025.11	**π*0.6+RECAP**	**引入强化学习，机器人从经验中自我进化**
2026.03	MEM	长短时记忆加持，支持15分钟长任务
2026.03	RL	引入RL Tokens (RLT)，快速提升最精细、最容易失败的操作阶段
2026.04	π0.7	组合式泛化能力

资料来源：Physical Intelligence公司官网，申万宏源研究

1.2.1 PI 提出VLA+流匹配架构具身模型

- **π0模型：2024年10月推出，架构为VLA+流匹配，解决了传统 VLA 模型只能输出离散 Token、控制频率低、动作不够丝滑的硬伤。实现了自回归语言大模型与连续动作生成网络的有机结合。**

π跨实体机器人数据集：
多种机器人本体（单臂、双臂、带轮子的移动底盘等）；丰富的现实家务与工业数据

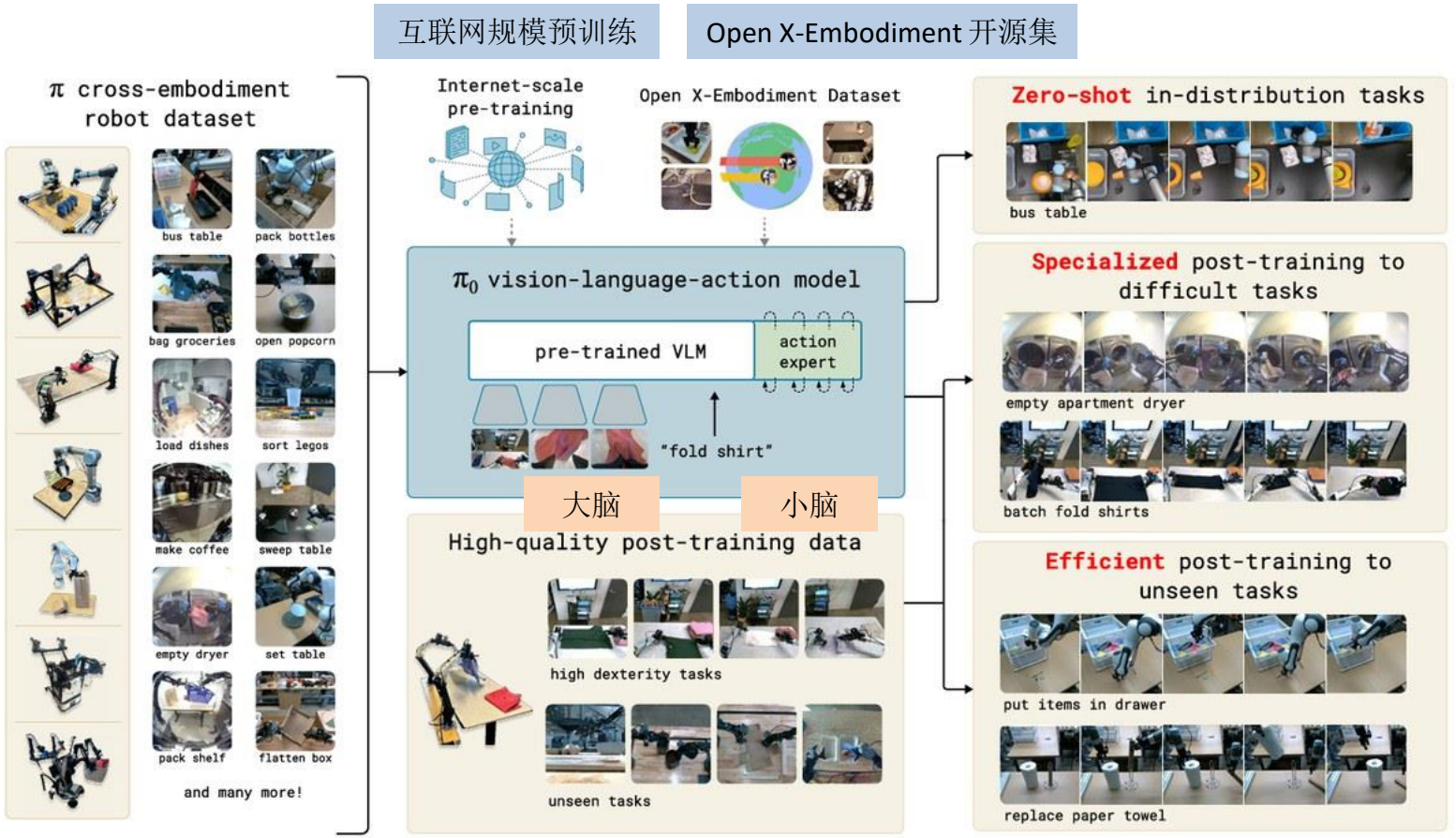


图2： π0模型的框架

①零样本基础能力：经过大规模预训练后，模型在面对与训练集环境相似的场景，不需要任何额外的训练或人类示范，直接就能上机以 Zero-shot（零样本） 的方式高成功率执行。

② 攻克极限高难度任务：面对结构极度复杂、具备重度非刚体特性的极难任务。方法：喂入底部框中的 High-quality post-training data（高质量后训练数据），包括高灵巧度任务（high dexterity tasks）的实机示范轨迹。经过专项微调后，机器人能完美攻克这类传统物理引擎根本无法解算的物理操作。

③ 高效的“未见任务”迁移：面对完全没学过的、全新的泛化任务。方法： 机器人展现出了强大的 Scaling Law（缩放定律）泛化红利。由于基座模型里已经封装了足够的物理常识，研发人员只需要补充极少量的（unseen tasks）实机数据进行“快闪式”微调，机器人就能在极短时间内（通常不足几小时）学会一套全新的复杂技能。

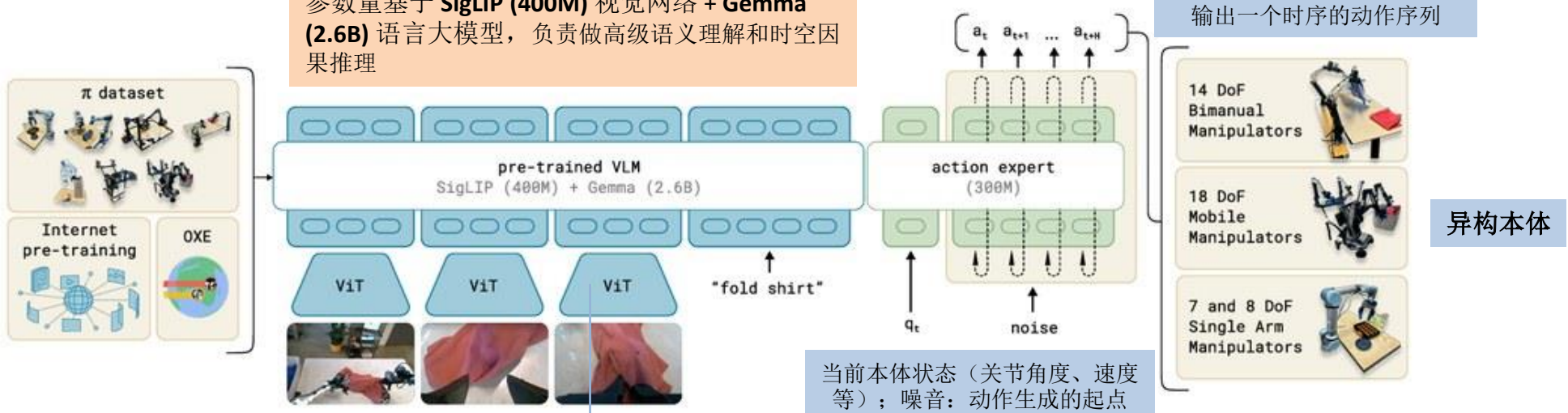
1.2.1 PI 提出VLA+流匹配架构具身模型

- **π0模型：VLA+流匹配架构，对外开源，发布后在行业内具有一定的引领作用。**

预训练混合数据：
π数据集
互联网预训练
OXE开源集

语义通识大脑：Pre-trained VLM
参数量基于 **SigLIP (400M)** 视觉网络 **+ Gemma (2.6B)** 语言大模型，负责做高级语义理解和时空因果推理

流匹配动作小脑：Action Expert
参数量 **300M**，Transformer 架构，输入语义特征，通过流匹配完成动作去噪，输出系列动作



①图像 Token 化（Vision）： 机器人的多视角摄像头,通过 ViT（Vision Transformer）进行特征提取，转化为图像 Token
②语言指令（Language）： 人类的自然语言指令通过标准的 Text Tokenizer 转化为文本 Token

流匹配（Flow Matching,）原理：通过神经网络去学习一个随时间变化的“速度场（Velocity Field）”，利用确定性的常微分方程（ODE），把一个简单的噪声分布直接“推”向复杂的真实数据分布。

图3： π0模型架构

资料来源：《π0: A Vision-Language-Action Flow Model for General Robot Control》，申万宏源研究

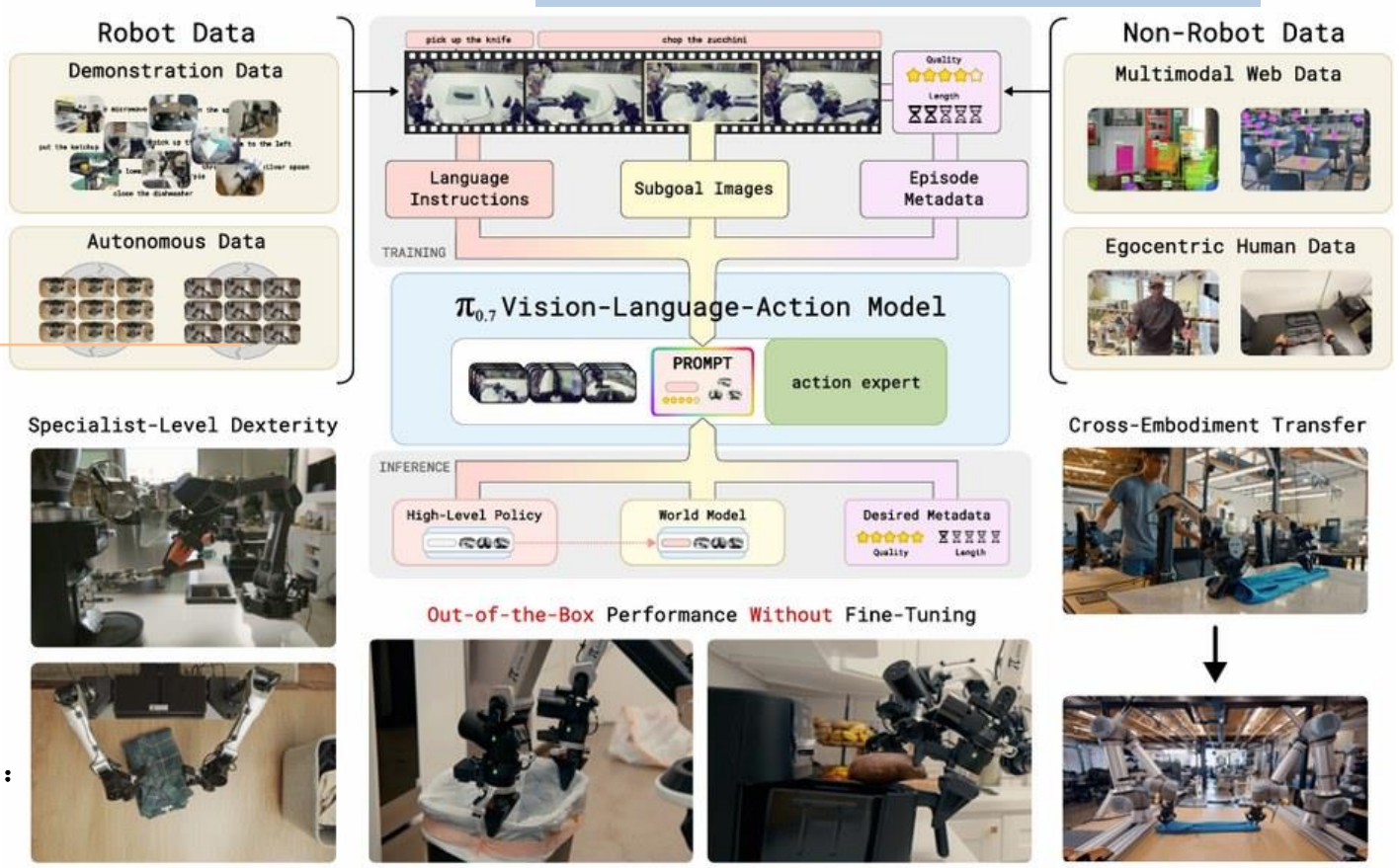
1.2.1 PI的最新一代具身模型，引入多元数据、世界模型

π0.7：引入多元化数据、数据打分机制、世界模型等

Robot Data
①Demonstration Data（示范数据）： 包含大量带有文本标注的高质量人类遥操作数据
②Autonomous Data（自主数据）： 机器人自己在环境里自主试错、探索和强化学习产生的数据

闭环推理流（INFERENCE）
①High-Level Policy 给出下一步的微观文本
②World Model 脑补出 Subgoal 画面
③Desired Metadata 注入满分标签，三者合一作为 Prompt，驱动Action Expert 输出电流

Language Instructions（语言指令）、Subgoal Images（子目标图像）、Episode Metadata（片段元数据）： 特别引入了 Quality（数据质量评分）和 Length（动作长度评分）的五星标签。



Non-Robot Data：Multimodal Web Data（互联网多模态数据）、Egocentric Human Data（第一人称人类动作数据）

能力表现

①灵巧度（Specialist-Level Dexterity）；

②极强的常识泛化与避障；

③跨实体无缝迁移（Cross-Embodiment Transfer）： 顶部的通识大脑学到的知识，可以直接在右侧展示的不同厂商、不同构型的多轴人形双臂机器人之间进行无缝技能迁移。

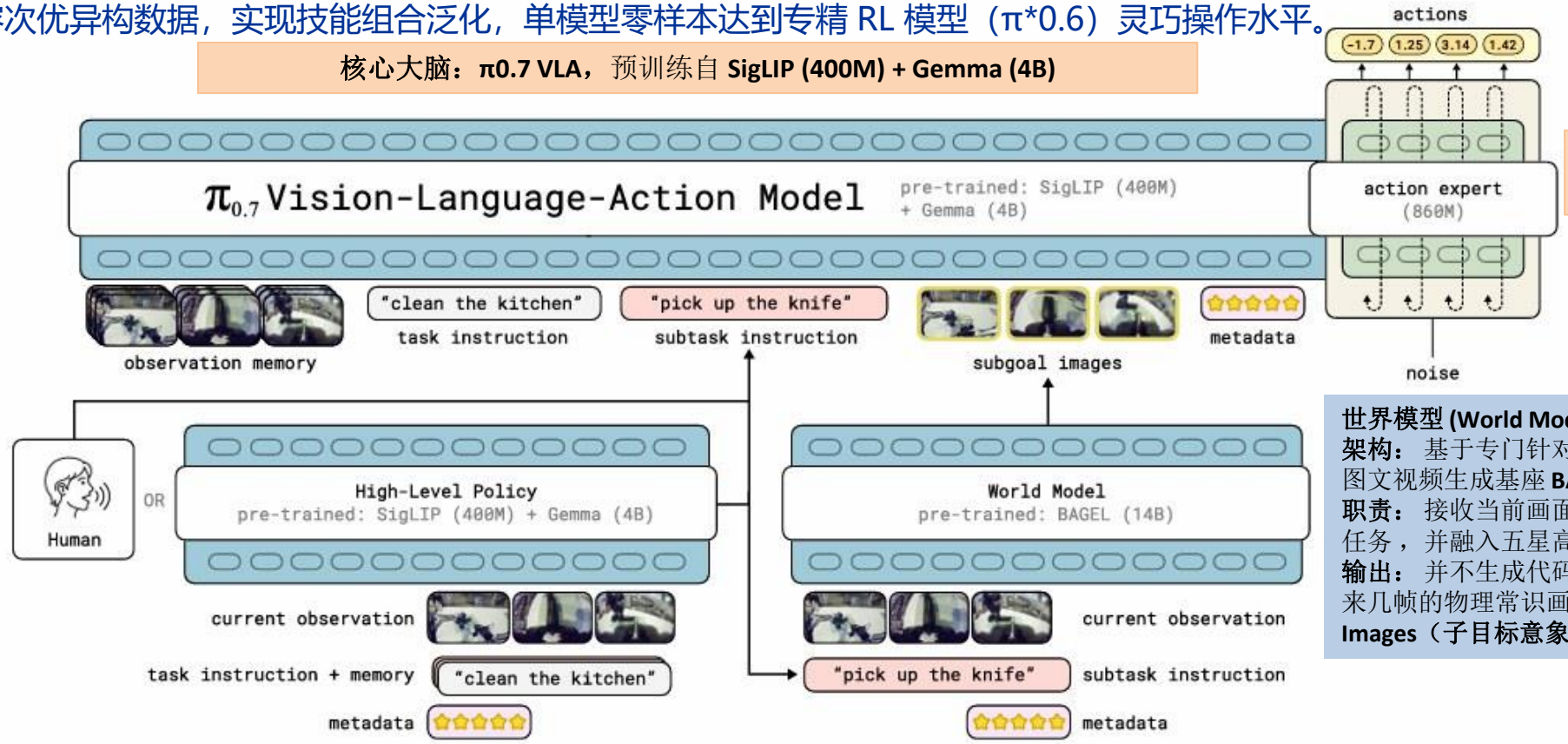
图4：π0.7模型的框架

1.2.1 PI的最新一代具身模型，引入多元数据、世界模型

π0.7模型架构：VLA+流匹配的底层架构延续。

- 最大的变化在于Token变化：observation memory历史视频流、task instruction任务目标、subtask instruction高层策略拆出的短期目标、subgoal images世界模型生成的未来参照图、metadata显式注入的五星高质量 Prompt。
- 性能：兼容次优异构数据，实现技能组合泛化，单模型零样本达到专精 RL 模型（π*0.6）灵巧操作水平。

核心大脑：**π0.7 VLA**，预训练自 **SigLIP (400M) + Gemma (4B)**



Action Expert：
860M

高层策略网络 (High-Level Policy)
职责：接收人类指令、当前多视角视频（current observation）、历史长短期记忆以及“五星”元数据标签。
输出：完成“测试时计算（Test-Time Computation）”，输出当前时间步的最优子任务文本

世界模型 (World Model)：
架构： 基于专门针对物理世界的开源图文视频生成基座 **BAGEL (14B)**。
职责： 接收当前画面和刚拆解出的子任务，并融入五星高评分质量条件。
输出： 并不生成代码，而是生成接下来几帧的物理常识画面——**Subgoal Images（子目标意象图）**

图5： π0.7模型架构

1.2.2 Google Deepmind是具身模型领域的重要参与者

1）谷歌RT-1：基于经典Transformer结构方案（2022年）

- 从机器人的相机中获取图像历史记录同时将以自然语言表达的任务描述作为输入，通过预训练的FiLM EfficientNet模型将它们编码为token，然后通过TokenLearner将大量标记映射到数量更少的标记中，实现标记压缩，最后经Transformer输出动作标记。其可以成功吸收来模拟环境和其他机器人的异构数据，不仅不牺牲在原始任务上性能，还提高了对新场景的泛化能力。

2） VLA模型RT-2：基于预训练LLM/VLM方案（2023年）

- 在视觉-语言模型（VLM）的基础上提出了视觉语言动作（VLA）模型，并在预训练的基础上进行联合微调得到实例化的RT-2-PaLM-E和RT-2-PaLI-X。它可以从网络和机器人数据中学习，并将这些知识转化为机器人控制的通用指令。PaLM-E和PaLI-X是两个已接受网络规模数据训练的视觉语言模型(VLM)，相当于赋予机器人规模足够大的数据库，使其具备识别物体和了解物体相关信息的能力。

3）RT-X：结合RT-1和RT-2模型，引入开源大型数据集训练（2023年）

- RT-X模型通过大规模、多样化的机器人学习数据集Open X-Embodiment上训练得到。其数据集由全球21家机构合作，涵盖了22种不同机器人类型的数据，包含了超过100万个片段，展示了500多项技能和在150000项任务上的表现。RT-X模型采用了基于Transformer的架构和算法，结合了RT-1和RT-2两个模型，其泛化、涌现能力得到了大幅提高。

4）Gemini Robotics 1.5系列（2025年）

- 谷歌推出让机器人拥有多模态理解能力的Gemini Robotics系列，6月推出Gemini Robotics On-Device，9月发布新一代通用机器人基座模型——Gemini Robotics 1.5系列，通过两个模型协同工作：

1.2.2 Google Deepmind首次提出VLA模型RT-2

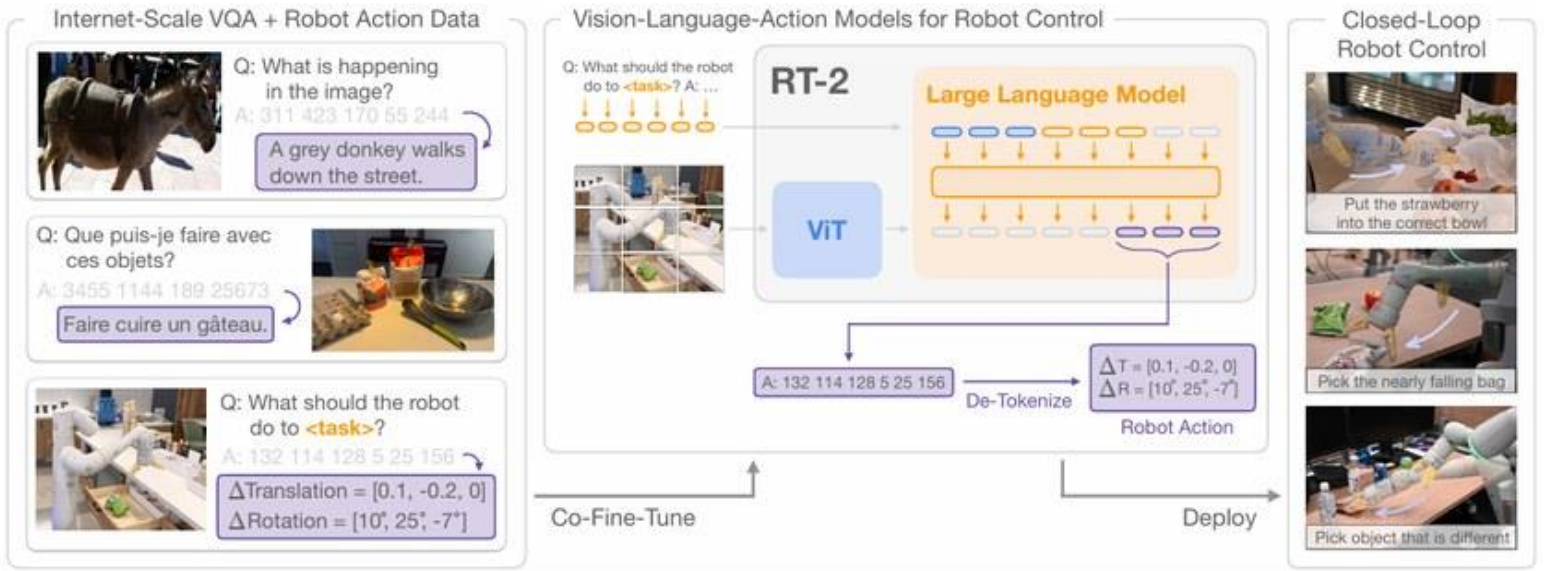
2023年7月由Google DeepMind首次提出VLA模型RT-2。

- 模型架构：将视觉输入、语言推理与动作输出端到端融合，把连续动作离散成 token，做分类自回归。
- 突破：第一次跑通了“把动作当语言写进大模型词表”的尝试，构建VLA架构；
- 局限性明显：输出的动作是离散的文本符号，动作不连贯；模型较大，无法做到快速行动

VLA for robot

视觉与文本输入： 机器人的实时摄像头：画面通过 ViT (Vision Transformer) 转化为视觉向量；人类的抽象指令转化为文本向量；**LLM：**融合后的 Token 喂给百亿级参数的语言大模型；**动作即语言（Token 映射）**在模型的词表（Vocabulary）里，给机器人的动作腾出“动作 Token”位置；**去 Token 化（De-Tokenize）：**模型输出这串数字后，通过底部的 De-Tokenize 环节，瞬间把文本数字恢复成真机的物理三维控制律

数据协同微调（Co-Fine-Tune）
网络问答、跨语言视觉推理、机器人轨迹控制，三种数据融合对模型进行协同微调



闭环控制部署（Closed-Loop Robot Control）：解算出的物理动作直接下发给真机，在非结构化环境里进行高泛化性的闭环操作

图6： RT-2模型架构

1.2.2 谷歌最新的具身智能模型研究Gemini Robotics 1.5

Gemini Robotics项目框架：上层Gemini多模态大模型为思考推理模型，下层动作模型VLA为动作执行模型

- 25年3月，谷歌推出让机器人拥有多模态理解能力的Gemini Robotics系列，6月推出Gemini Robotics On-Device，**9月发布新一代通用机器人基座模型——Gemini Robotics 1.5系列，通过两个模型协同工作。**

模型一：协调器Gemini Robotics-ER 1.5

输入：语音、文本、与环境反馈（图像、传感器数据）；

ER思考链：输出答案前，会消耗一定的“思考预算（Inference-time compute）”进行慢速的因果与空间逻辑推理

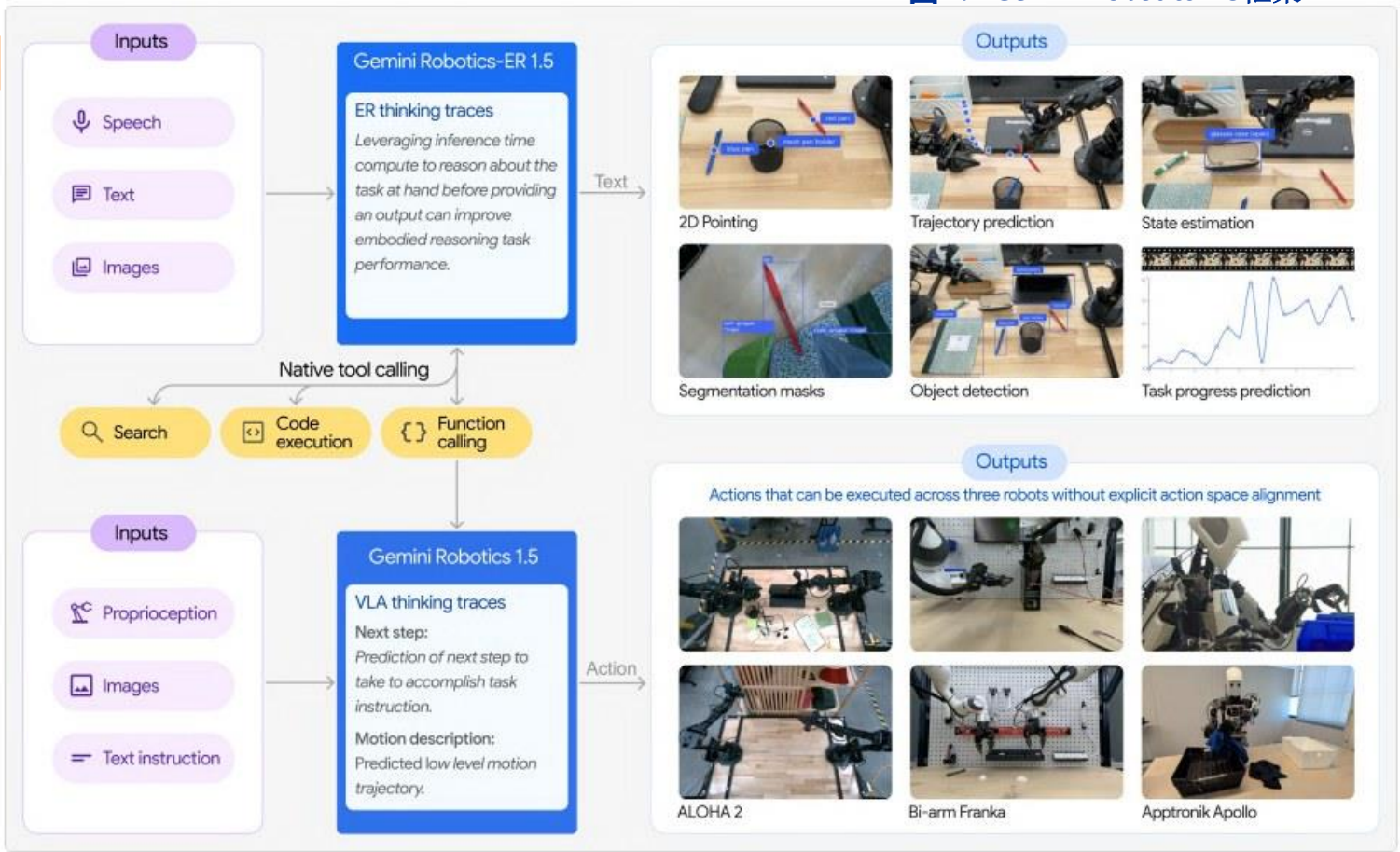
调用外部工具（如网页搜索天气数据）补充决策信息

模型二：动作模型Gemini Robotics 1.5

输入：图像、文字指令、Proprioception（机器人本体感受）机器人关节当前的角度、速度和力矩参数。

思考链：在输出具体的关节电信号前，会用自然语言生成一小段内部推理逻辑，显著提升了多步长任务的容错率。

图7： Gemini Robotics 1.5框架



输出：

2D Pointing / Object detection / Segmentation masks： 精确的物品像素级定位、边界框和掩膜（知道苹果和碗在空间的哪个绝对像素点）

Trajectory prediction： 宏观的机械臂运动轨迹预测。

Task progress prediction / State estimation：物品状态评估与任务进度自我监控（如评估倒水是否倒满、袋子是否快要掉落）。

输出：动作

异构本体的无缝运动迁移（Motion Transfer），支持ALOHA、Bi-arm Franka、Apollo 三种形态机器人直接控制，无需额外适配

资料来源：谷歌 DeepMind网站，《Gemini Robotics 1.5: Pushing the Frontier of Generalist Robots with Advanced Embodied Reasoning, Thinking, and Motion Transfer》，申万宏源研究

1.2.2 Gemini Robotics 1.5创新点

- **1）运动迁移，模型能从不同形态的机器人数据中学习，无需额外训练；**
  - **MT 的放大效应：**多形态数据本身已能提升性能，证明其能有效提取不同机器人的运动共性（如 “抓取” 动作的力控逻辑）；**跨大形态迁移能力；数据稀缺场景适配：**Bi-arm Franka 的原始训练数据量中等，引入多形态数据 + MT 后，任务泛化得分从 0.30（方案 3）提升至 0.50，解决了 “新机器人数据少、训练难” 的行业痛点。

- **2）“思考再行动”的VLA：在执行动作前生成思考轨迹和具体的动作步骤，提升多步骤任务的成功率、可解释现象和错误恢复能力；**
  - **隐式成功检测：**当机器人成功抓取黄色网球后，思考轨迹从 “拿起黄色网球” 自动切换为 “放入白色袋子”，无需额外传感器判断任务完成度；**自主错误恢复：**水瓶从右手滑落至左手附近时，下一轮思考轨迹立即更新为 “用左手拿起水瓶”，0.5 秒内完成纠错，避免任务中断；**场景几何理解：**放置软质衣物时，思考轨迹会包含 “调整 gripper 力度以避免衣物变形” “避开行李箱边缘防止碰撞” 等细节，证明模型对物理场景的深度认知。无 MT 机制。

- **3）推理升级：强化对物理世界的视觉-空间-时间推理能力。在实际测试中，模型的多形态和多任务泛化能力，以及长周期任务的能力，明显优于其他同类模型。**
  - **复杂指向能力：** 机器人定位物体关键部位、生成运动轨迹的基础能力; ER 1.5 的指向不仅精准，还能结合物理约束（如承重限制）、语义理解（如 “匹配袜子图案并指向”），为 VLA 提供精准的动作目标定位；**多形态泛化能力：**覆盖不同机器人与任务维度，从 “视觉 / 指令 / 动作 / 任务” 四个维度对比1.5版本和此前版本，1.5 版本在 “数据量少、形态复杂、任务新颖” 的场景中提升最显著；**长周期任务**：智能体系统协同效果。长周期任务（如 “整理桌面” “筛选过敏食物”）需融合规划、记忆、工具调用能力

1.2.2 Gemini Robotics 2

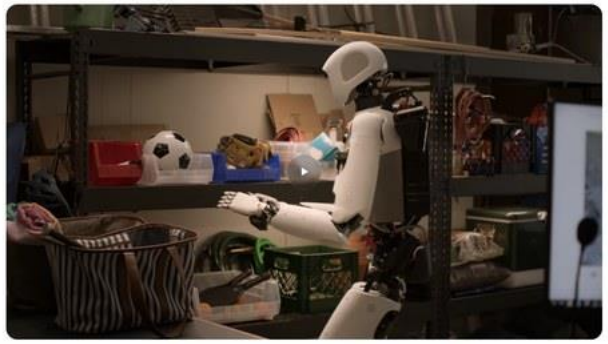
谷歌2026年7月发布的Gemini Robotics 2包括三个模型：

- 1）Gemini Robotics 2 为VLA 模型，可将视觉语言指令转化为动作指令；可控制各种全人形机器人+灵巧手；
- 2）Gemini Robotics ER 2 为具身推理模型，是VLM模型，支持实现人机交流、长序列规划、空间时序推理与多机协同调度；
- 3）Gemini Robotics On-Device 2 为端侧轻量化模型，本地化运行，低延迟实时控制；仅需不到200个样本、几小时即可适配新本体。

性能表现

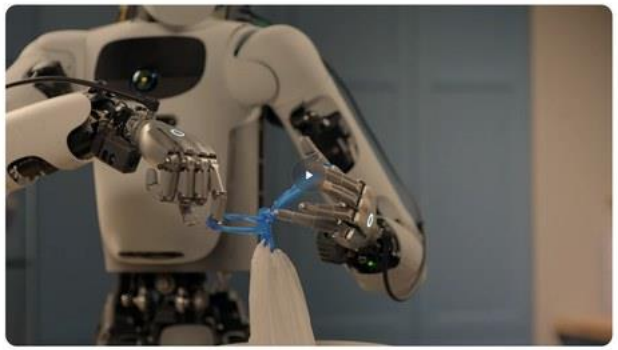
- 1）跨本体迁移：只需几个小时，就能适应任何双臂机器人，使其智能从手臂扩展到复杂的类人形身体；
- 2）高级灵巧度：机器人能够完成需要细腻技巧的精细动作，比如拧灯泡和打结；
- 3）全身控制：能控制从脚到指尖的整个身体，使机器人能完成从弯腰到伸手的全范围类人动作；
- 4）长序列任务：将深度空间推理和长视野规划结合，机器人能完成复杂陌生任务；支持两台机器人分工协作完成任务；
- 5）互动：能够理解并响应日常指令。

图8： Gemini Robotics 2的机器人性能表现



Intelligent whole-body control

Controls entire humanoid robots from feet to fingertips, translating intent into whole-body control to reach, bend, and balance.



Advanced dexterity

Controls complex humanoid hands and parallel grippers to unlock a new level of physical dexterity.

Multi-robot collaboration

Enables different types of robots to communicate and work together to solve complex workflows a single robot could not do alone.

1.2.3 英伟达Cosmos 3全模态物理AI模型

NVIDIA 正式推出 Cosmos 3：面向物理 AI 的开放前沿基础模型

AR subsequence（自回归序列/思考脑）
ViT 编码的离散视觉 Token、语言/指令 Token、以及特殊的终止符和开始生成符

Diffusion subsequence（扩散序列 / 生成脑）
由 VAE 编码的连续时空视觉 Token、音频 Token以及机器人的动作控制 Token。

通过注意力掩码（Attention Mask）来控制Reasoner和Generator的融合与隔离

共享多模态注意力机制（Shared Multimodal Attention）

输出：
左侧： 下一个时间步的离散语义 Token（Next-token predictions），指挥机器人下一步行为。
右侧： 经过流匹配（Flow-matching）预测出速度并彻底去噪后的干净 Token（Denoised tokens），直接还原成高保真的未来视频帧、碰撞声音或高频真机操作电流。

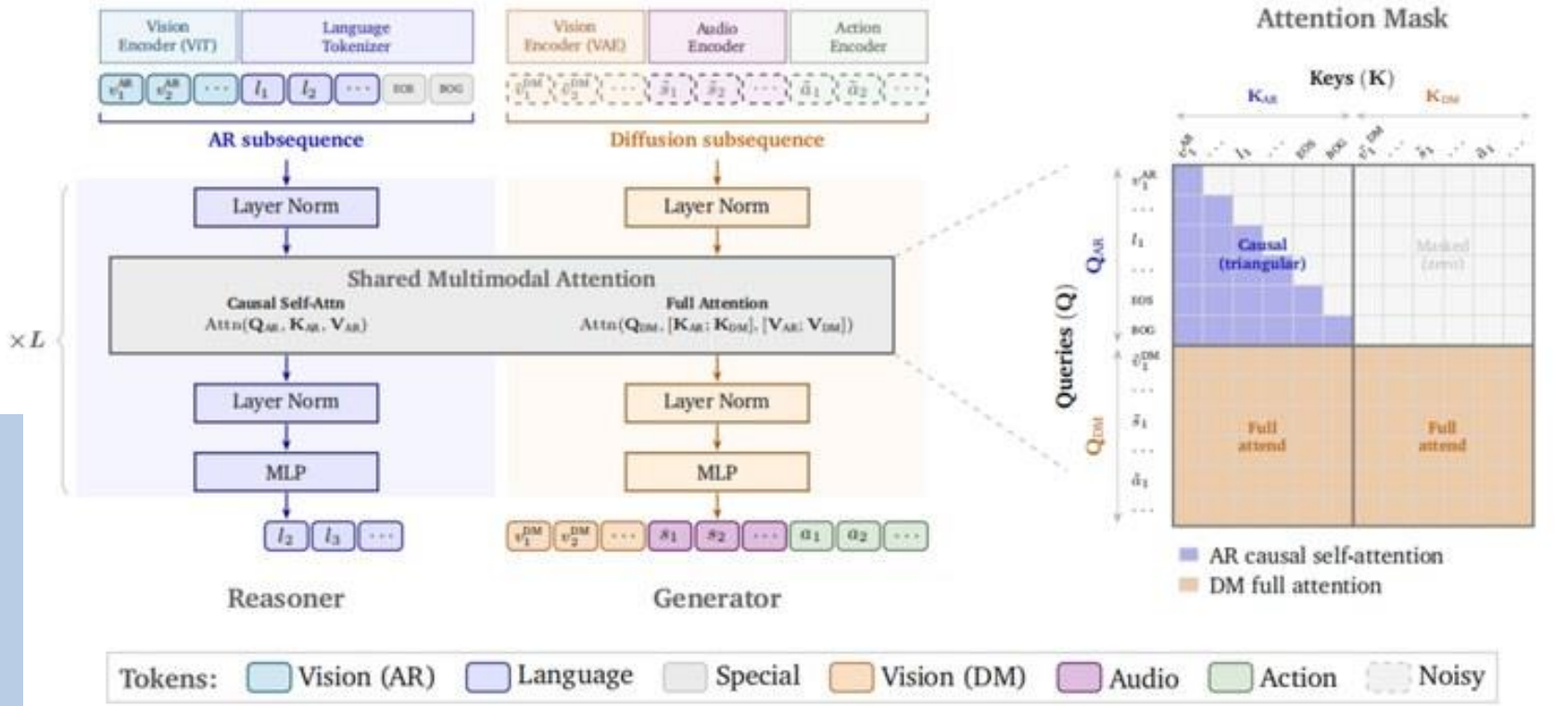


Figure 5: **Mixture-of-Transformers (MoT) architecture of Cosmos 3. Left:** a single transformer operates on

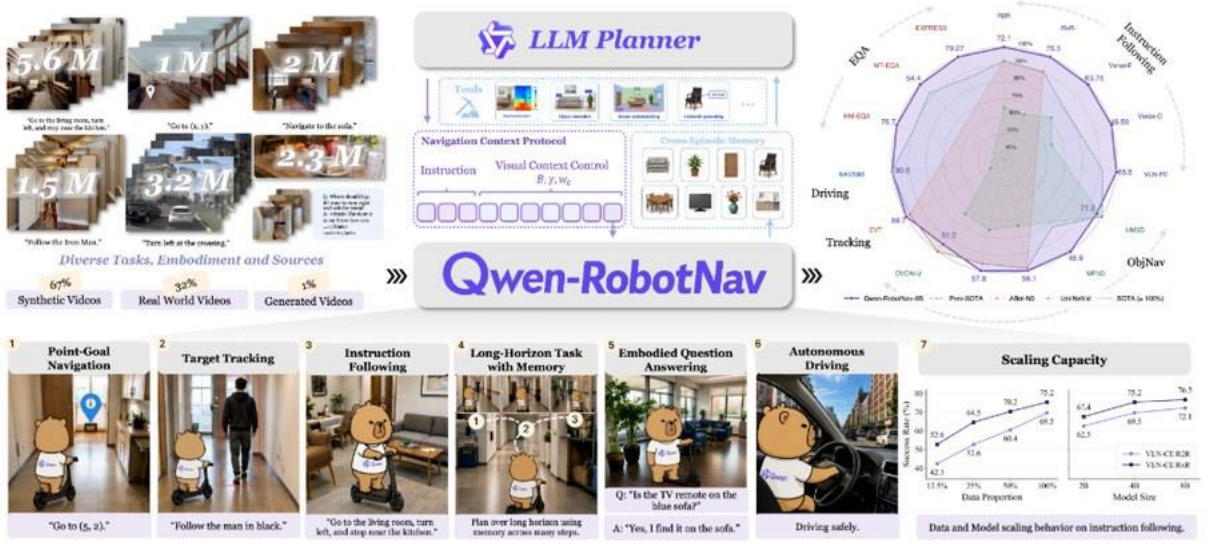
图9： 英伟达Cosmos3模型的MoT架构

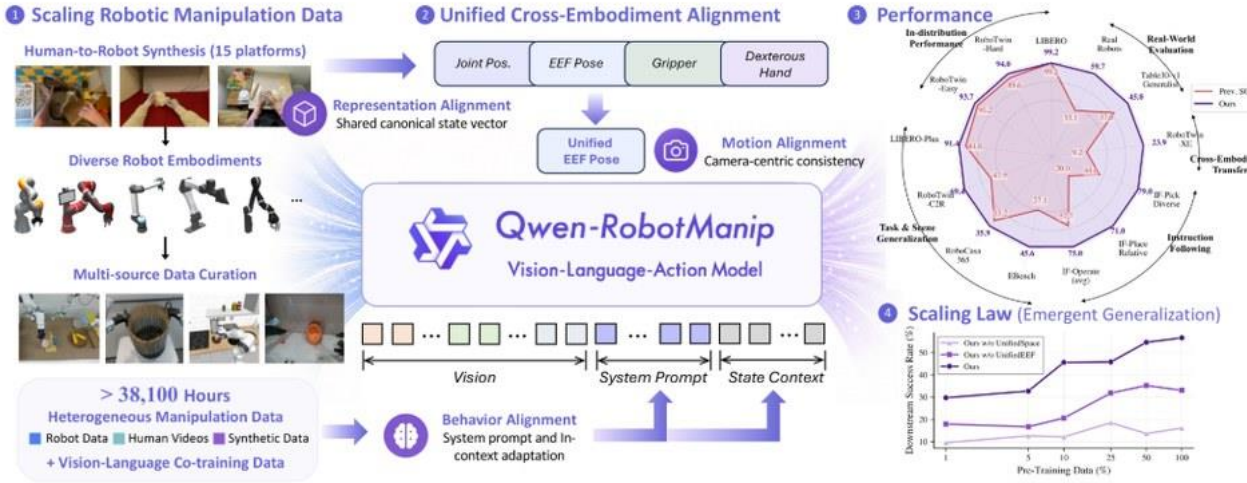
资料来源：英伟达官网，《Cosmos 3: Omnimodal World Models for Physical AI》，申万宏源研究

1.2.4 阿里Qwen-Robot Suite：迈向物理世界智能的基础模型套件

- **Qwen-Robot Suite 用三个基础模型Qwen-RobotNav、Qwen-RobotManip 与 Qwen-RobotWorld。**
  - Nav 通过自适应视觉分配策略将五类导航任务统一到单一模型；Manip 将异构机器人数据对齐到统一空间，实现大规模跨机型训练；World 以自然语言作为动作接口，将 20 余种机器人本体纳入同一世界模型联合训练。

图10：Qwen-Robot Suite三个基础模型

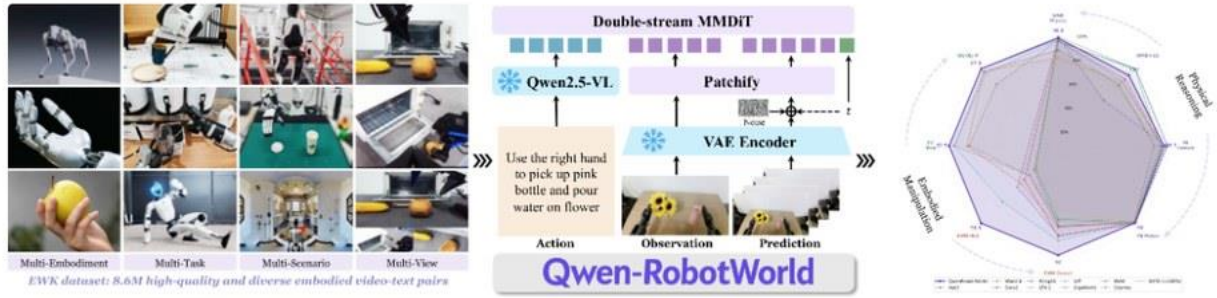




Qwen-RobotManip： 以 Qwen3.5-4B VL 为骨干、结合流匹配 DiT 动作头。

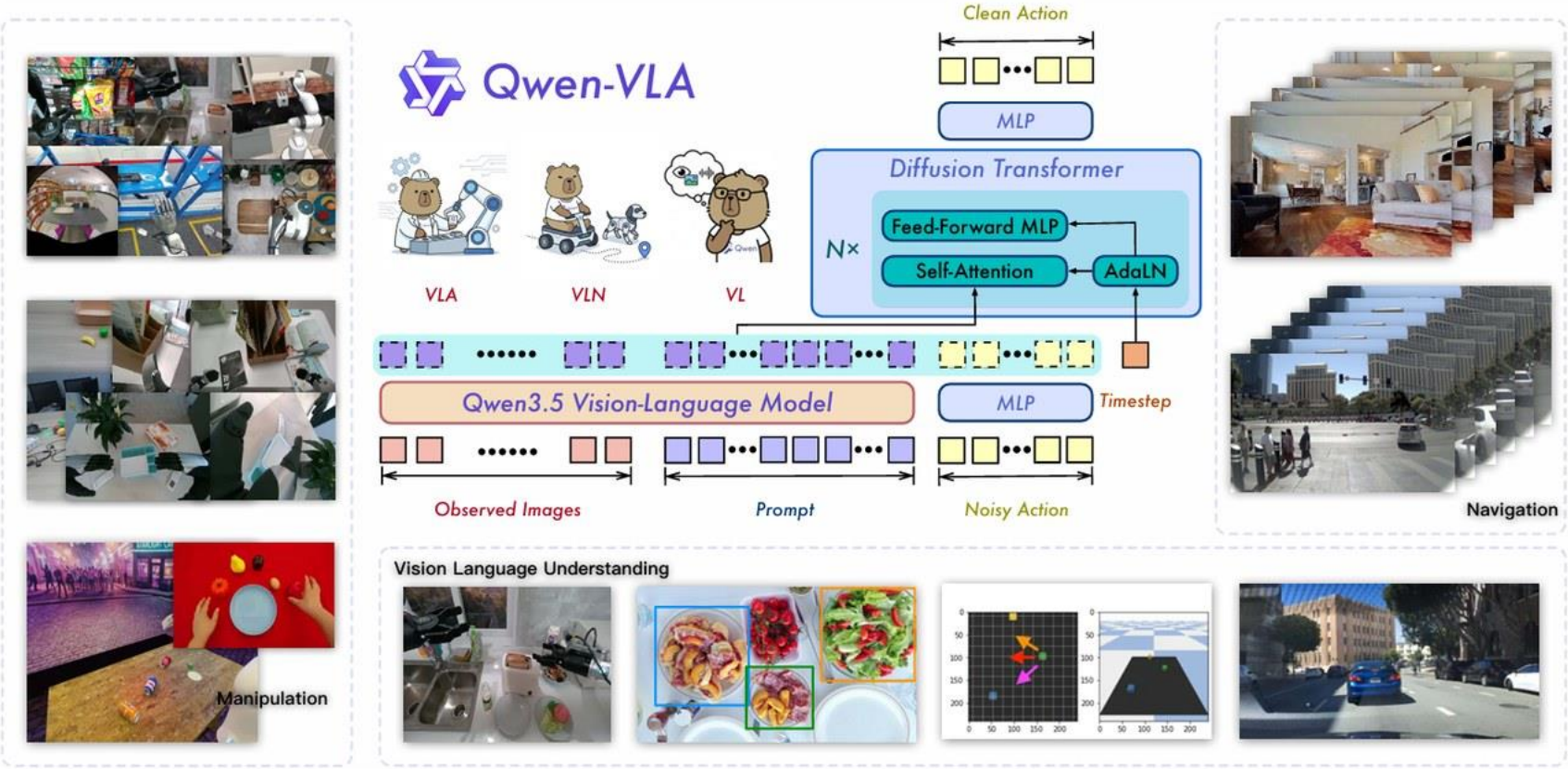
Qwen-RobotNav： 核心思路是将视觉分配策略本身参数化：任务模式选择导航行为（指令跟随、目标搜索、目标追踪、自动驾驶），可调节参数（视觉 token 预算、时间衰减、单相机权重、帧采样模式）决定视觉历史的编码方式。模型在 1,560 万条样本上训练，同时联合视觉语言数据以保留感知能力，一套权重统一五类导航任务。

Qwen-RobotWorld 通过直接学习世界的状态转移函数来解决这一问题：给定当前观测和一个自然语言动作，预测世界接下来将呈现的样子。



1.2.4 阿里Qwen-VLA：从看懂世界走向执行动作

- **Qwen-VLA 是一个通用视觉-语言-动作模型。**
  - 它基于 Qwen 多模态骨干模型，将视觉感知、语言理解和空间推理能力进一步扩展到连续动作生成和轨迹预测。从实验结果看，Qwen-VLA 展示了通用策略模型的潜力。单一模型可以横跨 LIBERO、Simpler、RoboCasa、RoboTwin 等操作评测集，并在多个任务上接近甚至超过专项策略模型。



四阶段训练

1）T2A（文本到动作预训练）：一句”拿起红色杯子”只有几个词，但对应的机器人动作是一段高维连续轨迹。Qwen-VLA 将这个过程看作一种从语言到动作的”解压缩”。在 T2A 阶段，我们冻结 VLM，不引入任何图像，只让动作解码器从语言指令和机器人形态描述中学习动作先验。

2）CPT（预训练）： 解冻 VLM 和动作解码器，在完整的多模态数据上联合训练。这一阶段将 T2A 习得的语言-动作先验对齐到具体视觉场景，同时让骨干模型适应具身感知，产出 Qwen-VLA-Base。

3）SFT（监督微调）：从 CPT checkpoint 出发，分为两条路线：多任务 SFT 在操作、导航、VQA 等异构任务上联合微调；真机 SFT 在内部遥操作数据上微调，用于真实机器人部署。

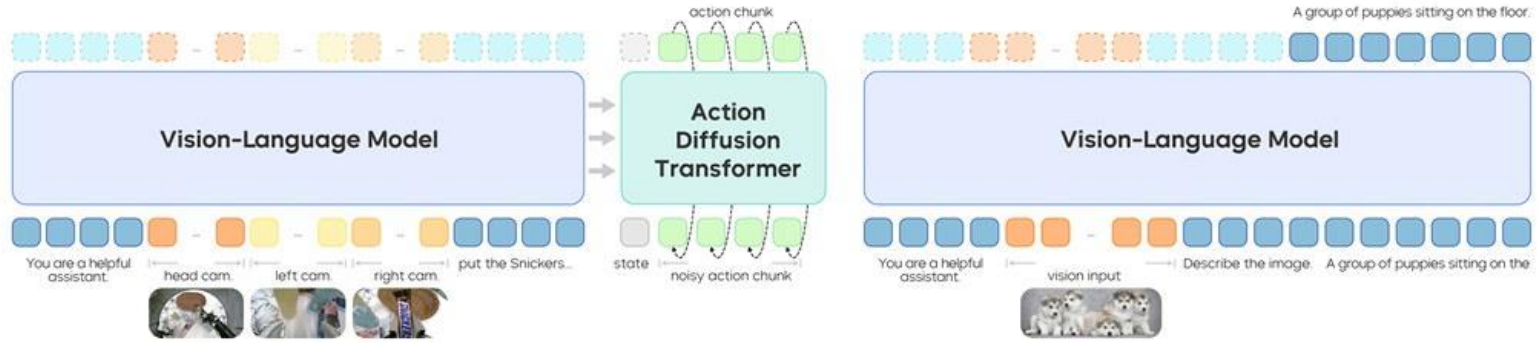
4）RL（强化学习）：从 SFT checkpoint 出发，在仿真环境中通过 PPO 直接优化闭环任务成功率，产出最终模型 Qwen-VLA-Instruct。RL 仅在 SimplerEnv 中进行，但实验表明收益可迁移到其他未见过的环境和机器人形态。

图11：Qwen-VLA模型框架

1.2.5 字节跳动seed发布GR-3与GR-RL两大具身模型

- **字节跳动Seed下设Robotics研究团队，持续发表具身智能相关研究。其模型架构为VLM+动作。**

GR-3模型：VLA模型



GR-RL是一种基于VLA策略，面向长周期灵巧操作任务的机器人学习框架。

GR-RL算法通过采用多阶段训练流程，实现了系鞋带任务中长期、灵巧且高精度的操作控制，该流程包含三个阶段：1）基于学习任务进度的离线过滤式行为克隆；2）简单而有效的动作增广技术；3）在线强化学习训练

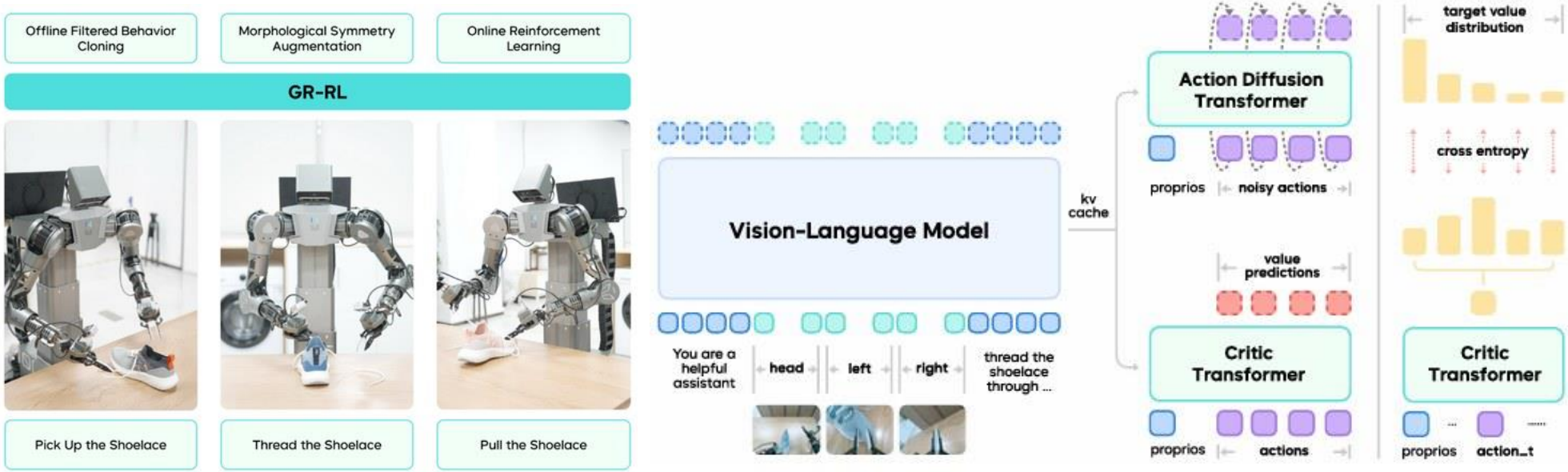
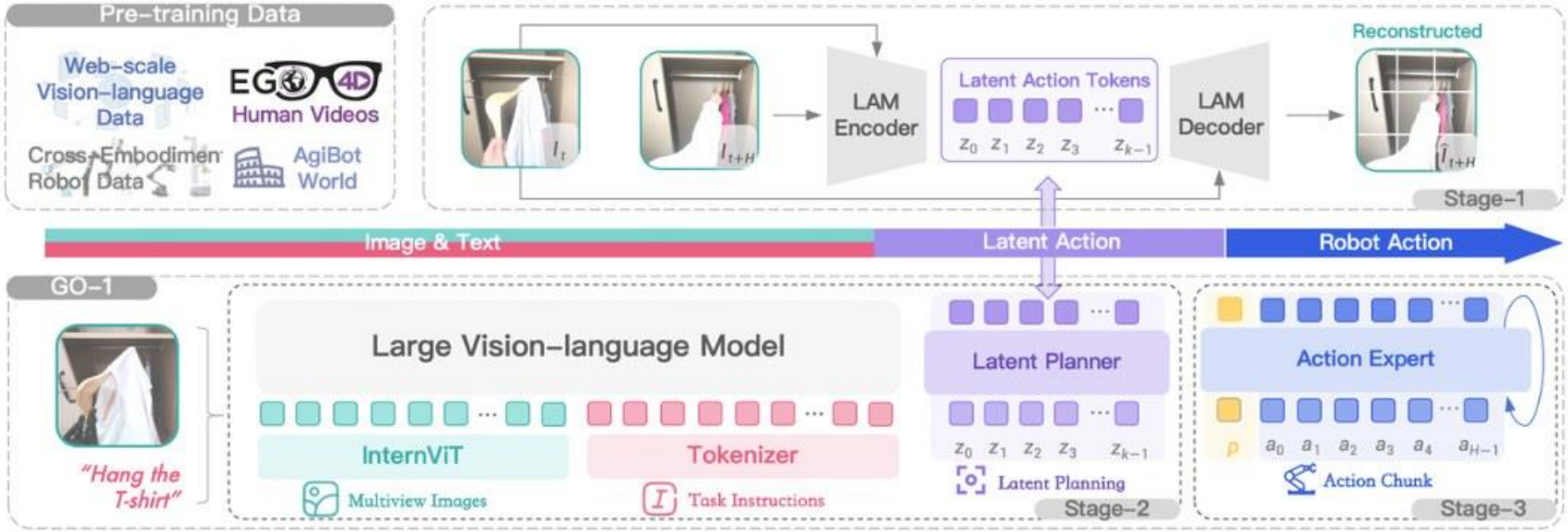


图12：字节GR3与GR-RL模型框架

1.2.6 智元布局具身智能模型和开源数据集

智元2025年3月发布智元启元大模型（Genie Operator-1）

- 该架构由VLM(多模态大模型) + MoE组成，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借助百万真机数据获得精细的动作执行能力。
- 2024年底，智元推出了 AgiBot World，包含超过100万条轨迹、涵盖217个任务、涉及五大场景的大规模高质量真机数据集。
- **2026年智元发布Go-2模型。**



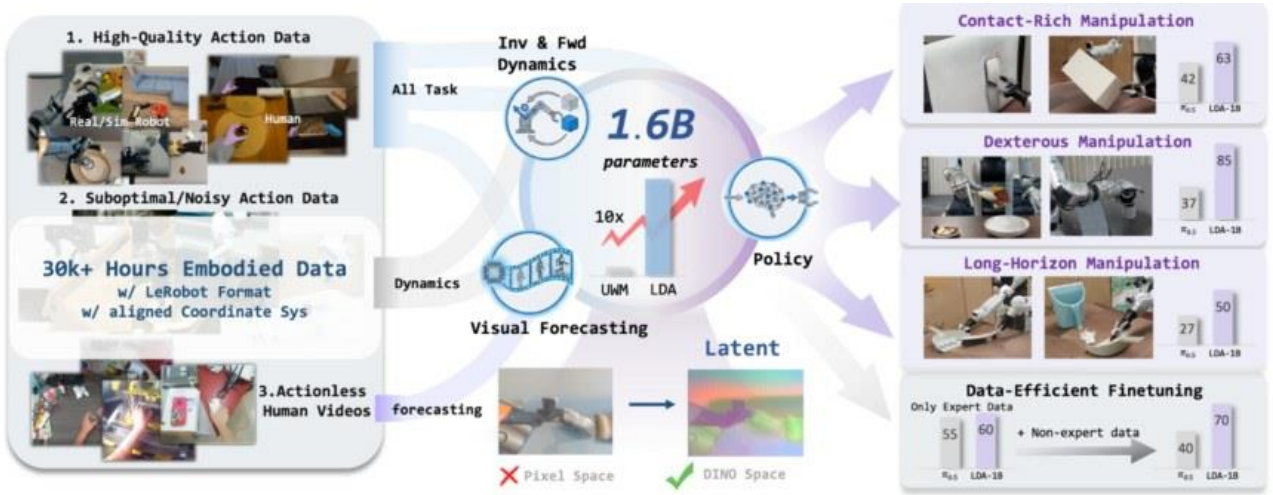
The framework of GO-1

图13：智元Go-1模型框架

资料来源：智元官网，申万宏源研究

1.2.7 银河通用发布WAM World-Action Model（世界-动作统一模型）

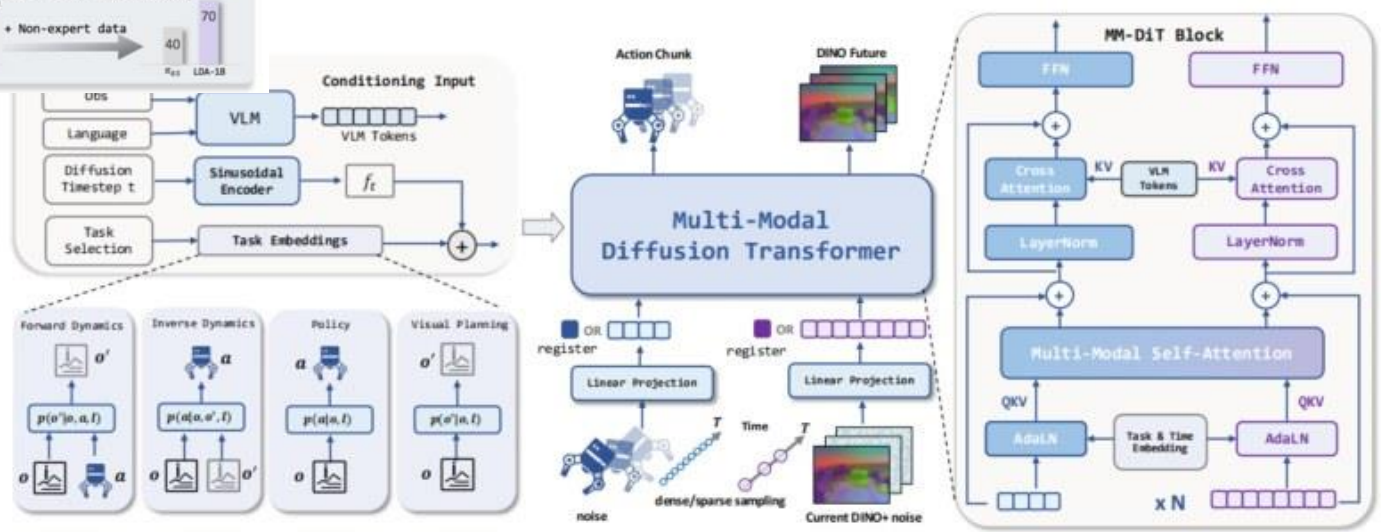
- **银河通用2026年4月发布LDA-1B模型，此前发布过GraspVLA、TrackVLA、GroceryVLA、NavFom等系列模型。**
  - LDA-1B是一个1.6B 参数一体化机器人基础模型（世界模型），数据来源包括高质量动作数据（真机、仿真、遥操）、次优/含噪声动作数据、无动作视频，共3万+小时异构数据，统一采用 LeRobot 数据格式、对齐坐标系。



注：1.6B参数指的是完整模型全部可训练参数量，包含MM-DiT 多模态扩散 Transformer 主干+输入投影层+输出预测头；1B指的是其中的 multimodal diffusion transformer（MM-DiT 主干）的参数规模。

图14：银河通用LDA-1B模型框架

核心模型为LDA-1B（Multi-Modal Diffusion Transformer，MM-DiT）



三大目标：

Policy（策略头）：输出机器人动作
Dynamics（动力学）：学习物理交互、物体运动规律
Visual Forecasting（视觉预测）：预测未来画面

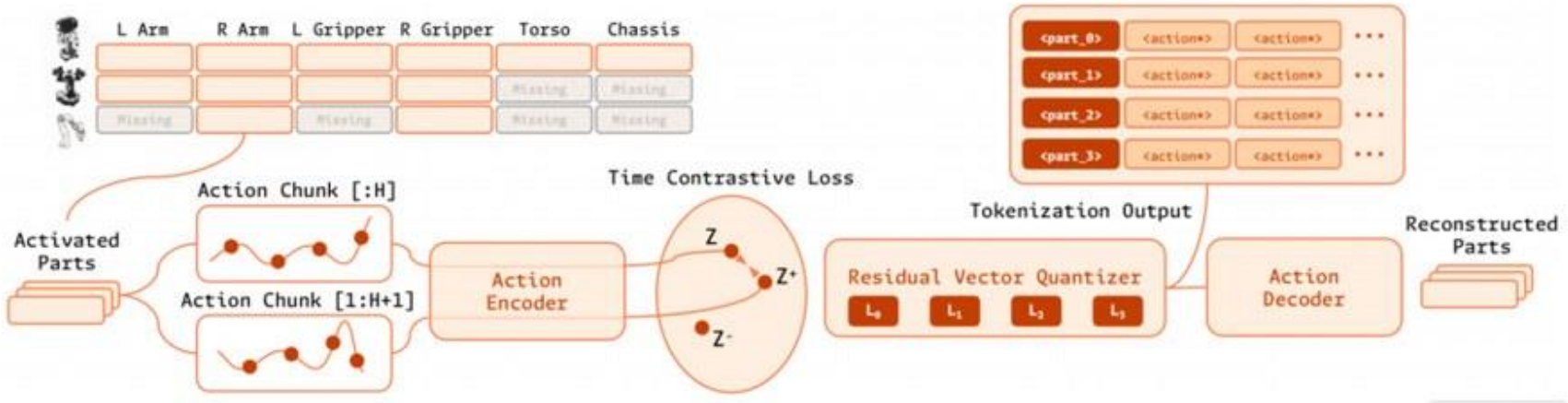
资料来源：银河通用官网，申万宏源研究

1.2.8 星海图选择VLA模型+真机数据为主的路线

星海图2025年8月发布VLA模型G0，2026年6月发布最新模型G0.5

- 当前主流范式为VLM+action DiT，两个模型相互独立；
- **G0.5架构创新：**采用统一自回归VLA架构，单一Decoder Transformer 主干，一套权重，视觉、语言、任务推理、动作生成全部统一到自回归框架；VLM 主干直接作为决策者，不再只是前置编码器。
- **三大核心能力**：1）统一异构动作编解码器：让一种“动作语言”覆盖所有机器人；2）原生动作思维链：让机器人不仅“边思考边行动”，还听得懂“怎么做”；3）时空注意力模块：为机器人注入上下文感知先验。

图15：星海图的模型G0.5中的异构动作编码器



1.2.9 腾讯布局世界模型和VLA模型

- **腾讯最新发布Hy-Embodied-RxBrain（联合文本推理 + 视觉想象的独立世界模型）和Hy-Embodied-0.5-VLA（执行端模型，接收观测 + 指令，输出机器人运动轨迹）。**

Hy-Embodied-RxBrain
文本推理 + 视觉想象融合在同一条规划序列内。**语言：**搭建任务抽象骨架，负责任务拆解、时序顺序、约束条件、决策逻辑（“要做哪几步”）；**视觉想象：**给文字规划落地，预测每一步执行后的世界状态、子目标画面（“做完这一步环境会长什么样”）

架构底座：采用 Modality-routed Mixture-of-Transformers（MoT，模态路由混合Transformer），单模型同时支持：文本、图像、视频的理解 + 生成。

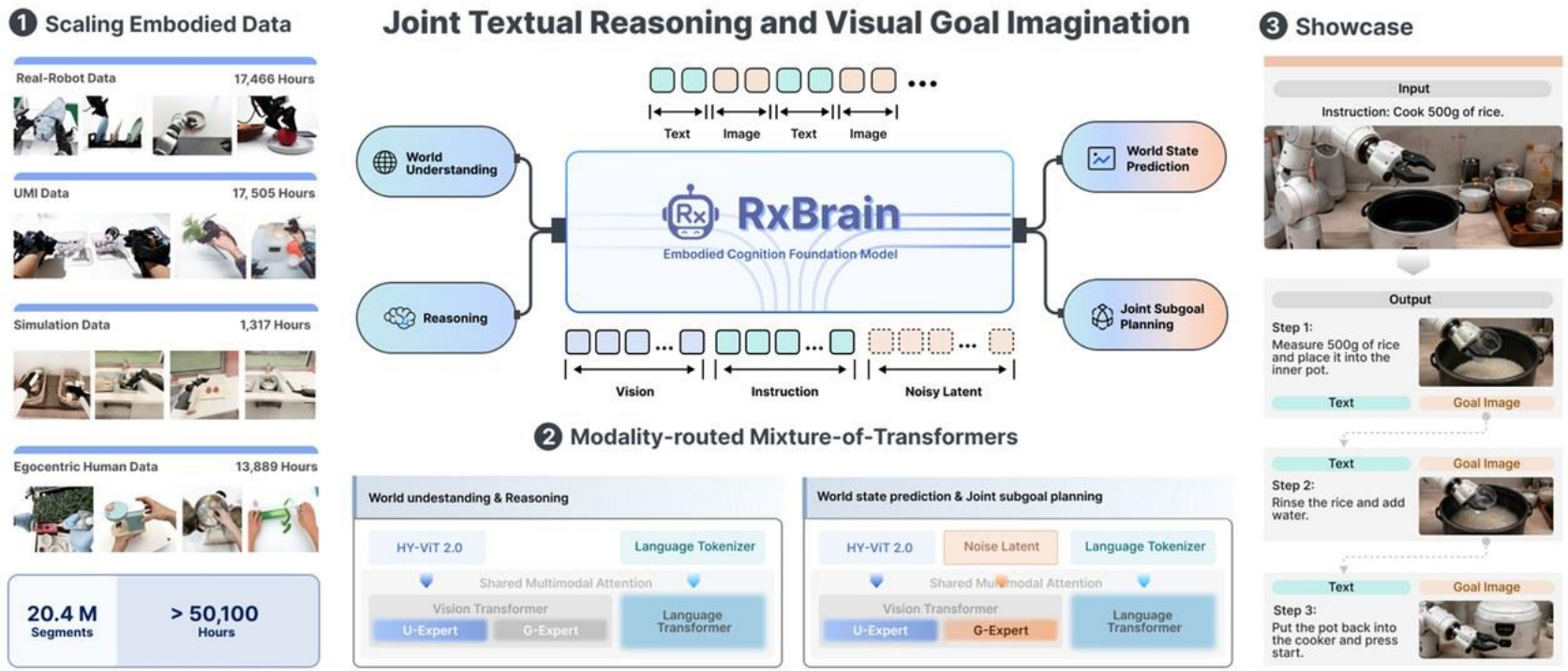


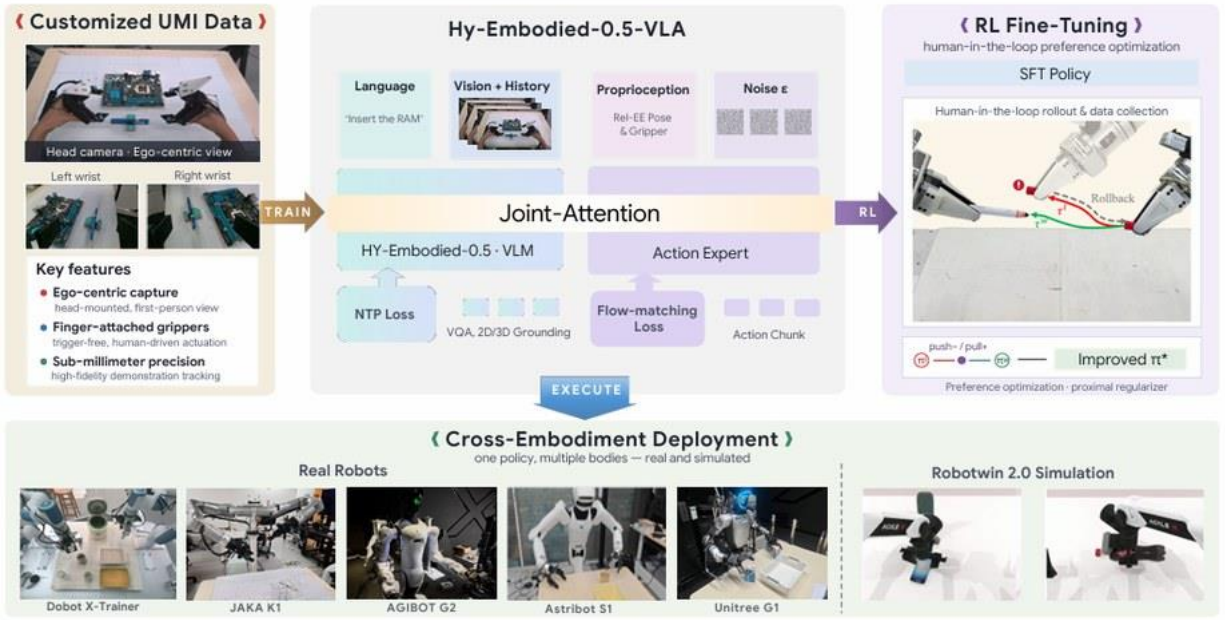
图16：腾讯Hy-Embodied-RxBrain模型架构

资料来源：《RxBrain: Embodied Cognition Foundation Model with Joint Language-Visual Reasoning and Imagination》，申万宏源研究

1.2.9 腾讯布局世界模型和VLA模型

- **腾讯最新发布Hy-Embodied-RxBrain（联合文本推理 + 视觉想象的独立世界模型）和Hy-Embodied-0.5-VLA（执行端模型，接收观测 + 指令，输出机器人运动轨迹）。**

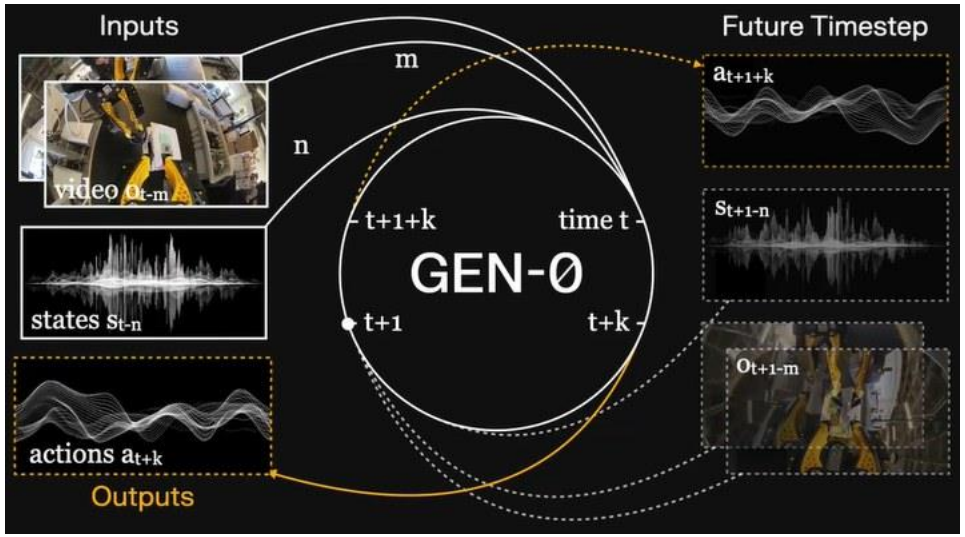
Hy-Embodied-0.5-VLA
提出一套通用机器人学习全栈体系：包括数据采集、模型结构、预训练+监督微调、离线强化学习后训练、真机部署



图： 腾讯Hy-Embodied-. -VLA模型架构

1.2.10 Generalist AI：自有数据集庞大，应用效果显著

- **Generalist AI 2024 年在美国加州 San Mateo 成立，目标构建面向物理世界的通用智能，打造通用机器人基础模型。**
  - **2025年11月发布GEN-0模型：**基于27万小时真实世界操控交互数据；实证机器人具身基础模型存在Scaling Law：随着真实物理交互数据规模提升，模型泛化能力持续可预测提升，类比大语言模型缩放规律；**能力：**快速适应新环境、新任务，具备基础物理常识；零样本跨机器人迁移初步验证。
  - **2026年4月发布GEN-1模型：**基于超50万小时高保真物理交互预训练数据集。实测结果大幅提升：1）任务成功率：同类任务平均成功率从64%提升至99%；2）执行速度提高三倍；3）数据效率：达成上述效果，每个目标任务仅需1小时真实机器人采集数据；4）涌现特性：动态环境自主故障恢复、意外场景即兴调整操作策略；5）泛化：单一模型同时支持桌面操作、仓储抓取、移动操纵等多类任务，可适应不同末端执行器。



图： GEN-模型架构

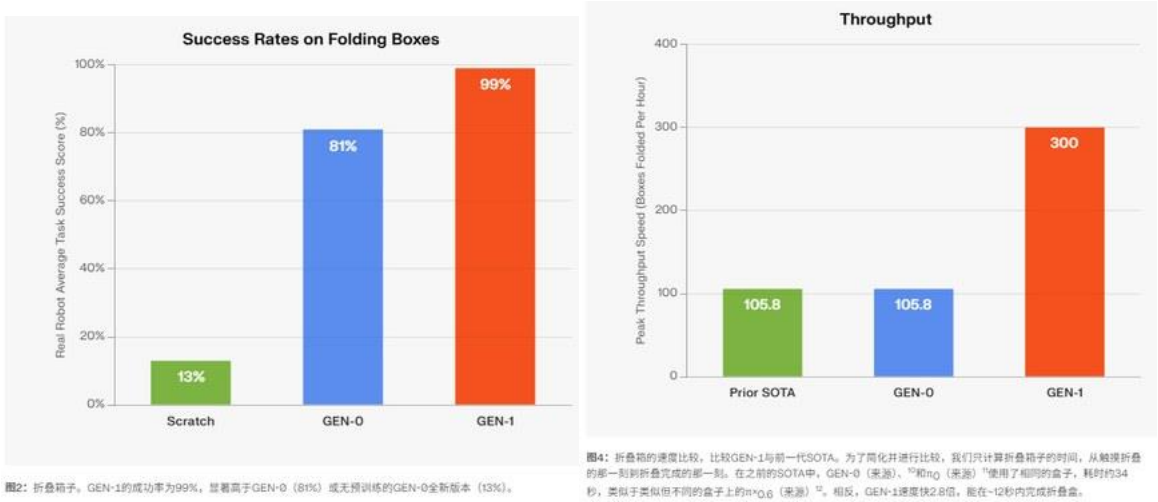


图2：折叠箱子。GEN-1的成功率为99%，显著高于GEN-0（81%）或无预训练的GEN-0全新版本（13%）。

图4：折叠箱的速度比较。比较GEN-1与前一代SOTA。为了简化并进行比较，我们只计算折叠箱子的时间，从触摸折叠的那一刻到折叠完成的那一刻。在之前的SOTA中，GEN-0（来源）、π_0（来源）[1]使用了相同的盒子，耗时约34秒，类似于类似但不同的盒子上的$\pi^{*}_{0.6}$（来源）[2]。相反，GEN-1速度快2.8倍，能在~12秒内完成折叠盒。

图： GEN-表现：可靠性和速度大幅提升

1.2.11 宇树科技：世界模型和VLA模型开源

- **宇树科技具备全栈硬件底座（G1人形、H1、Z1机械臂、四足B2），同时VLA模型和世界模型完整对外开源，主要面向宇树自己的机器人本体：**
  - **2025年发布预测式世界模型UnifoLM-WMA-0。**该模型具有两大功能：1） 仿真引擎：作为交互式模拟器生成合成数据用于机器人学习；2）策略增强：连接动作头，并通过预测未来与世界模型的交互过程，进一步优化决策性能。
  - **2026年1月发布VLA模型UnifoLM-VLA-0。**该模型基于开源的Qwen2.5-VL-7B骨干网，使用机器人多任务数据集进行预训练（340小时的开源真实机器人数据集），获得UnifoLM-VLM-0。VLM模型集成动作头，形成VLA模型。此外，宇树开发基于G1的12类任务的机器人数据集，该模型能够可靠地通过单一策略检查点完成全部12项任务。

LIBERO模拟基准测试

模型	LIBERO-空间	LIBERO-对象	自由人——进球	利贝罗-朗	平均值
UnifoLM-VLA-0	99.0	**100**	**99.4**	96.2	**98.7**
EO1	**99.7**	99.8	99.2	94.8	98.2
X-VLA	98.2	98.6	97.8	**97.6**	98.1
开放VLA-OFT	97.6	98.4	97.9	94.5	97.1
GR00T-N1.6	97.7	98.5	97.5	94.4	97.0
π0.5	98.8	98.2	98.0	92.4	96.9
MemoryVLA	98.4	98.4	96.4	93.4	96.7
InternVLA-M1	98.0	99.0	93.8	92.6	95.9
F1	98.2	97.8	95.4	91.3	95.7
π0.5-KI	98.0	97.8	95.6	85.8	94.3
π0	96.8	98.8	95.8	85.2	94.2
GR00T-N1	94.4	97.6	93.0	90.6	93.9
齿轮行动	97.2	98.0	90.2	88.8	93.2
π0 + 快速	96.4	96.8	88.6	60.2	85.5

注：表中所有数据均来源于原始论文、官方项目主页或学术论文。

图20：宇树科技UnifoLM-VLA-0在LIBERO模拟基准测试中的性能表现

1.3 模型架构：当前架构尚未收敛，各条路线持续迭代

- **路线1：VLM主干→ 独立动作生成头，是当前最主流的模型架构**
  - **代表：** PI π0.7、阿里 Qwen-VLA、字节GR-3、智元Go-1、腾讯Hy-Embodied-0.5-VLA、小米 U1等；**优势：**工程友好、训练稳定、易微调，最适合真机部署；VLM 可以复用开源大模型权重，降低训练成本；**短板：**基于当前观测输出动作，很难自主预判未来环境变化，长时序任务乏力。

- **路线2：分层快慢脑架构：上层模型负责高层任务推理、规划；下层动作模型负责动作输出**
  - **代表：**Gemini Robotics 1.5；**优势：**兼顾大模型推理能力与底层实时控制性能；架构灵活；**短板：**系统复杂度显著提升，需要解决跨模块特征对齐、多模块联合调优难题。

- **路线3：一体化自回归序列模型，把图像+文本+动作Token 放入同一个自回归Transformer**
  - **代表：**星海图 G0.5、Google RT-2；**优势：**统一范式，结构简单，天然支持思维链 CoT、文本规划 + 动作交织生成；**短板：**动作量化带来信息损失和精度缺失；推理延迟偏高；缺少显式世界预测能力。

- **路线4：一体化扩散/世界模型**
  - **代表**：银河通用 LDA-1B；**优势：**天然具备具身认知、环境预判，理论上限最高，适合长时序复杂任务；**短板：**训练难度极大、算力消耗高；真机微调困难，商业化落地进度慢。

主要内容

2.1模型训练的数据来源：真实数据、合成数据、ego数据、网络数据等

- **数据是当前具身智能的泛化面临的瓶颈之一。相对于多模态大模型可以利用互联网文字、图像、视频、音频等数据进行训练，具身智能大模型目前依然缺少可用于训练的高质量、大规模人类操作数据，且获取难度大。**
- **具身智能的数据来源包含互联网数据、仿真合成数据和真实数据等**

图21：具身智能模型的数据呈现金字塔架构

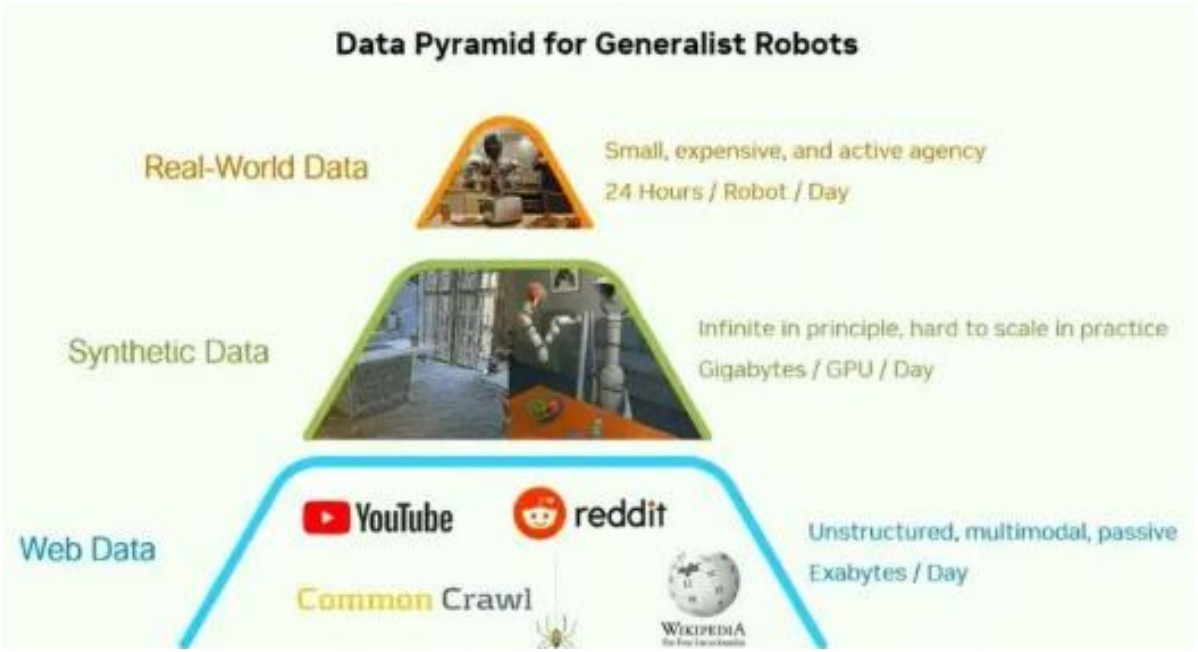


表4：不同数据来源的差异

采集方式	数据质量	采集成本与速度	核心瓶颈/挑战	典型代表
真机遥操作	高	极慢、成本昂贵	难以大规模量产(Scaling Law受限)	Mobile ALOHA, Stanford-UMI
仿真合成	中	极快、吞吐量无限	物理与视觉的逼真度(Sim-to-Real Gap)	NVIDIA Cosmos/Isaac,RoboHive
头戴第一人称 Ego 采集	中偏低	较快、模化，偏低	可众包规成本中等	人体 - 机器人本体鸿沟 Embodiment Gap，无原生机器人控制闭环数据	Ego4D、EPIC-KITCHENS、EgoMimic
互联网视频	低	资源丰富、速度快	体征映射难题(Embodiment Gap)、缺乏 Control Loop数据	/

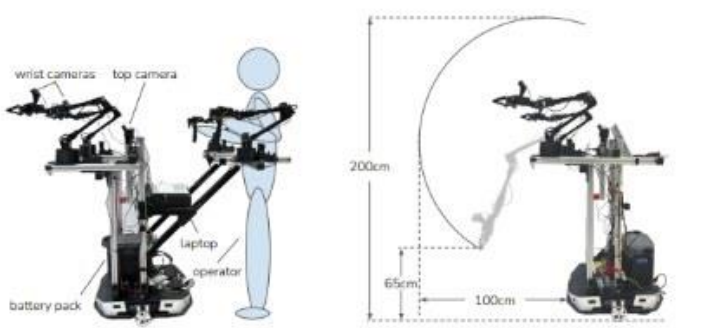
2.1.1 互联网数据：提供常识、逻辑和泛化能力

- **互联网数据（包括 YouTube 视频、大规模图文对、互联网维基百科等），作用在于：**
  - **提供物理常识：**通过这些海量被动视频，机器人能提炼出物理世界的因果律（例如：“玻璃杯掉地上会碎”、“水会往低处流”、“重物需要更强的支撑力”）
  - **视觉和语言泛化能力：**互联网数据（如 ImageNet、Open-Images、互联网文本）涵盖了人类社会各种光照、色调、角度、异形物体和语言表达。利用其预训练，机器人能够理解“不管是木头碗、不锈钢碗还是陶瓷碗，它们都是碗”，进而解锁零样本（Zero-shot）的任务泛化。
  - **长序列任务的语义逻辑：**当人类说“帮我泡杯咖啡”，机器人大模型能自动在脑海中将其拆解为：[走向咖啡机] -> [拿杯子] -> [放胶囊] -> [按下启动键]

2.1.2 真实数据采集：主从机械臂、VR/AR操作、外骨骼和人体动捕

- **主从机械臂式遥操作：操作人员控制机器人完成任务，机器人传感器实时采集数据。**
  - 主臂各关节配有高精度编码器，人类拖动主臂时，系统实时读取主臂的关节角度或末端位姿，通过正逆运动学（Kinematics）映射，将其作为控制目标发送给从臂，同时记录“图像+从臂状态+主臂输入”。涉及装备：主臂、从臂
- **外骨骼与人体动捕式遥操作**
  - 通过传感器（惯性或光学）捕获人类全身的骨骼节点数据（如重心的移动、脚底的接触力），利用运动重定向（Motion Retargeting）算法，将人类的生物结构动作等比例、安全地映射到机器人的机械结构上。涉及装备：惯性动作捕捉服、光学动作捕捉系统。**主要用于人形机器人（Humanoid Robots）的全尺寸、全身体态数据采集。**
- **VR/AR 虚拟现实遥操作：**
  - 人类操作员佩戴 VR 头显，在虚拟或增强现实环境中看到机器人摄像头的实时回传画面（或三维点云重建环境），通过手柄进行远程控制。适合远程、危险环境或跨地域的数据采集。涉及装备：VR/MR 头显及手柄、数据手套、渲染和仿真引擎

图22：惯性捕捉和光学捕捉方案



#Dofs	14 (arms) + 2 (base)
Weight	75kg
Size	80W * 84L * 140H (no leaders)
	90W * 135L * 140H
Payload	750g (per arm) 55kg (base)
Arm Repeatability	1mm
Arm Accuracy	5-8mm
Battery life	12 hours (1620Wh)
Max pulling force	100N at 100cm vertically
Rolling resistance	13N (vinyl floor)

图23：Mobile ALOHA采用主从臂遥操作方式

资料来源：《Mobile ALOHA: Learning Bimanual Mobile Manipulation with Low-Cost Whole-Body Teleoperation》，诺亦腾官网，申万宏源研究

2.1.2 真实数据采集-遥操作方式起源

- **斯坦福大学与UC伯克利大学团队在论文《Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware》(2023)提出具身智能的低成本遥操作范式：**
  - **ALOHA硬件系统：**设计了一套成本仅约2万美元的开源、低成本、双臂主从遥操作硬件系统（Leader-Follower）。人类操作员可以通过摆动“主臂”（Leader），灵活、精确地远程控制“从臂”（Follower）采集高难度的双手协同动作。
  - **ACT算法（Action Chunking with Transformers）：**针对模仿学习中“行为克隆（Behavioral Cloning）”容易出现的误差累积（Drift）问题，论文提出了一种新颖的生成式模型算法。它利用Transformer结构，不是逐帧预测下一个动作，而是“成块（Chunking）”预测一段连续的动作序列。
  - **实验效果：**仅通过10分钟（约50次）的人类遥操作数据采集，结合ACT算法，机器人就能以80-90%的成功率自主完成诸如：打开半透明调味杯盖子、给遥控器插拔电池、撕小胶带等极为精细的双手任务。

图24：ALOHA硬件系统：主从遥操作系统

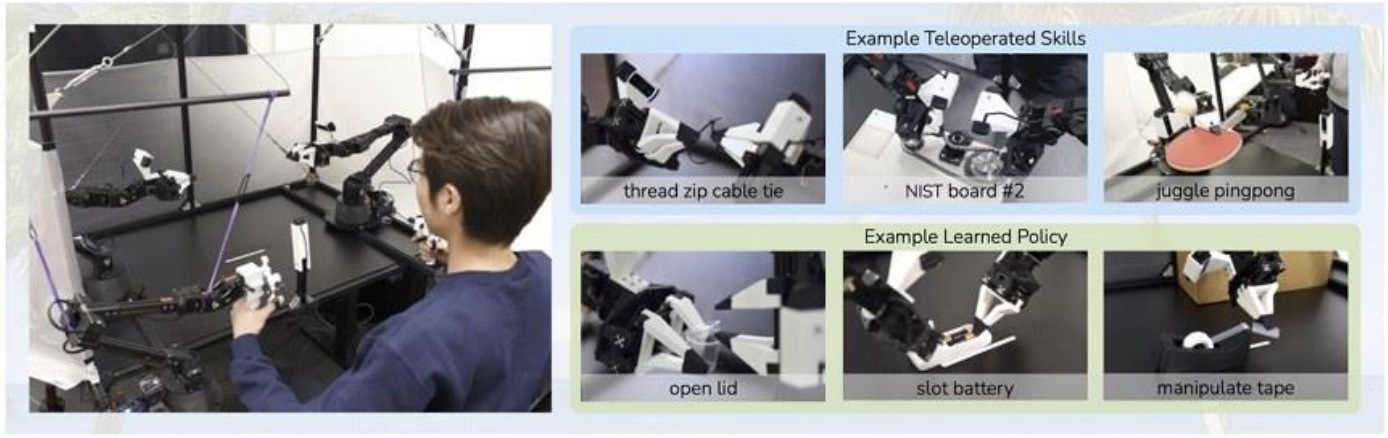


Fig. 1: *ALOHA : A Low-cost Open-source Hardware System for Bimanual Teleoperation*. The whole system costs <$20k with off-the-shelf robots and 3D printed components. *Left:* The user teleoperates by backdriving the leader robots, with the follower robots mirroring the motion. *Right: ALOHA* is capable of precise, contact-rich, and dynamic tasks. We show examples of both teleoperated and learned skills.

2.1.3 Ego数据：人类第一人称视角采集数据

- **Ego数据（一人称视角数据/主观视角数据）优势在于低成本、规模化、脱离机器人硬件、天然包含人类注意力机制、真实场景采集；缺点在于结构和视角不一致，数据转换难度大，力反馈和动态特征缺失，遮挡噪声等；**
  - 2021年：Meta联合全球科研机构推出Ego4D（23年推出Ego-Exo4D）大规模数据集，该数据集包含3670小时的日常活动视频，覆盖数百种场景（家庭、户外、工作场所、休闲娱乐等）
  - 2025-2026年，斯坦福大学 Chelsea Finn 团队利用可穿戴设备采集大量人类 Ego 视角的精细操作数据，通过自监督学习和跨具身技术，证明了“看人类做家务的Ego视频”可以显著提升机器人的泛化抓取能力；

图25：大规模第一视角数据集Ego4D和算法框架EgoZero

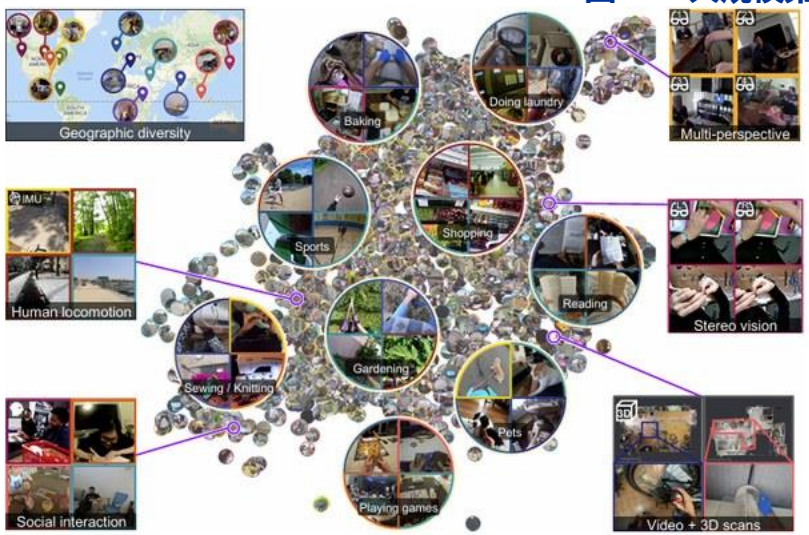


Figure 1. Ego4D is a massive-scale egocentric video dataset of daily life activity spanning 74 locations worldwide. Here we see a snapshot of the dataset (5% of the clips, randomly sampled) highlighting its diversity in geographic location, activities, and modalities. The data includes social videos where participants consented to remain unblurred. See https://ego4d-data.org/fig1.html for interactive figure.

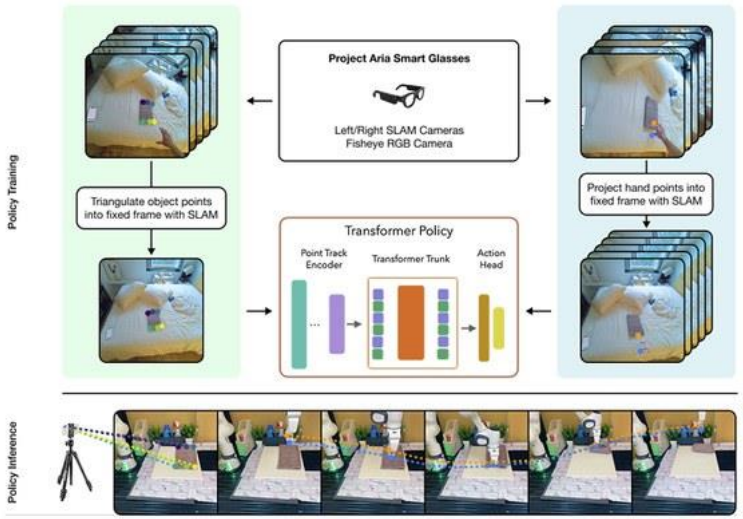


Figure 1: EGOZERO trains policies in a unified state-action space defined as egocentric 3D points. Unlike previous methods which leverage multi-camera calibration and depth sensors, EGOZERO localizes object points via triangulation over the camera trajectory, and computes action points via Aria MPS hand pose and a hand estimation model. These points supervise a closed-loop Transformer policy, which is rolled out on unprojected points from an iPhone during inference.

资料来源：《Ego4D: Around the World in 3,000 Hours of Egocentric Video》，《EgoZero: Robot Learning from Smart Glasses》，申万宏源研究

2.1.3 Ego数据：人类第一人称视角采集数据

- **Ego的使用情况和数据效果，参考英伟达发布《EgoScale: Scaling Dexterous Manipulation with Diverse Egocentric Human Data》**

1. 预训练阶段(Pre-training)
数据来源：包含20,854小时的人类第一人称视角视频（Egocentric Human Videos）。
主要任务：在这个阶段，基于流（flow-based）的视觉-语言-动作（VLA）策略首先在海量的人类视频上进行预训练。模型通过捕捉人类的手腕运动（wrist motion），并将其重新映射（retargeted）为灵巧手动作（dexterous hand actions），以此来学习人类物理交互的通用常识和运动先验。

2.中期训练阶段(Mid-training)
数据来源：包含50小时的人类数据+4小时的机器人数据，形成了人类与机器人对齐的游玩数据（Human-robot Aligned Play data）。
主要任务：将预训练阶段学到的“人类视觉与动作表征”，适配（adapt）到具体机器人的传感器输入和控制系统（robot sensing and control）中，从而跨越人和机器人的体征鸿沟（Embodiment Gap）。

3.后期训练与下游任务应用(Post-training)
通过前两个阶段得到的通用策略模型，在此阶段进行下游任务的部署和微调：
高效学习高灵巧任务(Efficient Learning on Highly Dexterous Tasks)：模型能够以极高的效率学习复杂的灵巧控制。
未知灵巧任务的单发泛化(One-shot on Unseen Dexterous Tasks)：得益于前期海量视频带来的泛化能力，模型甚至可以仅通过一次演示（One-shot），直接泛化去执行此前从未见过的全新技能：Unscrew Bottle：拧开瓶盖。Fold Shirt：折叠衣服。

图26：大规模第一视角人类视频对灵巧操作具备显著缩放效应

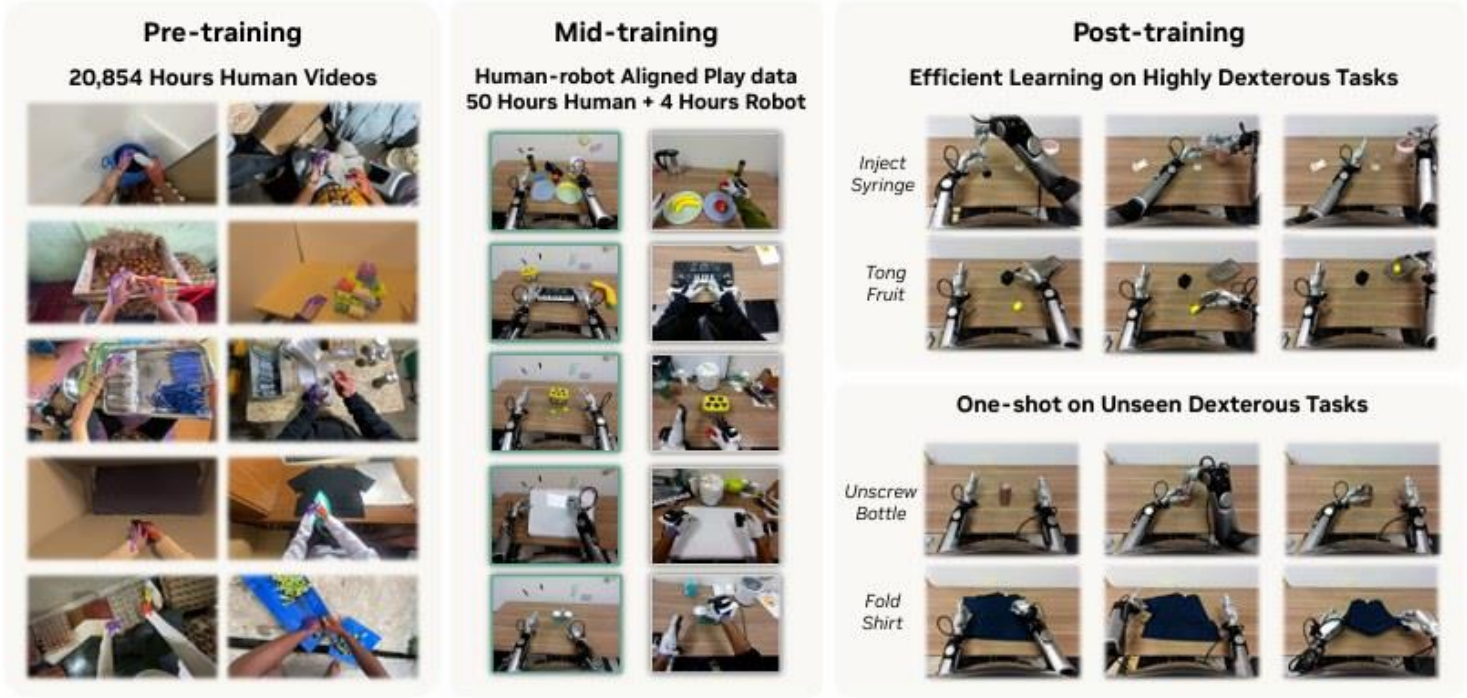


Figure 1: **EgoScale: Two-stage human-to-robot learning framework.** A flow-based Vision-Language-Action (VLA) policy is first pretrained on 20,854 hours of egocentric human videos using wrist motion and retargeted dexterous hand actions. A lightweight mid-training stage with aligned human robot play data (pairs highlighted with green and gray boundaries) adapts the representation to robot sensing and control. The resulting policy is post-trained on downstream tasks, enabling efficient learning of dexterous manipulation and one-shot generalization to unseen skills.

2.1.4 UMI数据：机器人动作端的执行数据

- **UMI数据：UMI（Universal Manipulation Interface通用操作接口），核心是用一只手持夹爪+腕部视角相机，采集与机器人末端同构的观测和动作数据。**
  - 数据采集设备是一个GoPro，主要采集两种模态视觉和IMU。 视觉捕捉腕部的图像，以及通过ArUco（一种类似于二维码的视觉标记）为标定测量的夹爪开合程度，基于ORB-SLAM3记录末端位姿。
  - **优势在于：**采集场景开放、本体无关、成本大幅降低
  - **缺点在于：**依赖视觉特征和slam稳定性、灵巧度受限、视角局限、力觉反馈缺失

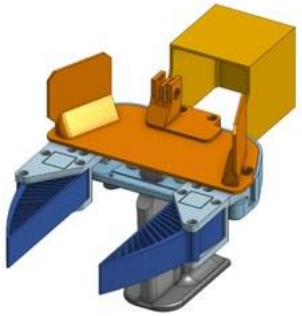


图27：手持机器人夹爪采集系统的数据采集演示

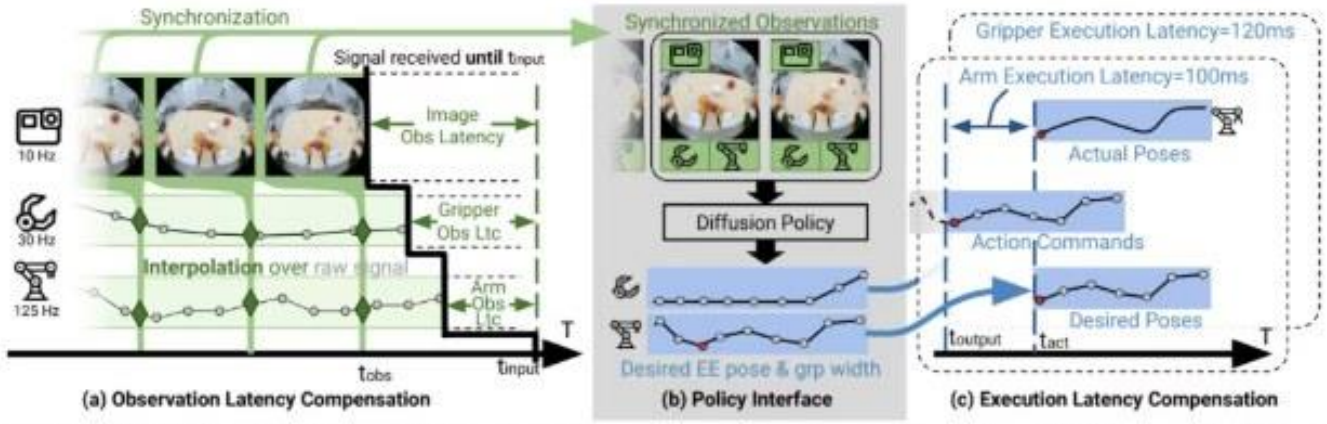


Fig. 5: **UMI Policy Interface Design.** (b) UMI policy takes in a sequence of **synchronized** observations (RGB image, relative EE pose, and gripper width) and outputs a sequence of **desired** relative EE pose and gripper width as action. (a) We synchronize different observation streams with physically measured latencies. (c) We send action commands ahead of time to compensate for robots' execution latency.

2.1.5 仿真合成数据：潜在规模大、边际成本低的最佳数据来源

- **仿真合成数据：依托物理仿真引擎在虚拟环境中自动生成的结构化数据，通过数字世界渲染相机图像、点云、物体位姿、机器人关节轨迹、力反馈、动作序列等信息；不需要真实物理机器人、真人现场采集，由程序自动批量生成。**
  - 优势：仿真数据不依赖人类时间的线性流逝，其规模呈现爆发式、指数级的增长。潜在数据规模极大、边际成本低，极端场景容易构建；可控环境，灵活域随机化。
  - 劣势：仿真物理、视觉和真实世界存在差异；纯仿真训练策略直接落地真机性能衰减；物体物理参数局限，如软体、布料、复杂接触摩擦仿真精度仍有上限；真实世界独有的随机噪声、材质特性很难完全复刻。
- **根据训练目标不同（如偏向视觉感知、强化学习运动控制或软体/流体交互），行业内主要使用以下几类工具：**

表5：仿真合成数据的工具

工具名称	主导/背景	核心优势与定位	适用场景
NVIDIA Isaac Sim / Isaac Lab	NVIDIA	基于 Omniverse 和 PhysX 5，具备顶级光追渲染与海量 GPU 并行吞吐，支持大规模场景生成。	自动驾驶、机械臂复杂装配、端到端视觉-动作（VLA）模型训练。
MuJoCo / MJX	Google DeepMind	经典的隐式欧拉接触力学求解器，MJX（JAX原生版本）极大弥补了其 GPU 并行算力。	人形/四足机器人步态控制（Locomotion）、高频接触的灵巧手抓取。
Genesis (Genesis World)	斯坦福/CMU/开源(2026最新)	基于 Taichi/CUDA 重新构建的统一多物理场引擎，速度比传统引擎快数十倍（高达数千万 FPS）。	刚体-软体-流体混合仿真、大模型自适应场景数据生成。
Gazebo / ODE	开源社区 / AWS	与 ROS/ROS 生态原生打通，符合工业机器人标准。	全栈机器人系统验证、导航、SLAM、激光雷达仿真。

资料来源：英伟达开发者官网，Google DeepMind网站，申万宏源研究

2.2 数据路线：多路线押注，融合方式为主

- **数据路线：**不同公司押注不同的数据路线，普遍采用多种数据类型、开源+内部数据集融合的方式。
- **其中：**1）真机遥操作数据仍是价值量最高的数据资产，即“数据金字塔”；2）ego数据凭借成本和规模化优势，已逐渐成为重要补充；3）以模型为投入中心的科技大厂，更侧重能够“scale up”的数据，例如英伟达主推仿真。

表6：机器人公司的数据路线和主要来源

公司	数据路线范式	代表模型	核心数据集
星海图	开源数据集 + 真机/仿真混合	G0 / G0 Plus / G0.5	Galaxea Open-World Dataset（自建开源全场景数据集）
小米	开源数据集 + 真机/仿真 + 跨模态迁移	Xiaomi-Robotics-0 / U0 / U1	开源机器人数据集 + 内部真机交互数据 + 自动驾驶迁移数据
阿里	内部真机数据 + 开源数据集	Qwen-VLA / Qwen-Robot Suite	达摩院内部机器人数据 + 开源数据集 + 长时序操作轨迹数据
字节	仿真数据 + 真机数据 + 强化学习	GR-3 / GR-RL	多模态机器人交互数据 + Seed 仿真数据 + 真机数据
智元	大规模遥操作真机数据 + 仿真数据	GO-1 / GO-2 / WITA-Omni	AgiBot World真机数据集（自建开源）+ 仿真数据
银河通用	大规模仿真数据 + 真机数据	GraspVLA / LDA-1B	大规模仿真抓取数据集 + 多机器人平台操作数据
千寻智能	开源数据集 + 内部真机 + 世界模型	Spirit v1.5 / Spirit v1.6	开源机器人数据集 + 内部真机数据 + 多场景交互数据
腾讯	第一人称视角数据 + 多模态交互数据	HY-Embodied-0.5-X	第一人称视角机器人数据 + 多模态交互数据
自变量	大规模机器人操作数据集（工业场景）	自变量具身VLA模型	大规模机器人操作数据集
谷歌DeepMind	开源联盟数据集 + 互联网视觉语言数据	RT-1 / RT-2 / Gemini Robotics 1.5	Open X-Embodiment（1M+ episodes，22种形态） + 互联网视觉语言数据 + 130k 自建操作数据
英伟达	仿真数据 + 遥操作 + 合成数据生成 + 人类视频学习	Isaac GR00T N1 / N1.5 / N1.6	大规模仿真数据 + 人类遥操作数据 + 第一人称视频 + 互联网数据
PI	大规模真实操作数据 + 接触丰富场景	π0	大规模机器人操作数据
SkillAI	多场景技能数据（模块化）	SkillAI 具身VLA模型	多场景机器人技能数据

2.3 数据采集环节的投资机会梳理

硬件：采集工具和传感器

- **遥操作套件：**以 Stanford ALOHA、UMI为代表的方案为例，相关硬件包括**高精度数据手套、VR 遥操作主端、手眼相机、机械臂、umi夹爪等。**组合成标准化、工业级量产、开箱即用的采集套件的硬件厂商具有价值。
- **传感器：**视觉传感器、力矩传感器、IMU和触觉传感器等，为数据采集的关键入口。

软件与算法层：物理仿真引擎、场景重建与运动学重定向

- **底层物理仿真引擎**（模拟碰撞、刚体动力学、接触力，输出机器人关节/末端 Action 信号；提供虚拟相机渲染观测图像）；
- **基于3D场景构建工具。**

数据集提供商和服务商

- **1）机器人数据采集与评测外包服务商：**这类公司自建机器人采集基地，专门为机器人公司批量生产、清洗并交付的轨迹数据集；
- **2）垂直领域物理数据公司：**虽然英伟达有 Cosmos 和 Isaac 平台，但针对特定垂直场景（如半导体超洁净室组装、精密汽车零部件打磨、医院长时序看护），依然需要行业定制化的物理求解器和程序化内容生成平台。

主要内容

1. 具身模型：负责决策的核心大脑
2. 数据采集：模型训练的关键燃料
3. 相关标的梳理
4. 风险提示

3. 相关标的梳理

- **人形机器人进入具身智能驱动的新阶段，投资机会分布于本体、模型、数据三大链条。本体零部件是硬件基础，量产落地带动减速器、伺服、传感器需求释放；具身大模型作为机器人智能核心，算法与仿真平台构筑软件壁垒；训练数据是模型迭代核心约束，真实示教采集+第一视角采集+仿真合成数据等多路线并行，数据供给端稀缺性持续提升。中长期建议沿三条主线布局：①人形机器人核心零部件厂商；②具身模型和本体厂商；③机器人数据采集硬件与数据服务相关标的。**

表7：重点公司估值表

公司代码	公司简称	相关业务布局	2026/8/5	归母净利润（百万）				PE		
			市值/亿元	25A	26E	27E	28E	25A	26E	27E	28E
688585	上纬新材	本体	644.5	41.1	/	/	/	1569	/	/	/
688322	奥比中光	视觉传感器+数采	450.1	127.9	295.7	499.4	750.9	352	152	90	60
688071	华依科技	IMU数采	31.7	-61.7	27.0	69.0	/	-	117	46	/
688507	索辰科技	仿真软件	155.5	31.5	76.4	101.2	112.3	494	204	154	139
688400	凌云光	光学动捕	222.8	161.4	362.8	358.8	470.4	138	61	62	47
603466	风语筑	数据采集	67.7	-18.4	108.7	124.2	150.1	-	62	55	45
02513	智谱	具身模型	4195.8	-1918.5	-4344.5	-4090.2	-1699.4	-	-	-	-
09880	优必选	本体+模型	387.6	-703.2	-269.3	62.2	567.7	-	-	623	68
02432	越疆	本体	99.0	-83.5	-105.3	-43.3	34.3	-	-	-	289
06600	卧安机器人	本体+模型	132.2	-24.7	10.9	76.0	167.5	-	1213	174	79
02590	极智嘉-W	本体+数据	120.5	25.5	189.2	428.4	542.5	473	64	28	22
00068	群核科技	仿真合成平台	139.9	-427.9	104.0	188.0	306.0	-	135	74	46
06651	五一视界	仿真合成平台	245.8	-181.8	-136.0	-6.0	60.0	-	-	-	410

资料来源：iFinD、申万宏源研究　注：盈利预测为iFinD一致预测；港股公司单位为港币，负值PE用-替代显示

主要内容

1. 具身模型：负责决策的核心大脑
2. 数据采集：模型训练的关键燃料
3. 相关标的梳理
4. 风险提示

4. 风险提示

1、技术迭代不及预期风险：

- 当前具身VLA模型、世界模型技术仍处于早期迭代阶段，普遍存在仿真与现实域偏差、人机本体迁移鸿沟等问题。模型泛化能力、实时推理效率难以适配复杂真实场景，若算法架构、训练范式突破速度放缓，将导致产品落地效果不及预期，行业商业化进程面临延期风险。

2、数据供给与质量瓶颈风险：

- 具身智能模型高度依赖高质量、场景匹配的机器人交互数据。目前行业真机标注数据稀缺、采集成本高昂，仿真数据存在真实性缺陷，第一人称Ego数据存在本体适配偏差。数据规模化、标准化能力不足，将持续制约模型训练效果，成为行业发展核心瓶颈。

3、商业化落地进度不及预期风险：

- 行业技术与数据投入成本高，短期难以实现盈利闭环，叠加场景碎片化、落地验证周期长，存在商业化落地进度慢、投入产出不及预期的风险。

信息披露

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、专业审慎的研究方法，使用合法合规的信息，独立、客观地出具本报告,并对本报告的内容和观点负责。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。

与公司有关的信息披露

本公司隶属于申万宏源证券有限公司。本公司经中国证券监督管理委员会核准，取得证券投资咨询业务许可。本公司关联机构在法律许可情况下可能持有或交易本报告提到的投资标的，还可能为或争取为这些标的提供投资银行服务。本公司在知晓范围内依法合规地履行披露义务。客户可通过compliance@swsresearch.com索取有关披露资料或登录www.swsresearch.com信息披露栏目查询从业人员资质情况、静默期安排及其他有关的信息披露。

机构销售团队联系人

华东团队	茅炯	021-33388488	maojiong@swhysc.com
华北团队	肖霞	15724767486	xiaoxia@swhysc.com
华南团队	王维宇	0755-82990590	wangweiyu@swhysc.com
华北创新团队	潘烨明	15201910123	panyeming@swhysc.com
华东创新团队	朱晓艺	18702179817	zhuxiaoyi@swhysc.com

股票投资评级说明

证券的投资评级：
以报告日后的6个月内，证券相对于市场基准指数的涨跌幅为标准，定义如下：
买入（Buy）：相对强于市场表现20%以上；
增持（Outperform）：相对强于市场表现5%～20%；
中性（Neutral）：相对市场表现在－5%～＋5%之间波动；
减持（Underperform）：相对弱于市场表现5%以下。
行业的投资评级：
以报告日后的6个月内，行业相对于市场基准指数的涨跌幅为标准，定义如下：
看好（Overweight）：行业超越整体市场表现；
中性（Neutral）：行业与整体市场表现基本持平；
看淡（Underweight）：行业弱于整体市场表现。
我们在此提醒您，不同证券研究机构采用不同的评级术语及评级标准。我们采用的是相对评级体系，表示投资的相对比重建议；投资者买入或者卖出证券的决定取决于个人的实际情况，比如当前的持仓结构以及其他需要考虑的因素。投资者应阅读整篇报告，以获取比较完整的观点与信息，不应仅仅依靠投资评级来推断结论。申银万国使用自己的行业分类体系，如果您对我们的行业分类有兴趣，可以向我们的销售员索取。
本报告采用的基准指数：沪深300指数（A股）、恒生中国企业指数（H股）、纳斯达克指数（美股）

法律声明

本报告由上海申银万国证券研究所有限公司（隶属于申万宏源证券有限公司，以下简称“本公司”）在中华人民共和国内地（香港、澳门、台湾除外）发布，仅供本公司的客户（包括合格的境外机构投资者等合法合规的客户）使用。本公司不会因接收人收到本报告而视其为客户。有关本报告的短信提示、电话推荐等只是研究观点的简要沟通，需以本公司http://www.swsresearch.com网站刊载的完整报告为准，本公司并接受客户的后续问询。本报告首页列示的联系人，除非另有说明，仅作为本公司就本报告与客户的联络人，承担联络工作，不从事任何证券投资咨询服务业务。

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、工具、意见及推测只提供给客户作参考之用，并非作为或被视为出售或购买证券或其他投资标的的邀请或向人作出邀请。本报告所载的资料、意见及推测仅反映本公司于发布本报告当日的判断，本报告所指的证券或投资标的的价格、价值及投资收入可能会波动。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。

客户应当考虑到本公司可能存在可能影响本报告客观性的利益冲突，不应视本报告为作出投资决策的惟一因素。客户应自主作出投资决策并自行承担投资风险。本公司特别提示,本公司不会与任何客户以任何形式分享证券投资收益或分担证券投资损失，任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。本报告中所指的投资及服务可能不适合个别客户，不构成客户私人咨询建议。本公司未确保本报告充分考虑到个别客户特殊的投资目标、财务状况或需要。本公司建议客户应考虑本报告的任何意见或建议是否符合其特定状况，以及（若有必要）咨询独立投资顾问。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。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所引致的任何损失负任何责任。市场有风险，投资需谨慎。若本报告的接收人非本公司的客户，应在基于本报告作出任何投资决定或就本报告要求任何解释前咨询独立投资顾问。

有信仰 敢担当
守正创新 追求卓越

申万宏源研究微信订阅号

申万宏源研究微信服务号

上海申银万国证券研究所有限公司
(隶属于申万宏源证券有限公司)

胡书捷
husj@swsresearch.com