



2025 年深度行业分析研究报告



目录

CONTENTS

- 01 从人脑看具身大脑
- 02 英伟达主导现有大脑方案
- 03 大脑能力技术路线
- 04 国内外厂商大模型进展
- 05 L4及L5具身智能应用展望

01

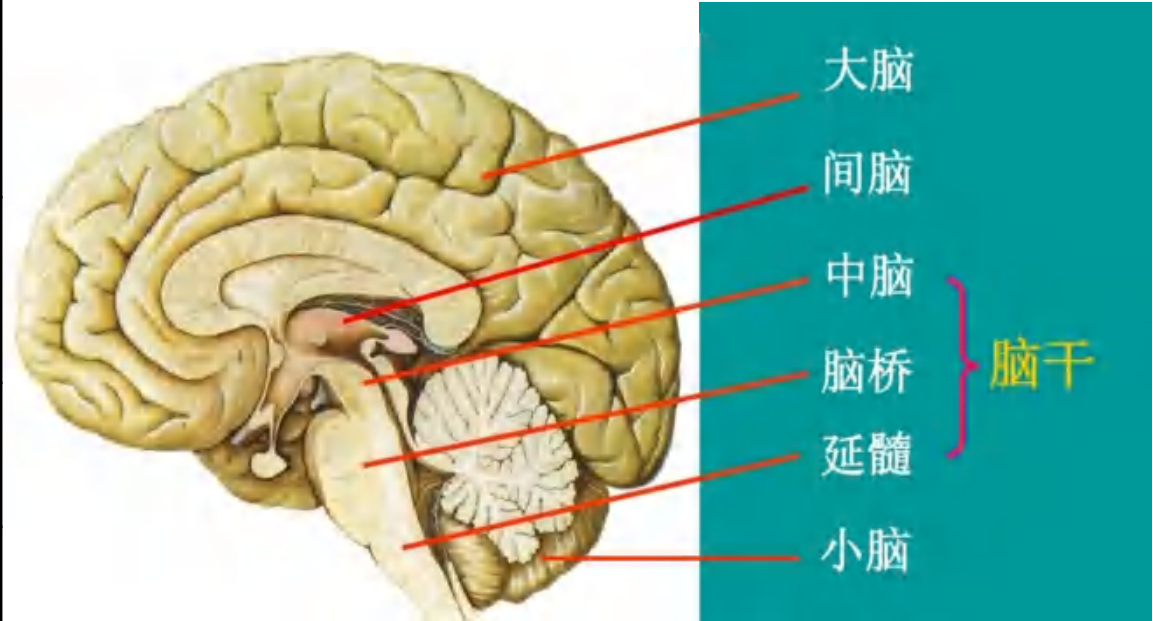
从人脑看
具身大脑

01

人脑结构拆解

了解机器人大脑之前我们不妨先了解一下我们人类的人脑。人脑由大脑、小脑、脑干以及间脑组成。其中大脑是人脑核心，其分为左右两个大脑半球，二者由神经纤维构成的胼胝体相连。脑干又可分为中脑、脑桥及延髓三部分。

部位	位置	功能
大脑	又称为端脑，位于脑的最上部，占人脑总体积的约85%，分为左右两个半球	高级认知： 负责思维、记忆、语言、决策、情感灯复杂功能 感觉与运动： 处理视觉、听觉、触觉等信息、并控制自主运动
小脑	位于大脑后下方，紧贴脑干后方，形似蝴蝶。	运动协调： 调节肌肉张力、协调精细动作（如鞋子、弹琴） 平衡与姿势： 帮助维持圣体平衡与空间定位 学习辅助： 参与运动技能学习
脑干	连接大脑与脊髓，由 中脑、桥脑、延髓 三部分组成。	生命维持： 控制呼吸、心跳、血压灯基本生命活动 信息中转： 船体大脑与脊髓之间的感觉和运动型号 反射控制： 管理咳嗽、吞咽、瞳孔反射灯原始反应
间脑	位于大脑半球深部，包裹在左右大脑半球之间。	丘脑： 感觉信息的中转，将视觉、听觉传递至大脑皮层 下丘脑： 调节体温、饥渴、睡眠周期，并控制内分泌系统



依照低层次感知到高层次感知逐个来看，人脑的**间脑与脑干**在机器人中对应的主要是（1）各类传感器及执行器自身部署的嵌入式驱动及算法。各类传感器及执行器包括旋转执行器、直线执行器、力传感器、温度传感器、姿态传感器等，此类部件一般内部配有MCU，其内部的嵌入式驱动及算法监控部件的各类状态，保证机器人部件的基本运作能力。（2）传递信息的线束及网关，起到各个控制器，传感器信息交互通联的作用。

部位	在人体中功能	在人形机器人中功能	机器人对应硬件
大脑	高级认知 ：负责思维、记忆、语言、决策、情感灯复杂功能 感觉与运动 ：处理视觉、听觉、触觉等信息、并控制自主运动	语义理解、环境信息理解、动作决策等	目前为机器人中央控制器担任此角色，但目前并未获得相应能力。后续可能在此基础上进一步增加硬件及算力
小脑	运动协调 ：调节肌肉张力、协调精细动作（如穿鞋子、弹琴等） 平衡与姿势 ：帮助维持身体平衡与空间定位 学习辅助 ：参与运动技能学习	动作学习模仿、复杂动作控制等	机器人中央控制器，即现有的机器人“大脑”
脑干	生命维持 ：控制呼吸、心跳、血压灯基本生命活动 信息中转 ：传递大脑与脊髓之间的感觉和运动信号 反射控制 ：管理咳嗽、吞咽、瞳孔反射灯原始反应	电源管理、通信网关控制、执行器控制器状态管理等	机器人各传感器，执行器，线束，网关
间脑	丘脑 ：感觉信息的中转，将视觉、听觉传递至大脑皮层 下丘脑 ：调节体温、饥渴、睡眠周期，并控制内分泌系统	电源管理、通信网关控制、执行器控制器状态管理等	机器人各传感器，执行器，线束，网关

人脑的**小脑**在机器人中对应的主要是动作学习模仿训练以及复杂动作的控制。而在机器人行业中，目前通常被称为机器人“大脑”，这主要是因为相对于工业机器人，具有“大脑”的人形机器人对复杂运动的学习掌握能力明显增强，比如近期宇树、众擎、波士顿动力等公司在视频中展示的人形机器人执行舞蹈，空翻高难度动作。至于对应人脑中的**大脑**的角色的硬件，目前人形机器人厂商多用中央控制器担任此角色，但对于高级认知，信息处理能力尚未建立。展望未来，机器人大小脑有望实现分离，大脑算力进一步加强，小脑专注运动控制。

部位	在人体中功能	在人形机器人中功能	机器人对应硬件
大脑	高级认知 ：负责思维、记忆、语言、决策、情感灯复杂功能 感觉与运动 ：处理视觉、听觉、触觉等信息、并控制自主运动	语义理解、环境信息理解、动作决策等	目前为机器人中央控制器担任此角色，但目前并未获得相应能力。后续可能在此基础上进一步增加硬件及算力
小脑	运动协调 ：调节肌肉张力、协调精细动作（如穿鞋子、弹琴等） 平衡与姿势 ：帮助维持身体平衡与空间定位 学习辅助 ：参与运动技能学习	动作学习模仿、复杂动作控制等	机器人中央控制器，即现有的机器人“大脑”
脑干	生命维持 ：控制呼吸、心跳、血压灯基本生命活动 信息中转 ：传递大脑与脊髓之间的感觉和运动信号 反射控制 ：管理咳嗽、吞咽、瞳孔反射灯原始反应	电源管理、通信网关控制、执行器控制器状态管理等	机器人各传感器，执行器，线束，网关
间脑	丘脑 ：感觉信息的中转，将视觉、听觉传递至大脑皮层 下丘脑 ：调节体温、饥渴、睡眠周期，并控制内分泌系统	电源管理、通信网关控制、执行器控制器状态管理等	机器人各传感器，执行器，线束，网关

02

英伟达主导现
有大脑方案

根据宇树、众擎、松延动力等厂商官网说明，可以发现目前主流机器人厂商都会将算力模组分为高低两个版本提供用户，其中低配的基础算力模组可以完成一些遥控的走路、跑步、跳跃等基础性动作，二次开发部署潜力较弱，起售价均在10万元以内。宇树、众擎使用英特尔 Core i5及N97作为基础算力模组CPU，其中i5为成熟民用CPU而N97为一款低功耗移动处理器，专为轻量级计算和嵌入式场景设计。如果极客开发者、院校培训、创业者等需要对机器人进行二次开发，则需使用高算力模组的高配版本，整机价格可能会有数倍的价差。目前，绝大部分厂商的高算力模组主要还是基于英伟达Jetson Orin平台进行开发。

厂商	宇树	宇树	宇树	众擎	松延动力
型号	GO2	G1	H1-2	PM01/02	N2
外形					
基础算力模组	8核CPU	8核CPU	英特尔 Core i5	英特尔 N97	未知
高算力模组	英伟达Jetson Orin Nano/NX	英伟达Jetson Orin NX	英特尔 Core i7或英伟达Jetson Orin NX	英伟达Jetson Orin NX	英伟达Jetson Orin
起售价 (不含高算力模组)	9997元	9.9万元起	未知	8.8万元起	3.99万元起

英伟达 Jetson系列是专为机器人和嵌入式边缘AI应用设计的计算平台，由Jetson模组、JetPack SDK和生态系统组成，加速软件开发。

Jetson系列的主要成员包括Jetson Nano、Jetson TX2和Jetson AGX Xavier，适用于不同的应用场景。Jetson Nano是最小的设备，配备了128核心GPU和四核ARM Cortex-A57 CPU。Jetson Xavier系列模组具有高达32 TOPS的AI性能，适用于自主机器的视觉测距、传感器融合、定位和地图构建等应用。



Jetson Thor

2024年GTC大会上，英伟达还发布了一款基于 NVIDIA Thor 系统级芯片（SoC）的新型人形机器人计算机 Jetson Thor。Jetson Thor 是一个全新的计算平台，能够执行复杂的任务并安全、自然地与人和机器交互，具有针对性能、功耗和尺寸优化的模块化架构。

该 SoC 包括一个带有 transformer engine 的下一代 GPU，其采用 NVIDIA Blackwell 架构，可提供每秒 800 万亿次8位浮点运算 AI 性能，以运行 GR00T 等多模态生成式 AI 模型。凭借集成的功能安全处理器、高性能 CPU 集群和 100GB 以太网带宽，大大简化了设计和集成工作。

波士顿动力 Atlas使用Jetson Thor



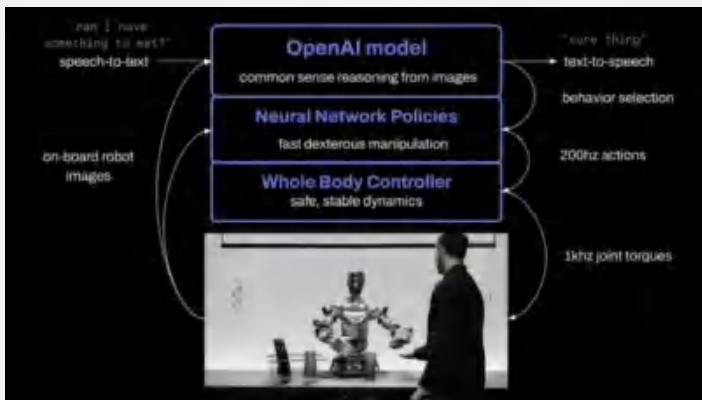
03

大脑能力
技术路线

具身智能的算法方案分为分层决策模型和端到端模型两种路线。

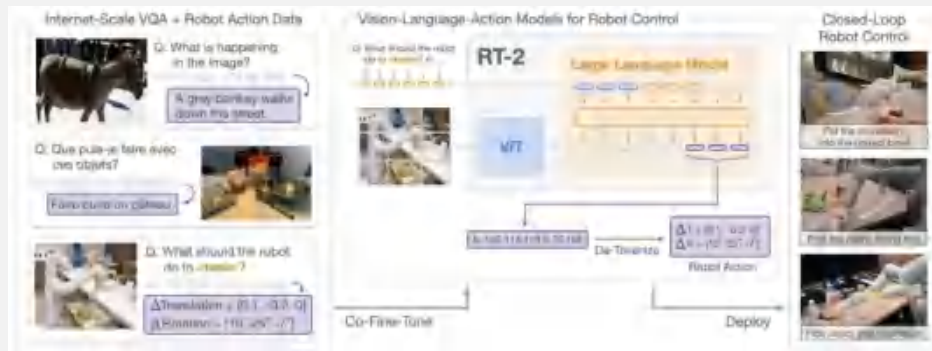
分层决策模型

以「Figure 01」为代表，将任务分解成不同层级，以多个神经网络训练，再以流程管线的方式组合。Figure 01顶层接入OpenAI的多模态大模型，提供视觉推理和语言理解；中间层神经网络策略作为小脑进行运动控制并生成动作指令；底层机器人本体接受神经网络策略的动作指令，进行控制执行。分层决策模型的缺点是：不同步骤间的对齐和一致性需解决。



端到端模型

以「Google RT-2」为代表，通过一个神经网络完成从任务目标输入到行为指令输出的全过程。首先在大规模互联网数据预训练视觉语言模型，然后在机器人任务上微调，结合机器人动作数据，推出视觉语言动作模型。RT-2不仅负责最上层的感知与规划，还参与中下层的控制与执行，打通了端到端的链路。端到端模型的缺点是：训练数据海量、消耗资源巨大、机器人执行实时性差。



具身智能的训练方法可分为模仿学习和强化学习两种路线。

模仿学习

模仿学习——

智能体通过观察和模仿专家（经验丰富的人类操作者或具有高级性能的系统）的行为来学习任务。

- 优势：可以快速学习专家策略，无需复杂的探索过程
- 劣势：学习到的行为策略受限于专家数据，对于未见过的情况泛化能力较差

强化学习

强化学习——

智能体通过与环境的交互来学习最佳行为策略，以最大化某种累积奖励。

- 优势：能够通过探索环境学习未知的策略；可以处理高度不确定和动态变化的环境
- 劣势：需要大量的探索和试错，学习过程缓慢；对于复杂任务，设计合适的奖励函数难度较高

具身智能的数据采集可分为基于仿真环境数据和基于真实世界数据两种路线。

基于仿真环境的数据采集（Sim2Real）

Sim2Real（Simulation to Reality）——
在仿真环境中学习技能和策略，并迁移到现实世界中。

- 优势：数据可大规模获取，成本低
- 劣势：对仿真器要求高，仿真环境与真实世界存在差异；迁移过程中存在性能下降

基于真实世界的数据采集

基于真实世界数据采集——直接从现实世界数据中学习，
包括本体采集、遥操作、动态捕捉、视频学习等方式。

- 优势：数据更真实可靠
- 劣势：数据少、泛化性差；通过机器本体和人采集，成本高、难度大、效率低

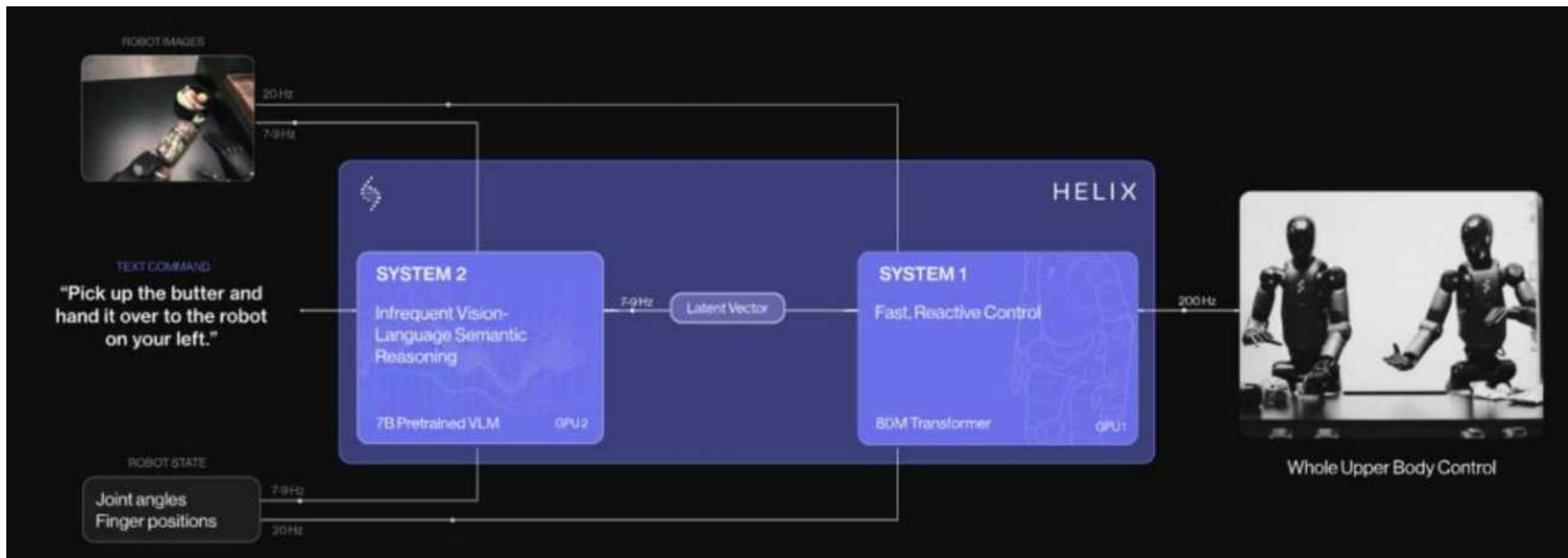
04

国内外厂商
大模型进展

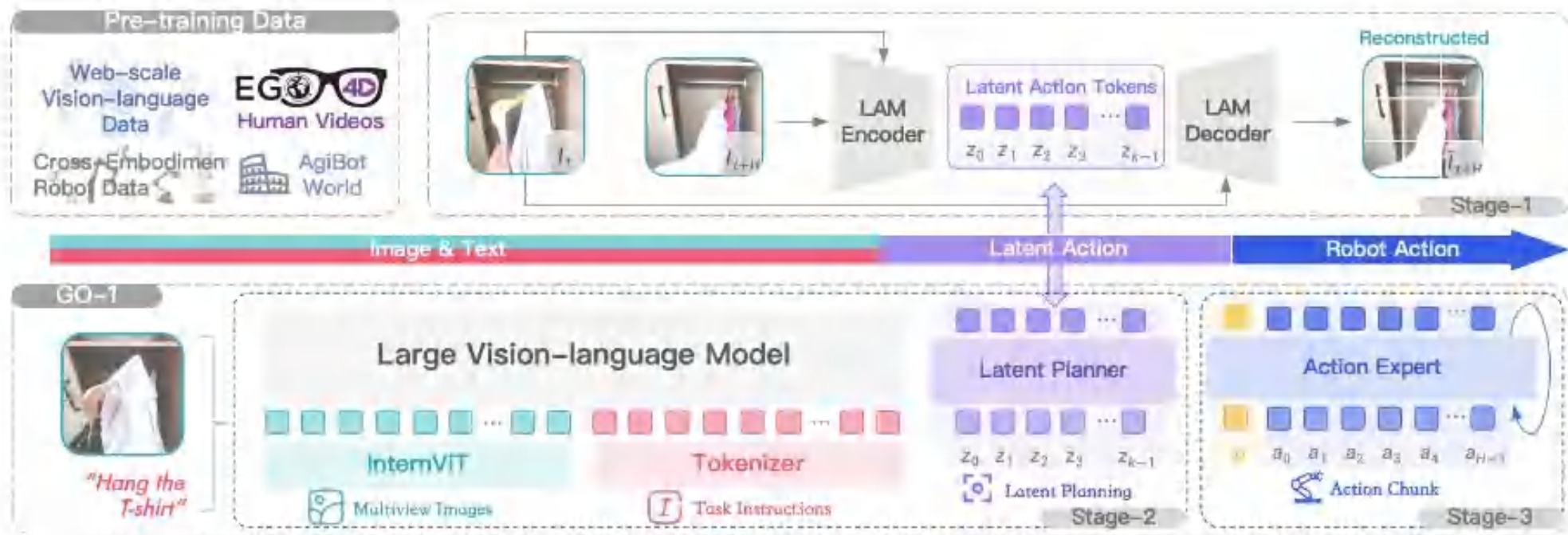
2025年1月，银河通用发布抓取基础大模型GraspVLA。GraspVLA的训练包括预训练及后训练。其中预训练完全基于合成大数据，训练数据达到了有史以来最大的数据体量——十亿帧「视觉-语言-动作」对，掌握泛化闭环抓取能力、达成基础模型；预训练后，模型可直接Sim2Real 在未见过的、千变万化的真实场景和物体上零样本测试，全球首次全面展现了七大卓越的泛化能力，满足大多数产品的需求；而针对特别需求，后训练仅需小样本学习即可迁移基础能力到特定场景，维持高泛化性的同时形成符合产品需求的专业技能。



2025年2月，Figure AI发布了人形机器人VLA通用大模型。为了解决视觉语言模型“通用、但不快速”，和机器人视觉运动策略“快速、但不通用”的矛盾，Figure通过建立一套互补的系统进行权衡。两套系统通过端到端训练以进行通信。其中，**系统2**是开源、开放权重的70亿参数量端侧互联网预训练视觉语言模型，用于理解场景和自然语言；**系统1**是一个8000万参数量的快速反应视觉运动策略，将系统2理解的语义转化为每秒200次的精确连续机器人动作。



3月10日，智元发布首个通用具身基座模型——智元启元大模型（Genie Operator-1），提出了Vision-Language-Latent-Action (ViLLA) 架构，该架构由VLM+ 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(动作专家)借助百万真机数据获得**精细的动作执行能力**，三者环环相扣，实现了可以利用人类视频学习，完成小样本快速泛化，降低了具身智能门槛，持续进化，将具身智能推上了一个新台阶。



The framework of GO-1

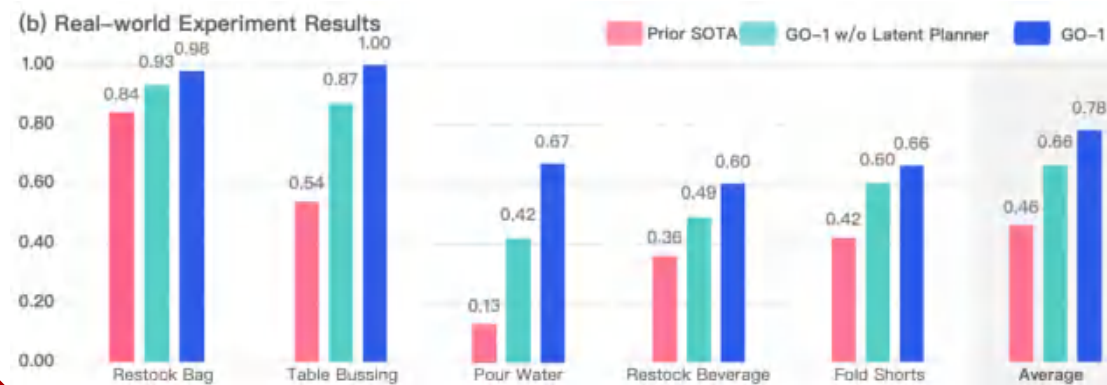
ViLLA通过预测Latent Action Tokens，弥合图像-文本与机器人执行动作之间的鸿沟

目前全球机器人高质量带动作标签的真机数据量仍然有限，远少于互联网规模的数据集。智元采用Latent Actions（隐式动作）来建模当前帧和历史帧之间的隐式变化，然后通过Latent Planner，预测这些Latent Actions，从而将异构数据源中真实世界的动作知识转移到通用操作任务中。

通过ViLLA创新性架构，在五种不同复杂度任务上测试 GO - 1，相比已有的最优模型，GO - 1成功率大幅领先，平均成功率提高了32%(46%→78%)。其中“Pour Water”（倒水）、“Table Bussing”（清理桌面）和“RestockBeverage”（补充饮料）任务表现尤为突出。

此外智元还单独验证了ViLLA架构中Latent Planner的作用，可以看到增加Latent Planner可以提升12%的成功率(66%→78%)。

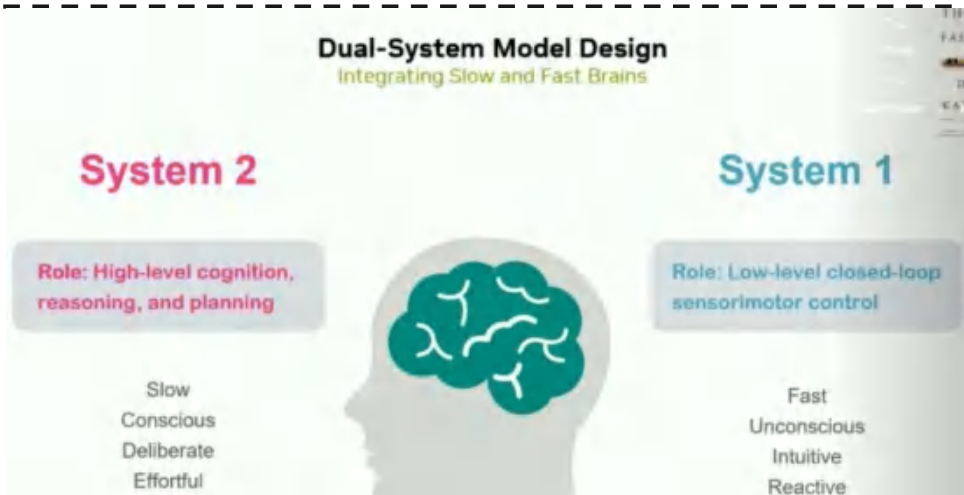
采用Diffusion Model作为目标函数来建模低层级动作的连续分布。Action Expert结构设计与Latent Planner类似，也是与VLM主干网络共享相同的Transformer结构，但使用两套独立的FFN和Q/K/V/O投影矩阵，它通过Denoising Process（去噪过程）逐步回归动作序列。



(a) Experiment Setups for Policy Evaluation

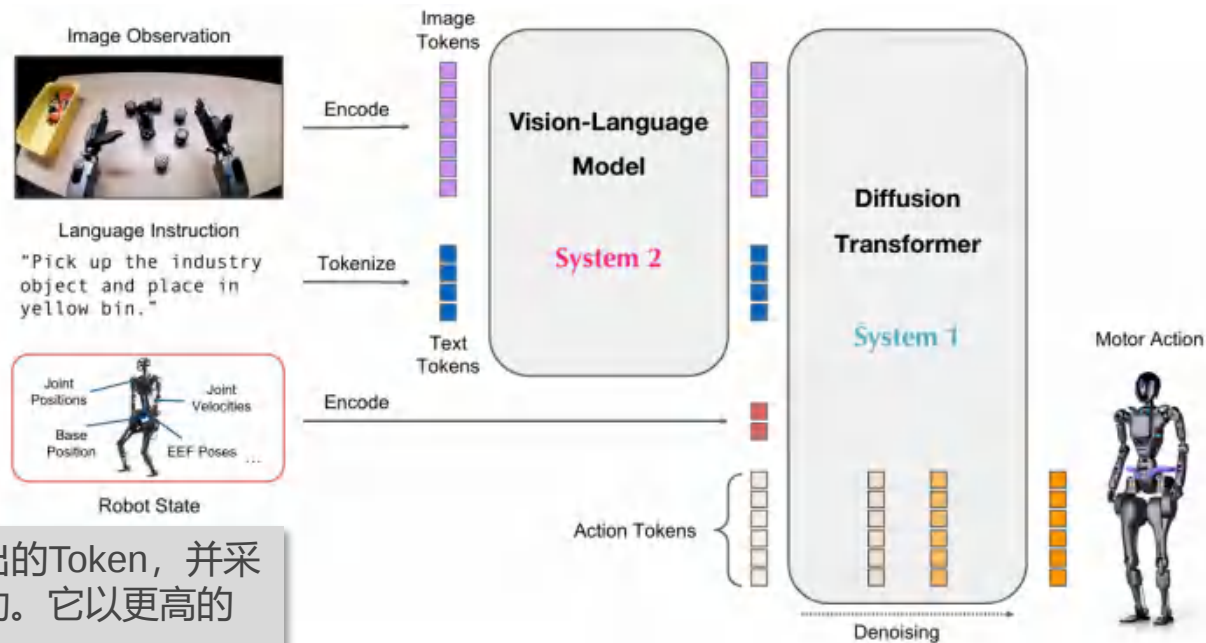


2025年英伟达GTC大会上，英伟达发布了通用机器人基座大模型GR00T N1。GR00T N1 灵感来自于人类认知处理，是一个具有双系统架构的视觉-语言-动作（VLA）模型。视觉-语言模块（系统2）通过视觉和语言指令来理解环境。随后的扩散转换器模块（系统1）实时生成流畅的运动动作。这两个模块紧密耦合，并进行端到端联合训练。英伟达使用真实机器人轨迹、人类视频和合成生成的数据集的异构混合物来训练 GR00T N1，解决了具身智能所面临的数据稀缺和多样性不足的问题，从而提升机器人在复杂环境中的操作能力。英伟达表示，GR00T N1 在多个机器人实施例的标准模拟基准上优于最先进的模仿学习基线。



System1是一个扩散转换器（DiT）作为动作模块。它交叉关注 VLM 输出的Token，并采用特定于实施例的编码器和解码器来处理可变状态和动作维度以产生运动。它以更高的频率（120Hz）产生闭环电机动作。

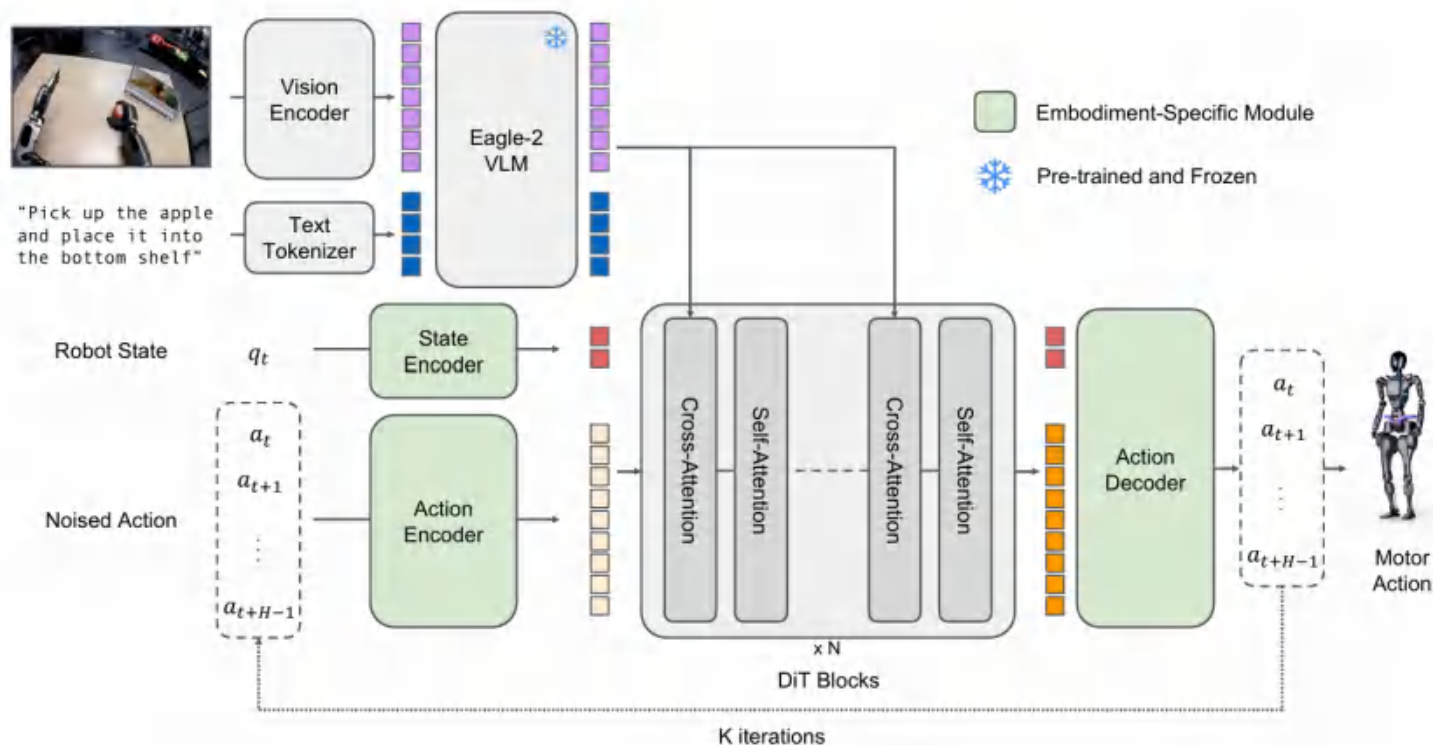
System 2是一个预先训练的视觉语言模型（VLM），它处理机器人的视觉感知和语言指令，以解释环境并实现任务目标，它可以在英伟达 L40 GPU 上以 10Hz 运行。



VLM模块主要功能是从图像和文本指令中提取特征，并为后续的动作生成提供上下文信息。GR00T N1使用Eagle-2 VLM作为其基础，该模型在大规模互联网数据上进行了预训练。

输入处理：VLM模块接受图像观测和文本指令作为输入。图像被编码为 224×224 的分辨率，经过像素重排后生成64个图像token嵌入。这些嵌入与文本嵌入结合，形成一个丰富的多模态表示。

特征提取：通过对图像和文本的共同编码，模型能够提取出具有高度相关性的特征。在策略训练过程中，任务的文本描述和可能的多张图像会以对话格式传递给VLM，从而获得形状为（批量大小 $\times$ 序列长度 $\times$ 隐藏维度）的多模态特征。



DiT模块负责处理机器人的状态和动作生成。它采用了一种变体的扩散变换器，结合了去噪步骤的自适应层归一化。

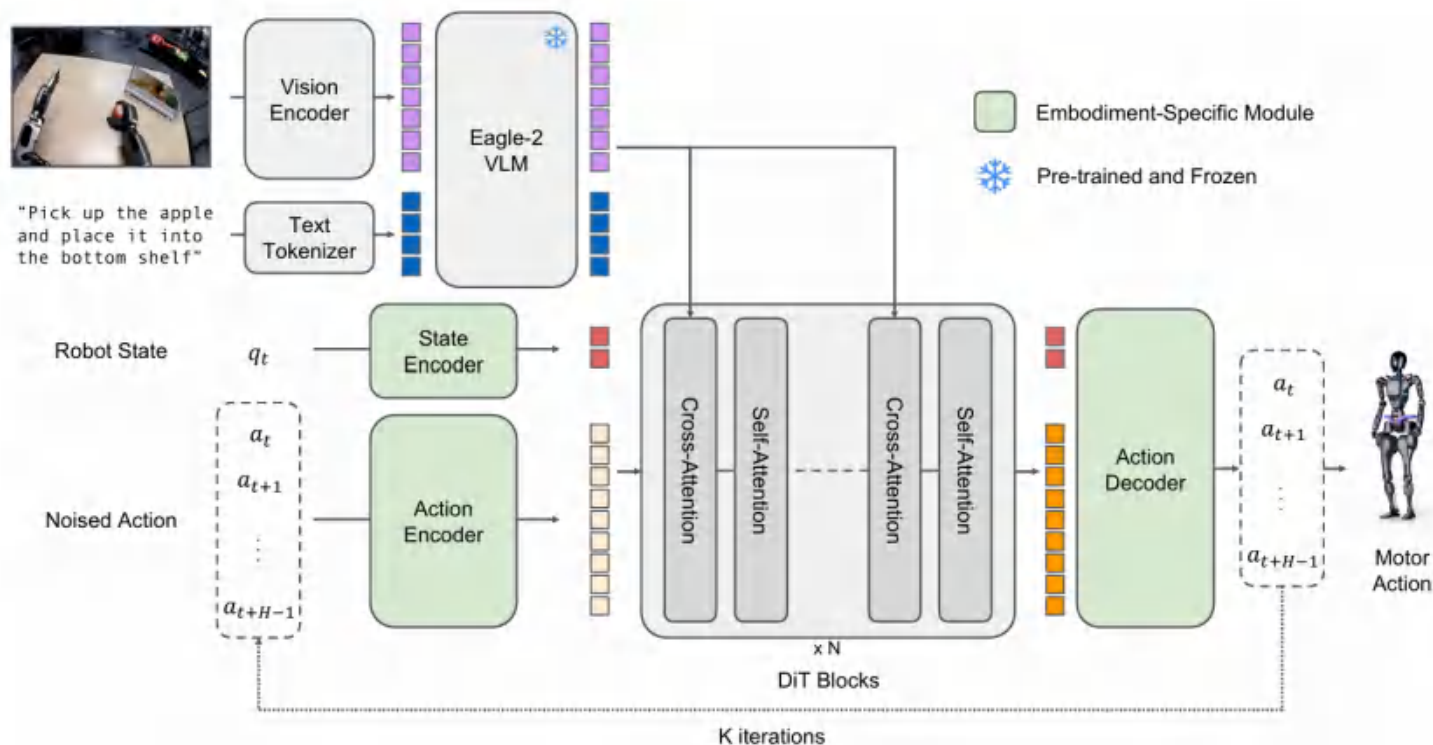
状态和动作编码器：针对不同机器人构型的状态和动作，GR00T N1为每种构型使用一个多层感知机（MLP）进行投影，确保所有输入映射到一个共享的嵌入空间。

流匹配机制：GR00T N1使用流匹配（Flow-Matching）技术，通过迭代去噪来生成动作。该模型接收噪声动作、机器人的状态编码、图像token和文本token作为输入。

交叉注意力和自注意力：在DiT中，自注意力模块处理噪声动作token嵌入和状态嵌入，而交叉注意力模块则允许模型根据VLM输出的视觉-语言token嵌入进行条件化。

这样的设计使得模型能够充分利用视觉和语言的信息来指导动作生成。

动作解码器：在最后的DiT模块之后，使用一个特定于构型的动作解码器（另一个MLP）。



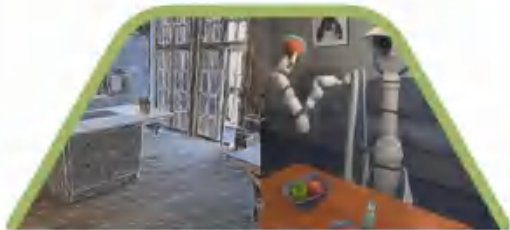
与自动驾驶的训练获取较为简单相比，通用机器人模型数据规模少且存在“数据孤岛”问题，因此无法建立互联网级统一数据集。对此，英伟达将视觉语言动作（VLA）训练语料库构建成一个数据金字塔，整合异构数据源，构建覆盖不同抽象层次的训练数据体系。这个数据金字塔共分为三层：（1）大量网络数据和人类视频构成金字塔的底层；（2）通过物理模拟生成和 / 或借助现成神经模型增强得到的合成数据形成中间层；（3）在实体机器人硬件上收集的真实世界数据则构成顶层。金字塔的底层提供广泛的视觉和行为先验知识，而顶层确保模型能在实体机器人执行任务时落地应用。

Real-World Data



顶层：来自各种构型的机器人在真实环境中的操作数据，数据收集成本高，但对于模型的准确性至关重要。特点是此层用于确保模型在真实环境下的执行能力。

Synthetic Data



中层：来自视频生成及基于物理模拟器生成的轨迹数据及机器人操作轨迹数据。其特点是数据量中等，成本低。如使用多模态LLM对视频进行过滤，并重新标注。最终生成视频数据相当于原始数据的约10倍。

Web Data & Human Videos



底层：来自互联网及人类活动的视频。其特点是数量多成本低。特点是通过大规模互联网数据对视觉语言模型进行预训练，提供了丰富的语义信息。

05

L4及L5具身 智能应用展望

今年两会上，小鹏汽车董事长、CEO何小鹏认为机器人可分为五个智能等级：L1级（完全由人控制）、L2级（基础辅助智能）、L3级（具身智能和训练监督）、L4级（自成长智能）和L5级（完全自主智能）。当前，人形机器人产业正朝着L3级迈进，这一阶段的机器人能够在大量场景中独立运行，但在复杂情况下仍需人工监督。何小鹏预计，到2026年，具备L3初阶能力的人形机器人将进入适度规模的商业化量产阶段。结合何小鹏对于机器人分级以及我们对于机器人能力程度的拆解，我们将具身机器人分级进一步细化。

	L1	L2	L3	L4	L5
等级分类	完全由人控制	基础辅助智能	具身智能和训练监督	自成长智能	完全自主智能
代表产品	传统机械产品	工业机器人、AGV、扫地机器人	现有人形机器人	尚未出现	尚未出现
拥有的能力	被人类完全操控的能力	基于人类制定规则进行执行的能力，需要人类监督	基于人类指定的规则及训练的成果进行一定的自主运行的能力，复杂情况下仍需人工监督	基于人类指定的规则及训练的成果进行一定的自主运行的能力，少量人工监督，具有自成长能力	无需人类监督，能够在人类规则下自主智能化运行，并且具有自成长能力，具有情感化交流能力
对于大脑能力的需求	解析指令，执行指令	感知环境，解析规则，自动化执行	多模态感知理解，解析规则，自动化解析执行复杂的运动	多模态感知理解，解析规则，自动化解析执行复杂的运动，自主迭代优化能力	多模态感知理解及执行，自主迭代，情感化交流能力，成为机器与AI Agent的结合体
信息/指令获取端口	遥控器，控制器	遥控器，指令集，摄像头、雷达等传感器	遥控器，指令集，文本，摄像头，雷达，麦克风等各类传感器	遥控器，指令集，文本，摄像头，雷达，麦克风等各类传感器	遥控器，指令集，文本，摄像头，雷达，麦克风等各类传感器
应用场景	工业制造、科研机构	工业制造、家庭清扫、特种应用	工业制造、科研机构、特种应用	工业制造、科研机构、公共服务、特种应用	工业制造、科研机构、公共服务、特种应用、大众生活

L4级具身智能预计3-5年内落地，场景还是以B端为主，主要玩家为本体厂商。随着L5级具身的到来，C端应用落地。机器人核心能力将从**理解世界做任务**向**理解人性作伴侣**转变，互联网大厂依靠独特数据优势将在L5级具身智能竞赛中取得更多优势。



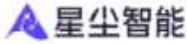
L4级具身智能机器人-理解世界做任务

- **与人类关系：**硅基佣人或硅基工人
- **应用场景：**B端为主，少量C端。主要为工业制造，商业服务等。如餐厅服务员，清洁工，制造车间工人等
- **智能化能力：**在工商业特定活动范围内独立自主完成工作，具有一定的特定场景下的泛化学习能力
- **行动能力：**运行环境封闭固定，轮式或者轮足式为主，达到安静稳定与运行效率的平衡
- **主要厂商：**现有机器人本体厂商为主，互联网大厂为辅
- **应用落地时间：**未来3-5年内落地，并在相当一段长时间内占据具身智能主流



L5级具身智能机器人-理解人性做伴侣

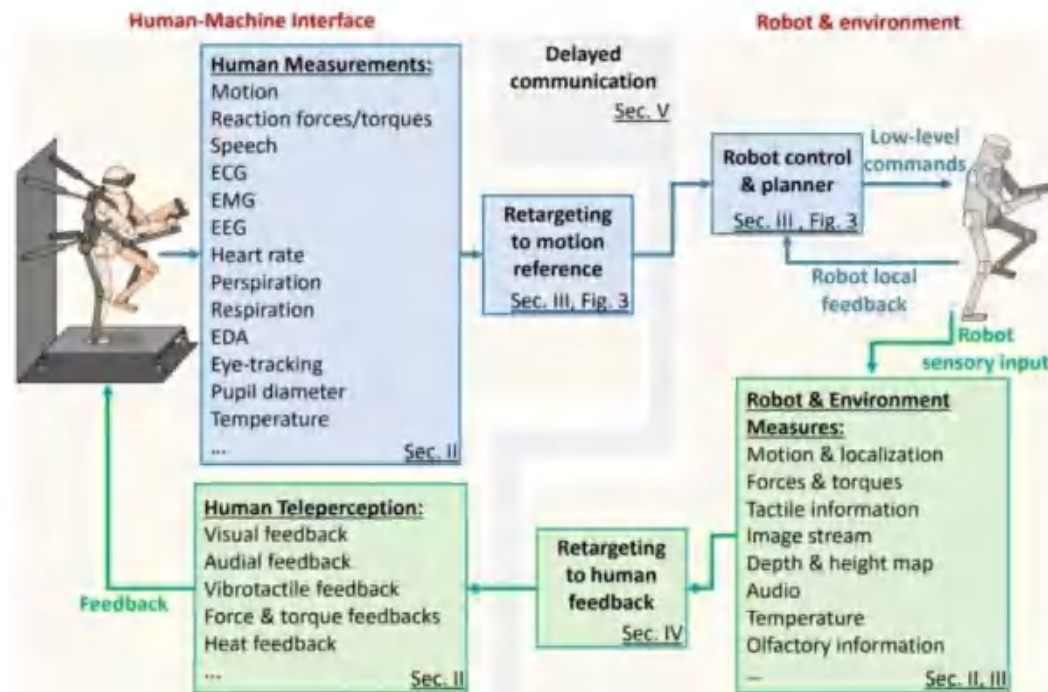
- **与人类关系：**硅基家人，人类理想中的人形机器人
- **应用场景：**C端为主，B端为辅。主要为家用机器人，在家庭内担任家务助理及个人事务助理
- **智能化能力：**在家庭内高度复杂场景完成自主工作，并熟悉理解家庭成员个人信息，成为家庭实体AI Agent
- **行动能力：**轮式及足式各一半，如无需外出，轮式即可
- **主要厂商：**互联网大厂依靠数据优势介入C端市场，本体厂商或接入互联网大厂能力或为互联网大厂代工
- **应用落地时间：**未来5-10年内落地

公司	星尘智能 	银河通用 	普渡科技 	星海图 
型号	S1	G1	闪电匣 Arm	R1
外形				
官方定位	全能型AI机器人	可泛化操作的具身大模型机器人	类人形服务机器人	全尺寸双臂仿生人形机器人
性能参数	身高：170cm 体重：80kg 臂展：194cm 续航：4~6小时 单臂负载：5kg 单臂自由度：7	身高：173cm 体重：85kg 臂展：190cm 续航：10小时 单臂负载：5kg 垂向作业空间：0-2.4m	身高：144cm 臂展：160cm 显示屏幕：10.1寸触屏 续航：8h（空载） 舱体载重：15kg 单臂自由度：7 建图方式：VSLAM+激光SLAM	身高：170cm 单臂负载：3.5kg 单臂自由度：6 垂向作业空间：0-2m 传感器：7*相机+1激光雷达
产品特点	丰富工具链，可视化开发界面， 自主仿真平台支持 支持VR遥操控，高效数据采集， 多模态融合，真实环境感知	教育科研、工业物流、家庭康养、 零售药店等领域应用 具有大脑大模型，理解三维场景， 与人自然语言理解，并将长线程任 务分解，自主决策所需操作	楼宇畅行，双手可按电梯，刷闸机 交互友好，AI多模态交互 人机安全，触觉传感器+柔顺力控 算法，确保人机安全	360°全方位传感器 支持同构及遥操作，全身力反馈， 全关节映射，支持本体端力反馈 至遥操端

具身机器人遥操作，是指在相关机器人控制中把人类操作包含在控制回路中，任何的上层规划和认知决定都是由人类用户下达，而机器人本体只是负责相应的实体应用。当机器人处理复杂的感知和大量任务时，在快速做出决策和处理极端情况时，遥操作远远优于智能编程。目前遥操作已广泛应用在医疗领域、极端环境探索如太空与深海场景、反恐防爆应用场景，以及基于工业机械臂的自动化生产中。我们认为遥操作将有助于加速具身智能所欠缺的数据采集环节，此外在实际商业化领域，通过遥操作，也能在恶劣环境下取代人类，同时降低企业用工成本。

按照遥操作主端设备的不同，主流遥操作方式可以分为以下几类：

主端方案	特点
VR设备	VR设备自带的对手柄和手势的识别功能，能够实现机器人在笛卡尔空间中的位姿映射，以及灵巧手的指关节精确映射。这在操作任务上有着比较明显的优势。然而，这种基于视觉+模型的方式容易在视线受遮挡时产生识别或预测错误。
动捕设备	使用动捕设备，可以实现对机器人关节空间运动的映射，在具有冗余自由度机械臂的避障，人型机器人的仿人运动时有着明显的优势。动捕设备有惯性、电磁、以及光学三种方案。光学动捕精度高，但对场地要求严格，需要布置多视角相机，且价格高昂。
机器人	以斯坦福ALOHA为代表，用相似的主动臂直接做关节空间的映射到从动臂，这种操作方式下机器人的运动与人的感受高度一致，劣势在于需要主动臂，设备需要更大的空间，并且需要几乎两倍于机器人成本。
外骨骼	使用外骨骼进行遥操作，一方面可以直接做关节空间位置映射，另一方面可以将机器人的关节受力反馈给人。这种方式下，无论是精度还是操作人员的感受都会比较好，但一套成熟的、有丰富反馈的外骨骼设备价格十分昂贵。并且需要额外考虑外骨骼主动发力对人体潜在的安全风险。
其他末端控制	除此之外，还有许多能够用来做机械臂笛卡尔空间位姿映射的设备，简单的比如手柄、手机、键盘、3D鼠标等等，这些更多是实验室级别的解决方案，无法应对双臂、人型等机器人平台。



2024年5月，特斯拉发布了Optimus人型机器人的演示视频中，展示了二代Optimus学会了分装电池，并且比以前走得更快更远更稳。特斯拉表示，Optimus人型机器人的训练数据都是来自穿戴VR头显的人类训练员。可以看到，完整的系统集成了VR头显、传感器、手套、动捕服和相关软件。通过利用VR头显，Optimus人型机器人可以1:1地复刻映射人类训练员的动作，而软件可以实现第一人称视频的实时传输和精确控制输出，同时保持极低延迟。另外，这种解决方案具有可扩展的远程数据收集的优势，因为来自世界各地的注释人无需前往现场就可以为项目作出贡献。特斯拉官方 Optimus X 账号在 2023 年 5 月曾预告过该职位的性质，根据推测，特斯拉目前人类训练员数量大约为50人以上。

特斯拉机器人人类训练员岗位信息

岗位要求

- 必须每天行走 7 小时以上，同时携带最多 30 磅（13.6 公斤）的重物；
- 身高在 5'7"到 5'11"（约 170 到 180cm）之间，而 Optimus 的身高是 5'8"（约 173cm）；
- 长时间佩戴和操作动捕服和 VR 头显；
- 持续的手 / 眼协调和精细操作、身体协调、动觉意识和上下楼梯；
- 全天站立、坐下、行走、弯腰、弯曲、伸展、蹲下和扭动；
- 灵活安排工作时间：白班 / 夜班和 1 个周末 + 必要时加班。

岗位职责

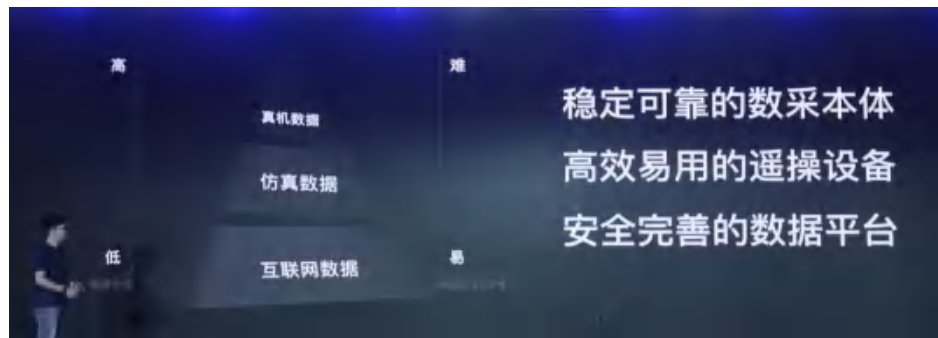
- 每天走预定的测试路线进行数据采集；
- 穿着动捕服和 VR 头显，根据项目要求执行指定的动作和操作；
- 启动 / 停止记录设备并执行小型设备和软件调试；
- 提供有关设备性能的反馈；
- 分析和报告轮班期间采集的数据；
- 上传采集的数据并撰写每日报告，详细说明观察结果和问题；
- 确保分配的设备处于正确和安全的工作状态并安全运输到各个采集地点。



2025年初，智元机器人在上海建了一座“数据采集工厂”。在这座3000平方米的数采厂中，不同主题的房间占据了工厂的绝大部分面积，每个房间都还原了现实生活的物件布局，机器人就在不同的场景中执行不同任务。为了让机器人快速学习技能，智元为它们安排了一对一的数据采集员，数据采集员也需要肢体协调，动作标准。数据采集员们手持设备，手把手地控制机器人完成抓、握、放等动作。有时他们也会头戴VR设备，更精准地让机器人模仿学习人类动作。现在智元数据采集工厂投放了近百台机器人，日均采集3-5万条数据。目前，智元数据采集工厂模拟了家庭、零售、服务业、餐饮、工厂五个场景。在基础能力建设方面，智元退出了智元具身智能数据系统AIDEA，其涵盖了广泛的数据采本体，遥操设备以及“数据采集-数据标注-数据管理-模型训练-模型评测-模型部署-数据回传”等全链路的数据平台。



智元具身智能数据系统 AIDEA



智元数采工厂

在具身智能技术发展的关键阶段，数据要素已成为驱动该领域突破的核心性基础。Meta首席技术官安德鲁·博斯沃思近期指出，当前数据资源尚无法有效支撑具身智能系统实现泛化性精细动作控制，特别是人类与生俱来的类人无意识动作基元智能（如器具抓握、设备操作等），其实现难度可能远超高层认知智能的开发挑战。

在此背景下，遥操作技术及运动捕捉系统构成了现阶段技术研发的关键支撑体系。基于技术演进路径分析，若L5级具身智能体（具备完全自主环境适应能力）可在未来十年实现突破，遥操作与运动捕捉技术将发挥双重战略价值：**短期维度**上，该技术可加速具身智能系统的商业化进程，通过虚实融合的混合增强模式实现关键场景落地；**长期维度**上，其将演进为具身智能生态的基础设施级技术，持续赋能物理交互能力的进化迭代。基于此，本研究提出以下两类遥操作技术商业化路径的可行性分析。

遥操及数采作商业化路径一：众包数采平台

数据平台通过众包形式以低廉价格获取海量数据，并服务各家具身本体企业

众包数采平台

数采任务认领下发

数采完成数据上传

众包训练员

依托地域发展差异及就业时间差异，采用兼职众包方式让训练员在生产生活中进行数据采集

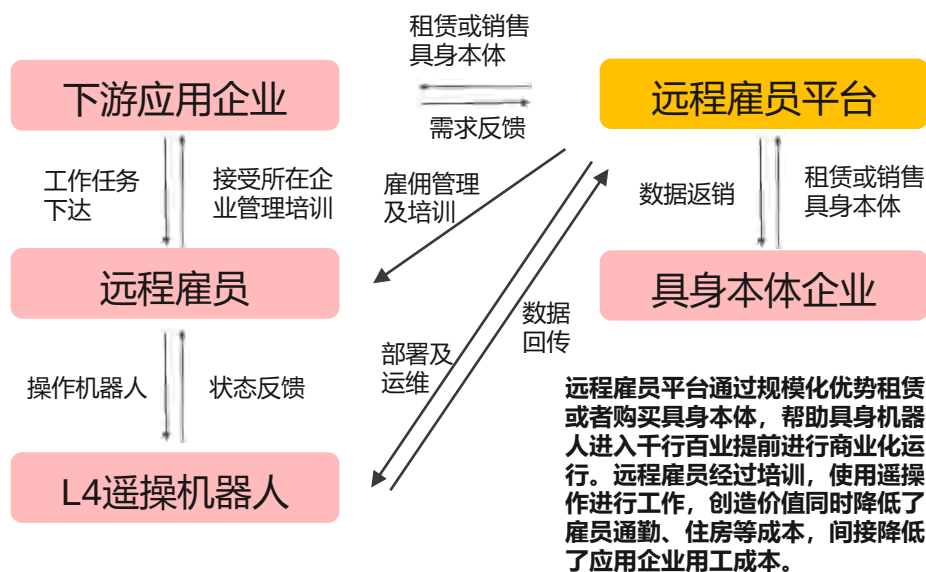
数采需求受理

数采需求受理

具身本体企业

具身智能企业通过低廉价格获取优质实战环境真实数据，并以此优化模型算法

遥操及数采作商业化路径二：远程雇员平台



06

风险提示