人形机器人系列专题之具身模型

RT-2→OpenVLA→π系列的架构演进

行业研究 · 行业专题

汽车 · 汽车零部件

投资评级：优于大市（维持）

证券分析师：唐旭霞
0755-81981814
tangxx@guosen.com.cn
S0980519080002

证券分析师：孙树林
0755-81982598
sunshulin@guosen.com.cn
S0980524070005

证券分析师：傅嘉良
0755-22940076
fujialiang@guosen.com.cn
S0980526100002

请务必阅读正文之后的免责声明及其项下所有内容

核心观点

- **科技大厂加速布局具身模型，通用模型能力持续向物理世界延伸。**基础模型厂商、科技巨头依托在通用模型开发、多模态理解、算力平台与模型工程化方面的积累，持续向具身智能领域布局拓展，为机器人处理开放式指令和复杂场景提供更强基础。Google以Gemini Robotics ER 2与Robotics 2.5衔接高层推理和动作执行；NVIDIA以Cosmos、GR00T 2和Isaac Lab推进世界预测、动作生成与仿真训练协同；OpenAI GPT-6 Astra已被机器人公司用于探索机器人动作生成及高层规划；Anthropic Fable 5.2在仿真双臂评测和真机操作中能力明显提升；Apple通过HALO衔接任务规划、全身运动生成与运动跟踪，探索人形机器人分层自主执行。我们认为，科技大厂持续入局有望加速具身智能技术突破，大厂的通用模型、算力平台与工程化能力为具身模型研发提供更强支撑，有望加快模型迭代及机器人应用落地。
- **VLA模型将视觉观测与语言指令转化为机器人动作，成为具身模型重要技术路线。**VLA是视觉-语言-动作模型，将环境理解和任务目标转化为动作输出，当前已成为具身模型的重要架构。VLA模型可分为一段式和分层架构，一段式架构在统一模型内完成理解与动作预测，强调统一表征和联合学习；分层架构通过子任务或中间目标连接高层规划与低层执行，强调任务分解与模块协同。
- **通用模型具有较强的高层规划能力，推动具身模型S2/S1/S0分层协同。**面向复杂、长时序任务，分层协同有助于兼顾高层推理与实时响应，提升技能复用和执行调整能力。通用大模型在语义理解、任务分解与复杂推理方面的能力，为System 2层理解任务目标、规划操作步骤和调整执行策略提供支撑；System 1层将高层目标转化为具体动作，System 0层负责实时运动控制，形成从通用认知到物理执行的协同链路。我们认为，随着通用模型能力持续提升，高层规划与动作执行分工的价值将进一步显现，推动S2S1S0的分层架构在复杂、长时序任务中加快应用，兼顾任务泛化、技能复用与实时响应，拓展机器人自主作业的能力边界。
- **π系列VLA模型分层架构逐步完善，高层规划与动作执行分工更加清晰，提升复杂任务自主执行能力。**π系列是Physical Intelligence推出的代表性VLA模型，我们深度研究了π模型的持续迭代过程，梳理其在模型架构、训练方法与任务能力上的演进：π0采用视觉语言主干+动作专家架构，分别负责理解场景与指令、生成机器人动作；π0.5增加子任务预测，将任务拆解与动作生成结合；π*0.6在训练环节引入价值模型，帮助动作策略学习更有效的操作；π0.7进一步配置独立的高层规划模型与世界模型，通过子任务指令和目标图像引导动作生成。我们认为，π系列呈现出分层架构逐步完善的演进方向，高层规划与动作执行的分工更加清晰，并通过世界模型提供目标引导，兼顾复杂任务规划与具体动作生成，提升机器人自主执行能力。
- **风险提示：**具身智能模型进展不及预期、具身机器人出货量不及预期、具身数据采集进度低于预期、具身模型技术路线变化、行业竞争加剧。

请务必阅读正文之后的免责声明及其项下所有内容

目录

01 具身模型：科技大厂加速布局，通用模型向具身拓展

02 架构解析：一段式与分层架构，复杂任务凸显分层协同价值

03 VLA模型发展脉络：从RT-2到π系列的架构与能力升级

04 风险提示

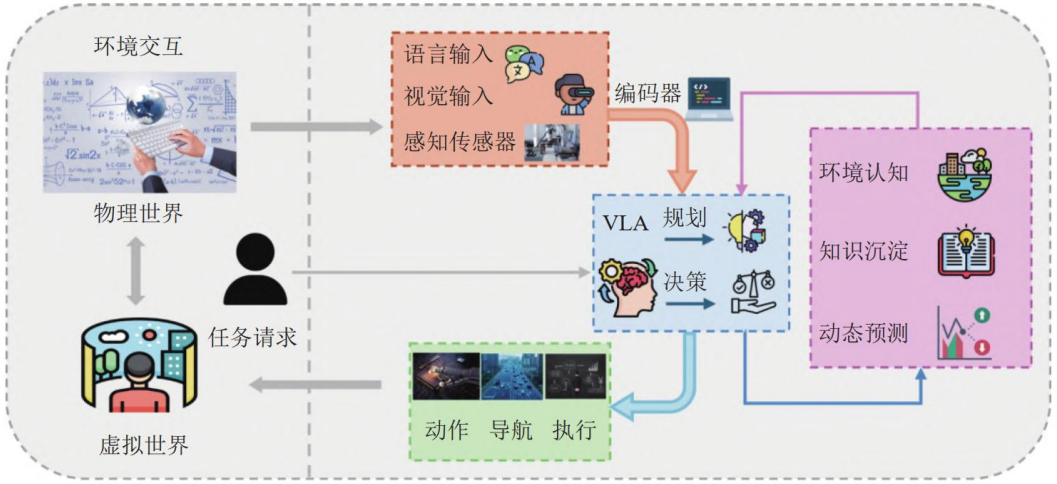
请务必阅读正文之后的免责声明及其项下所有内容

具身模型：机器人智能化的核心支撑，VLA成为重要技术路线

国信证券 GUOSEN SECURITIES

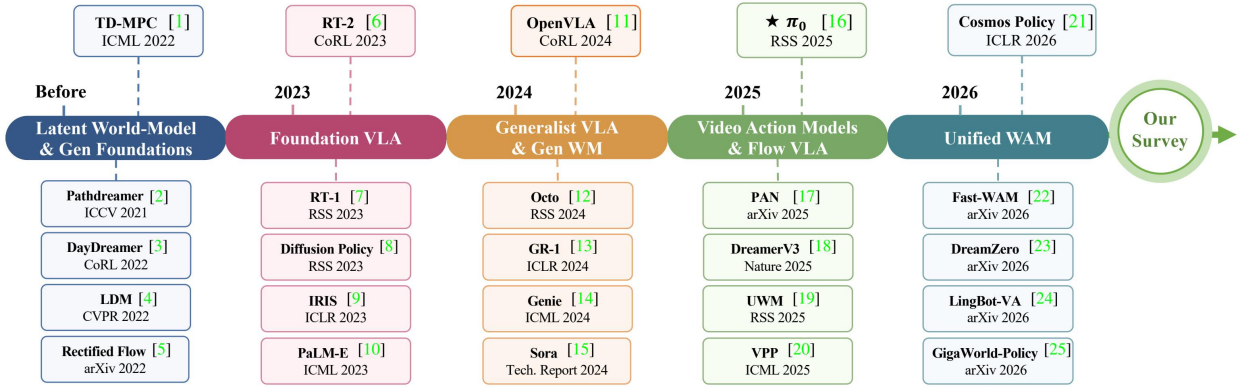
- **具身模型是推动机器人从传统自动化执行向自主感知、认知与决策升级的核心技术，也是实现机器人智能化的关键。**作为机器人系统的认知与决策中枢，具身模型承担环境语义理解、自然语言解析、长时序任务规划及多模块协同调度等高阶认知任务和控制功能，核心在于将人类开放式语言指令与多模态传感器获取的环境信息相结合，转化为可执行的任务规划与机器人动作，实现从人类意图理解到物理世界交互的完整闭环。
- **VLA模型（Vision-Language-Action，视觉-语言-动作）成为具身模型的重要技术路线之一**。随着ChatGPT等大语言模型展现出强大的语义理解与推理能力，研究者开始探索将大语言模型（LLM）的语义推理、视觉语言模型（VLM）的环境感知与机器人控制系统的动作生成能力进行深度融合。2023年，Google DeepMind提出Robotic Transformer 2（RT-2），将视觉、语言与机器人动作统一编码为Token，并纳入同一Transformer架构，将机器人动作预测转化为自回归序列建模任务，实现了视觉语言预训练知识向机器人控制能力的迁移，显著提升了机器人对陌生物体和新任务的零样本泛化能力，推动了视觉-语言-动作（VLA）架构的发展。

图1：具身智能系统架构



资料来源：刘国华等，《跨模态交互驱动的具身智能：视觉-语言-动作融合模型、数据与平台综述》，《集成技术》，2026，15(3)：1-43；国信证券经济研究所整理

图2：具身模型发展历程



资料来源：Zuxing Lu等，《World-Action Models for Robot Learning and Control: A Survey》，arXiv，2026，1－19；国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

大厂布局：科技大厂积极入局具身，通用模型向物理世界加速拓展

科技大厂加速布局具身智能，技术投入覆盖通用推理、动作生成及仿真训练。我们认为，大厂积累的多模态能力与开发工具有望降低机器人技能开发成本，真实交互数据和持续运行经验将进一步提高模型在复杂场景中的任务完成能力。

- OpenAI的GPT-6 Astra被银河通用等机器人公司用于接入机械臂，尝试直接生成动作及协同VLA控制。
- Google以Gemini Robotics ER与Robotics 2.5衔接高层推理和动作执行，利用Gemini 4 Argon辅助数字孪生构建和仿真实验。
- NVIDIA以Cosmos、GR00T 2和Isaac Lab构建模型与训练工具体系。
- Anthropic展示Fable 5.2在机器人操作中的能力。
- Apple通过HALO探索视觉感知、全身运动生成和运动跟踪的分层系统。

表1：部分科技大厂的具身智能模型与技术布局

公司	代表模型或项目	具身方向与公开进展
Google	Gemini Robotics ER、Robotics 2.5、Argon	高层推理与动作模型协同，通用模型辅助仿真研发
NVIDIA	Cosmos、GR00T 2、Isaac Lab	世界预测与动作生成结合，配套仿真训练、数据处理及评测工具
Anthropic	Fable 5.2	展示通用模型参与真机操作，并公布仿真双臂机器人评测结果
OpenAI	GPT-6 Astra	银河通用团队研究其直接动作生成，以及与π0.5协同控制的效果
Apple	HALO	通过任务规划、视觉条件运动生成和运动跟踪，实现人形机器人全身动作

资料来源：Google DeepMind、NVIDIA、Anthropic、Apple；Xuchuan Chen等，《Systematically Exploring the Capabilities of GPT-6 Astra as Embodied Policies》，arXiv，2026，1-46；国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

OpenAI GPT-6 Astra：空间理解与具身推理能力提升

- **通用模型空间理解与具身推理表现提升，为机器人理解环境和任务提供基础。**在宇树公布的评测中，GPT-6 Astra在空间指代定位、放置位置定位和具身推理问答三项测试中的得分分别达到79.6、69.0和77.7，较GPT-5.6-sol分别提升36.5%、35.0%和19.9%，相比前代模型体现出更强的目标定位、空间关系判断与操作场景理解能力。
- **我们认为，基础模型厂商的持续入局有望加速具身模型研发与应用拓展。**Astra在空间理解与具身推理评测中的进步，体现了通用模型为机器人提供更强认知能力的潜力，推动具身技术加速迭代。机器人企业可利用基础模型已有的知识、理解与推理能力，减少任务识别和规划环节的重复开发，并结合自身动作模型与控制系统，加快新技能开发和场景适配。基础模型的持续迭代有望进一步带动具身模型进步。

表2：GPT模型在空间理解与具身推理评测中的表现

评测	GPT-5.6-sol	GPT-6-Astra	提升
RefSpatial 空间指代定位	58.3	79.6	+36.5%
Where2Place 物体放置判断	51.1	69.0	+35.0%
ERQA 具身推理问答	64.8	77.7	+19.9%

资料来源：宇树项目官网，国信证券经济研究所整理

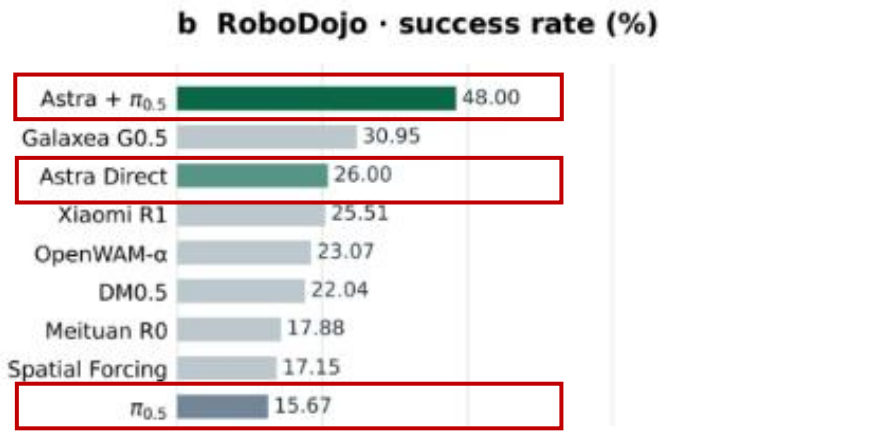
请务必阅读正文之后的免责声明及其项下所有内容

OpenAI GPT-6 Astra：展现机器人操作能力，与VLA协同提升复杂任务表现

国信证券 GUOSEN SECURITIES

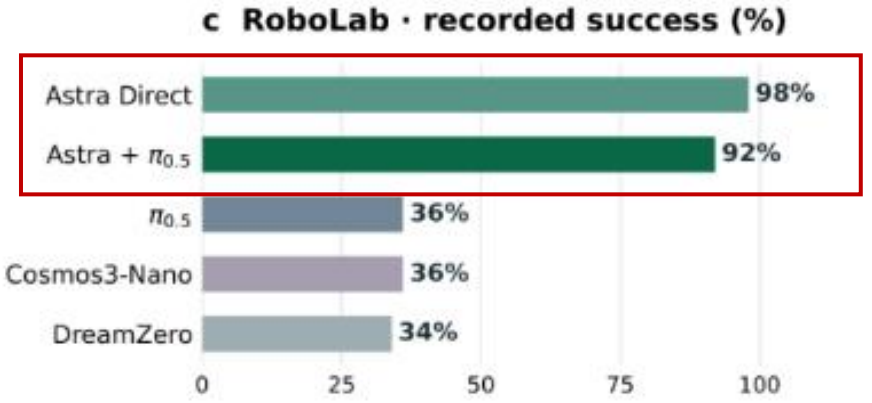
- **RoboDojo实验：通用大模型Astra与VLA模型有互补价值。**RoboDojo实验设置了两种Astra控制机器人的方案，包括Astra Direct直接生成机器人动作、Astra与机器人VLA模型π0.5协同（π0.5提供候选动作，Astra结合视觉信息、机器人状态及执行历史，**判断动作合理性并进行局部修正**）。结果显示，Astra Direct在50次任务中成功13次，成功率为26%，48条有效评分记录的平均得分为37.81，在测评模型中排名第3。**Astra+π0.5的成功次数提高到24次，成功率达到48%**，平均得分升至62.60，排名上升至第1。
- **RoboLab实验：通用大模型Astra具备零样本泛化能力。**RoboLab实验选取了10类任务、合计50次评测，**GPT-6 Astra Direct在50次实验中成功49次，成功率98%；Astra+π0.5完成46次，成功率92%，成功率远超其他模型**。同时，**Astra没有针对任务进行额外训练**，模型根据当前画面、机器人状态和自然语言任务描述，持续判断下一步应该如何操作，模型能够识别目标、理解空间关系，将语言要求转化为动作，并在环境变化后继续调整策略。

图3：GPT-6 Astra在RoboDojo中的任务成功率



资料来源：Xuchuan Chen等，《Systematically Exploring the Capabilities of GPT-6 Astra as Embodied Policies》，arXiv，2026，1-46；国信证券经济研究所整理

图4：GPT-6 Astra在RoboLab中的任务成功率



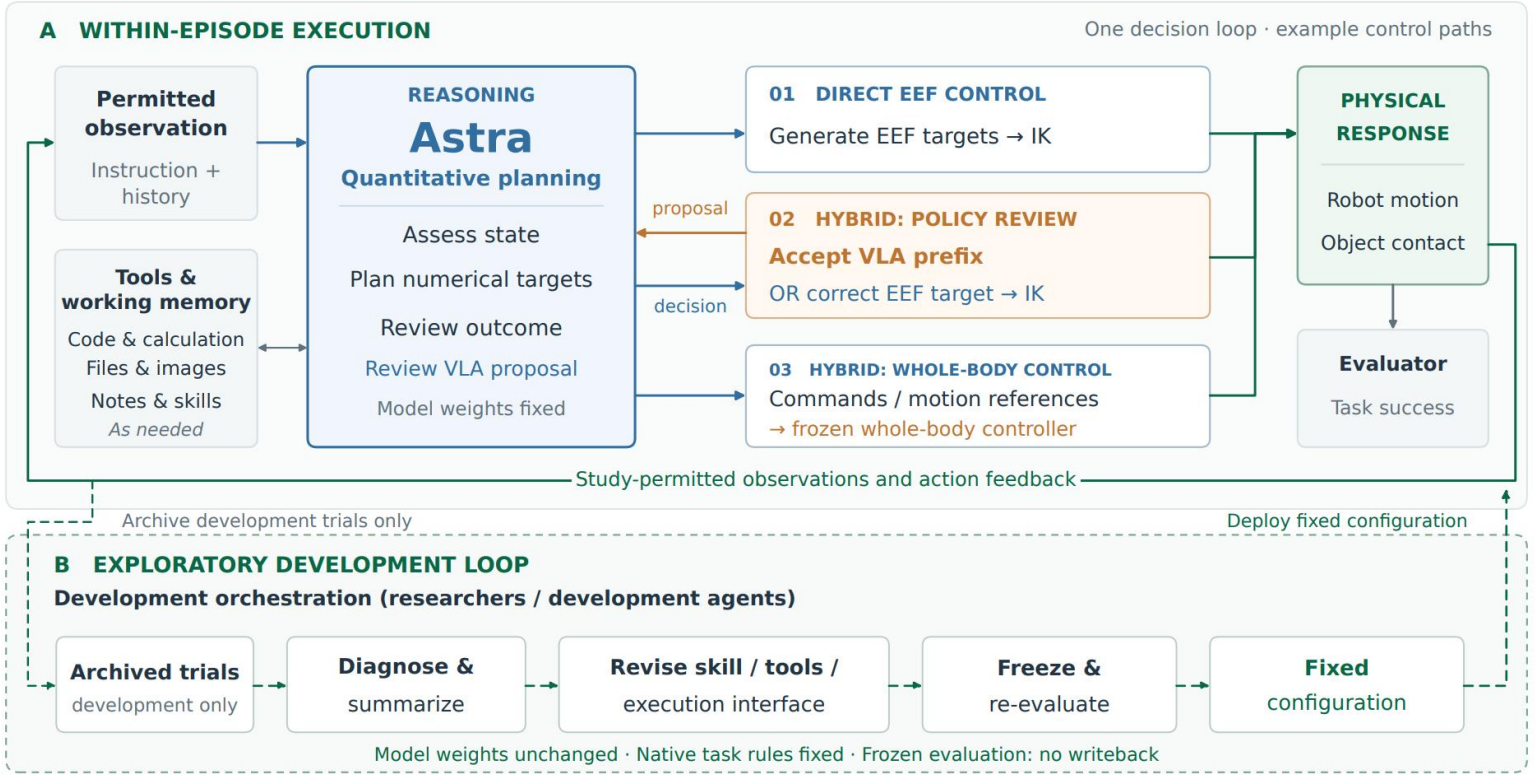
资料来源：Xuchuan Chen等，《Systematically Exploring the Capabilities of GPT-6 Astra as Embodied Policies》，arXiv，2026，1-46；国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

OpenAI GPT-6 Astra：Astra与π0.5协同机制

- **第一步：**V+L→关节角度。π0.5先提出未来50步动作（关节角度）。
- **第二步：**关节角度→位姿。调用正运动学的计算模块，把50步翻译成Astra容易审查的末端位姿。
- **第三步：**Astra分别检查上一步结果、下一段动作方向。输入是任务要求和当前进度、上一步执行前后的图像、机器人状态和执行记录、π0.5提出的动作。输出是对上一步结果的判断、下一段动作的判断、执行选择、执行参数（末端位置、朝向等等）。
- **第四步：**Astra从三种执行方式中选一种，执行动作。
- **第五步：**执行后，astra和π0.5都基于新状态进入下一轮。

图5：GPT-6 Astra与π0.5的协同控制流程



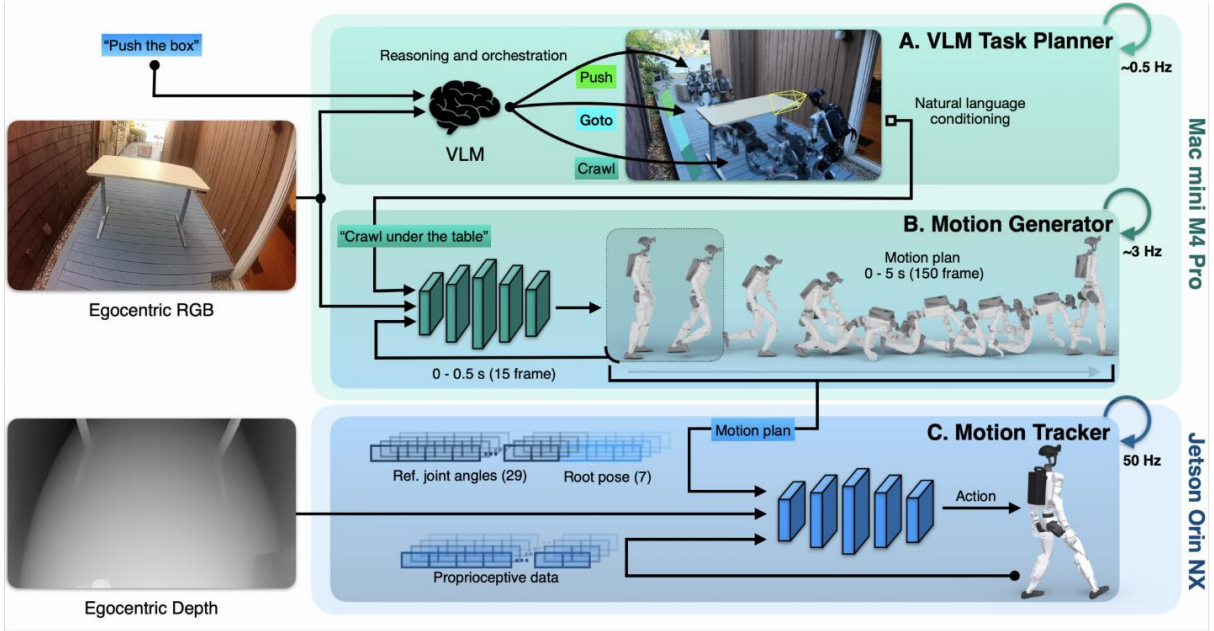
资料来源：Xuchuan Chen等，《Systematically Exploring the Capabilities of GPT-6 Astra as Embodied Policies》，arXiv，2026，1-46；国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

苹果：提出HALO分层模型架构，核心展示locomotion能力

- **模型架构：**采用分层架构，VLM任务规划器＋VLM－DiT运动生成器＋强化学习运动跟踪器。任务规划器默认采用Qwen3.6-35B-A3B，负责任务理解、子任务拆解和执行进度判断；运动生成器使用Qwen3.5-2B理解图像与子任务文字，通过DiT生成全身参考运动，承担主要的VLA功能；底层运动跟踪器结合参考运动、深度图、关节状态和IMU，持续调整关节目标，处理平衡、接触和执行偏差。从功能分工看，可类比为S2任务规划、S1运动生成、S0运动控制。
- **训练数据：**核心数据在仿真中生成，人工设计任务模板、路径和关键姿态→利用Kimodo等生成动作→利用Infinigen生成房间→将动作与目标物体放入场景→渲染机器人第一视角画面。核心配对语料包含约16.3万条片段、296.3小时数据，覆盖6类技能、10条训练指令。主要数据生成使用250张H100，耗时约48小时。
- **和Astra+π0.5的区别：**Astra+π0.5的范式主要完成的任务是细粒度的操作控制（manipulation），比如用机械臂完成抓取、摆放等任务；而苹果HALO主要是移动控制（locomotion），比如行走、俯身、爬行、坐下等。

图6：HALO分层架构及感知、规划与运动控制流程



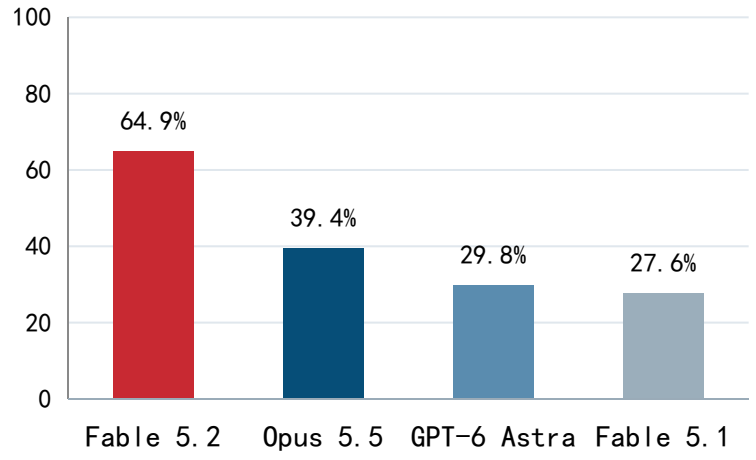
资料来源：Siyuan Hu等，《Perceptive Humanoid Motion from Simulation》，arXiv，2026，1-20；国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

Anthropic Claude Fable5.2：具身任务表现显著提升，输出提速、调用降本

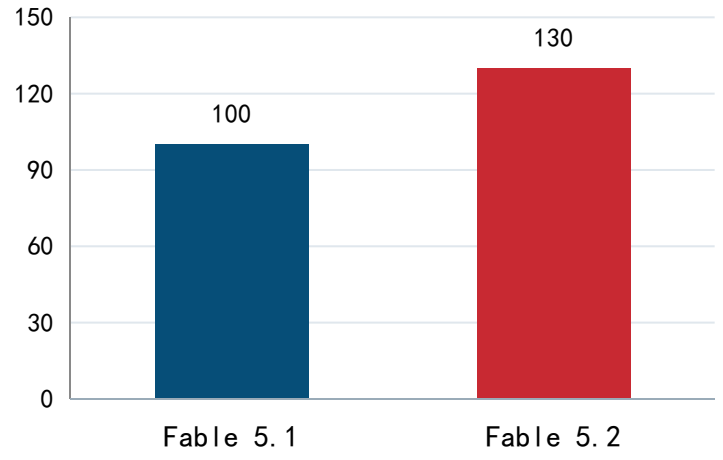
- **任务表现：具身任务表现显著提升。**①仿真评测中机器人操作平均得分显著提升，在Anthropic内部开展的RoboDojo Sim评测中，Fable 5.2根据摄像头图像控制仿真双臂机器人，平均得分达到64.9%，较Fable 5.1的27.6%提高37.3个百分点，为GPT-6 Astra的29.8%的约2.18倍。参与比较的前沿模型采用同一套运行框架及各自最高推理强度，Fable 5.2使用Max配置。②真机操作能力强。Anthropic同时展示了Fable 5.2控制实体机器人完成解谜方块、拧开瓶盖和绘制机器人图案，官方称这些任务未经明确的专项训练。
- **输出速度：文本输出速度最高提升30%，生成效率进一步提高。**Fable 5.2相较Fable 5.1的输出生成速度最高提升30%。
- **使用成本：标准API输入、输出价格均下降20%，模型调用成本进一步降低。**Fable 5.2的标准输入价格由Fable 5.1的每百万tokens从10美元降至8美元，标准输出价格由50美元降至40美元，两项降幅均为20%。
- **模型运作方式：**Fable 5.2以摄像头图像作为视觉输入，依据视觉信息形成操作决策，再驱动机器人执行动作。在RoboDojo Sim测试中，控制对象为仿真双臂机器人，参与比较的前沿模型使用同一套harness（运行与评测框架）。

图7：RoboDojo Sim平均得分（%）



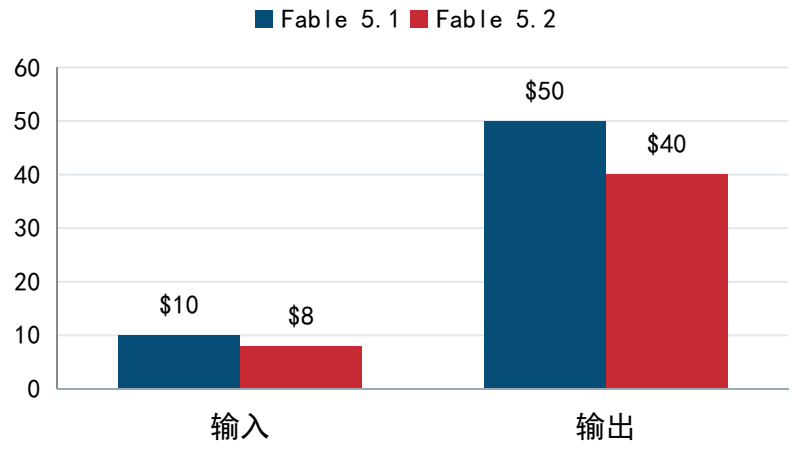
资料来源：Anthropic官网，国信证券经济研究所整理

图8：文本输出速度（Fable 5.1＝100）



资料来源：Anthropic官网，国信证券经济研究所整理

图9：标准API价格（美元/百万tokens）



资料来源：Anthropic官网，国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

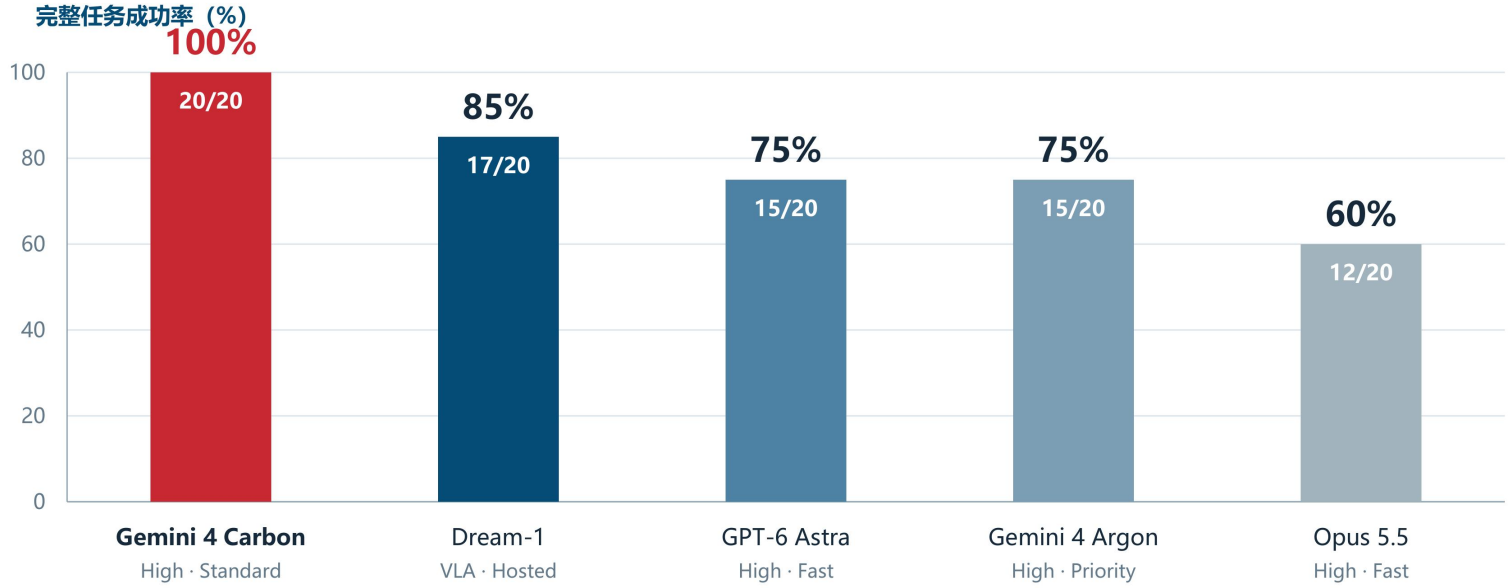
谷歌Gemini 4 Carbon：通用大模型向具身拓展

国信证券 GUOSEN SECURITIES

Gemini 4 Carbon基于通用多模态模型Argon进一步迭代，面向更复杂的推理任务，通常使用更多推理tokens。

- **输入输出：**读取多视角画面和机器人状态，输出带数值参数的动作工具调用。在已公开的叠碗运行记录中，输入包括任务指令、顶部及左右腕三路摄像头图像，以及关节位置、末端位姿、夹爪状态和估计关节扭矩。Carbon输出工具调用，**指定机械臂末端的目标位置、姿态和夹爪开合**；执行系统通过逆运动学转换为关节运动，并将更新后的观测回传模型。
- **具身任务表现：**控制真实双臂，叠碗实测20次全部成功。在Robocurve的RoboDojo-RC Tier 1评测中，Carbon在“三个碗堆叠”任务上取得20/20、100%的完整成功率，Argon为15/20、75%。覆盖六项任务、合计120轮试验时，Carbon的平均进度分为54.5%，较Argon的48.2%提高6.3个百分点；按任务完整完成统计，Carbon成功29轮，完整成功率约24.2%。
- **执行效率：**通过多轮观察与动作调整推进任务，真机操作耗时仍较长。含模型推理、工具执行和机械臂运动的端到端任务耗时：Carbon的叠碗任务平均耗时约27.7分钟/轮；六项任务合计120轮的平均耗时约32.1分钟/轮。

图10：RoboDojo-RC三碗堆叠任务的完整成功率



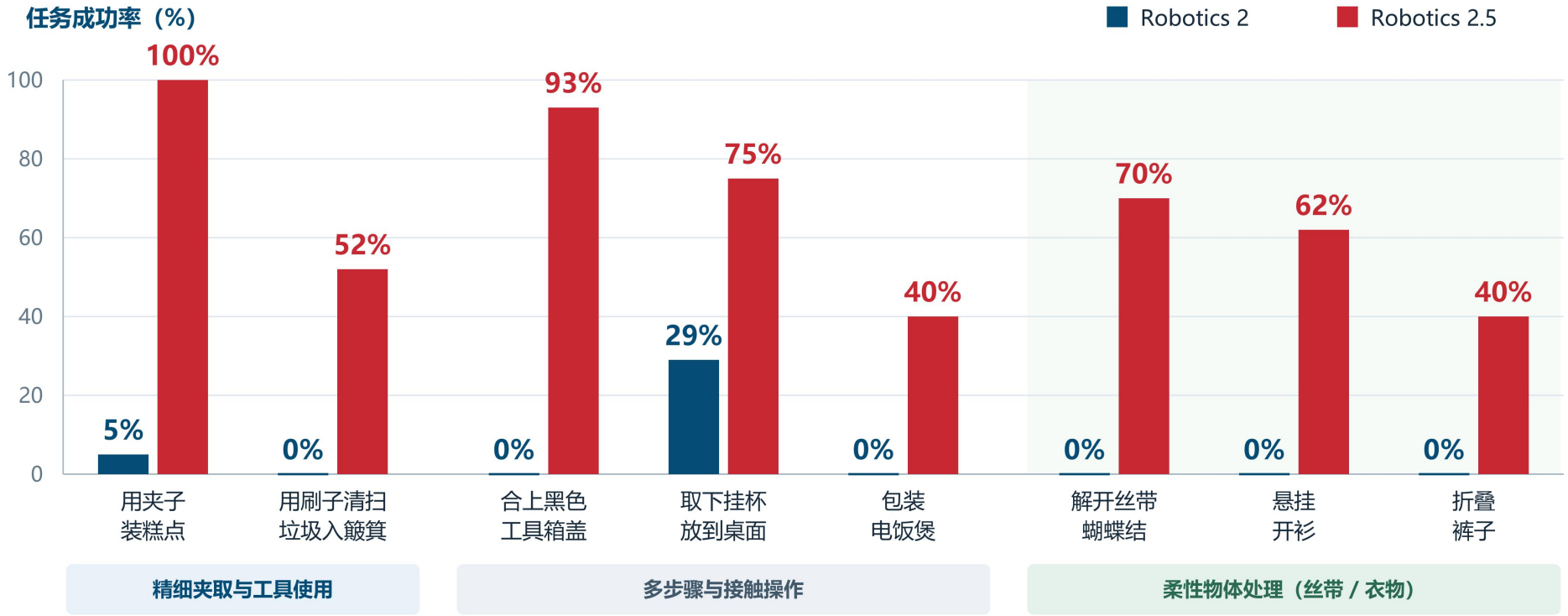
资料来源：Robocurve，国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

谷歌Gemini Robotics 2.5：精细操作与柔性体接触能力提升

- **任务表现：**覆盖200多项机器人任务，推进灵巧操作与工业应用。Google展示了使用夹具取放食物、系蝴蝶结、移液等操作，**精细操作和软物体处理方面有明显提升**。在奔驰生产线的概念验证中，Apptronik Apollo 2人形机器人完成汽车零部件分拣与装盘，成功率达到96%。
- **执行效率：**任务成功率与完成速度同步提升。相对此前基线，Google报告不同类型任务的成功率明显提高，任务完成速度提高20%—30%。这些指标对应机器人操作任务的表现与执行效率。

图11：Gemini Robotics 2与2.5的典型任务成功率对比



资料来源：Google DeepMind，国信证券经济研究所整理

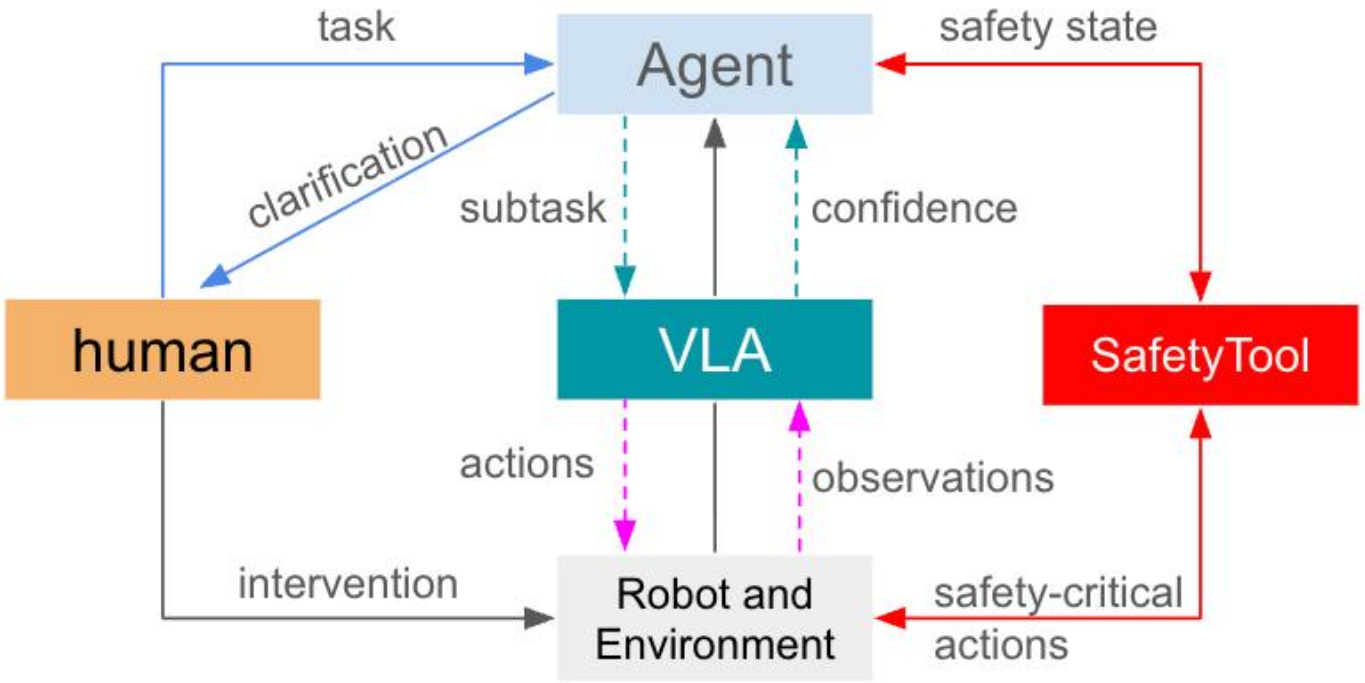
请务必阅读正文之后的免责声明及其项下所有内容

谷歌Gemini Robotics 2.5：精细操作与柔性体接触能力提升

国信证券 GUOSEN SECURITIES

- **模型架构：**专用VLA动作模型，与高层具身推理模型协同。Gemini Robotics 2.5承担视觉—语言—动作转换，将视觉信息和语言指令转为机器人运动控制。Gemini Robotics ER负责理解环境、组织多步任务和跟踪执行进度，2.5负责具体动作执行。
- **输入输出：**视觉与语言进入动作模型，输出面向机器人的运动控制。上层ER输出任务计划和子任务，2.5结合视觉观测执行相应操作。其适用机器人覆盖双臂平台及人形机器人，能够控制平行夹爪、灵巧手和全身运动。

图12：Gemini Robotics 2系列架构



资料来源：Google DeepMind，国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

目录

01 具身模型：科技大厂加速布局，通用模型向具身拓展

02 架构解析：一段式与分层架构，复杂任务凸显分层协同价值

03 VLA模型发展脉络：从RT-2到π系列的架构与能力升级

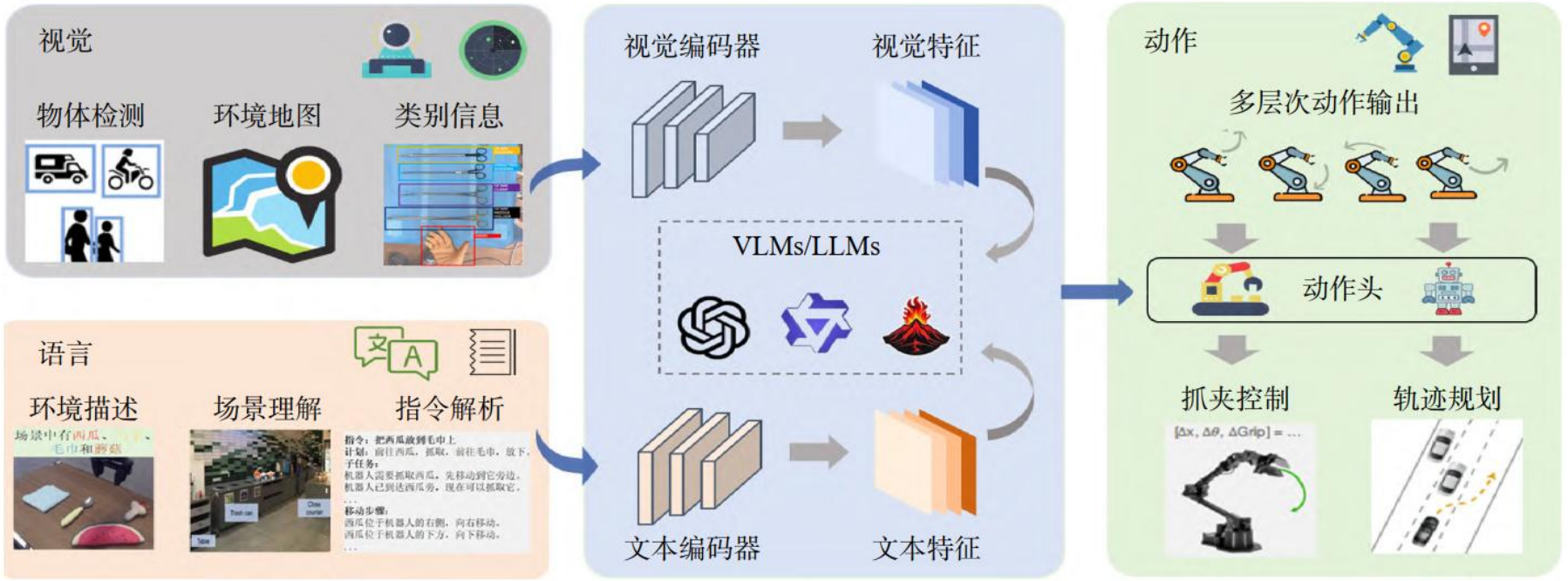
04 风险提示

请务必阅读正文之后的免责声明及其项下所有内容

VLA模型总体架构：从视觉语言指令到动作输出

- **VLA模型：**通过视觉编码器、语言编码器与动作解码器的协同处理，融合来自机器人本体、外部环境及语言交互的多源感知信息，逐步生成从低阶控制命令到高层轨迹规划的动作策略。
- **核心组件：**视觉编码器、语言编码器与动作解码器，三者协同工作，完成从多模态输入至机器人动作的映射。视觉编码器负责自高维图像或点云中提取紧凑的语义特征；语言编码器将自然语言指令转换为机器可理解的语义向量；动作解码器进一步将融合后的多模态表征转化为具体动作。

图13：VLA模型架构



资料来源：张慧等，《VLA架构下的智能体演化：从机理建构到应用拓展》，《科技导报》，2025，43(20)：48-61；国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

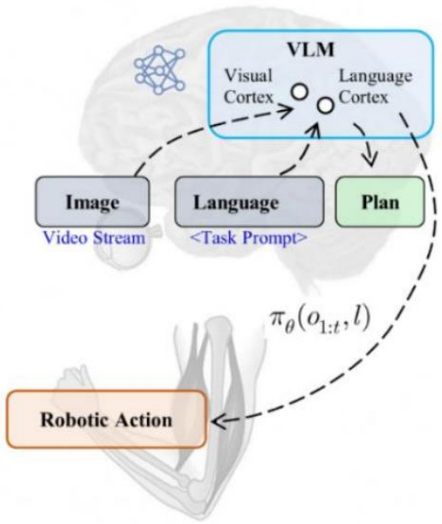
VLA模型分类：一段式与分层架构

- **模型架构：**VLA模型可从规划与执行的组织方式划分为一段式架构和分层架构。前者在统一模型内完成多模态理解与动作预测，后者通过中间目标连接高层规划与低层执行。
- **架构取舍：**一段式架构强调统一表征与联合学习，分层架构强调任务分解与模块协作。具体选择需要结合任务时长、实时响应要求、训练数据条件及系统集成成本。

图14：一段式与分层VLA架构对比

一段式架构

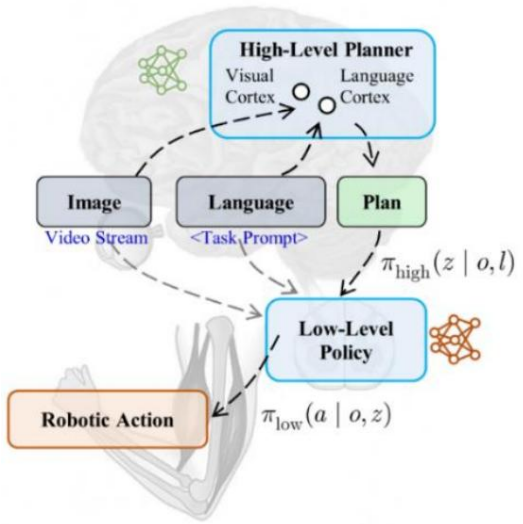
在统一模型中完成理解与动作预测



- **输入输出：**输入视觉观测与语言指令，通过统一模型预测机器人动作；动作可表示为离散词元或连续控制目标。
- **运作方式：**视觉、语言与动作共享模型表征，通过动作学习将场景理解与操作行为关联起来，RT-2、OpenVLA等采用这一思路。
- **优势与挑战：**整体流程紧凑，便于多模态联合学习；复杂长任务中的误差累积、推理过程解释及大模型推理开销是优化重点。

分层架构

高层确定操作目标，低层生成执行动作



- **高层规划：**根据视觉观测与语言指令，生成子任务、目标位置或运动轨迹等中间结果，明确当前操作目标。
- **低层执行：**接收高层目标与最新环境观测，将目标转化为机器人动作，并结合执行反馈持续调整。
- **优势与挑战：**任务分工清晰，便于多步骤规划与技能复用；系统性能取决于层间接口、信息传递及规划与执行的协同效率。

资料来源：Jingjing Pei等，《The evolution of end-to-end Vision-Language-Action (VLA) architectures in robotics》，Robot Learning，2026，1-36；国信证券经济研究所整理

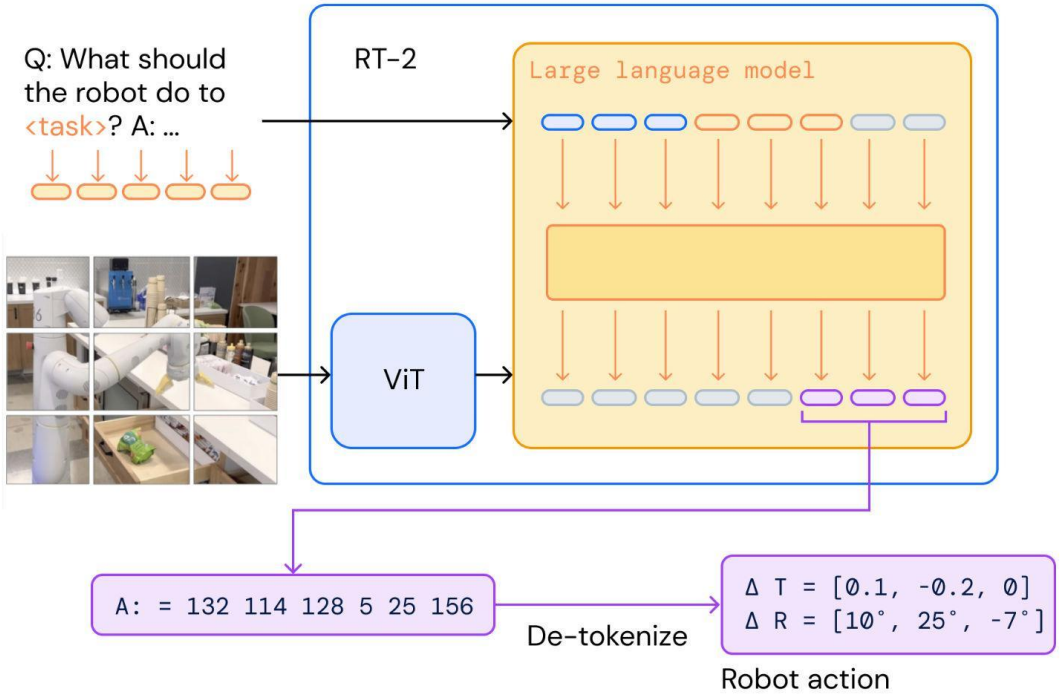
请务必阅读正文之后的免责声明及其项下所有内容

一段式架构：RT-2视觉语言模型直接预测动作

核心机制：让视觉语言模型学会逐个预测机器人动作Token

RT-2将机器人动作编码为文本Token，在统一模型中完成视觉理解、指令理解与动作预测。模型以预训练的PaLI-X或PaLM-E为骨干，输入机器人相机图像和自然语言任务指令；图像经ViT编码后，与文本信息共同输入模型，以自回归方式逐个生成**动作Token**。

图15：RT-2模型可直接输出动作token



资料来源：Anthony Brohan等，《RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control》，arXiv，2023，1-26；国信证券经济研究所整理

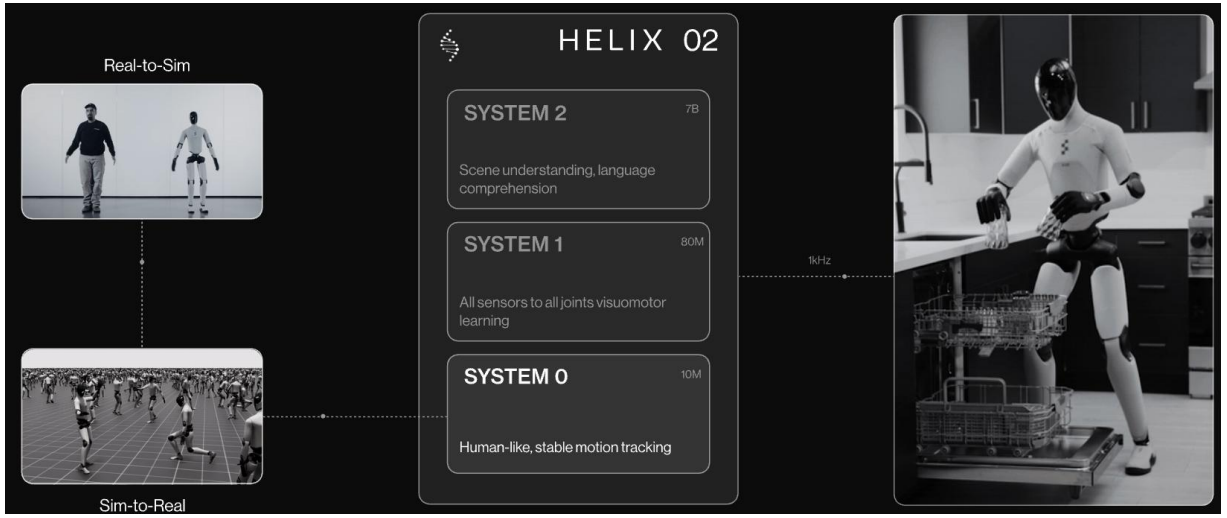
请务必阅读正文之后的免责声明及其项下所有内容

分层架构：S2S1S0分工明确，分别进行任务理解规划、动作策略生成、动作执行

Helix 02：S2进行高层任务规划，S1输出全身关节目标，S0输出关节级执行器命令。

- **System 2：高层语义理解、任务推理与规划层**，相当于每台机器人的“专属大脑”，负责理解语言指令、识别场景并拆解任务，输出语义目标，引导S1生成动作。搭载**多模态语言模型**，Helix01版本中的S2频率为7-9 Hz，具备强大的环境感知与指令解析能力，负责记忆管理、收集指令、信息理解、任务规划（拆解子任务及顺序）、做出决策。
- **System 1：VLA动作策略层**，根据推理结果、摄像头、触觉传感器和机器人状态生成下一步动作，针对机器人部位下达精准的**目标位姿指令**，以200 Hz的频率生成全身关节目标，每次预测都会生成一组高频动作序列段并交由System 0执行，且采用全异步执行模式——上一组动作序列段执行的同时，系统已完成下一组序列段的生成准备。
- **System 0：全身运动控制层**，Figure的S0架构是一个1000万参数的神经网络，输入是全身关节状态和基座运动，输出是1k Hz的关节级执行器命令。主要任务是完成System 1下达的**位姿目标**，同时通过求解机器人所有**关节的运动状态**，持续维持机器人的动态平衡。Figure的S0直接从超过1000小时的关节级重定向人体运动数据集中学习如何协调力、调整姿势、保持平衡。

图16：Helix 02模型架构



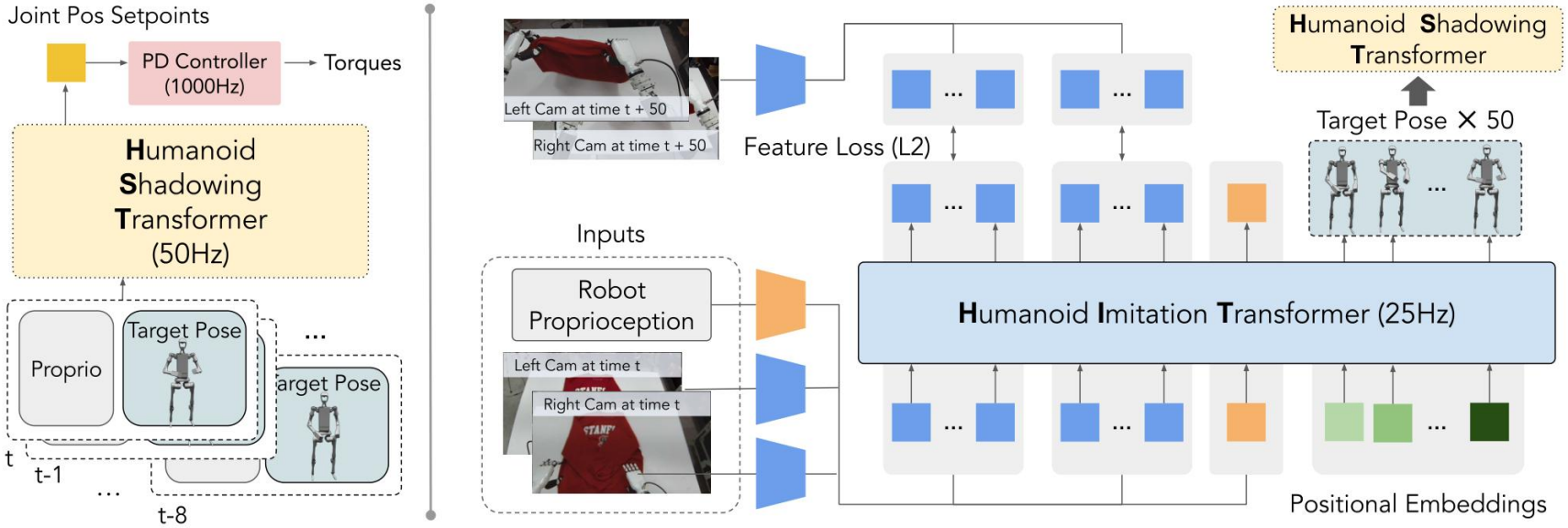
资料来源：Figure官网，国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

S0全身运动控制：HumanPlus的HST策略与PD控制器

- **输入：**①机器人的实际身体状态：关节角、关节速度、躯干倾斜角、躯干角速度，以及上一轮动作等；②希望机器人完成的参考动作，具体包括参考关节角，以及身体姿态、移动和转向目标，可以由HIT提供。
- **S0层HST：**经过训练的全身运动控制网络，读取参考动作和身体状态历史，计算这一轮的关节控制指令，每秒运行约50次。
- **输出：**身体关节的角度。控制器读取角度设定值，再结合关节实际角度和速度，计算各关节要施加的力矩，传导到电机，从力矩转为电流，控制电机工作。

图17：HumanPlus的HST运动跟踪与HIT技能策略架构



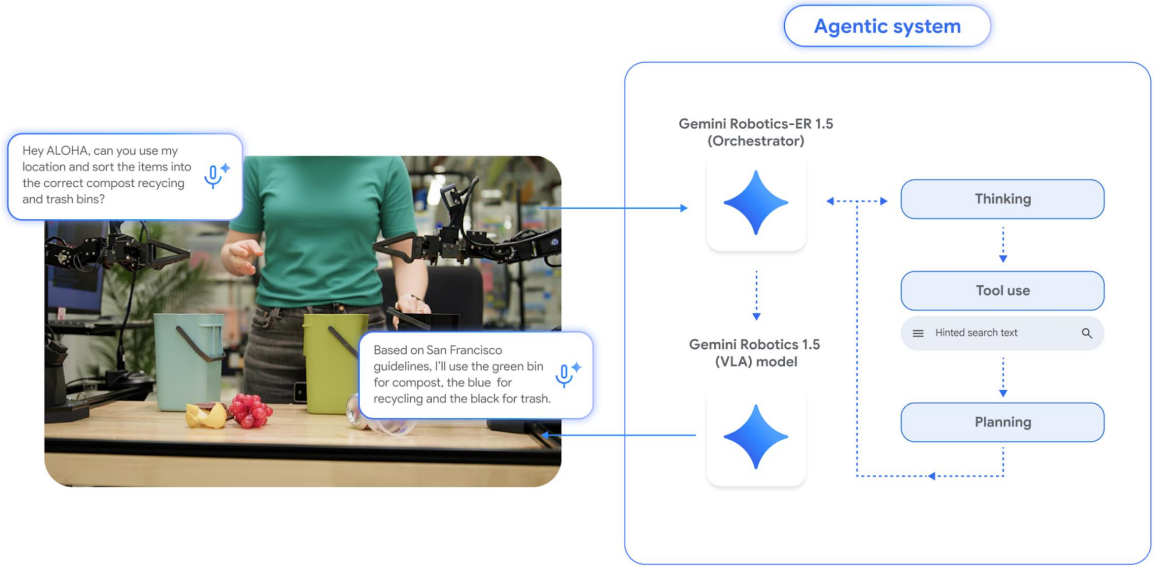
资料来源：Zipeng Fu等，《HumanPlus: Humanoid Shadowing and Imitation from Humans》，arXiv，2024，1-17；国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

基础模型能力提升，凸显复杂任务中的分层协同价值

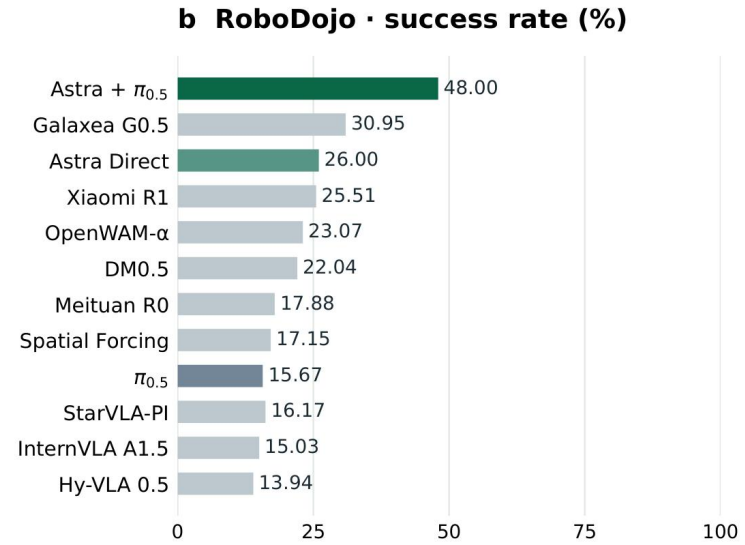
- **基础模型的知识与推理能力，为机器人高层规划提供更强基础。**基础模型在语言理解、空间推理与工具调用方面的积累，有助于机器人理解开放式任务、组织操作步骤，并根据环境变化调整计划。Google的Gemini Robotics ER负责理解用户指令、规划多步骤任务及跟踪执行进度，再与VLA动作模型协同完成操作。通用模型的认知能力与机器人已有动作技能形成互补，为多步骤任务、技能复用和执行纠错提供支持。
- **通用模型与动作策略的协同已展示互补价值，当前阶段通用模型更适合优先承担S2层的高层决策。**在10项RoboDojo仿真任务中，Astra与π0.5协同的成功率达到48%，高于Astra直接控制的26%。**我们认为，随着基础模型的任务理解、规划与反馈判断能力增强，高层模型有望更有效地组织和调用动作技能，推动分层架构在复杂、长时序任务中的应用。**面向复杂、长时序任务，可由通用模型负责目标理解、任务分解、进度判断及按需纠偏，S1动作模型负责连续动作生成，S0控制器负责实时跟踪与执行。分层分工与调整高层调用频率，有望兼顾推理能力、执行效率和使用成本。

图18：Gemini Robotics 1.5高层推理与动作模型协同框架



资料来源：Google DeepMind，《Gemini Robotics 1.5 brings AI agents into the physical world》，2025；国信证券经济研究所整理

图19：RoboDojo仿真基准任务成功率对比（%）



资料来源：Xuchuan Chen等，《Systematically Exploring the Capabilities of GPT-6 Astra as Embodied Policies》，arXiv，2026，1-46；国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

目录

01 具身模型：科技大厂加速布局，通用模型向具身拓展

02 架构解析：一段式与分层架构，复杂任务凸显分层协同价值

03 VLA模型发展脉络：从RT-2到π系列的架构与能力升级

04 风险提示

请务必阅读正文之后的免责声明及其项下所有内容

RT-2→OpenVLA→π系列的架构演进

- **RT-2：**首次将动作当作文本token处理；
- **OpenVLA：**通过开源模型与微调降低适配门槛；
- **π0：**奠定基本架构，以VLM为骨架，外接一个Action Expert，通过流匹配一次生成一段动作，提升生成速度，实现连续动作生成；
- **π0.5：**开始在数据方面下功夫，将一类动作数据扩展为五类数据联合训练。模型先输出子任务，再输出具体动作，将高层规划与低层控制连接起来；
- **π*0.6：**引入了强化学习，让模型能够从自主运行产生的数据中学习。经过一轮轮迭代，模型可以完成连续13小时制作咖啡，或在工厂中装载实际使用的纸箱等复杂任务；
- **π0.7：**数据方面，将大量次优数据、强化学习产生的数据以及人类视频全部纳入训练，并通过细粒度标注将它们转化为输入的一部分。同时，接入了世界模型，由世界模型生成子目标图，并将其与原始数据一起提供给模型。最终，π0.7生成的模型无需进行任何针对性微调，就能够承接此前各个专家模型完成的任务。

图20：从RT-2、OpenVLA到π系列的VLA模型演进



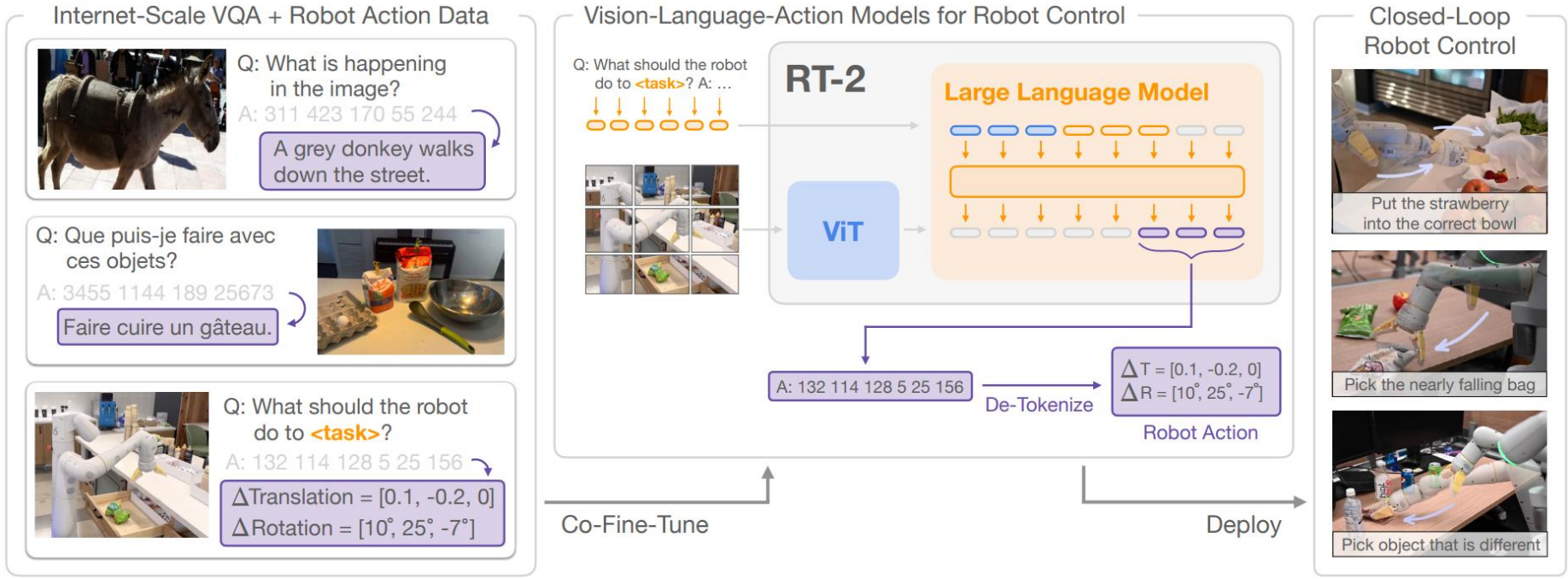
资料来源：Anthony Brohan等，《RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control》，arXiv，2023，1-26；Moo Jin Kim等，《OpenVLA: An Open-Source Vision-Language-Action Model》，arXiv，2024，1-37；Kevin Black等，《π0: A Vision-Language-Action Flow Model for General Robot Control》，arXiv，2024，1-17；Kevin Black等，《π0.5: a Vision-Language-Action Model with Open-World Generalization》，arXiv，2025，1-19；Ali Amin等，《π*0.6: a VLA That Learns From Experience》，arXiv，2025，1-18；Bo Ai等，《π0.7: a Steerable Generalist Robotic Foundation Model with Emergent Capabilities》，arXiv，2026，1-25；国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

RT-2：VLA一段式模型架构

- **RT-2：**首次将动作当作文本token处理，模型将动作离散成若干整数，其中一个表示结束或无动作，另外三个表示末端位置的增量，还有三个表示旋转增量。
- **模型架构：**RT-2使用一个现成的视觉语言模型Google自己的PaLM-E，输入“机器人接下来该做什么”这样的文本指令，模型像回答问题一样输出八个数字，只不过这些数字代表机器人动作，而不是自然语言。训练时不需要新增任何参数，将网络上的图文问答数据与机器人数据混合训练。
- **存在问题：**输出速度慢，采用LLM架构，当时使用的55B模型每秒只能输出一到三个动作，1~3Hz，实际无法使用。

图21：RT-2机器人模型架构



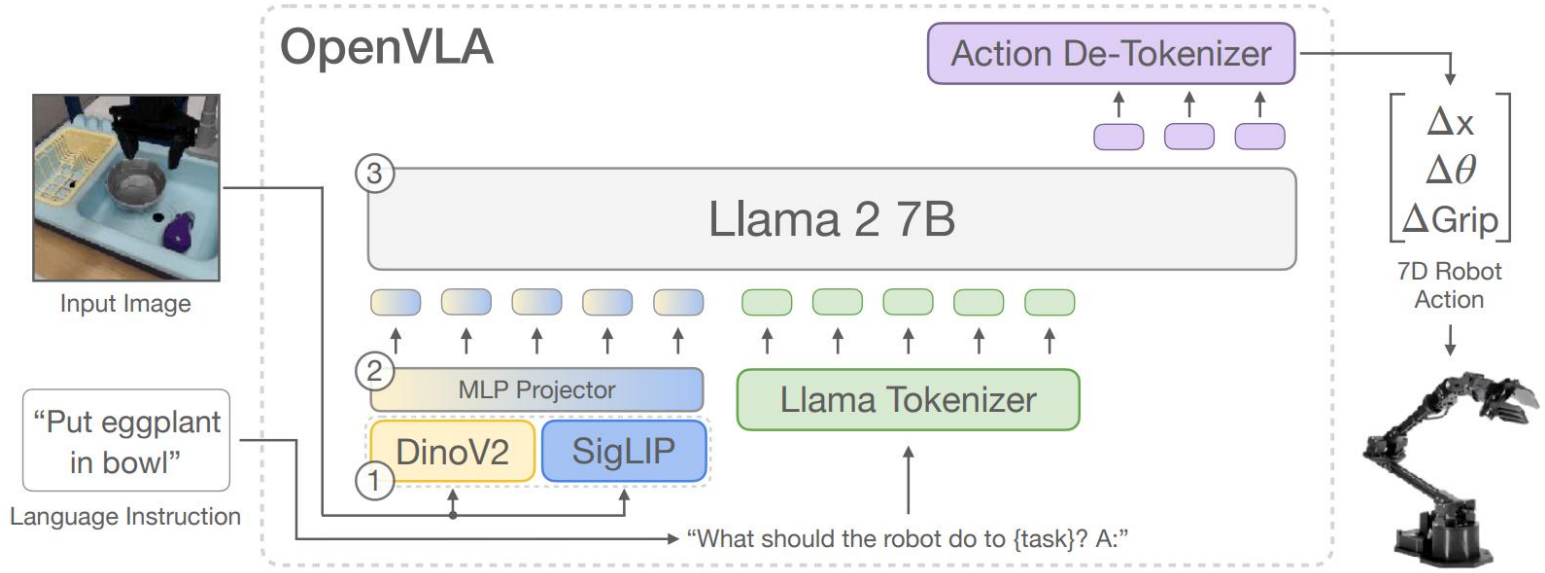
资料来源：Anthony Brohan等，《RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control》，arXiv，2023，1-26；国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

OpenVLA：开源VLA，降低模型部署与任务适配门槛

- 与RT-2相比，模型架构上增加视觉编码器：使用两个视觉编码器，SigLIP和DinoV2，将二者拼接起来。语言模型采用当时备受关注的Llama2 7B。输入是图片和指令，输出由动作Token组成。
- OpenVLA使用了约97万条轨迹训练，宣称使用64张A100显卡训练了14天。在Bridge数据集上的成功率达到70.6%，超过了RT-2所使用的55B大模型。
- RT-2和OpenVLA存在的问题：①两者都采用**语言模型的推理架构，每次推理只输出一步动作**，因此推理速度很慢，导致控制频率低，最高只有6Hz左右。对于一些能够以50Hz运行的机械臂速度不够用。②模型只输入单张图像，没有本体状态，也没有记忆或历史信息，可能是受到了当时计算资源和大语言模型能力的限制。

图22：OpenVLA模型架构



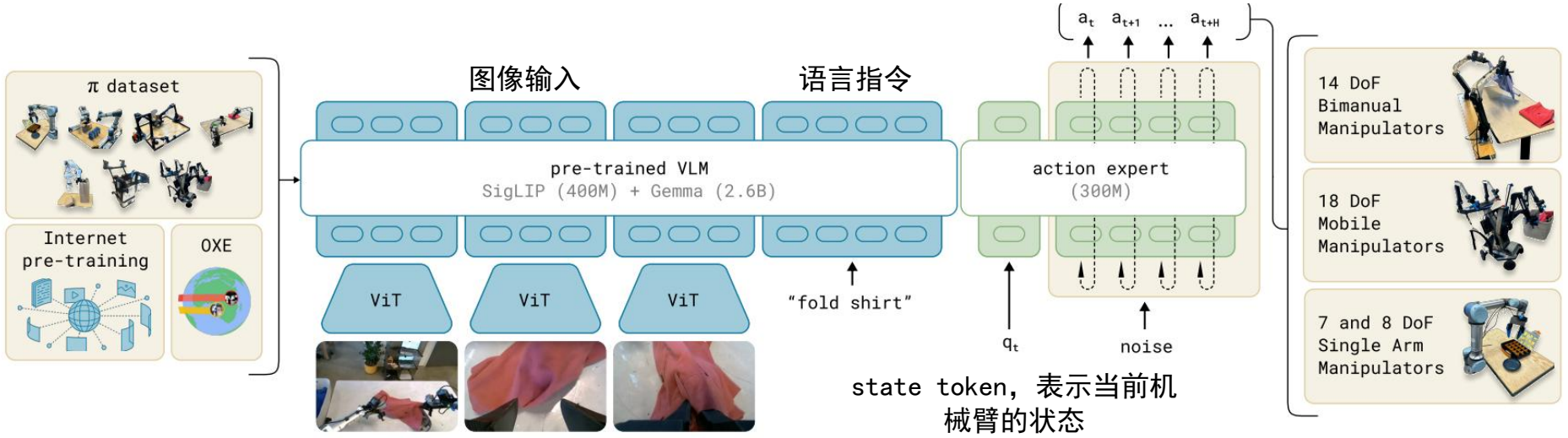
资料来源：Moo Jin Kim等，《OpenVLA: An Open-Source Vision-Language-Action Model》，arXiv，2024，1-37；国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

π0：分层VLA架构，基于流匹配提升动作输出效率

- **数据：**π0的一个训练样本对应一个时间序列，一个时间步后面连接50个动作。模型采集了9亿300万个时间步的数据，加起来共有一万多小时，包含七种机器人构型和六十八类任务。
- **架构：**与之前的模型相差不大。左侧是Google开源的VLM Gemma，采用SigLIP作为图像编码器（四亿参数）。三个图像经过编码后，与文本混合并一起送入Gemma，这一部分与之前的做法差别不大。最重要的区别是为Gemma增加了**Action Expert**，3亿参数，以前的范式更像语言模型，通常是逐个token、逐个动作向外生成；Action Expert以流匹配为基础，**一次生成全部50个动作**。
- **训练模式：**以前具身模型的训练模式通常只有一步，π0借鉴LLM的训练方式，先进行大规模预训练，再进行后训练。**①预训练阶段**把全部数据一次性投入训练，尽可能覆盖更多任务、更多场景和更多机器人，但数据本身并不十分干净。**②后训练阶段**会筛选出一部分高质量数据，专门用于微调模型，让模型执行任务时更加流畅、更加稳定。因为高质量数据中完全没有错误，而预训练数据中可能包含许多错误，模型可以**根据错误学习如何恢复到正确动作**。

图23：π0模型架构



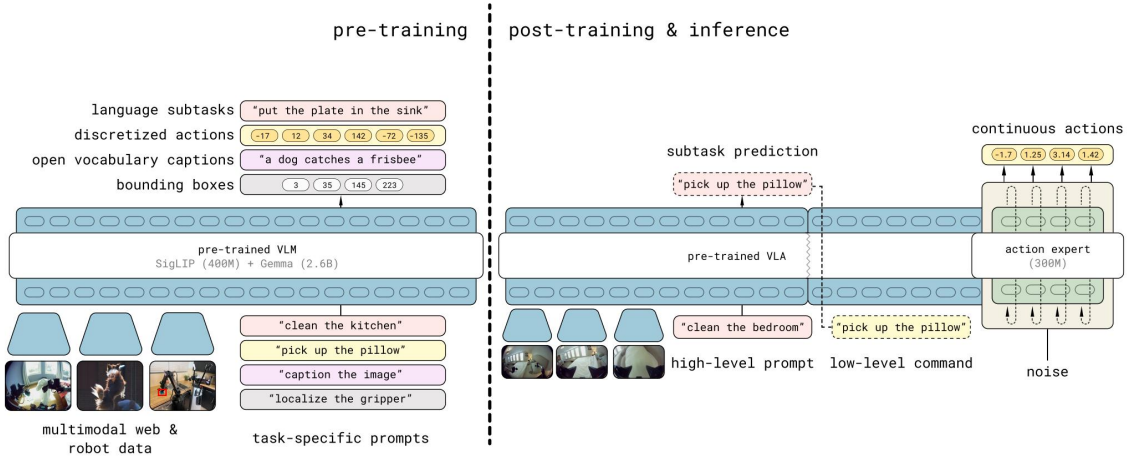
资料来源：Kevin Black等，《π0: A Vision-Language-Action Flow Model for General Robot Control》，arXiv，2024，1-17；国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

π0.5：分层VLA架构，分为预训练与后训练阶段

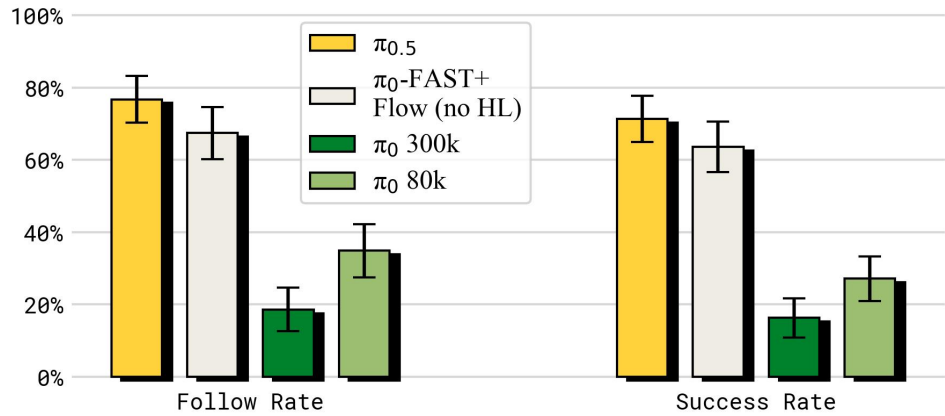
- **π0的问题：**①所有评测和训练数据都在相同的环境中采集，没有验证模型的泛化能力。②π0的动作大多是原子动作，整体相对简单。③Action expert所采用的流匹配机制在训练阶段速度较慢（推理阶段速度快）。
- **π0.5要解决的问题：**①**只给出一句话**，机器人就能在一个训练时完全没见过的家庭或其他环境中，自动完成任务，一次任务时长扩展到10~15分钟。②提高训练和推理中动作生成的速度（**大规模预训练需要采用离散化token提高训练速度；后训练阶段再用流匹配，用少量数据进行精调，提高推理速度**）。
- **架构：**
  - **预训练：**输入侧增加了一类信息：以往会单独输入的本体状态被全部离散化为文本Token，与图像、指令等信息一起放入VLM中统一编码。输出侧，VLM能够输出文本，用于**将高层任务生成子任务**、图片描述等。同时，VLM还能输出离散动作Token，加快预训练速度。预训练阶段完全转变为一种全Token化的语言模型范式，所有内容都可以逐Token学习和预测。
  - **后训练：**骨干模型、Action Expert与π0相同，区别在于π0.5只在后训练阶段加入Action Expert。本体状态被放到前面Token序列中，与其他信息一起编码。使用一批高质量数据训练Action Expert，从而提高最终动作输出的速度和精度。

图24：π0.5模型架构



资料来源：Kevin Black等，《π0.5: a Vision-Language-Action Model with Open-World Generalization》，arXiv，2025，1-19；国信证券经济研究所整理

图25：π0.5指令遵循率与任务成功率对比



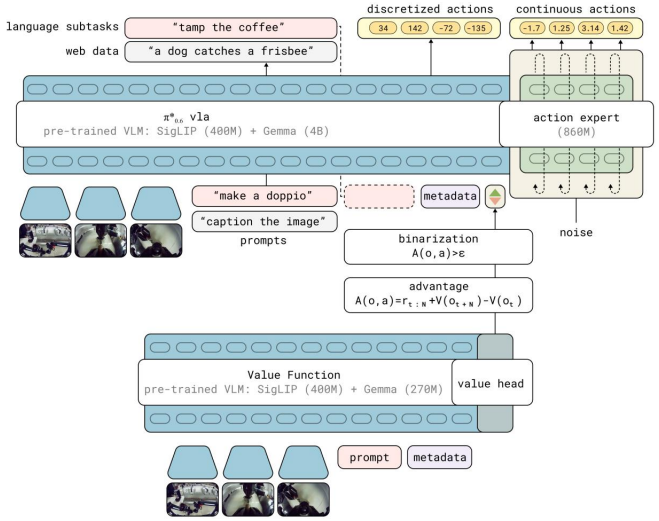
资料来源：Kevin Black等，《π0.5: a Vision-Language-Action Model with Open-World Generalization》，arXiv，2025，1-19；国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

π*0.6：引入强化学习，从自主运行与人工纠正中改进策略

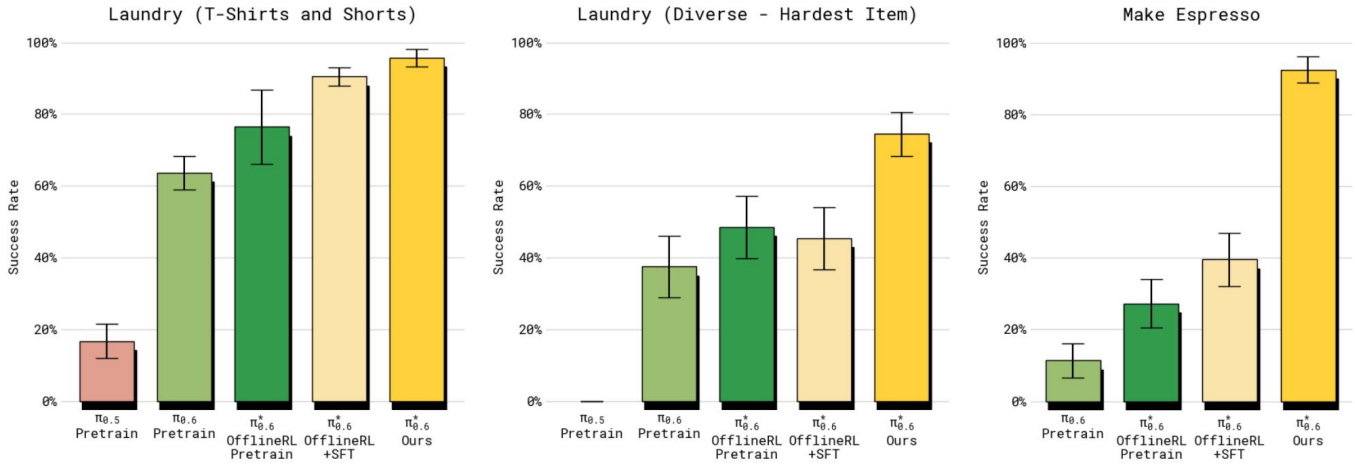
- **π0~π0.5的问题：**模仿学习，训练数据来自遥操作，有一个很强的隐藏假设：数据中出现的每一个动作都是正确的，都值得学习，所有样本一律平等。**问题1：复合误**差，模型自己运行时，每一步都会产生一点小误差。如果只偏离了前几步，这些偏差可能仍在示范数据覆盖的分布之内，因为模型见过类似的画面，能够纠偏回来。但误差会不断累积，到了某一步，累积后的状态会逐渐变得陌生，可能会失败。**问题2：**模型的上限是人类水平。
- **模型架构：**π*0.6采用Gemma 3 4B骨干和8.6亿参数的动作专家，生成连续机器人动作。π*0.6通过RECAP引入强化学习，结合示范、自主运行和人工纠正数据，利用价值模型估计动作优势，并将优势标签作为策略的训练条件。部署时，策略在正优势条件下生成动作，以提升任务完成质量与执行效率。
- **任务成功率：**一台机器人可以连续工作13个小时，不间断地制作浓缩咖啡；可以前往一个从未去过的陌生家庭，并操作此前从未见过的衣服。可以连续工作两个多小时，全程不需要任何人工干预。可以进入工厂，组装实际使用的包装纸箱。在难度最高的几个任务上，比π0.5失败率也降低了约一半。

图26：π*0.6模型架构



资料来源：Ali Amin等，《π*0.6: a VLA That Learns From Experience》，arXiv，2025，1-18；国信证券经济研究所整理

图27：π*0.6任务成功率



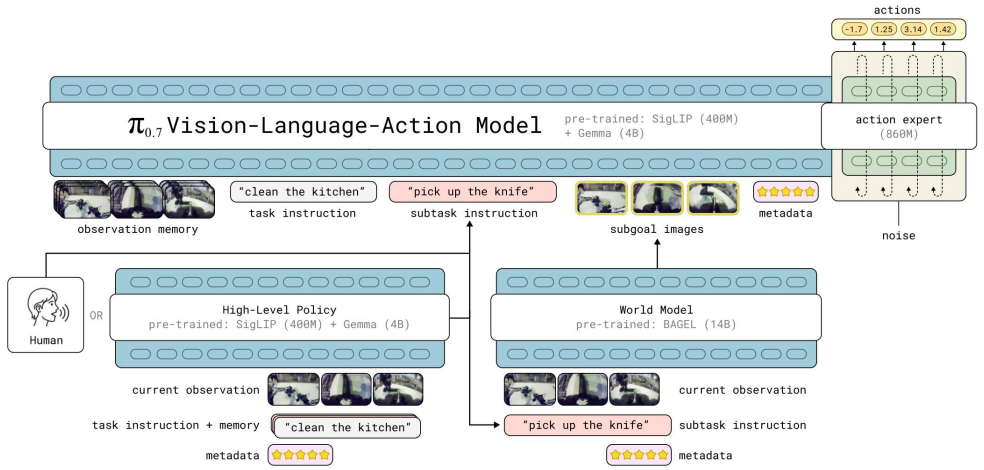
资料来源：Ali Amin等，《π*0.6: a VLA That Learns From Experience》，arXiv，2025，1-18；国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

π0.7：分层VLA架构，引入高层策略S2和世界模型

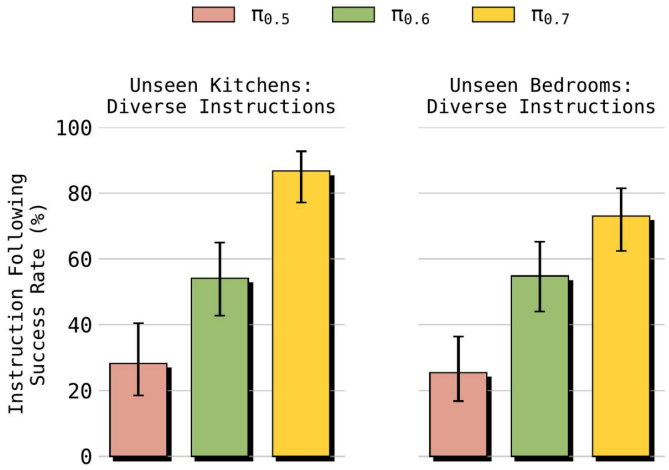
- **高层策略（High-Level Policy）：高层任务理解与子任务规划。**输入当前观察、整体任务、历史记忆与元数据，输出下一条子任务指令。子任务既可由高层模型生成，也可由人类逐步提供，用于指导后续动作。
- **世界模型（World Model）：**生成期望达到的视觉目标。基于BAGEL，依据当前观察、子任务和元数据，生成近未来的多视角子目标图，将物体布局、机械臂与物体的相对关系等空间目标具体化，为动作生成提供可选的视觉引导。
- **视觉语言主干（VLM）：**整合观测与任务信息。将当前及历史图像、本体状态、任务与子任务、可选子目标图、元数据和控制模式编码为上下文表示。历史观测提供场景变化信息，元数据描述任务时长、执行质量与错误情况，控制模式指定采用关节控制或末端控制。
- **动作专家（Action Expert）：**生成连续机器人动作。动作专家包含8.6亿参数，通过注意力读取主干上下文，结合带噪动作与去噪时间步，采用流匹配迭代生成未来50步动作，输出关节目标或末端位姿及夹爪指令。视觉语言主干与动作专家共同构成VLA动作模型。

图28：π0.7模型架构



资料来源：Bo Ai等，《π0.7: a Steerable Generalist Robotic Foundation Model with Emergent Capabilities》，arXiv，2026，1-25；国信证券经济研究所整理

图29：π0.7在陌生家庭环境中的指令遵循成功率



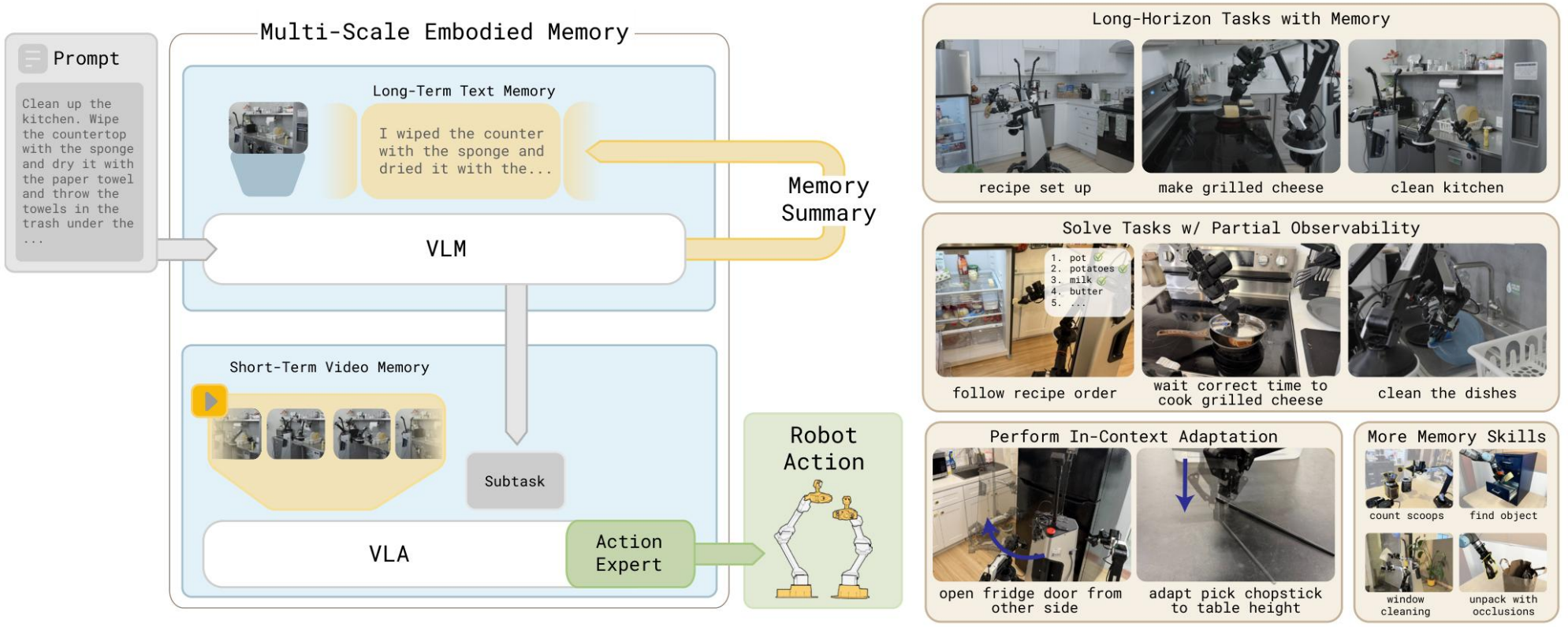
资料来源：Bo Ai等，《π0.7: a Steerable Generalist Robotic Foundation Model with Emergent Capabilities》，arXiv，2026，1-25；国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

π0.7：多尺度记忆机制提升具身模型上下文处理能力

- π0、π0.5和π*0.6只能处理当前看到的这一帧画面，无法记住之前任务。π0.7采用分层记忆架构（短期稠密记忆+长期稀疏记忆）。短期记忆放在底层，使用一个经过改造的视觉编码器，把最近几秒的画面压缩进当前帧的表示中。长期记忆放在高层，高层会输出一段更新后的语言摘要，如“已经检查过这个台面了”或“已经放过盐了”。下一步执行时，模型会读取这段摘要，并对其进行更新，维护长期记忆。

图30：MEM多尺度记忆架构



资料来源：Marcel Torne等，《MEM: Multi-Scale Embodied Memory for Vision Language Action Models》，arXiv，2026，1-15；国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

π0.7：多维标注与视觉子目标提升数据利用效率

- **模型输入逐步丰富，从描述任务拓展到明确操作要求。**在视觉观测、任务指令和本体状态基础上，π0.5加入子任务信息，π*0.6加入优势条件，π0.7进一步引入视觉子目标、执行时长、质量、错误标记及控制模式，让模型获得更具体的任务目标和操作指引。
- **训练时区分操作表现，执行时指定期望行为。**π0.7对数据进行质量、时长和错误等标注，区分不同操作过程的表现，使示范、自主执行及失败经验都能用于训练；执行时，模型根据期望质量、时长等条件生成动作，并可通过视觉子目标明确物体应达到的位置和状态。
- **我们认为，更丰富、细致的输入条件有助于将多样化经验转化为可复用的模型能力。**通过明确“完成什么任务、达到什么状态、采用何种方式”，模型能够更有针对性地学习和生成动作，为技能组合、陌生场景适应及复杂任务执行提供支撑。

图31：π系列模型数据信息结构



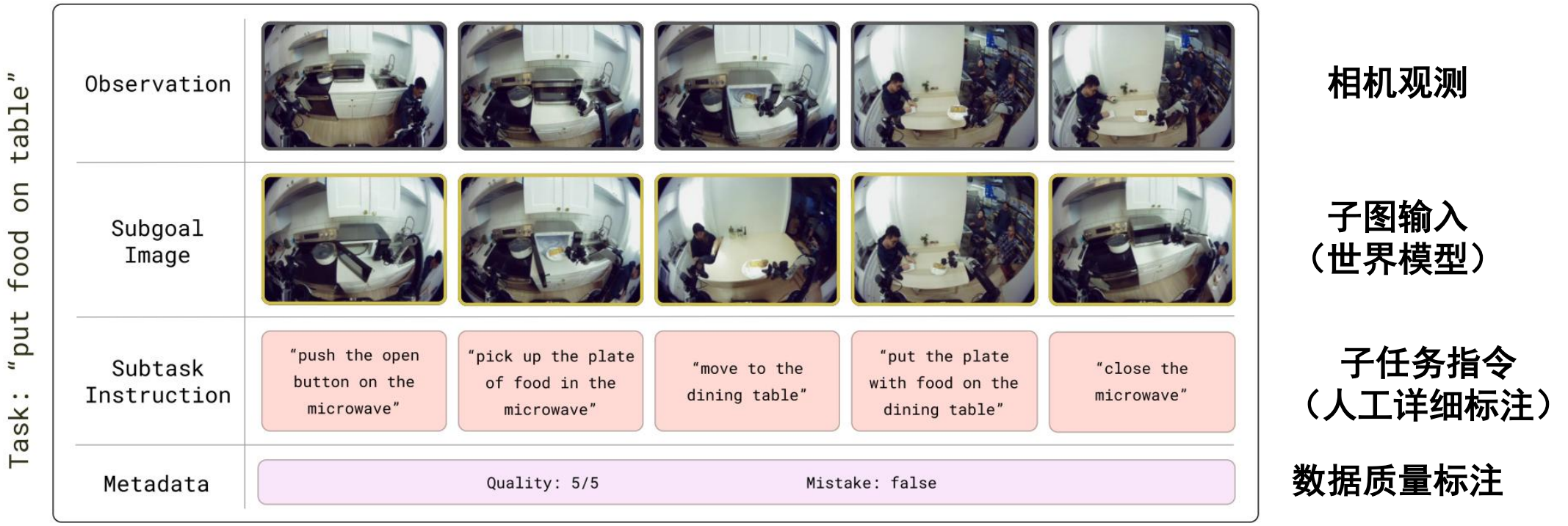
资料来源：Kevin Black等，《π0: A Vision-Language-Action Flow Model for General Robot Control》，arXiv，2024，1-17；Kevin Black等，《π0.5: a Vision-Language-Action Model with Open-World Generalization》，arXiv，2025，1-19；Ali Amin等，《π*0.6: a VLA That Learns From Experience》，arXiv，2025，1-18；Bo Ai等，《π0.7: a Steerable Generalist Robotic Foundation Model with Emergent Capabilities》，arXiv，2026，1-25，国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

π0.7：典型训练样本与数据标注

- **π0.7进行了详细的数据标注：**π0.7在数据中加入了π*0.6在强化学习时采用的数据，其中包含较多失败数据，以及人类中途接管的轨迹等。
- **世界模型负责输出子目标图：**会根据现有的观测，预测机器人执行完当前动作后，未来可能出现的世界状态。作用是直接用一张图告诉模型，执行完接下来这一段动作后世界的变化。

图32：π0.7训练样本：观测、视觉子目标、子任务与元数据



资料来源：Bo Ai等，《π0.7: a Steerable Generalist Robotic Foundation Model with Emergent Capabilities》，arXiv，2026，1-25；国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

π系列演进脉络：分层架构逐步完善，泛化与执行能力持续提升

回顾π系列的模型演进，高层规划与动作执行的分工逐步清晰，训练方式从学习示范拓展至利用自主执行经验，共同提升模型执行复杂任务的能力。

- **架构演进：高层规划与动作生成的分工逐步清晰。**π0形成“视觉语言主干+动作专家”的基础架构；π0.5加入子任务预测，将任务拆解与动作生成结合；π0.7进一步设置独立的高层规划模型，并通过世界模型生成可选目标图像，为动作执行提供更明确的指引。
- **训练升级：从学习示范拓展至利用自主执行经验。**π*0.6通过RECAP结合示范、自主操作与人工纠正，改善执行可靠性和效率；π0.7通过质量、速度等条件描述，帮助模型学习不同操作方式，将更多执行经验用于通用模型训练。
- **能力拓展：机器人逐渐能够自主完成更复杂的任务，从掌握多种操作技能，走向陌生环境适应与技能组合。**π0.5增强新环境中的任务执行能力，π0.7进一步展示指令遵循、技能组合与跨机器人迁移能力。

表3：π系列模型的研究重点、核心方法与能力演进

模型	研究重点	核心方法	重点能力
π0	建立通用动作基础	预训练VLM＋流匹配动作专家；多机器人、多任务训练	生成连续动作，学习多种操作技能
π0.5	增强陌生环境泛化	异构数据联合训练；高层子任务预测与动作执行结合	在训练中未出现的家庭执行长时序任务
π*0.6	提高执行可靠性与效率	RECAP结合示范、自主经验及人工纠正，通过优势条件改进策略	减少任务失败，提高作业吞吐量
π0.7	增强可引导性与技能复用	继承记忆机制，加入丰富指令、元数据及可选视觉子目标	更灵活地遵循指令，组合技能并开展跨本体迁移

资料来源：Kevin Black等，《π0: A Vision-Language-Action Flow Model for General Robot Control》，arXiv，2024，1-17；Kevin Black等，《π0.5: a Vision-Language-Action Model with Open-World Generalization》，arXiv，2025，1-19；Ali Amin等，《π*0.6: a VLA That Learns From Experience》，arXiv，2025，1-18；Bo Ai等，《π0.7: a Steerable Generalist Robotic Foundation Model with Emergent Capabilities》，arXiv，2026，1-25；Marcel Torne等，《MEM: Multi-Scale Embodied Memory for Vision Language Action Models》，arXiv，2026，1-15；国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

目录

01 具身模型：科技大厂加速布局，通用模型向具身拓展

02 架构解析：一段式与分层架构，复杂任务凸显分层协同价值

03 VLA模型发展脉络：从RT-2到π系列的架构与能力升级

04 风险提示

请务必阅读正文之后的免责声明及其项下所有内容

风险提示

一、具身智能模型进展不及预期。

二、具身机器人出货量不及预期。

三、具身数据采集进度低于预期。

四、具身模型技术路线变化，行业竞争加剧。

请务必阅读正文之后的免责声明及其项下所有内容

免责声明

国信证券 GUOSEN SECURITIES

国信证券投资评级

投资评级标准	类别	级别	说明
报告中投资建议所涉及的评级（如有）分为股票评级和行业评级（另有说明的除外）。评级标准为报告发布日后6到12个月内的相对市场表现，也即报告发布日后的6到12个月内公司股价（或行业指数）相对同期相关证券市场代表性指数的涨跌幅作为基准。A股市场以沪深300指数（000300.SH）作为基准；新三板市场以三板成指（899001.CSI）为基准；香港市场以恒生指数(HSI.HI)作为基准；美国市场以标普500指数(SPX.GI)或纳斯达克指数（IXIC.GI）为基准。	股票投资评级	优于大市	股价表现优于市场代表性指数10%以上
		中性	股价表现介于市场代表性指数±10%之间
		弱于大市	股价表现弱于市场代表性指数10%以上
		无评级	股价与市场代表性指数相比无明确观点
	行业投资评级	优于大市	行业指数表现优于市场代表性指数10%以上
		中性	行业指数表现介于市场代表性指数±10%之间
		弱于大市	行业指数表现弱于市场代表性指数10%以上

分析师承诺

作者保证报告所采用的数据均来自合规渠道；分析逻辑基于作者的职业理解，通过合理判断并得出结论，力求独立、客观、公正，结论不受任何第三方的授意或影响；作者在过去、现在或未来未就其研究报告所提供的具体建议或所表述的意见直接或间接收取任何报酬，特此声明。

重要声明

本报告由国信证券股份有限公司（已具备中国证监会许可的证券投资咨询业务资格）制作；报告版权归国信证券股份有限公司（以下简称“我公司”）所有。本报告仅供我公司客户使用，本公司不会因接收人收到本报告而视其为客户。未经书面许可，任何机构和个人不得以任何形式使用、复制或传播。任何有关本报告的摘要或节选都不代表本报告正式完整的观点，一切须以我公司向客户发布的本报告完整版本为准。

本报告基于已公开的资料或信息撰写，但我公司不保证该资料及信息的完整性、准确性。本报告所载的信息、资料、建议及推测仅反映我公司于本报告公开发布当日的判断，在不同时期，我公司可能撰写并发布与本报告所载资料、建议及推测不一致的报告。我公司不保证本报告所含信息及资料处于最新状态；我公司可能随时补充、更新和修订有关信息及资料，投资者应当自行关注相关更新和修订内容。我公司或关联机构可能会持有本报告中所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或争取提供投资银行、财务顾问或金融产品等相关服务。本公司的资产管理部门、自营部门以及其他投资业务部门可能独立做出与本报告中意见或建议不一致的投资决策。

本报告仅供参考之用，不构成出售或购买证券或其他投资标的要约或邀请。在任何情况下，本报告中的信息和意见均不构成对任何个人的投资建议。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。投资者应结合自己的投资目标和财务状况自行判断是否采用本报告所载内容和信息并自行承担风险，我公司及雇员对投资者使用本报告及其内容而造成的一切后果不承担任何法律责任。

证券投资咨询业务的说明

本公司具备中国证监会核准的证券投资咨询业务资格。证券投资咨询，是指从事证券投资咨询业务的机构及其投资咨询人员以下列形式为证券投资人或者客户提供证券投资分析、预测或者建议等直接或者间接有偿咨询服务的活动：接受投资人或者客户委托，提供证券投资咨询服务；举办有关证券投资咨询的讲座、报告会、分析会等；在报刊上发表证券投资咨询的文章、评论、报告，以及通过电台、电视台等公众传播媒体提供证券投资咨询服务；通过电话、传真、电脑网络等电信设备系统，提供证券投资咨询服务；中国证监会认定的其他形式。

发布证券研究报告是证券投资咨询业务的一种基本形式，指证券公司、证券投资咨询机构对证券及证券相关产品的价值、市场走势或者相关影响因素进行分析，形成证券估值、投资评级等投资分析意见，制作证券研究报告，并向客户发布的行为。

国信证券
GUOSEN SECURITIES

国信证券经济研究所

深圳

深圳市福田区福华一路125号国信金融大厦36层

邮编：518046　　总机：0755-82130833

上海

上海浦东民生路1199弄证大五道口广场1号楼12楼

邮编：200135

北京

北京西城区金融大街兴盛街6号国信证券9层

邮编：100032