

具身智能：从数字空间走向物理世界

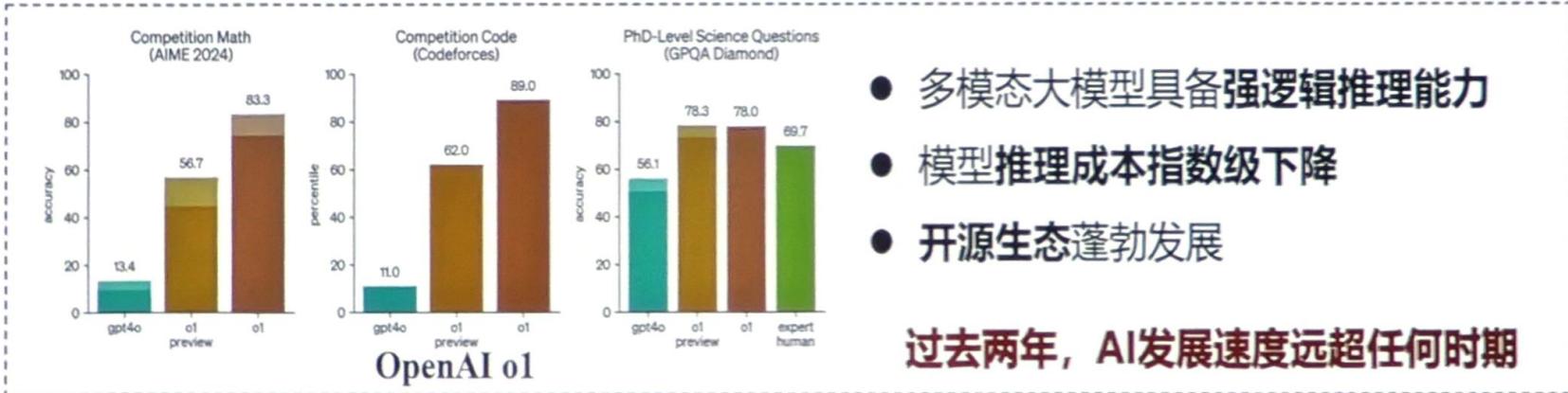
林惊

鹏城实验室 具身智能研究所

中山大学 人机物智能融合实验室

2024年9月

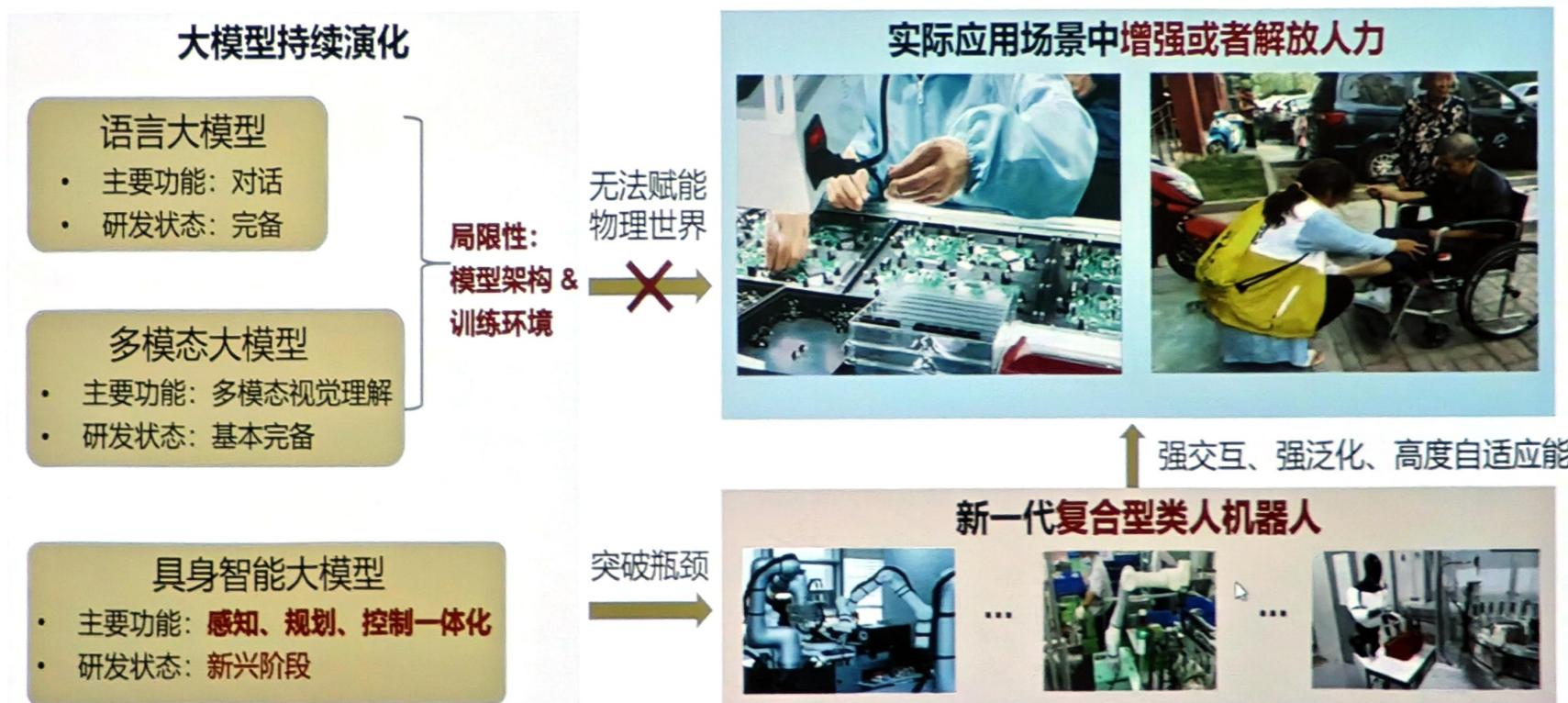
一、研究背景：从数字空间到物理世界



AGI的最终形态不在于某个超级模型，而是对齐数字空间和物理世界

一、研究背景：具身智能

■ 具身智能：突破传统机器人大规模应用瓶颈的关键技术，实现通用人工智能的必经之路



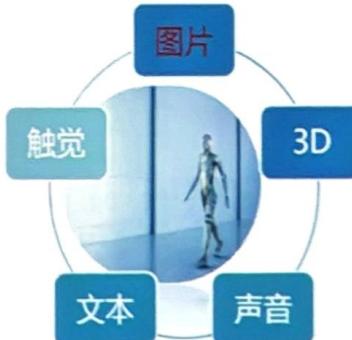
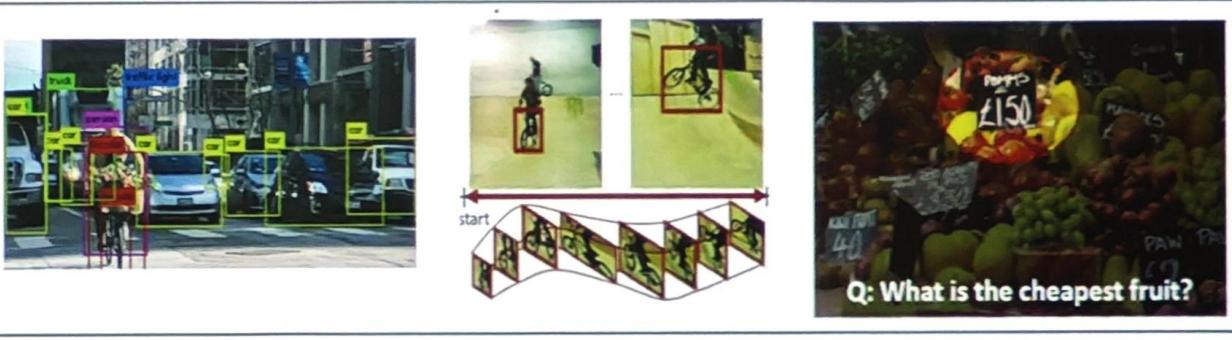
一、研究背景：具身智能

离身智能: 模型**被动感知数字空间**, 无法直接改变环境并作用于物理世界。

促进



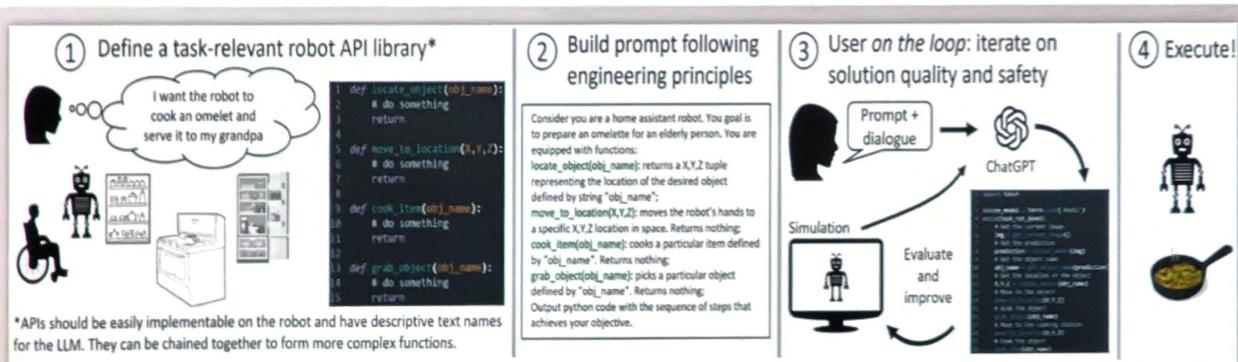
具身智能: 智能体**主动理解物理世界**, 通过适应性行为和自主学习来完成任务。



工业、家居、服务机器人

一、研究背景：具身智能

- **具身智能：**人工智能领域前沿研究方向,受到国际顶级院校和研究机构的广泛关注



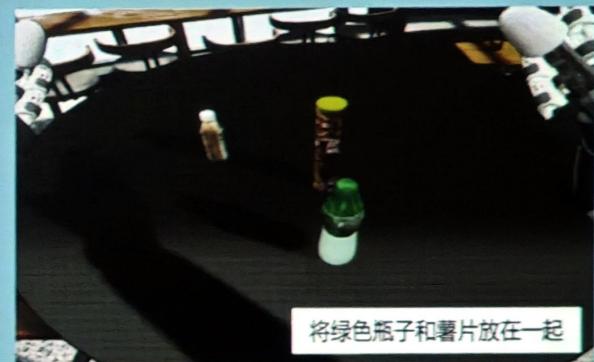
2023年2月，微软发布ChatGPT应用于机器人的研究成果，用自然语言指令控制机器人



2023年7月，谷歌发布机器人大模型RG-2，
机器人执行“捡起灭绝的动物”指令



2023年7月，李飞飞团队发布VoxPoser，实现零样本的日常操作任务轨迹合成，图为机器人执行指令“打开上面的抽屉，小心花瓶！”



2024年腾城实验室发布国内首个全物理引擎的导航操控一体化仿真平台和发布的操纵大模型在多级别操纵任务上的准确率比字节的GR-1高23.1%，比谷歌的RT-1高19.5%

一、研究背景：具身智能的挑战

空间推理



长程任务规划

请帮我把卫生间洗手台上的布洗干净，并帮我把厨房冰箱上的面包放到微波炉。



Human

机器臂现在能把右侧水杯抓起来吗?

GPT-4o

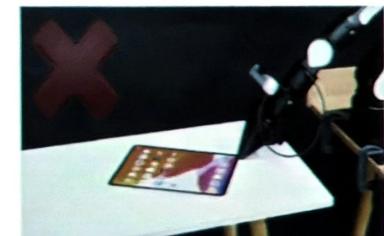
可以，因为机器臂的手已经**抓住了盘子**...



- 错误预测深度信息
- 错误关联空间关系

空间推理能力不足

复杂操纵



思维链推理

思考1:机器人到达卫生间
步骤1:找到卫生间
思考2:已经到达卫生间，
需要接近洗手台 ...

GPT-4o推理

1. 找到卫生间
2. 拿起布 (不在洗手池附近)
3. 打开水龙头
4. ...



- 无法正确推理物理位置关系

长程任务规划可解释性低

- 工业产品种类不同
- 机器人种类不同抓取失败
- 环境干扰不同抓取失败

复杂操作泛化能力弱

三、关键技术

案例：帮助老年人收拾厨房

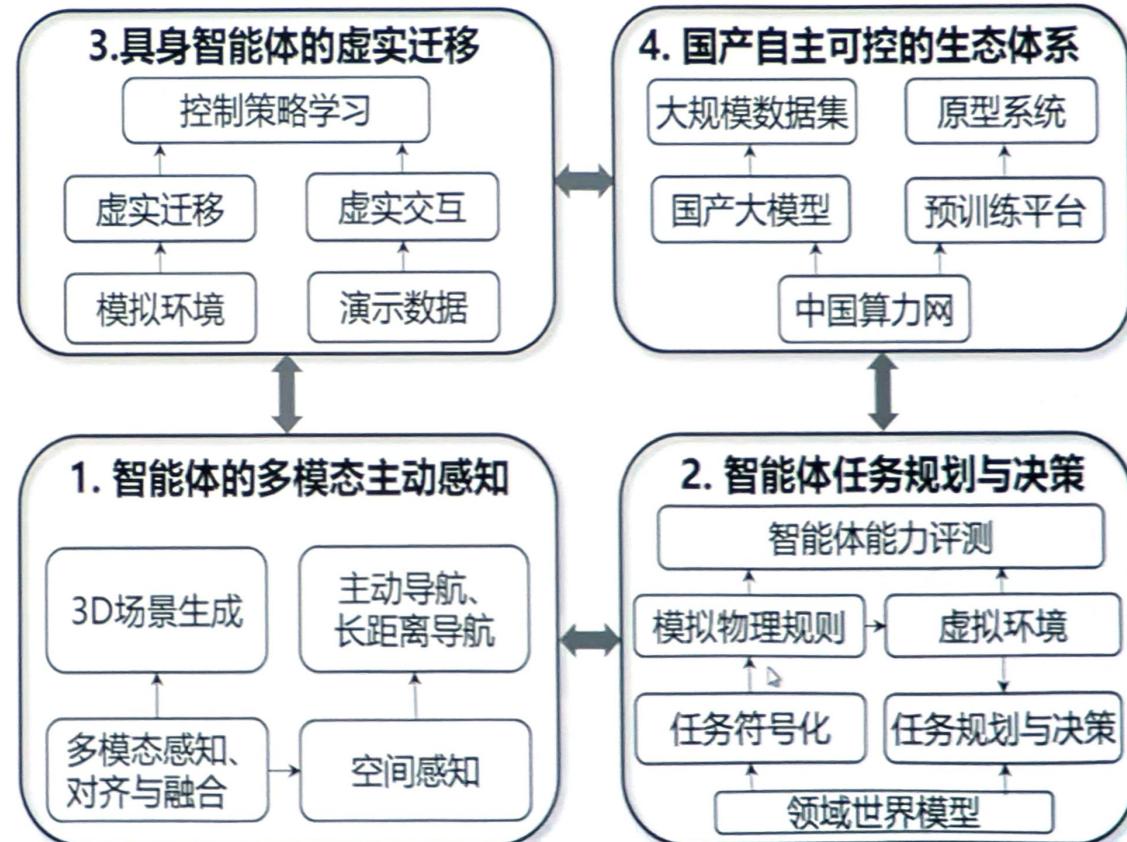


问题：

1. 厨房里面有什么物体？
2. 需要收拾哪些物体？收拾的步骤是什么？
3. 怎样操纵双臂去收拾这些物体？

挑战：

1. 具身智能体需要对开放式环境进行高效感知
2. 具身智能体需要对任务进行全面理解和规划
3. 具身智能体需要进行精准动作控制操作物体



提纲

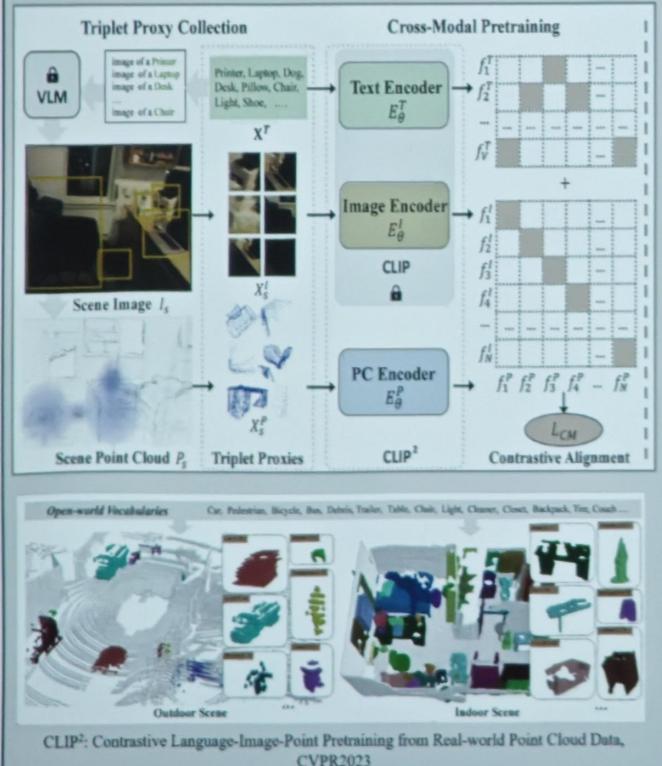
- 一、开放式多模态环境主动感知
- 二、知识引导的具身任务规划与决策
- 三、自适应具身智能体的虚实迁移
- 四、国产自主可控的具身智能生态体系

Part I: 开放式多模态环境主动感知

1. 开放式词汇引导的零样本视觉理解

2. 具身智能体主动环境探索感知

视觉-语言大模型驱动的零样本识别



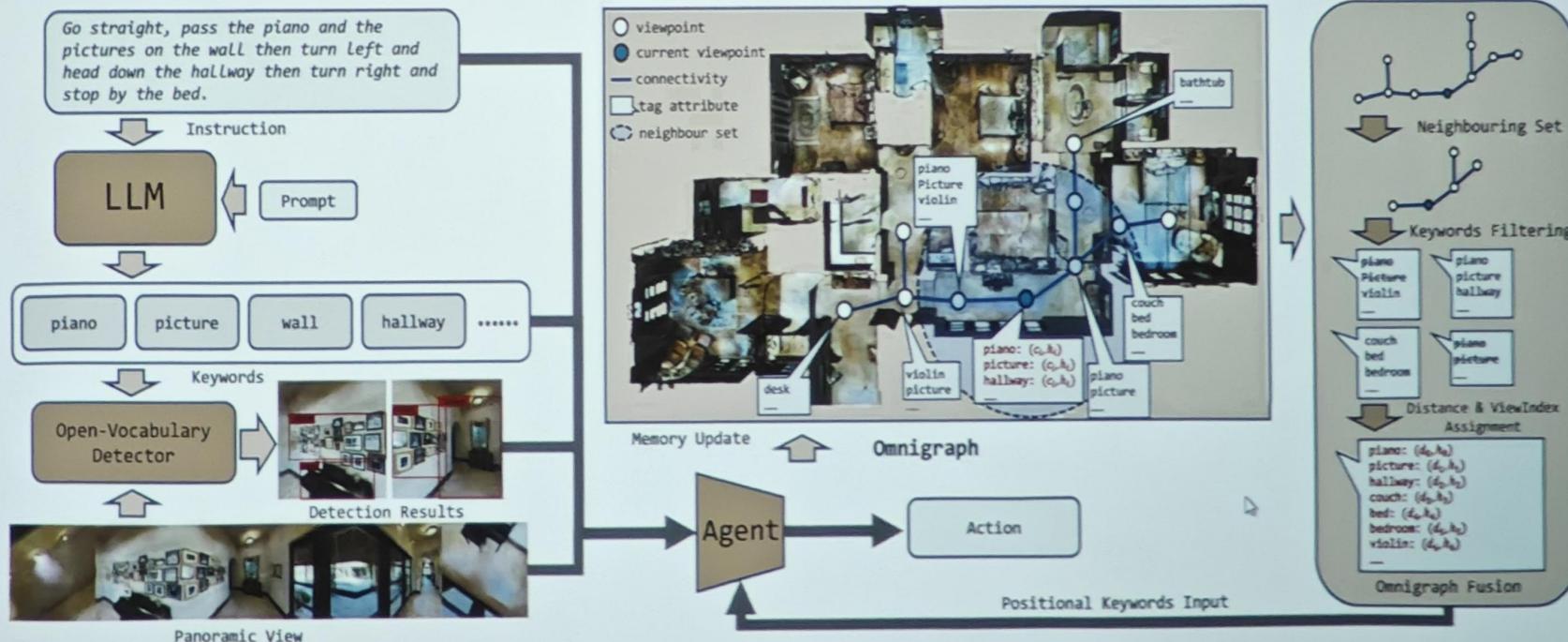
第三人称 视角

第一人称 视角



Part I: 开放式多模态环境主动感知

3. 复杂场景的多模态结构化表征学习，融合多模态信息与全局语义地图构建动态记忆模块，实现对动态未知环境的精准感知与可靠任务执行。



提纲

一、开放式多模态环境主动感知

二、知识引导的具身任务规划与决策

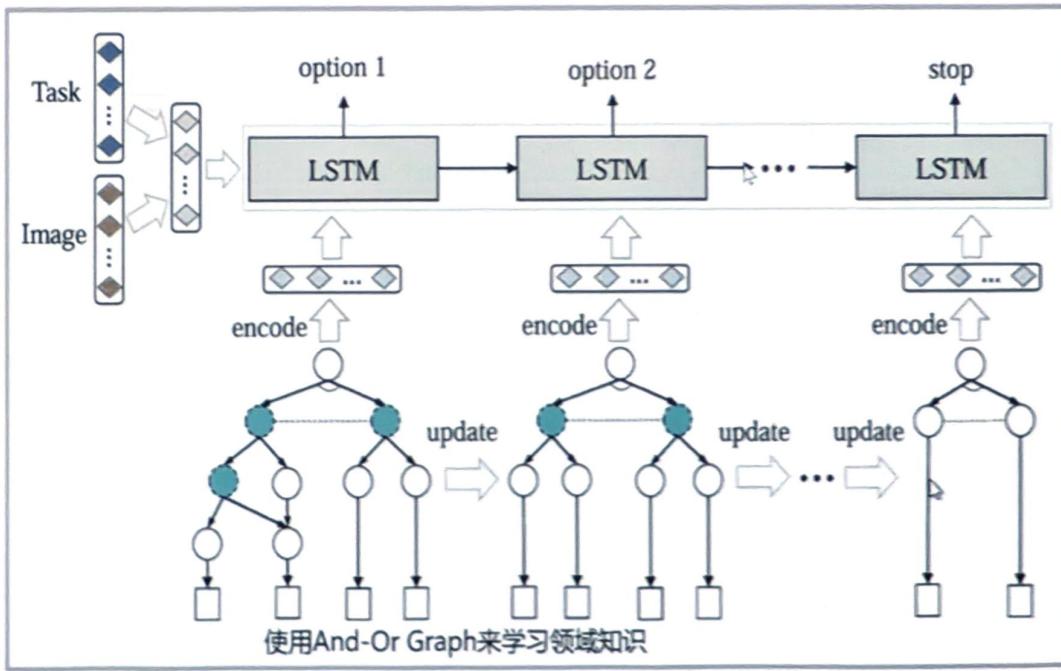
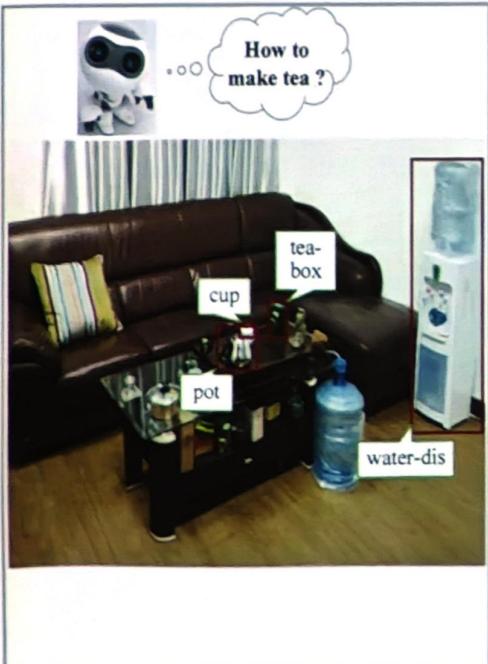
三、自适应具身智能体的虚实迁移

四、国产自主可控的具身智能生态体系

Part II: 知识引导的具身任务规划与决策

1. 领域知识驱动的具身任务规划，引入丰富领域知识进行精准任务理解，融合具体场景视觉语义信息，自适应生成任务关联原子动作序列，进而高效完成复杂任务。

- 问题定义：给定具体场景S和任务T，生产在该场景下完成该任务的原子动作序列



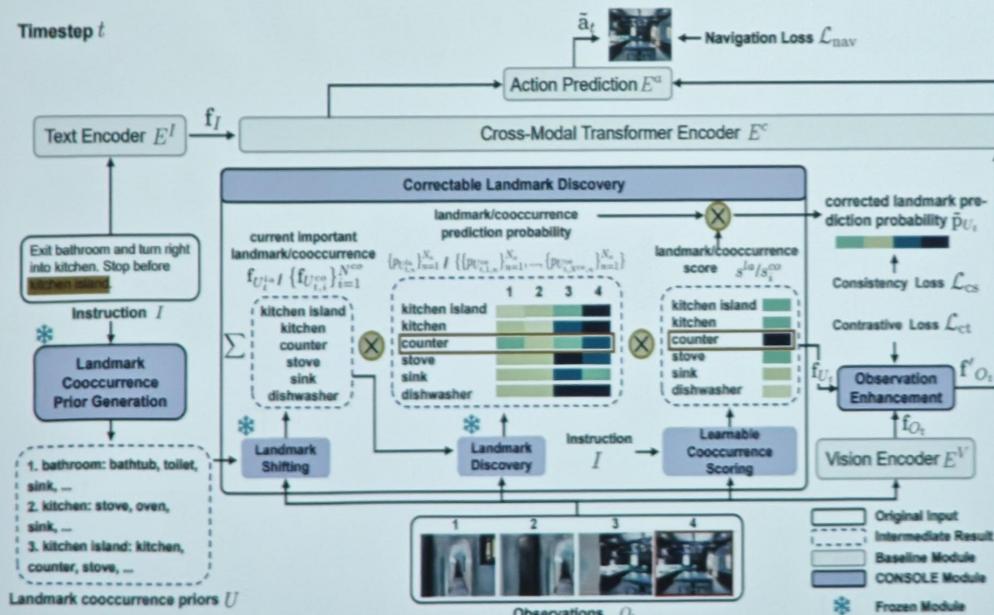
A1: {move to, tea-box}
A2: {grasp, tea-box}
A3: {open, tea-box}
A4: {grasp, tea-box}
A5: {put into, cup}
A6: {move to, water-dis}
A7: {pour into, cup}

A1: {move to, tea-box}
A2: {grasp, tea-box}
A3: {open, tea-box}
A4: {grasp, tea-box}
A5: {put into, cup}
A6: {grasp, pot}
A7: {pour into, cup}

The world's first 10K Best Paper Diamond Award

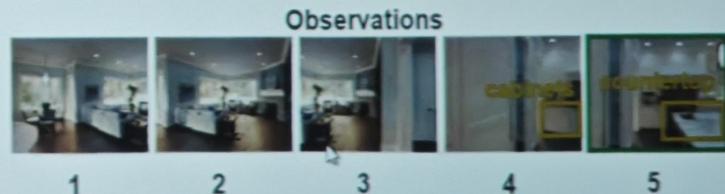
Part II: 知识引导的具身任务规划与决策

2. 大模型驱动的复杂任务规划，利用大语言模型ChatGPT和视觉模型CLIP解决视觉语言导航中的地标发现问题，通过动态调整地标重要性实现精准导航决策。



在R2R数据集上验证，超过法国国家信息与自动化研究所团队 DUET 3%

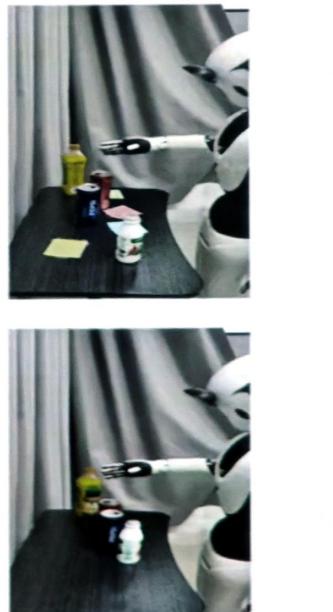
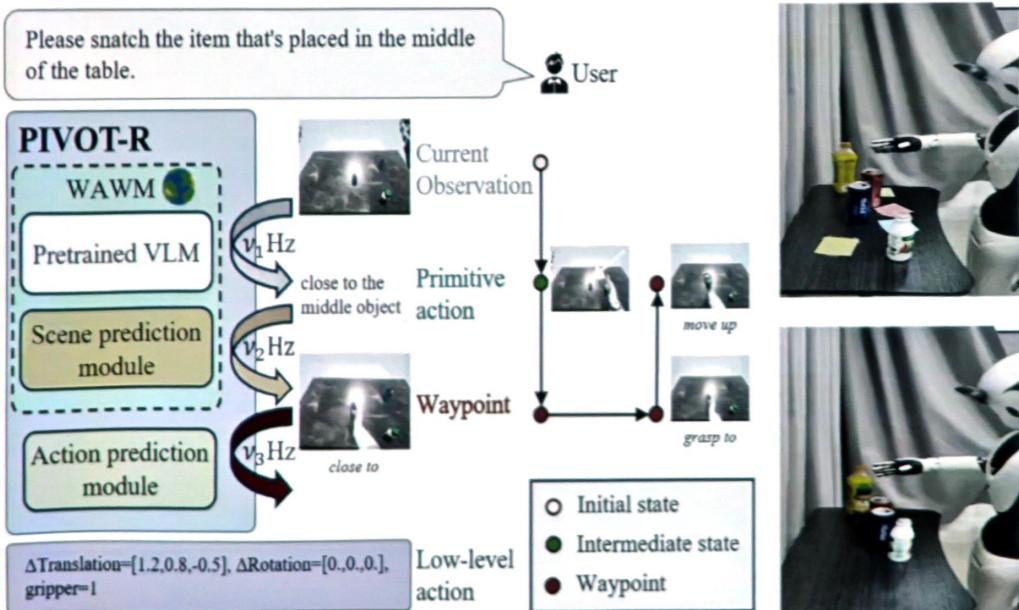
Turn around and walk across the **kitchen**. When you reach the stove, turn left and exit the room into the dining room. Walk past the table and stop at the other end.



Correctable Landmark Discovery Via Large Models for Vision-Language Navigation, TPAMI2024

Part II: 知识引导的具身任务规划与决策

3. 融合世界模型的具身操纵，提出路径点感知世界模型，提升模型的空间推理能力，实现多模态大模型、世界模型与轻量级动作决策模型异步分层执行。



在真机上验证，并取得最优的效果，超过谷歌RT-1 13.0%

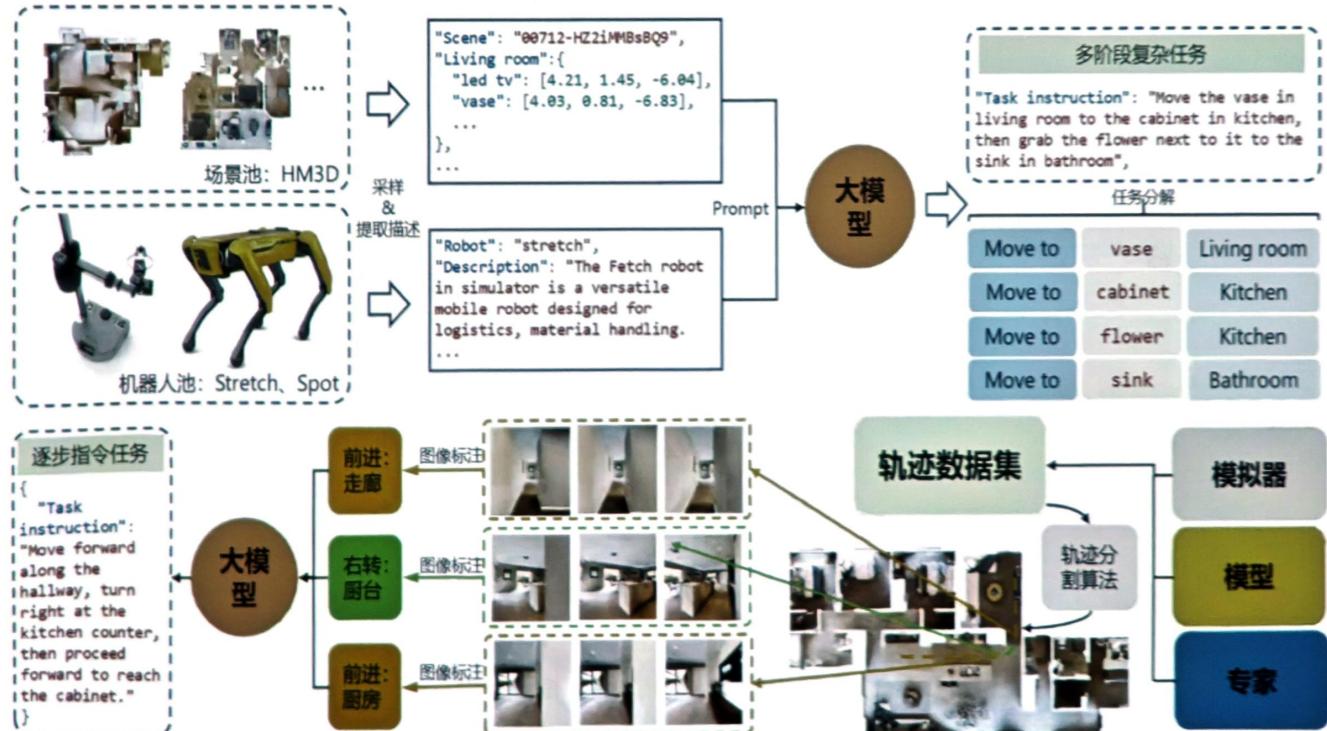
Model	Pick up	Put on	Push to	Mean
Octo	34.72	27.78	4.17	22.22
RT-1	40.28	22.22	19.44	27.31
GR-1	26.39	29.17	8.33	21.30
Surfer	41.67	29.17	31.94	34.26
PIVOT-R	54.17	41.67	25.00	40.28



PIVOT-R: Primitive-Driven Waypoint-Aware World Model for Robotic Manipulation, 2024

Part II: 知识引导的具身任务规划与决策

4. 长程复杂任务导航数据集NavGen，基于两类机器人视角采集数据，包含216个复杂场景，2164个长程推理复杂任务，为验证具有强泛化性以及长程规划能力的导航模型提供了平台。



场景和任务数超过
斯坦福Behavior-1k

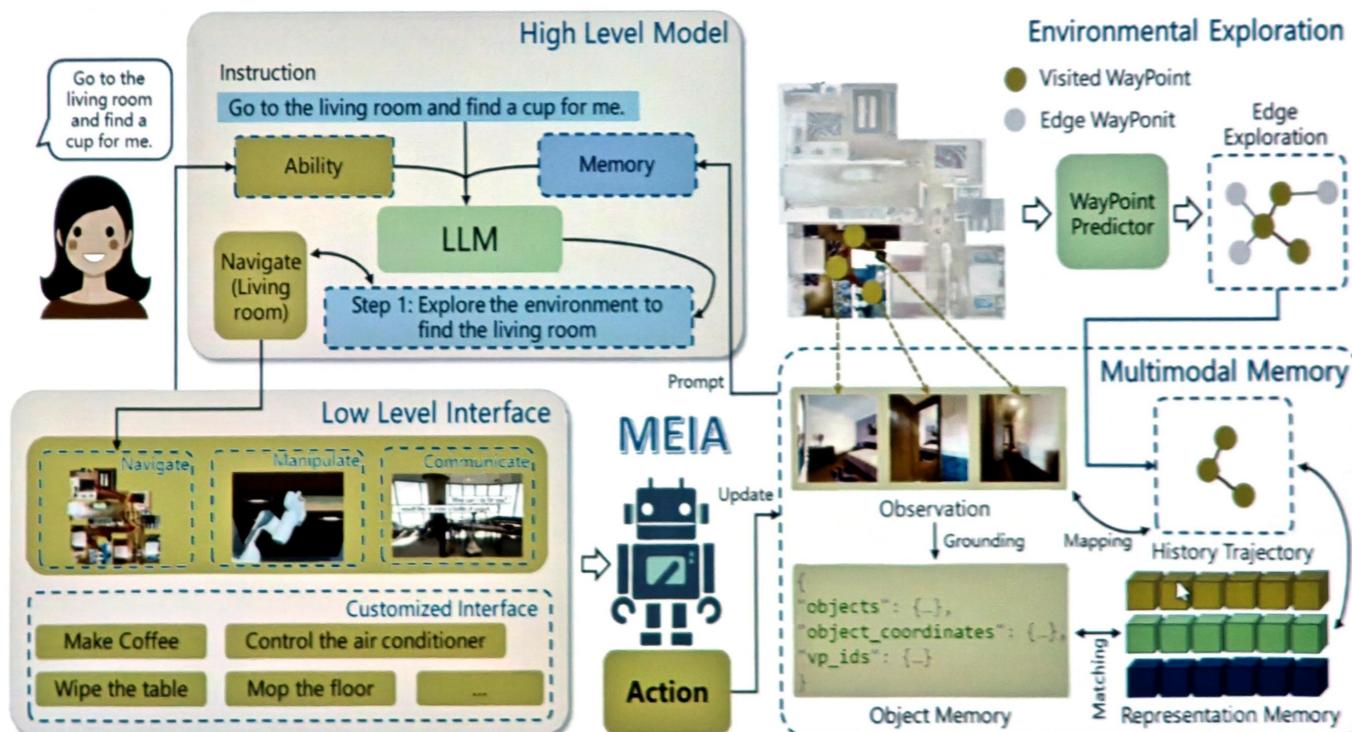
数据集	场景	任务
R2R	90	21567
ALFRED	120	25743
VLN-CE	90	4475
Behavior-1k	50	1000
NavGen	216	2164

提纲

- 一、开放式多模态环境主动感知
- 二、知识引导的具身任务规划与决策
- 三、自适应具身智能体的虚实迁移
- 四、国产自主可控的具身智能生态体系

Part III: 自适应具身智能体的虚实迁移

1. 通用具身智能体，统一了高层任务规划模块、下层控制接口以及多模态记忆模块，能够实现导航、问答、操纵与自定义任务。



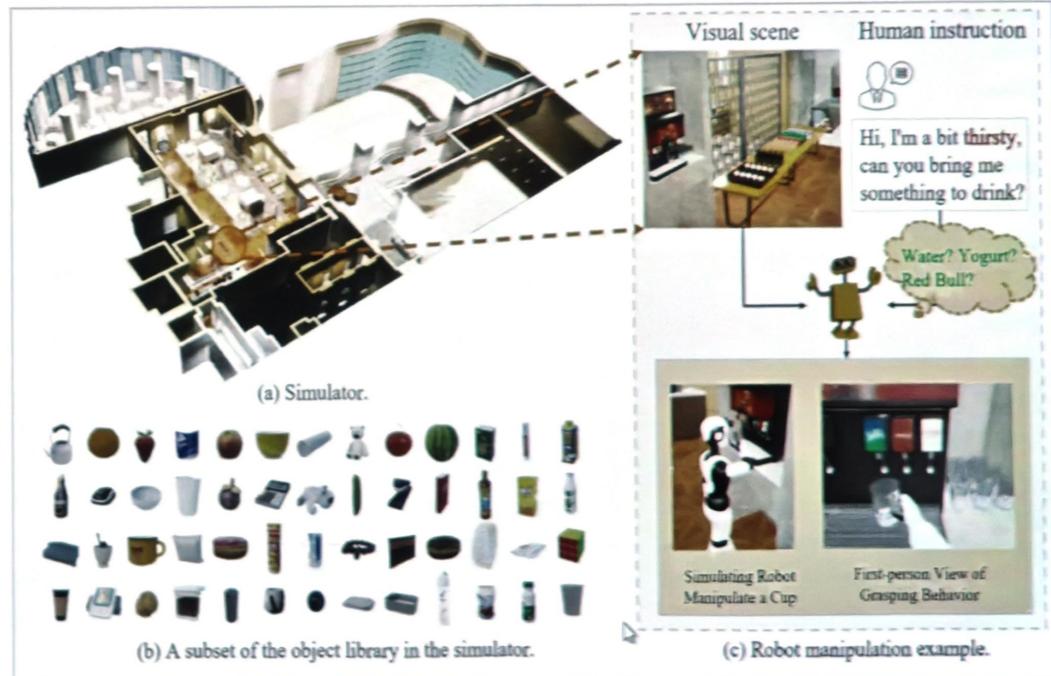
MEIA: Multimodal Embodied Perception and Interaction in Unknown Environments, arxiv 2024

Part III: 自适应具身智能体的虚实迁移

2. 细粒度可编辑可交互的平行数字空间，支撑具身智能体的高效训练。



↑
三维环境重建



↑
细粒度环境编辑

↑
具身交互

Part III: 自适应具身智能体的虚实迁移

3. 具身仿真平台 InfiniteWorld Simulator, 支持具身智能体的高质量示教数据采集与技能演练。



室外场景

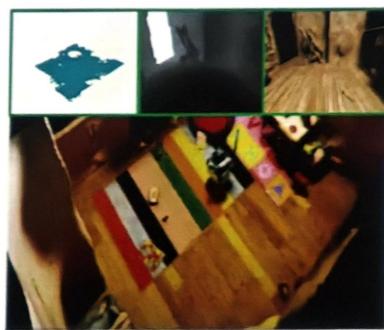


室内场景

即开即用的大规模室内外场景，包含医院、办公室、仓库、工厂、城市等海量场景。



扫描重建



领先的3D扫描与重建技术，助力现实场景的便捷虚拟化。



文本场景生成

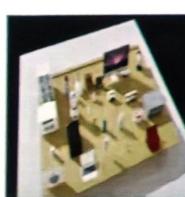


场景编辑

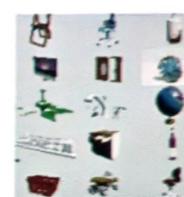


风格变换

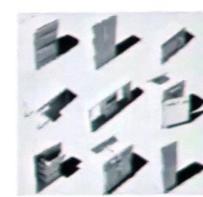
- 文本场景生成+物体编辑/替换，实现场景的无限扩增
- 200+纹理/背景/材质自由替换，场景量×200+



刚性物体



铰接物体



可控交互式
物体生成



机器人家族



导航与抓取



机器人家族

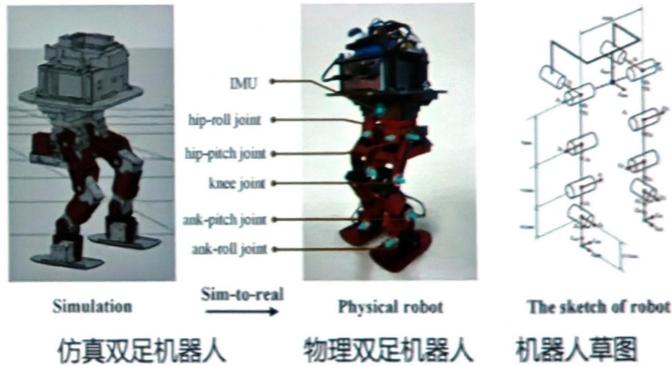
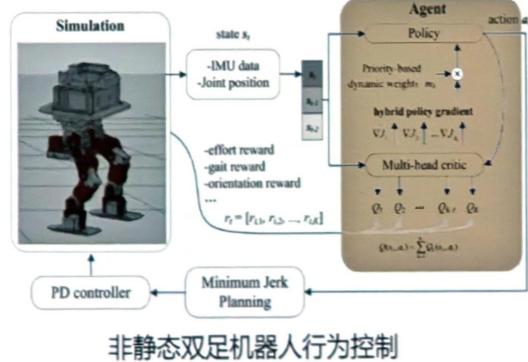


导航与抓取

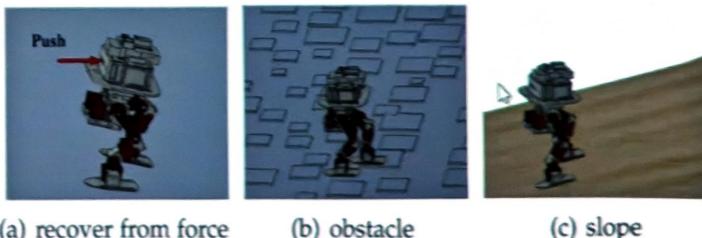
- 10W+高质量物体资产
- 无限的可控式交互物体生成与文本物体生成
- 支持90+种多样的机器人
- 丰富的机器人交互方式

Part III: 自适应具身智能体的虚实迁移

4. 机器人行为控制策略高效学习，提出一种面向非静态双足机器人运动控制的新型奖励自适应强化学习方法，设计了一种动态混合测量梯度来优化控制测量学习。在Gazebo模拟器中构建了双足机器人，并成功地转移到了物理机器人上。



HDPG 策略被成功地转移到了物理机器人上



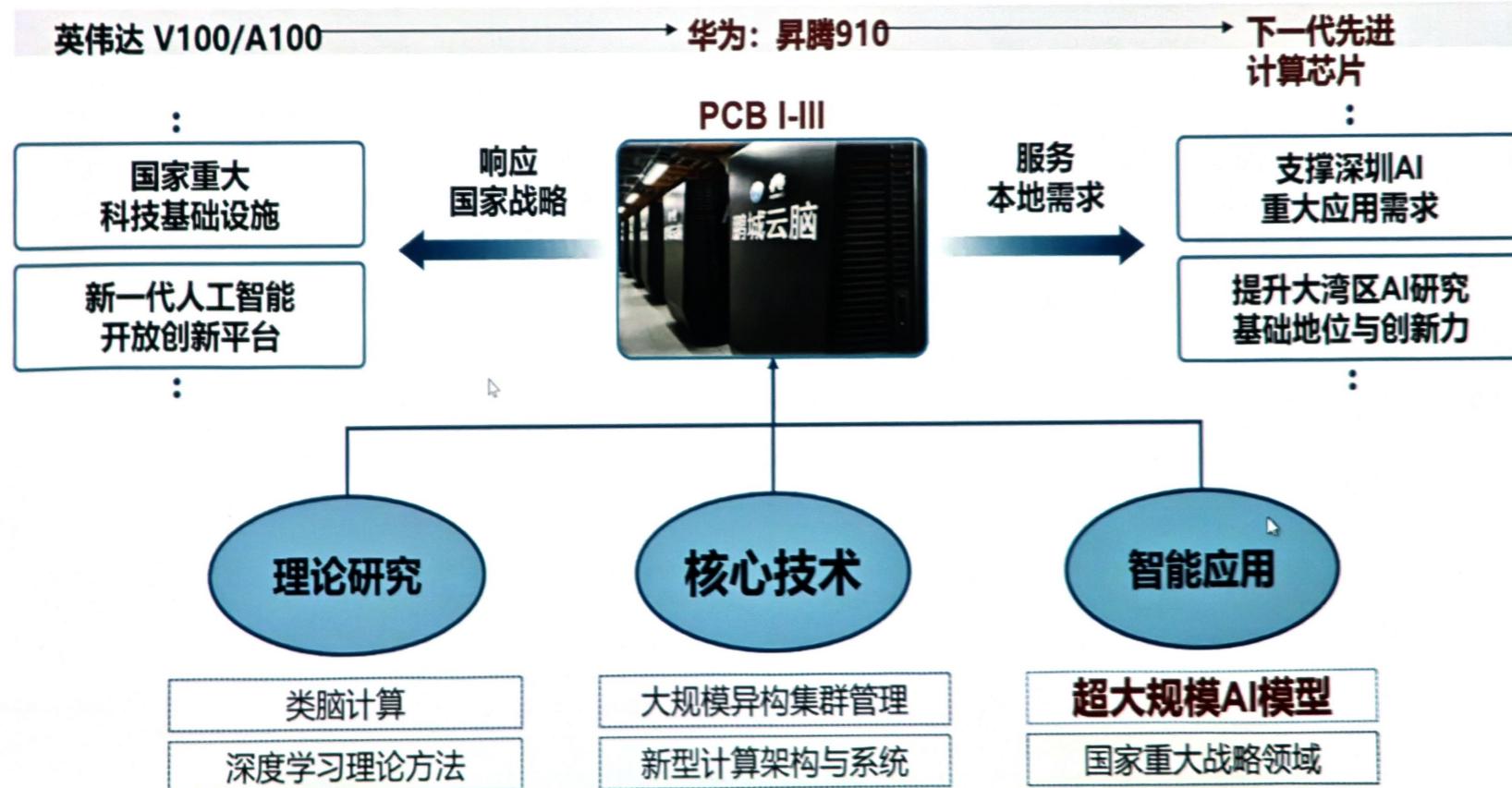
双足机器人运动任务的示意图

提纲

- 一、开放式多模态环境主动感知
- 二、知识引导的具身任务规划与决策
- 三、自适应具身智能体的虚实迁移
- 四、国产自主可控的具身智能生态体系

Part IV: 国产自主可控的具身智能生态体系

➤ 鹏城云脑：超大规模中国算力网，为国产化大模型训练提供充足算力支撑



Part IV: 国产自主可控的具身智能生态体系

■ 跨机器人跨场景的大规模预训练平台，构建与真实场景同步的机器人模拟技能库，支持多种机器人形态和原子操作，生成海量演示数据，实现跨机器人跨场景的大规模预训练。



细粒度可编辑可交互的具身平行数字空间

中国算力网

■■■ 国产自主可控的具身智能应用支撑平台

已联合搭建多套具身智能硬件平台，共同构建具身智能创新应用生态。

1

移动操作机器人



2

双臂机器人



3

安防巡检、自主操控



全球首个具身智能数据标准，最大规模数据集

优势

300多万条轨迹数据，340多个场景，30多种机器人

ARIO: All Robots In One
https://imaei.github.io/project_pages/ario

任务种类及复杂度方面超越Open X-Embodiment

统一采集标准

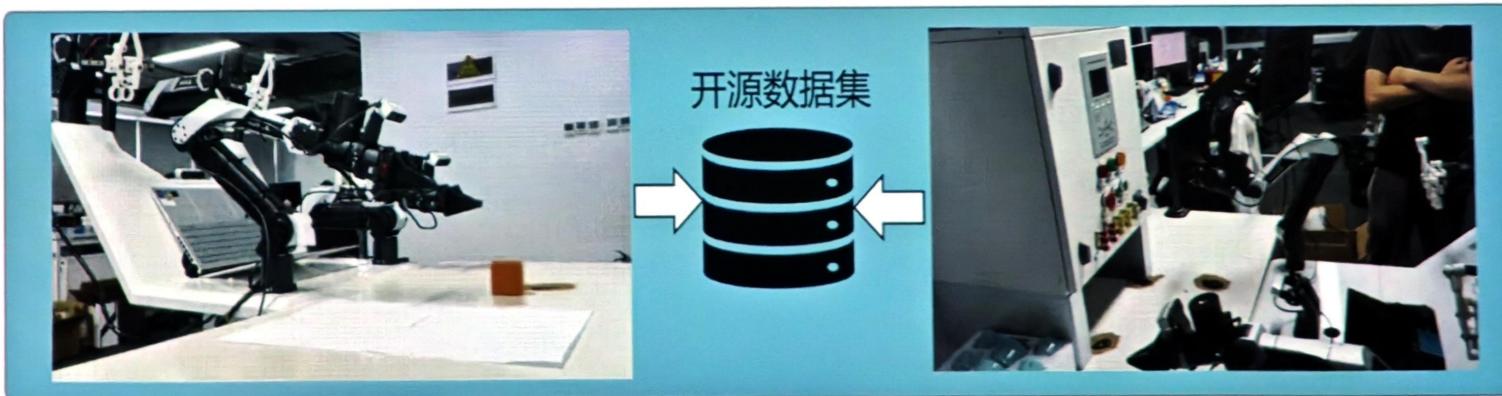
支持5种模态

大规模场景

跨机器人平台

海量复杂任务

数据



联盟



达闼科技

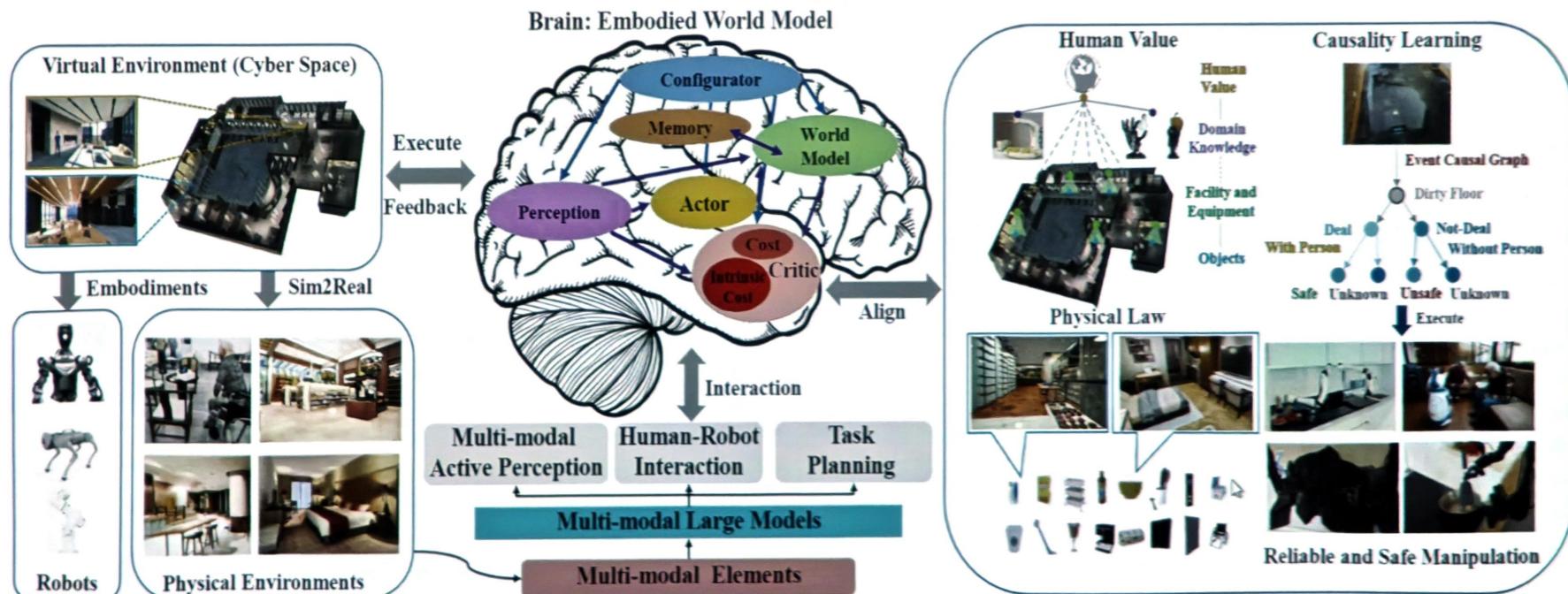


智元机器人 ...

多模态大模型时代首篇具身智能综述

Aligning Cyber Space with Physical World: A Comprehensive Survey on Embodied AI

- 调研400篇文献，全面总结具身机器人、模拟环境、感知、交互、智能体、Sim2Real核心技术



具身智能资源仓库: [https://github.com/HCPLab-SYSU/Embodied AI Paper List. \(500 Star+\)](https://github.com/HCPLab-SYSU/Embodied AI Paper List. (500 Star+))

《多模态大模型——新一代人工智能技术范式》

--刘阳、林惊 著



本书架构

本书以深入浅出的方式全面地介绍了多模态大模型的核心技术与典型应用，并围绕新一代人工智能技术范式，详细阐述了因果推理、世界模型、超级智能体与具身智能等前沿技术。

大模型基础理论与结构	第1章 大模型全家桶		
多模态大模型核心技术	第2章 多模态大模型核心技术		
最具代表性的多模态基础模型	第3章 多模态基础模型		
多模态大模型典型应用	第4章 多模态大模型的应用		
面向AGI的前沿技术	第5章 多模态大模型迈向AGI		
大模型典型结构	多模态大模型核心技术	典型多模态基础模型	面向AGI的前沿技术
BERT ViT GPT系列 ChatGPT ChatGLM 百川大模型	提示语 提示学习 上下文学习 微调 思维链 RLHF	CLIP BLIP LLaMA VideoChat SAM PaLM-E	因果推理 世界模型 超级智能体 具身智能
典型应用	视觉问答、AIGC、具身智能		

新一代人工智能技术范式 面向、可解释、可信的多模态大模型



鹏城实验室

多智能体与具身智能研究所

共创、共享、共建！