大模型推动的通用机器人创新

Generalist Robot Innovation Driven by Large Models

王 芃(中国联合网络通信集团有限公司,北京 100033)
Wang Peng(China United Network Communications Group Co.,Ltd.,Beijing 100033,China)

摘 要:
在由ChatGPT引爆的新一轮人工智能浪潮下,大语言模型、推理模型、多模态语言模型、扩散模型等技术快速发展,相关的大模型技术也从离身智能向具身智能延伸,加快推动通用机器人领域的技术创新。分析了通用机器人产业的发展趋势和机遇,梳理总结了对通用机器人发展最为重要的大脑、小脑、数据、本体等4个方面的进展,以近2年业界若干重要的具身模型为例详细剖析了具身基础模型、具身推理模型、视觉语言动作模型等技术的发展脉络,最后对未来具身智能技术发展方向进行了展望。

关键词:
通用机器人;具身智能体;具身基础模型;视觉语言动作模型;具身推理模型
doi:10.12045/j.issn.1007-3043.2025.07.002
文章编号:1007-3043(2025)07-0008-07
中图分类号:TP18
文献标识码:A
开放科学(资源服务)标识码(OSID):

Abstract:
Amidst the new wave of artificial intelligence ignited by ChatGPT, technologies such as large language models, reasoning models, multimodal language models, and diffusion models have achieved rapid development. The relevant large model technologies have also extended from disembodied intelligence to embodied intelligence, accelerating the innovation of generalist robots. It analyzes the development trends and opportunities of the generalist robots industry, reviews and summarizes the progress in the four most important aspects for the generalist robots, namely the brain, cerebellum, data, and the robot itself. Taking several important embodied models over the past two years as examples, it provides a detailed analysis of the development course of technologies such as embodied foundation models, embodied reasoning models, and visual-language-action models. Finally, it gives an outlook of the embodied intelligence technology development.

Keywords:
Generalist robot; Embodied agent; Embodied foundation models; Vision-language-action models; Embodied reason models

引用格式:王芃. 大模型推动的通用机器人创新[J]. 邮电设计技术,2025(7):8-14.

1 通用机器人产业发展趋势和重要机遇

具身智能是可通过物理实体与环境交互并能从经验反馈中实现智能增长和行动自适应的智能系统,是将人工智能应用于物理世界的关键使能技术,也被认为是实现通用人工智能的基础路径。

具身智能机器人从形态上可分为固定基座型机器人、轮式机器人、履带机器人、四足机器人、人形机器人、仿生机器人等。虽然以工业机械臂和AGV为代表的传统专用机器人已在一定范围内应用于面向单一任务的结构化环境中,但是产业亟需更具通用智能、具备任务泛化能力、可高质量完成操作任务和运动控制的通用机器人。从训练数据构建难度、潜在市场规模、基础设施兼容性等维度来看,人形/类人形的形态是最适宜的通用机器人形态。经过多年发展,机器人的运动控制算法已从基于运动学/动力学模型的经典控制算法和以模型预测控制(Model Predictive Control,MPC)为代表的优化控制算法演进到基于深度

收稿日期:2025-06-10

强化学习的模仿学习算法，人形/类人形的机器人已可在遥控下较好地进行平衡行走或慢跑，或表演一套预设的舞蹈或武术动作，甚至进行前后空翻等高动态的动作；灵巧手在刚性物体抓取放置或柔性物体叠放等简单的操作类单点技能上已取得长足进步并展现了一定的环境适应性和针对各类物体的技能泛化性。然而，现阶段人形/类人形机器人在场景理解、任务规划及技能泛化、精细操作等方面距离通用机器人的愿景还有相当大的差距[1-5]。借鉴自动驾驶的技术分级方法，可针对移动、操作和交互等主要任务场景，将感知、决策、行动、技能等要素按照泛化能力、自适应性、智能水平等维度进行评估，把具身智能智能化等级分为五级。根据这个分级方法，现今具身智能技术大致处于L1~L2。工信部人工智能标准化技术委员会正在组织该分级分类标准的研究制定工作。

在技术发展上，随着大模型技术向具身智能领域的迁移应用和关节电机等机器人技术的逐步成熟，通用机器人已确定成为具身智能技术创新的重要领域。在产业发展上，据预测，到2029年中国人形机器人市场规模将从2024年的30亿元扩大到750亿元，到2030年有望达到3 000亿元。国内人形机器人已成为全球最重要的产业力量，在CES2025上，在黄仁勋身后的14款人形机器人中，有6款来自中国企业。在国家布局上，2023年11月，工信部发布了《人形机器人创新指导意见》。2024年5月，工信部与上海市政府成立“国家地方共建人形机器人创新中心”。2024年10月，工信部与北京市政府成立“国家地方共建具身智能机器人创新中心”。在商业闭环上，国内人形机器人厂商正积极探索面向娱乐表演、迎宾导引、零售物流、食品加工等场景的产业化规模应用。

面对通用机器人和具身智能的产业机遇，运营商可通过与国家级实验室和领先高校的产学研联合创新来夯实具身智能基础能力，发挥与下一代网络、大数据、网络安全等领域融合创新的协同优势，借助具身智能基础模型、端边云智能协同计算、机器人联网服务、机器人技能训练服务、机器人数据和运营管理服务等进入面向机器人产业的算网数智市场，培育和抢占随着机器人普及而拉动的增量联网通信市场，探索Robot as a Service模式下的运营服务体系，成为通用机器人产业重要的赋能者、聚合者和运营者，使具身智能成为推进网络向新、技术向新、服务向新的重要的驱动力。

2 通用机器人架构和关键技术

智能体（Agent）是一种能够感知环境并采取行动以实现特定目标的自主实体。近2年，大语言模型和多模态语言模型的快速发展进一步扩展了智能体在实际场景中的应用。当这些基于大语言模型和多模态语言模型的智能体嵌入在物理实体中，它们可以有效地将其能力从虚拟空间转移到物理世界，成为具身智能体（Enbodied Agent）（见图1）。

通用机器人作为一种高阶具身智能体，需要能够

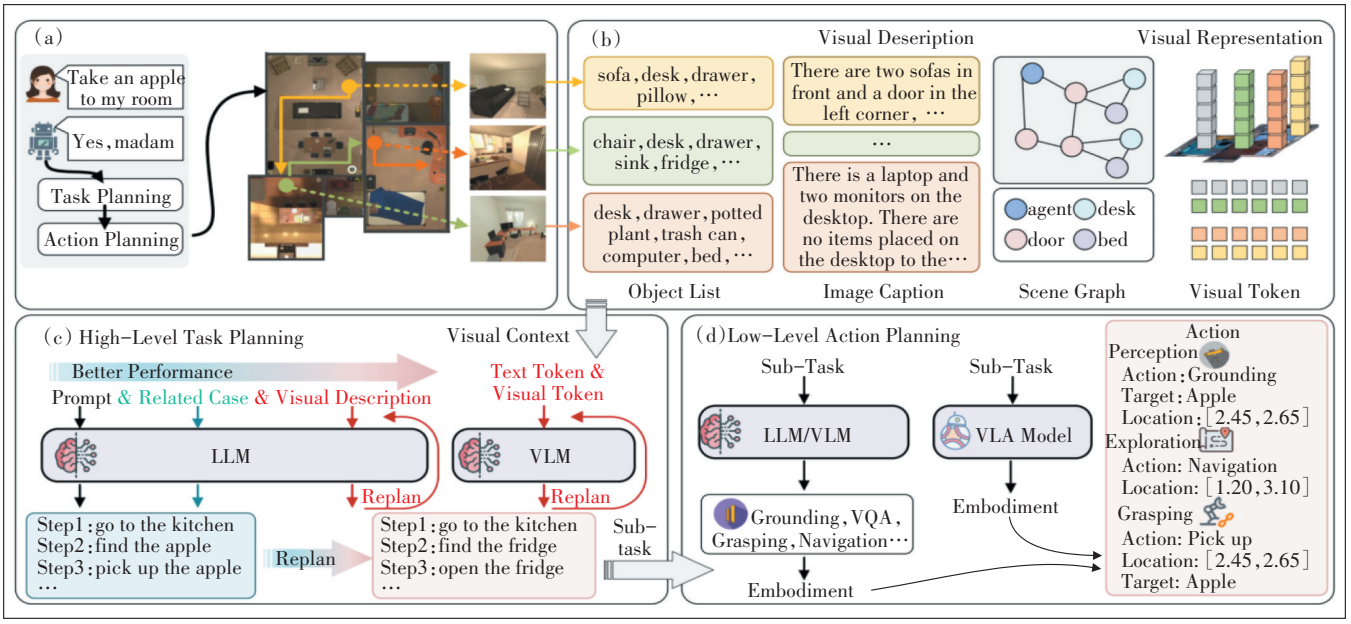

图1　基于大模型的具身智能体的整体架构[1]

根据预设目标或输入指令在开放物理世界中完成面向各类任务的感知、决策、规划、执行等工作。为了完成上述工作，具身智能体通常涉及2个层面的功能：一是在高层推理层完成自然语言交互、场景理解、目标/任务规划功能，将抽象且复杂的任务分解为具体的子任务，这称为高层次具身任务规划（Task Planning）；二是在感知和规划层，结合目标识别和定位、场景内的自我定位、移动行走等功能，完成动作规划，通过具身感知和具身交互模型或视觉语言动作模型[2-4]（Vision-Language-Action Models，VLA）逐步实现这些子任务，这称为低层次具身动作规划（Action Planning）。

正如人脑是由枕叶、颞叶、顶叶、额叶、运动皮层、小脑等不同脑区分别来实现视觉感知、听觉感知、空间导航、决策规划、动作规划、运动协调等功能，通用机器人也可由不同软件模块/模型来实现感知理解、机器人导航、推理规划、抓取与操作、运动控制等任务，而操作类任务则是大模型技术应用于具身智能领域的“皇冠”问题。本文后续章节也将主要围绕操作类任务讨论具身智能关键技术及具身基础模型。

借鉴人类的动作控制机制，在具身智能操作领域自然会想到也使用分层模型的思路，“大脑+小脑”模型成为通用机器人完成操作类任务的主流范式。大语言模型或视觉语言模型（即VLM）作为“具身大脑”负责根据用户意图、外部环境感知等输入，对长程任务进行任务规划、拆解为元技能序列等高层指令提供给“具身小脑”。具身大脑应具备“一脑多形”的能力，可以适配控制多种机器人本体，输出的控制指令一般为文本形式，内容中除了目标物体和应实施的动作技能外，还可包括关于物体可操作区域的信息或总体轨迹路径要求。“具身小脑”负责根据高层指令、本体状态、外部感知等输入来直接输出机器人的技能级的动作轨迹规划。“具身小脑”应支持“一机多能”的实现，可控制底层软硬件完成各项元技能的细节实现，其中VLA模型已成为“具身小脑”的主流实现方式。DeepMind发布的RT系列模型是VLA领域最为重要的前期工作，其中RT-2[6]展示了利用VLM中互联网规模的先验知识可有效提升模型对未见过任务的泛化能力，RT-X验证了使用多样化的跨本体数据进行训练可有效提升模型性能，RH-T提出的Language Motion理念对推进快慢系统模式的VLA模型以及如何利用人类操作示范数据有重要启发作用。本文第3章将对具身大脑和小脑模型的发展脉络进行进一步详细剖析。

在大模型的数据驱动范式下，具身数据集成为具身智能领域创新的核心驱动力或构建护城河的关键要素。针对操作类任务，通过机器人遥操作等方式收集的真实机器人轨迹数据（真机数据）是价值最高的训练数据。例如，Open X-Embodiment（简称OXE）是DeepMind牵头采集的开源数据集，包括针对室内和厨房环境下以抓取放置类任务为主的超百万条真机数据，覆盖了单臂和双臂机器人等22种本体。由于真机数据的采集需要大量人力、时间和采集设备的投入，使用仿真数据或模型生成数据（统称合成数据）用于模型训练成为具身智能领域的重要方式。业内已普遍借助生成式大模型的生成能力，构建面向“环境—任务—本体—轨迹”的跨模态数据生成方法和融合体系，形成专门的具身数据生成平台。此外，人类自身的操作视频中也蕴含着人类动作先验知识，利用互联网规模的人类操作视频也是解决人形机器人数据稀缺问题的重要方法。人类数据、合成数据、真机数据这3类数据的规模逐渐减小，而数据价值逐渐增大，构成了具身训练数据的“数据金字塔”3层架构。数据集以及相关数据构建流水线工具和平台也成为各具身智能训练场建设的重要内容。

在机器人本体方面，产业资本等力量正推动高功率密度关节电机、高力矩密度减速器、高能量密度动力电池、智能能源管理、触觉和听觉等高精度传感器、专用计算芯片等基础部组件加速产品迭代，推动机器人本体成本逐步降低。在本体开源方面，已出现了“Open Loong”“天工”“Isaac”等国内外开源的机器人硬件平台。根据通用机器人软件的功能属性和控制输出/交互频率，可将软件架构栈由高层至低层划分为4个层级。第1层级是高层推理层，输出/交互频率一般为1~5 Hz；第2个层级是感知和规划层，输出/交互频率为30 Hz左右；第3层级是实时底层控制框架层，主要功能包括基于MPC或深度神经网络策略的动作控制算法以及关节扭矩控制器、位置控制器和速度控制器等，输出/交互频率在100~1 000 Hz；第4层级是硬件抽象层，包括摄像头/麦克风、IMU、触觉传感器、执行器、位姿传感器等硬件的驱动程序。针对以上软件运行要求，国内外芯片厂商也逐步推出专用计算芯片来更好支撑机器人端侧计算需求。例如，基于Blackwell GPU架构的英伟达Jetson雷神芯片，内嵌了可原生支持FP4和FP8精度计算的Transformer加速引擎，提供多GPU实例功能、可实现在一个GPU上隔离运行具有

不同时延要求的应用。

大脑、小脑、数据和本体成为推动通用机器人发展的四大支柱，未来的发展还需要在高质量的机器人数据集构建、人类操作数据的高效利用、基于触觉的精细操作、复杂环境的认知、长程任务的规划执行、跨机器人协作、因果关系发现以及持续学习等方面进一步突破。

3 具身基础模型和VLA模型

类比离身智能领域由专用小模型向通用大模型的发展趋势以及大语言模型带动的基础模型概念，为了推动由专用机器人向通用机器人的发展，以情景、任务、物体、本体的全维度通用为最终目标的具身基础模型的概念也被提出。从发展脉络上看，具身基础模型在功能范畴上已从最初的简单任务规划延伸至复杂长程任务规划和高效地动作规划；在模型架构上，从模块化分层模式演进为端到端一体化快慢系统紧耦合模式，从离散动作Token的输出演进为连续动作块的生成；在训练数据上，从单本体数据到跨本体数据，从真机数据延伸至合成数据和人类数据。

在任务规划方面，大语言模型出现后就有Inner Monologue和ReAct等工作在尝试利用大语言模型的内在逻辑和知识将复杂任务拆分为可管理的步骤，但是由于缺乏实际视觉信息，复杂环境下拆解的子任务常会偏离实际情景，导致任务执行失败。因此直接使用能够在隐空间捕获视觉细节并对齐视觉特征和语义特征的VLM进行任务规划成为主流技术选择。随着OpenAI o1和DeepSeek R1[7]等推理模型的涌现，面向具身智能领域的具身推理模型也随之发展起来并成为实施任务规划的新选择。不同于一般推理模型相对擅长解决复杂数学或编程问题，具身推理模型则是强化提升了物理常识推理和具身推理能力，可以更好地理解空间、时间、基本物理规律等知识，加强了处理复杂感官输入、预测行动效果、遵循物理规律以及根据交互进行学习的能力。例如，英伟达在GTC 2025上发布的Cosmos Reason1是基于完成视频预训练和通用监督微调的VLM的基础上，进一步使用具身领域数据进行监督微调和基于GRPO算法的强化学习而得到的，可以根据视频输入对物理世界进行感知和理解、通过长思维链思考过程生成文本形式的决策输出[8]。

在动作规划方面，一种典型方案是向负责任务规划的语言模型或VLM模型提供已训练好的运动策略模型的定义和描述，由任务规划以工具调用的方式通过API实施相应动作。Figure AI的Figure 01机器人采用的就是这种方案，由VLM负责任务规划并调用负责动作规划的神经网络策略模型，神经网络策略模型以200 Hz的频率生成24自由度的动作指令作为底层以1 000 Hz频率运行的全身控制器（Whole Body Control，WBC）的执行锚点。上述分层规划范式下，任务规划和动作规划模型可分别开发、测试、优化，具有更大的系统灵活性，但独立的运动策略模型会增加控制时延，在实时任务中会影响反应时间和效率。因此随着大模型技术的成熟，使用VLA模型进行动作规划已成为重要趋势。VLA模型是基于预训练的VLM再使用具身数据集进行训练而得到的，利用了VLM中通过互联网规模的预训练数据而捕获的先验知识，可有效提升模型面向各类机器人任务的泛化能力；此外，与传统运动策略模型相比，VLM模型也可更方便地接收具身大脑以文本形式提供的任务规划指令。

2024年6月，由斯坦福大学等机构联合研发并开源的OpenVLA[9]是一个具有70亿参数的视觉语言动作模型（见图2），作为类RT-2的开源平替实现，在业界具有重要影响力。OpenVLA将动作预测问题处理为一个“视觉—语言”任务，并将机器人连续的动作映射为语言模型分词器中的离散化的token，从而使模型能够根据输入的观察图像和自然语言任务指令以预测下一个Token的方式来预测生成表示动作的Token，并通过De-Tokenizer转化为面向机器人末端执行器的动作指令。在OpenVLA模型中，模型将动作视为语义向量空间中一种新的语言进行处理，就像手语就是一种可以用动作表达语义的语言，机器臂的移动也是表达了某种语义。OpenVLA使用了从OXE数据集中选取的97万条真实单臂机器人操作轨迹进行的微调训练，通过量化优化可在英伟达4090 GPU上以约6 Hz的速度进行推理运行，也可通过参数高效微调来快速

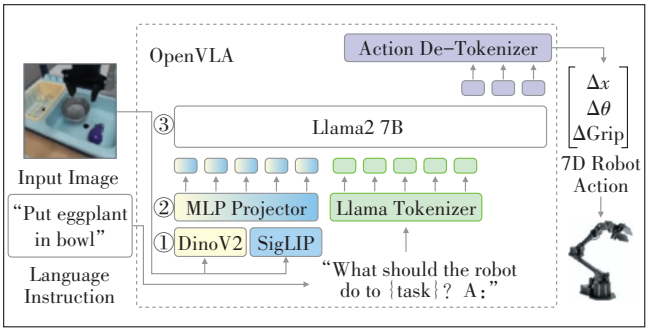

图2　OpenVLA模型架构[10]

适应新的机器人本体。

为进一步提升模型性能，VLA模型在模型架构上也在不断优化和创新。例如，在模型输入中加入本体状态（proprioceptive state），使用基于双向注意力机制的并行解码来提升模型推理速度，使用L1自回归或条件去噪扩散模型直接输出连续动作，也有Hybrid VLA等工作尝试在动作输出方面深度融合基于扩散模型的方式和基于Next Token Prediction的自回归方式。此外，自从ACT模型引入动作块（Action Chunking）方法后，后续VLA模型常会使用此方法，在预测生成动作Token时，不再仅预测下一个Token，而是每次都预测多个Token，可类比于DeepSeek V3[10]所使用的Multi-Token Prediction（MTP）方法。由于大部分具身操作任务的完成方法不是唯一的，训练数据可能会包括多种可行的运动轨迹，因此希望模型能够更好地表达这种多峰（Multi-Modal）的动作分布。主流方法是基于条件变分自编码器（Conditional VAE）或者使用流匹配（Diffusion Policy/Flow Matching）这些基于扩散模型的生成模型来表达这种分布[11]。Physical Intelligence于2024年10月发布的π_0模型[12]被认为是第一个使用流匹配方法生成高频动作块的VLA模型，能够生成面向复杂长程移动操作等场景的精准和流畅的连续动作，推动了VLM与扩散模型技术在具身领域的融合。π_0使用预训练的3B参数规模的PaliGemma作为VLM骨干网络，再增加一个300M参数的动作专家专门用于处理机器人状态和动作输入/输出（见图3），使用来自开源数据集和9亿条自采的跨本体的真机数据以多个目标来训练模型。

具身基础模型架构演进的一个重要里程碑是一体化快慢系统架构的出现。Figure.AI于2025年2月发布的VLA模型Helix被认为是第1个通过快慢系统架构实现上身全身控制的通用抓取任务模型，慢系统是一个7B参数的开源VLM，以7~9 Hz的频率输出隐空间的语义向量提供给快系统；快系统是一个仅具有0.08B参数的视觉运动策略模型，以200 Hz频率快速输出35个自由度的动作指令（见图4）。Helix通过标准的自回归损失函数进行快慢系统的统一端到端训练，快系统向慢系统的梯度后向传播也是通过隐空间的语义向量进行，以保证2个系统的联合优化。

视觉语言动作模型的泛化性提升在于利用的互联网规模的语义知识，然而面对机器人真实操作轨迹数据相对缺乏和仿真数据质量不高的现状，产业界又进一步考虑如何有效利用人类操作数据，并将其泛化应用到机器人操作。智元启元大模型GO-1是智元机器人在2025年3月发布的一款开源的通用具身基础模型，它有效利用人类操作示范进行学习。Go-1采用

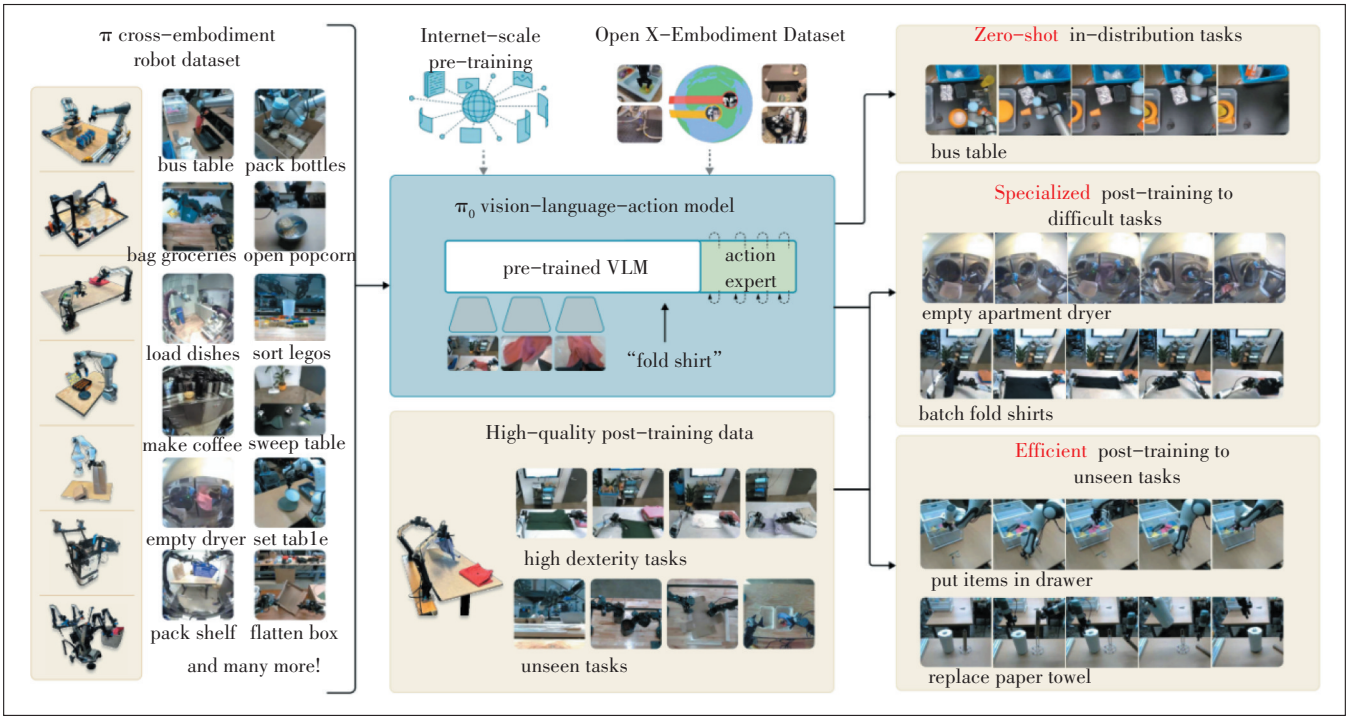

图3 π_0模型总体架构[12]

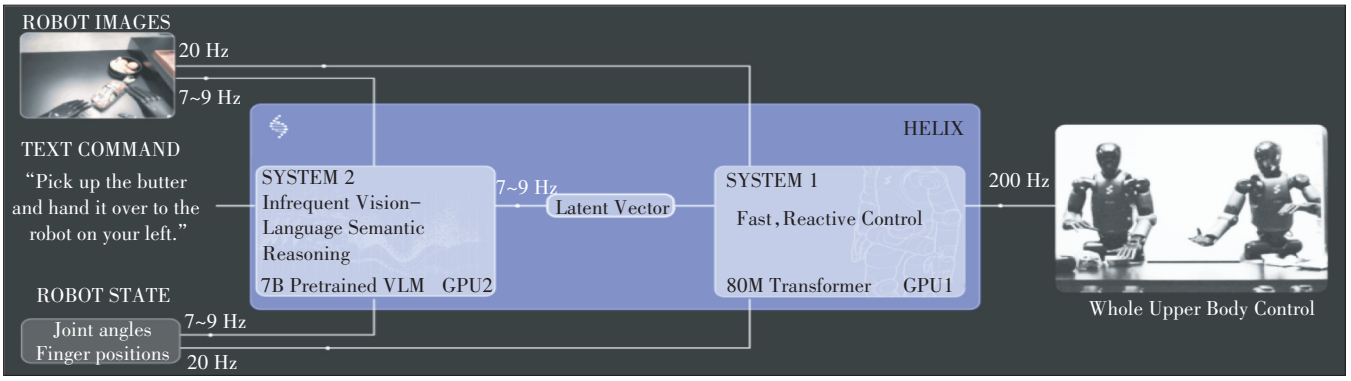

图4　Helix模型架构

了视觉语言隐空间动作架构(见图5),由VLM主干网络和2组混合专家(MoE)构成。隐式规划器(Latent Planner)作为一组专家,基于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,动作专家(Action Expert)作为另一组专家,基于百万量级机器人真实操作轨迹数据获得精细的机器人动作执行能力。

Groot N1是英伟达在GTC 2025上发布的一款面向人形机器人的开源具身基础模型,在合成数据构建和机器人本体适配方面进行了进一步创新。英伟达以"数据金字塔"的理念组合真机数据、合成数据(仿真数据和由视频生成模型生成的生成数据)和人类数据来构建训练数据集。在合成数据方面,使用基于真机数据微调的视频生成模型低成本地生成了10倍于真机数据的模型生成数据,同时使用DexMinicGen生成了78万条仿真轨迹。为了解决合成数据和人类视频数据缺乏动作标签的问题,训练了专门的VQ-VAE模型给人类视频数据和模型生成数据添加隐空间动作标签,训练了专门的逆动力学模型给仿真数据生成"伪动作"标签。在模型架构上,Groot N1也采用了端到端的快慢系统架构(见图6),并在模型中引入状态编码器、动作编码器和动作解码器以适配不同机器人本体(如傅里叶的GR-1、1X的Neo等机器人)。

虽然VLA模型在快速演进,但现阶段VLA模型在未见过物体、环境、任务、动作、本体方面的泛化性还有待改进,具身数据集的构建技术、面向预训练和后训练的训练数据组合方式以及训练数据与模型参数的最佳匹配关系也需要进一步的探索。现阶段VLA模型是基于模仿学习方法,无法避免模仿学习的局限和问题,可探索在动作规划层面VLA与强化学习结合,通过操作过程中的"在线进化"来提升模型性能。

4　具身基础模型和通用机器人发展展望

具身基础模型与世界模型内在的有机融合也被认为是有希望通往AGI的路线。具身基础模型一方面要在更丰富的具身数据集支持下继续推进模型架构上的创新,探索具身智能领域的Scaling Law,进一步提升全维度的泛化性;另一方面具身基础模型向上需融合群体智能相关的架构和协议,实现跨本体多机器人

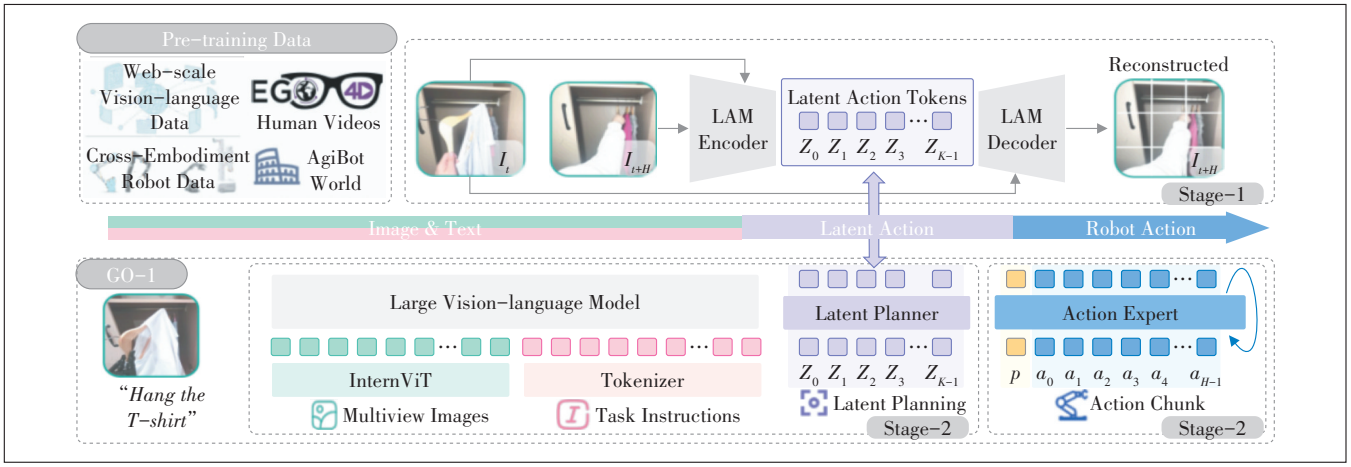

图5　Go-1模型架构

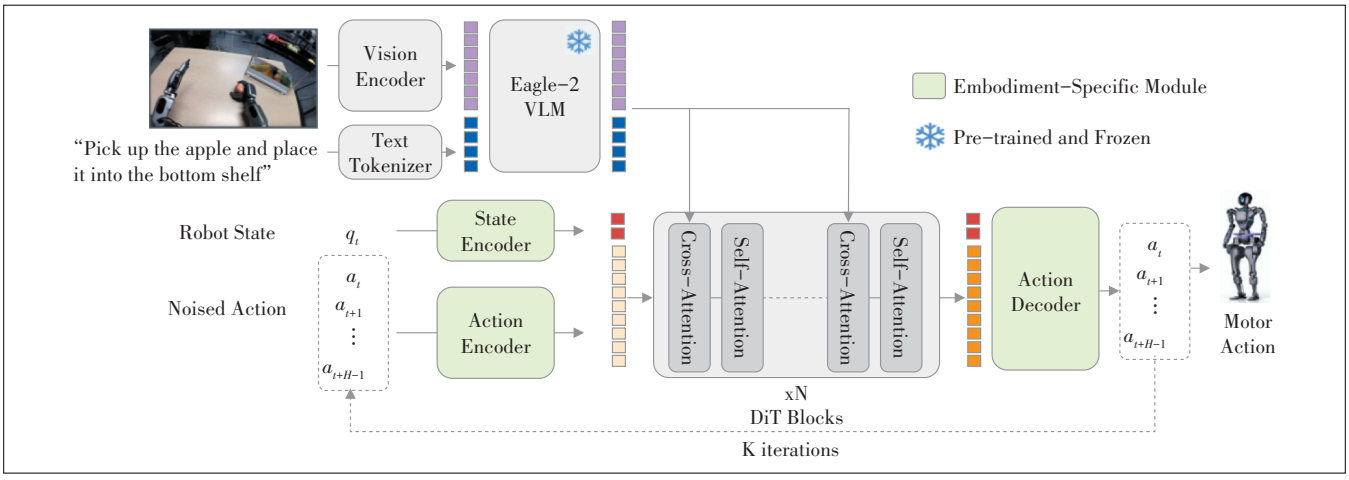

图6 Groot N1模型架构[13]

协同控制以及机器人间记忆、状态和技能的动态共享，向下需更高效地适配各类机器人本体和更为复杂的新任务，助力通用机器人的专业化技术技能提升，加快产业落地应用。

通用机器人及AGI的通用性更应强调机器人技能的“后装”可塑性而不是“预装”全面性，数据驱动范式是否能够推动人工智能走向真正的通用还存在较大疑问，学术界在探索通过赋予智能体完备的认知架构和价值体系来发展AGI，推进机器学习范式由“数据驱动”到“价值驱动”的迁移，由“大数据、小任务”方式到“小数据、大任务”的迁移。笔者相信对智能本质的认知及AGI的实现需要依赖脑科学研究的深入，以及在此基础上以系统科学的范式通过认知论、演讲论和调控论的研究路线来理解微观海量复杂神经连接（无论人类中枢神经系统、现有人工神经网络或新兴的脉冲神经网络）在宏观层次呈现智能的构建原理、涌现机制和调控手段。

参考文献：

[1] LIU Y, CHEN W X, BAI Y J, et al. Aligning cyber space with physical world: a comprehensive survey on embodied AI[EB/OL]. [2025-05-21]. https://arxiv.org/abs/2407.06886.

[2] SAPKOTA R, CAO Y, ROUMELIOTIS K I, et al. Vision-language-action models: concepts, progress, applications and challenges[EB/OL]. [2025-05-21]. https://arxiv.org/abs/2505.04769.

[3] MA Y, SONG Z X, ZHUANG Y Z, et al. A survey on vision-language-action models for embodied AI[EB/OL]. [2025-05-21]. https://arxiv.org/abs/2405.14093.

[4] 白辰甲，许华哲，李学龙. 大模型驱动的具身智能：发展与挑战[J]. 中国科学（信息科学），2024，54（9）：2035-2082.

[5] WADEKAR S N, CHAURASIA A, CHADHA A, et al. The evolution of multimodal model architectures[EB/OL]. [2025-05-21]. https://arxiv.org/abs/2405.17927.

[6] BROHAN A, BROWN N, CARBAJAL J, et al. RT-2: vision-language-action models transfer Web knowledge to robotic control[EB/OL]. [2025-05-21]. https://arxiv.org/abs/2307.15818.

[7] GUO D Y, YANG D J, ZHANG H W, et al. DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning[EB/OL]. [2025-05-21]. https://arxiv.org/abs/2501.12948.

[8] AZZOLINI A, BAI J J, BRANDON H, et al. Cosmos-reason1: from physical common sense to embodied reasoning[EB/OL]. [2025-05-21]. https://arxiv.org/abs/2503.15558.

[9] KIM M J, PERTSCH K, KARAMCHETI S, et al. An open-source vision-language-action model[EB/OL]. [2025-05-21]. https://arxiv.org/abs/2406.09246.

[10] LIU A X, FENG B, XUE B, et al. DeepSeek-V3 technical report[EB/OL]. [2025-05-21]. https://huggingface.co/papers/2412.19437.

[11] LIU S M, WU L X, LI B G, et al. RDT-1B: a diffusion foundation model for bimanual manipulation[EB/OL]. [2025-05-21]. https://arxiv.org/abs/2410.07864.

[12] BLACK K, BROWN N, DRIESS D, et al. π0: a vision-language-action flow model for general robot control[EB/OL]. [2025-05-21]. https://arxiv.org/abs/2410.24164.

[13] BJORCK J, CASTAÑEDA F, CHERNIADEV N, et al. GR00T N1: an open foundation model for generalist humanoid robots[EB/OL]. [2025-05-21]. https://arxiv.org/abs/2503.14734.

作者简介：

王芃，高级工程师，博士，主要从事人工智能和具身智能研发、科创生态合作等工作。