

具身智能中的 VLA 技术及其应用

隋 伟

地瓜机器人技术副总裁





源自智能驾驶，通往通用智能



赋能智能驾驶

让生活更安全、更美好

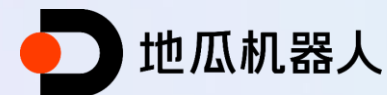
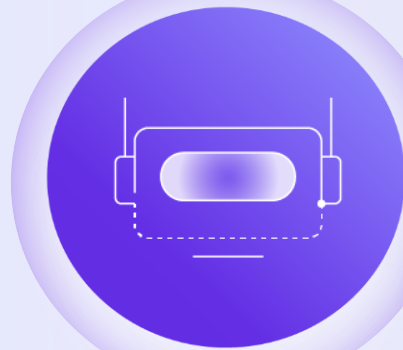


计算架构

开发平台

智能算法

产业生态



加速机器智能进化
促进人机和谐伴生

共通的三大计算任务

环境感知

人机交互

决策控制



为机器人创造更简单的全链路开发基础设施



目录

01 具身智能中VLA的现状和挑战

02 VLA的主流架构

03 VLA的数据方案

04 VLA模型的量化部署

05 前景和展望

06

📍 北京

QCon

全球软件开发大会

会议时间：4月10-12日

- 大模型赋能 AIOps
- 越挫越勇的大前端
- 云上业务架构演进
- 多模态大模型及应用
- 海外 AI 应用创新实践

📍 北京

AiCon

全球人工智能开发与应用大会

会议时间：6月27-28日

- 端侧智能
- AI Agent
- 多模态大模型
- 金融+大模型

📍 上海

QCon

全球软件开发大会

会议时间：10月23-25日

- AI Agent
- 大模型训练推理
- 端侧 AI
- 搜推深度融合
- 数智金融

4月

5月

6月

8月

10月

12月

📍 上海

AiCon

全球人工智能开发与应用大会

会议时间：5月23-24日

- 通用大模型
- AI Agent
- 垂直领域应用
- 模型可解释性

📍 深圳

AiCon

全球人工智能开发与应用大会

会议时间：8月22-23日

- 模型效率与部署
- 多模态大模型
- 大模型安全
- 智能硬件

📍 北京

AiCon

全球人工智能开发与应用大会

会议时间：12月19-20日

- 通用大模型
- 智能硬件
- LMOPs
- 具身智能

极客邦科技 2025 年会议规划

促进软件开发及相关领域知识与创新的传播



参会咨询



查看会议

01

VLA的现状和挑战

具身智能：堪比“计算机诞生”级的颠覆式创新

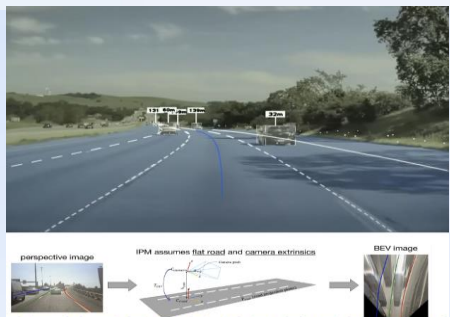


具身智能：堪比“计算机诞生”级的颠覆式创新



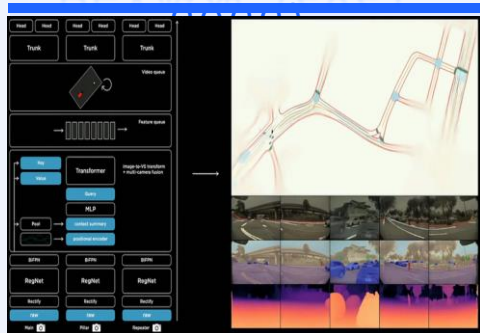
具身智能中的技术演变

模块化 <2021



- 2D感知结果通过规则化后处理转换到3D空间

BEV感知 (2021-)



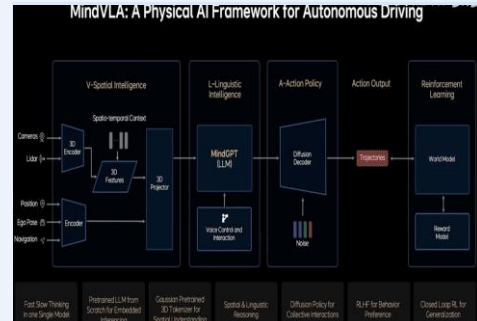
- 感知结果直接输出到planning的空间，减少后处理
- 为端到端奠定基础

端到端 (2023-2024)



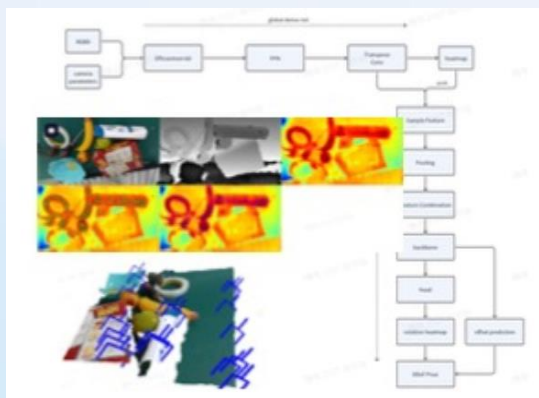
- 更多的learning based 更少的rule based
- 减少了模块间的信息损失
- 拟人化的效果，scaling law 得到验证

VLA (2025-)

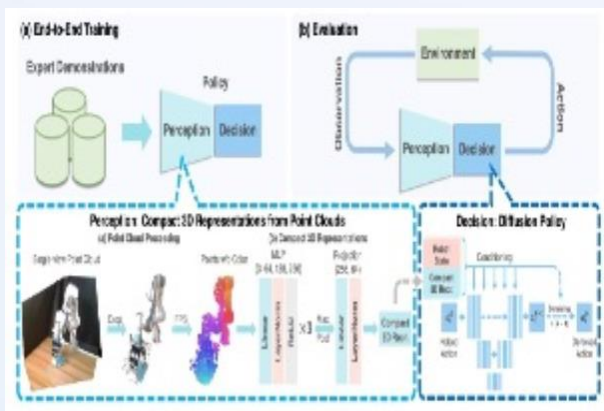


- 利用预训练模型的通用理解能力，解决cornercase 问题
- 智能驾驶开始具备思考能力

Detect and Grasp



Imitation Learning



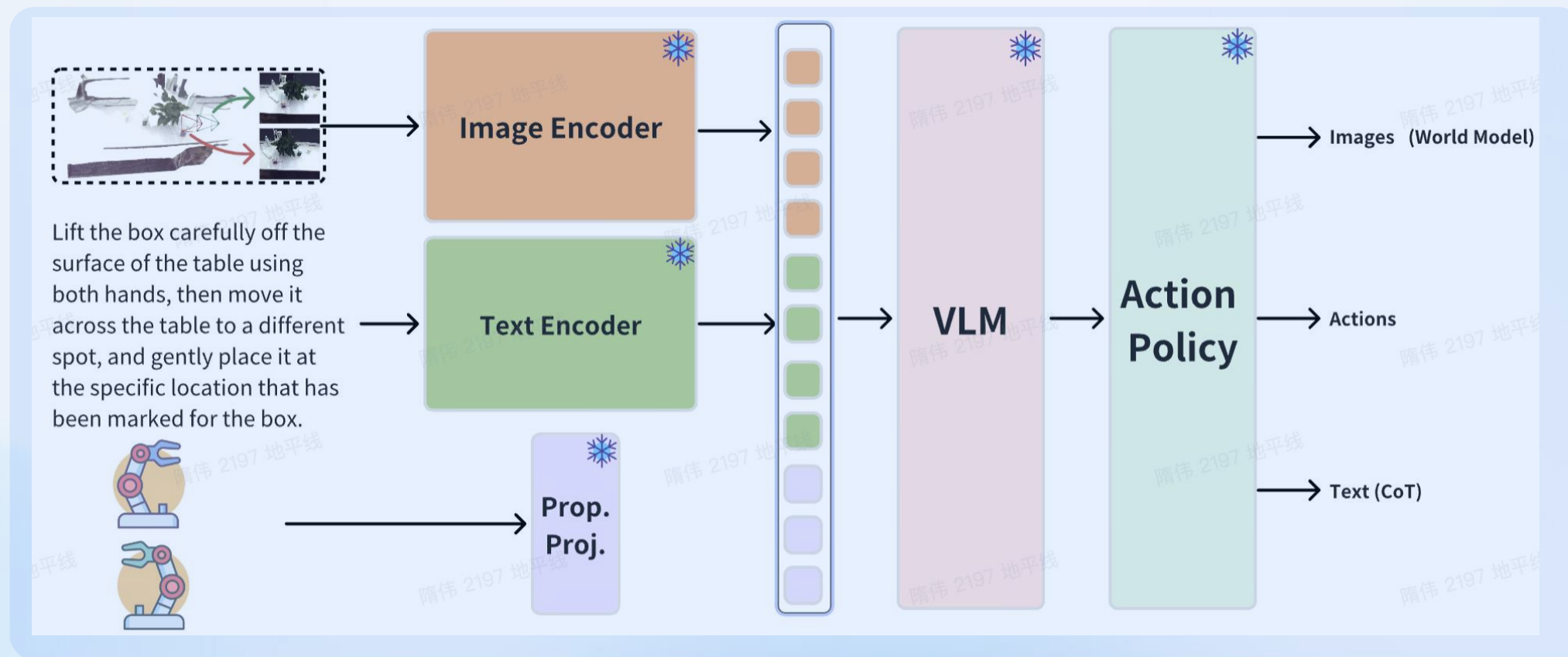
VLA



- 场景泛化性
- 任务泛化性
- 本体泛化性

VLA的模型结构

- VLM在LLM的基础之上增加视觉输入，在互联网上海量的数据训练，具备了通用“常识”能力
- VLA在VLM的基础之上，增加了Action Policy模块，将VLM的特征映射到Action，输出机器人的关节角度或者轨迹
- 具身领域代表性的工作有OpenVLA、Pi-0、Pi-0.5、GrootN1等



VLA的各种尝试

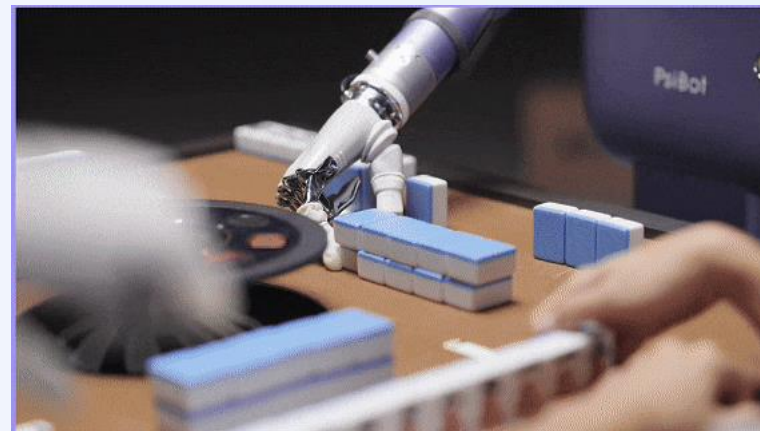
叠衣服



倒水



打麻将



微波炉热菜



收纳



做香囊



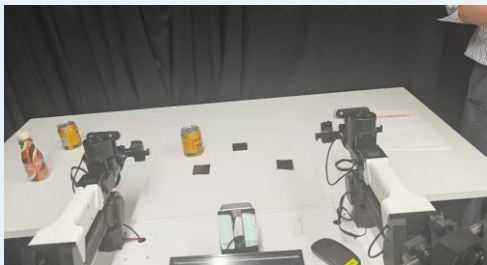
机械臂演示的高效分拣与精准操作能力正是破局关键

VLA操作模型的性能现状

1 泛化能力和通用能力非常有限

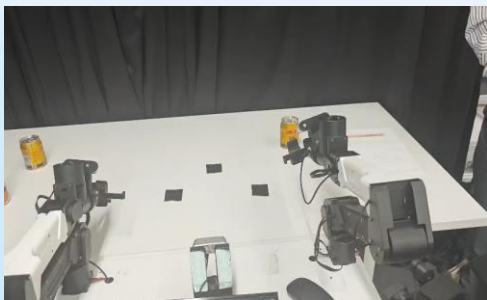
- VL和A的数据分布存在显著差异，L起不到作用，反而导致模型难以学习
- 硬件和模型的限制，VLA很难完成精细的任务

正常数据



Success

饮料放到了远处



Failed

背景发生变化



Failed

其它饮料瓶干扰



Failed

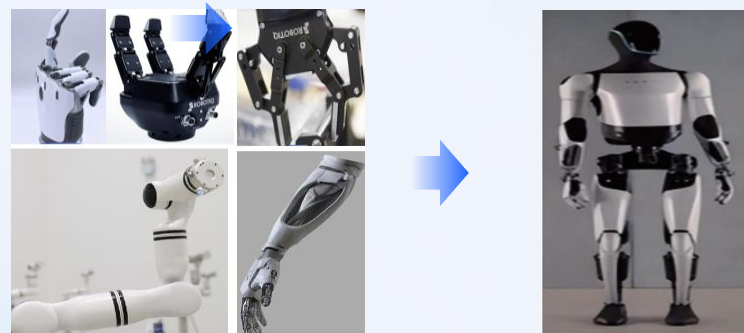
VLA的性能还处在初级阶段

2 当前的数据规模不足以发挥VLA的性能

- VLA需要海量的高质量、多样性数据，目前的条件不具备
- VLA算力要求高，相比VA更适合作为落地方案

	安全等级	控制精度	自由度	场景复杂度	数据量
智能驾驶	极高	厘米级	3	场景单一，但强交互博弈	千万clips级别，对应10w+小时
具身智能	极高	毫米级	30+	场景复杂	百小时级别

3 硬件结构没有标准化，影响数据规模

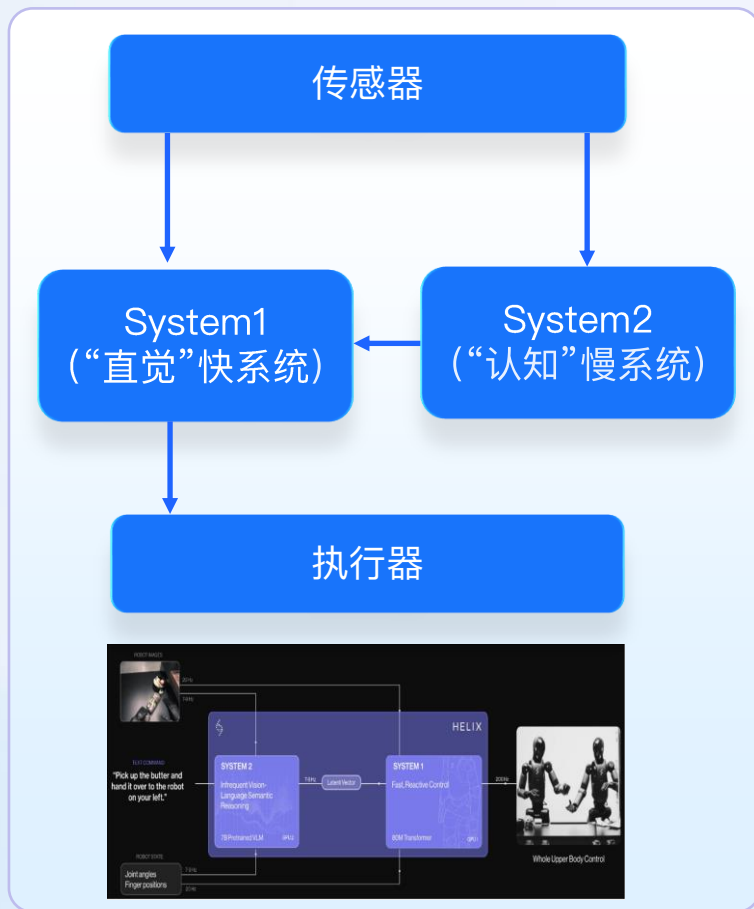


02

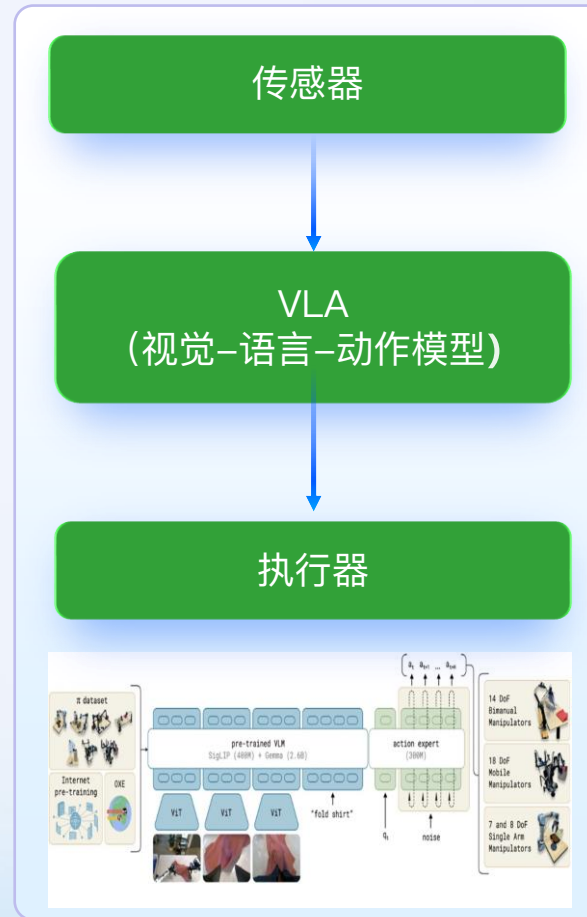
VLA的主流架构

一段式架构vs分层式架构

分层式：非全程可求导

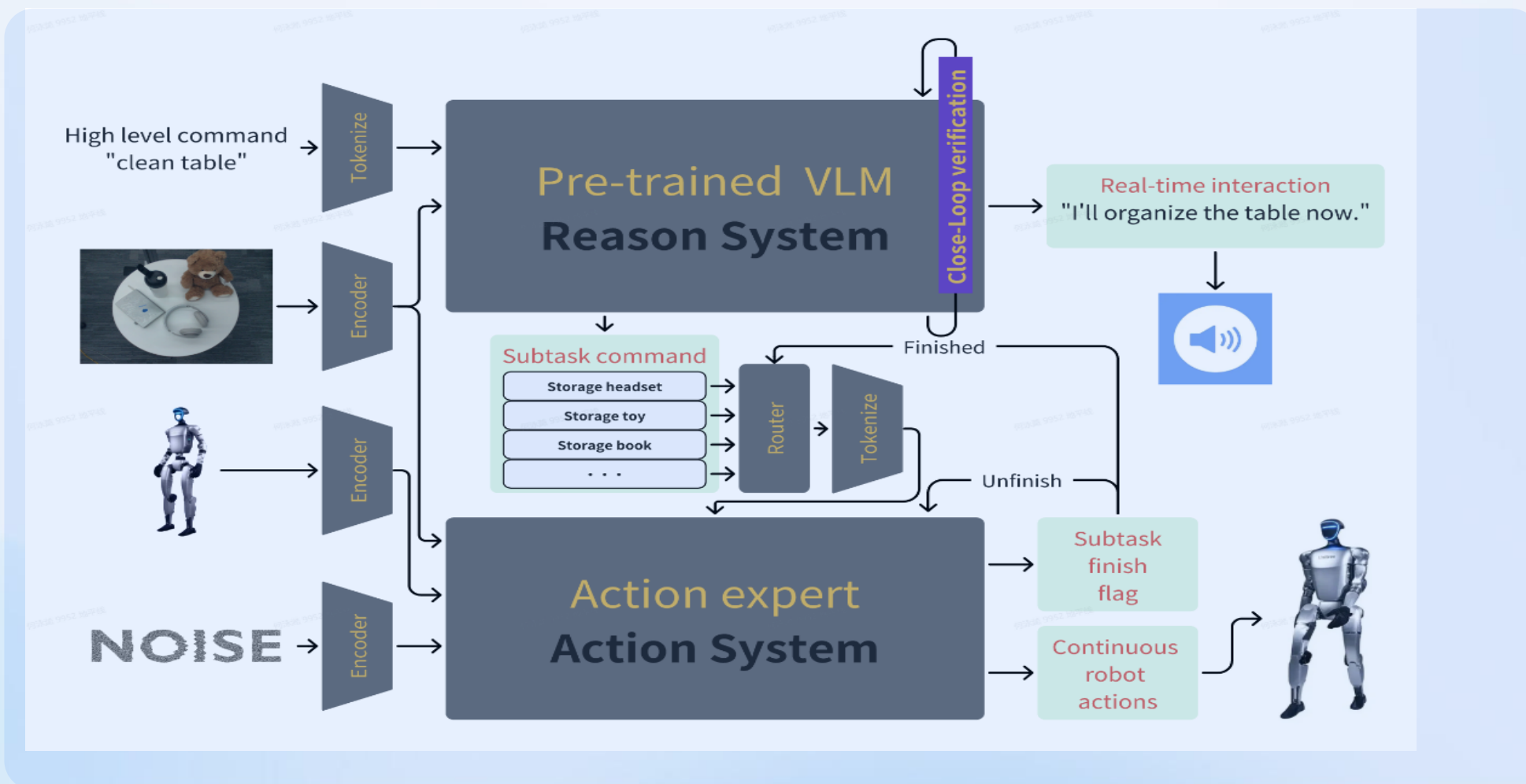


完全端到端：全程可求导



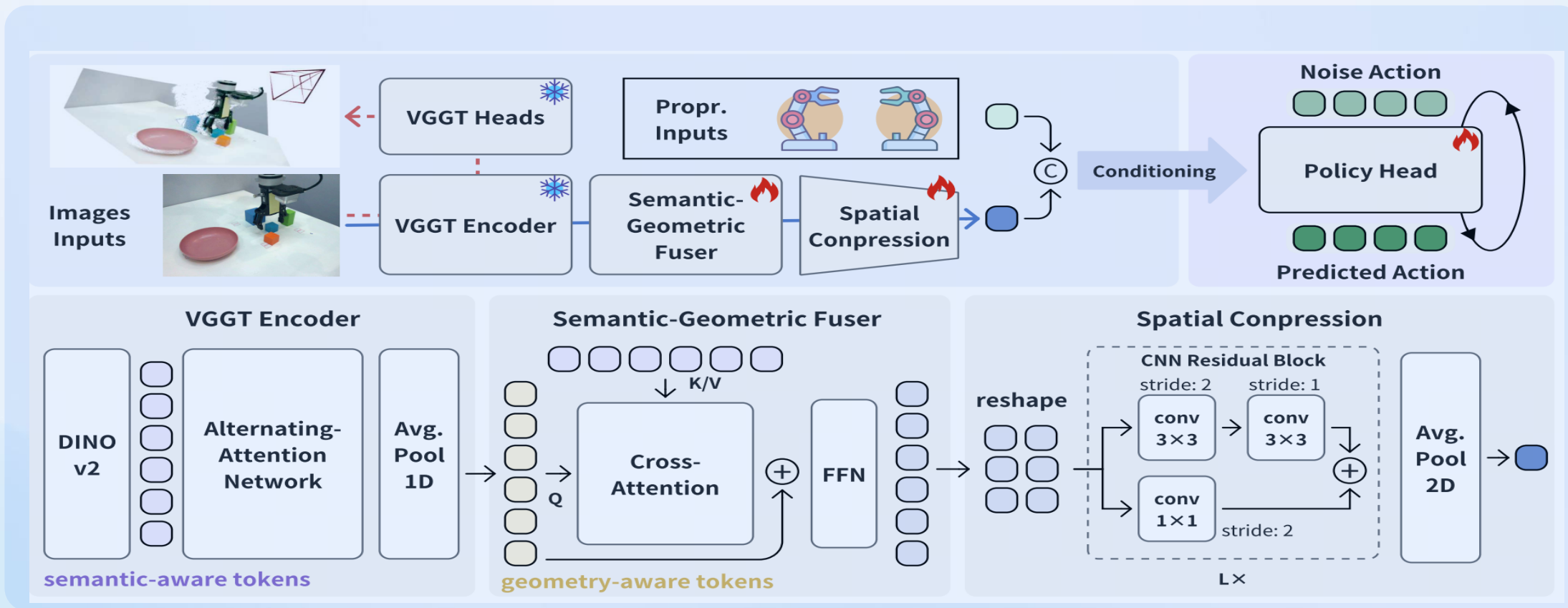
● 分层式架构：目前最具备落地可行性的方案

充分利用VLM的通用“常识”进行任务规划，通过动作原子完成复杂长程任务



低成本纯视觉VA方案构建动作原子库

- 纯视觉方案性能超过RGBD方案\泛化性和鲁棒性超过当前的VLA
- 3D-ware 预训练可明显提升任务的成功率



VO-DP: Semantic-Geometric Adaptive Diffusion Policy for Vision-Only Robotic Manipulation

Zehao Ni^{1,2,5,6*}, Yonghao He^{2*▲}, Lingfeng Qian^{2*}, Jilei Mao², Fa Fu², Wei Sui², Hu Su⁴,
Junran Peng^{3†}, Zhipeng Wang^{1,5,6}, Bin He^{1,5,6†}

¹ National Key Laboratory of Autonomous Intelligent Unmanned Systems,
² D-Robotics,

³ University of Science and Technology Beijing,

⁴ State Key Laboratory of Multimodal Artificial Intelligence System (MAIS),
Institute of Automation of Chinese Academy of Sciences.

⁵ Frontiers Science Center for Intelligent Autonomous Systems,

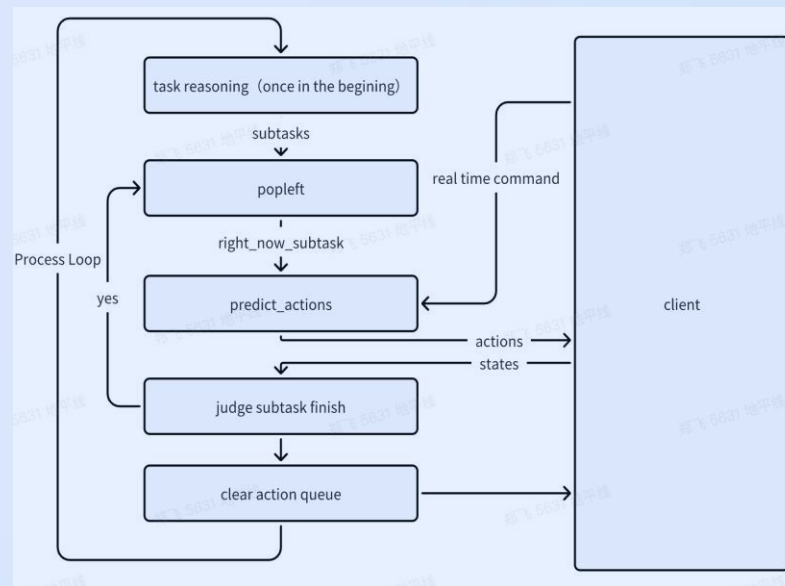
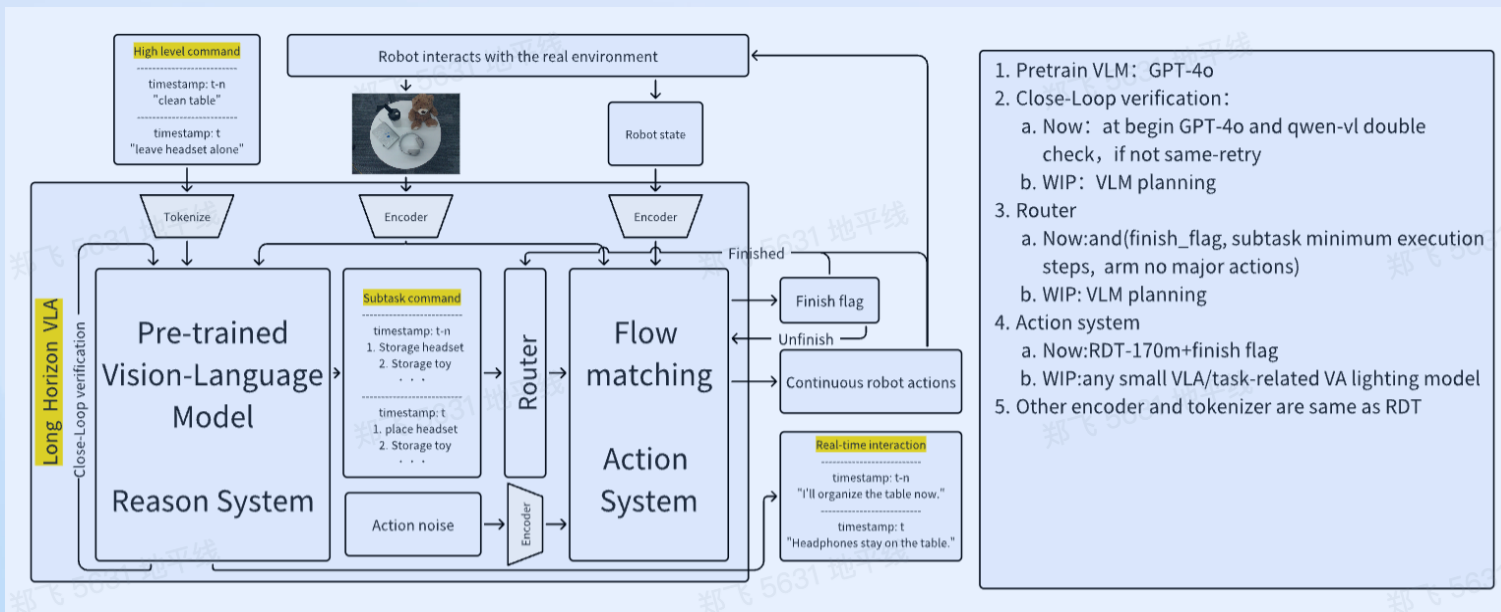
⁶ Shanghai Institute of Intelligent Science and Technology, Tongji University,

* Equal contribution, ▲ Project leader, † Corresponding author

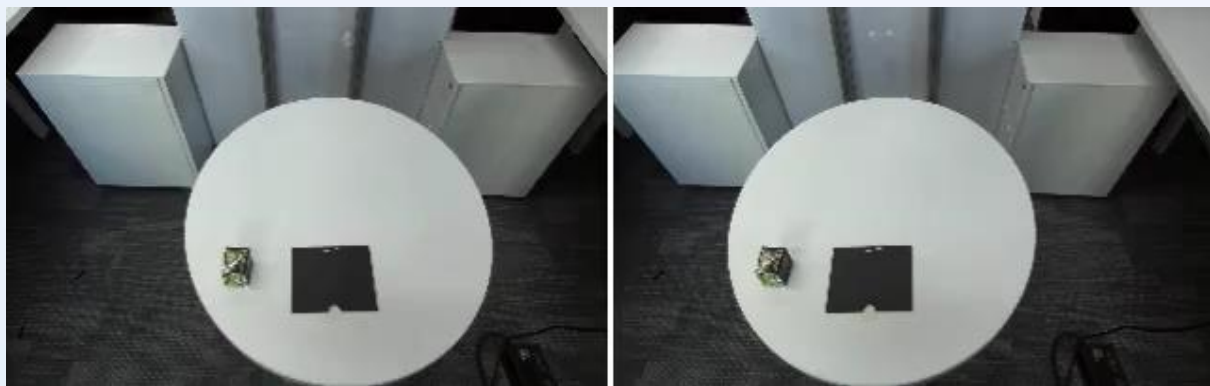


Agentic-VLA

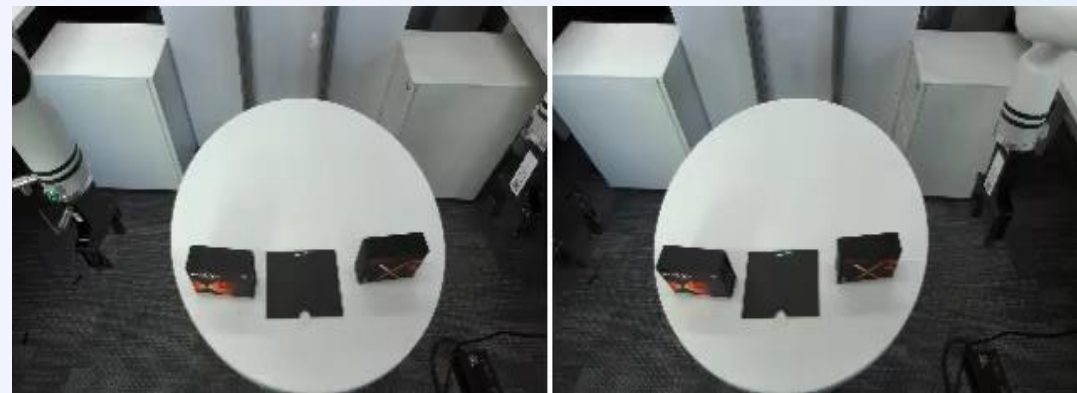
- ✓ Agentic VLA将只能完成单一任务的VA算法，通过智能体相关的技术提升成能够完成长程复杂任务的VLA算法
- ✓ Agentic VLA具备如下的关键能力：自然语音交互、复杂任务拆解和规划、VA调用和自我纠错等
- ✓ Agentic VLA核心依赖的中枢为一个强大的VLM模型，采用MCP的技术方案将所有能力进行串联，具备良好的可拓展性和灵活性



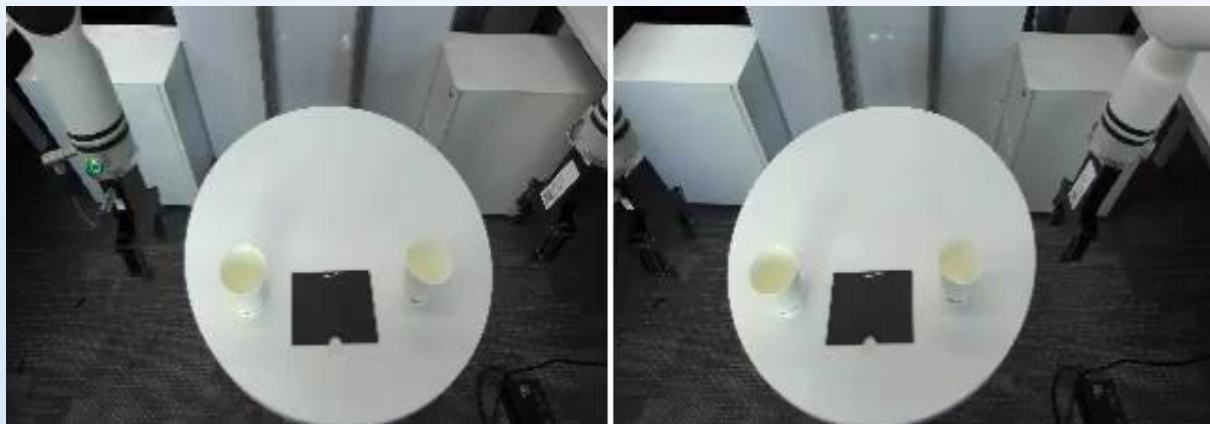
Agentic-VLA



把积木从左手给到右手



把一个盒子叠加到另一个盒子上



整理桌面桌面



把插头从插排里拔出来

03

VLA的数据方案

● 遥操作

惯性动补设备



光学动补设备



外骨骼数据采集



● 仿真

仿真在具身智能中起到的作用

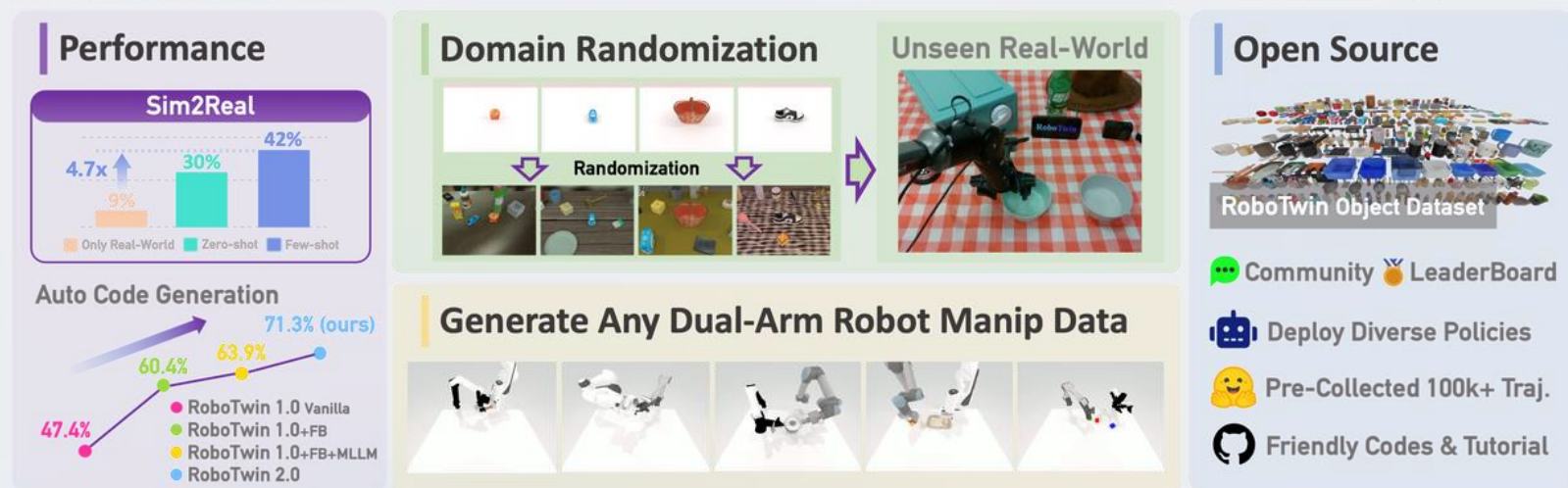
- 闭环学习/测试
- 数据生产

一个仿真器需要考虑哪些要素

- 丰富的资产
- 物理仿真
- 传感器仿真
- 模型支持
- ...

目前主流的仿真平台

- RoboTwin 2.0
- RoboVerse
- DISCOVERSE



DISCOVERSE:面向复杂真实世界的高保真多尺度仿真器

场景级高保真: 采用laser-scanned 3DGS方案, 对3DGS引入强几何正则, 对于真实世界中的大规模、非朗伯表面、精细结构、弱/重复纹理等各类复杂场景均能鲁棒地实现高质量Real2Sim

物体级高保真: 通过光照估计&PBR, 实现场景与物体的高质量联合渲染

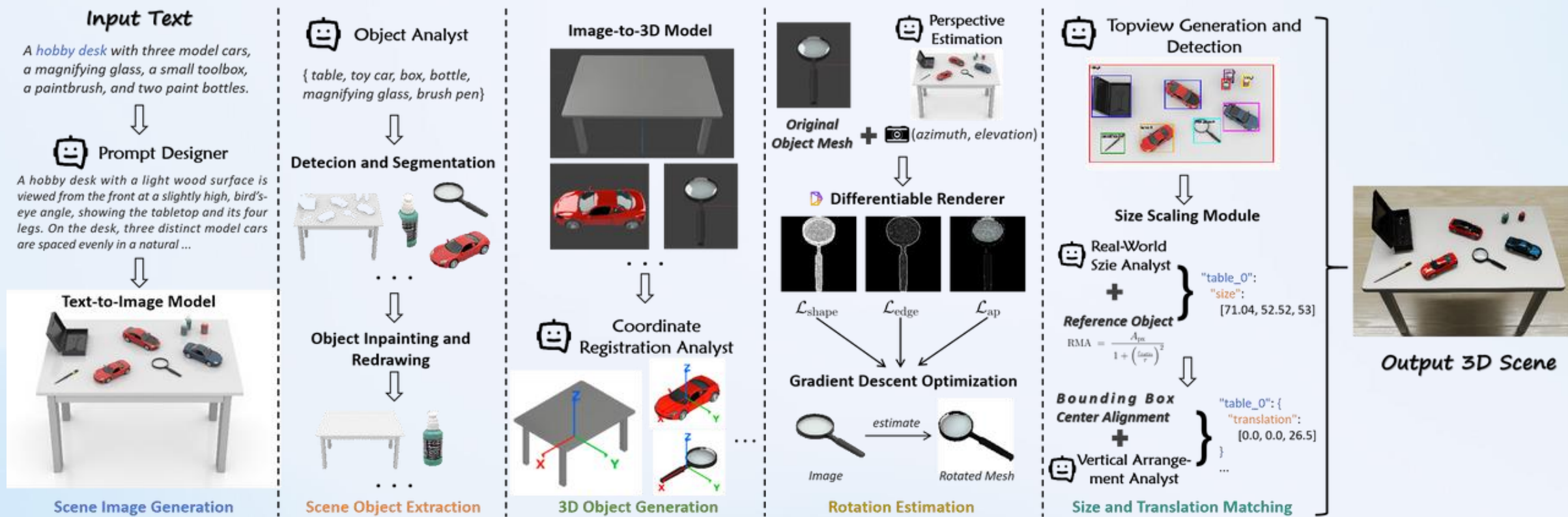


<https://github.com/TATP-233/DISCOVERSE>



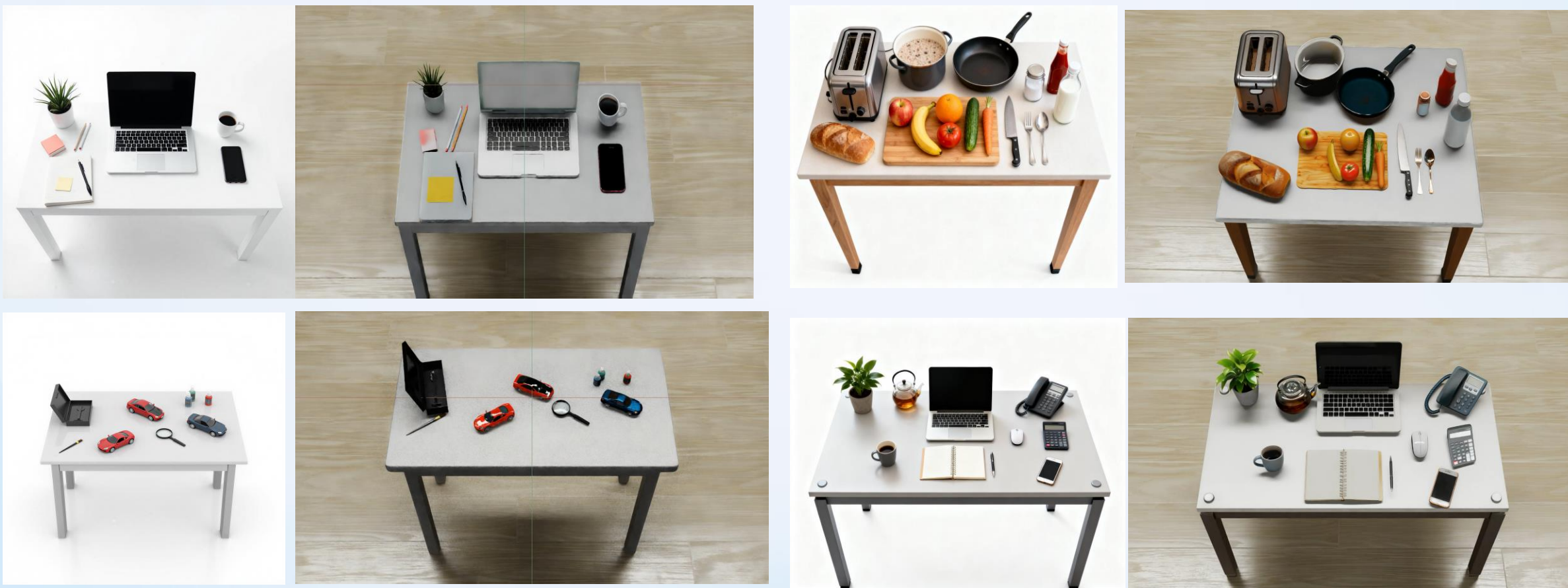
TabletopGen: 生成式桌面资产重建方案

我们聚焦一个很细分的任务：生成各种桌面场景，物体以刚体为主且可交互，因为目前绝大多数的操作任务都是在不同类型的桌子上进行pick&place



● TabletopGen: 生成式桌面资产重建方案

数字表亲：与数字孪生不同，数据表亲不需要完全一比一重建，但是会确保场景中出现需要的物体

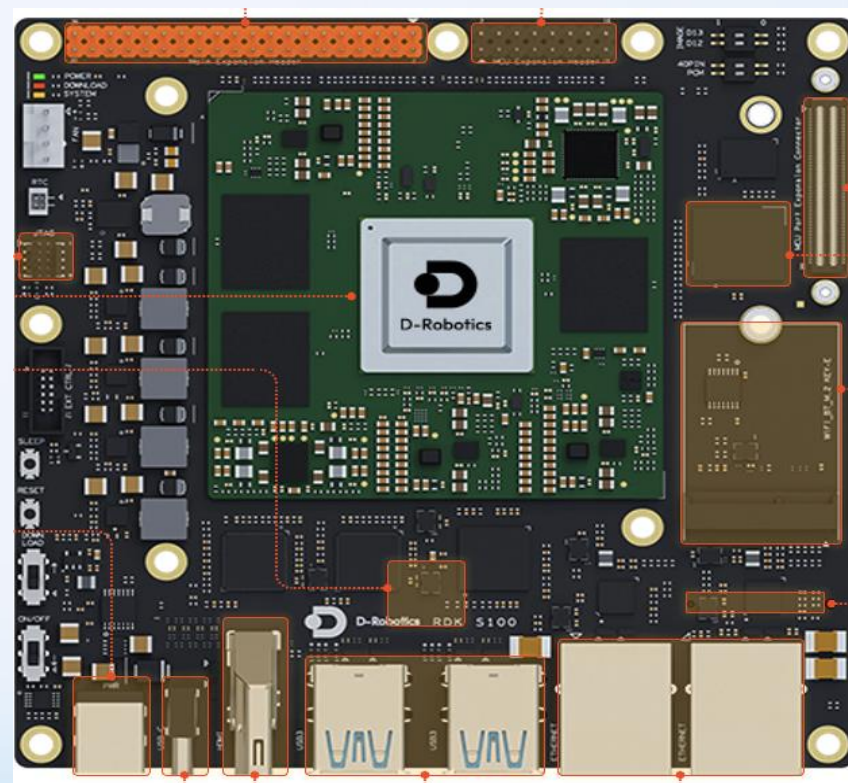


04

VLA模型的量化部署

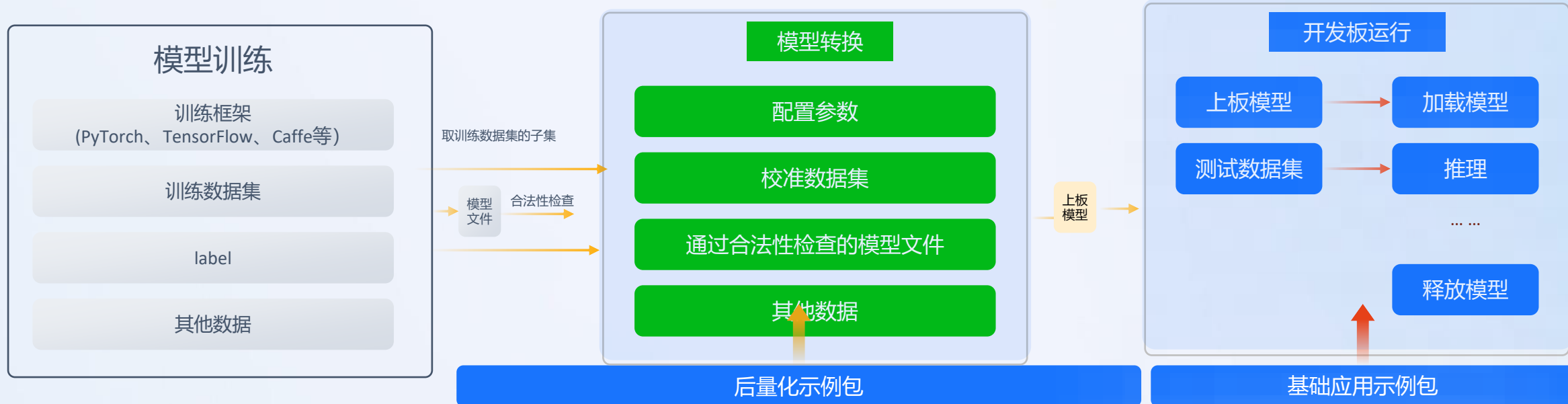
VLA大模型量化

- 模型的训练基本都是在GPU上，推理时为了提升推理效率，降低内存占用，通常会将模型量化成低bit
- 1B 模型量化成int8, 对应大小为1Gb, 相比较float数据模型可以减少4倍内存



模型量化部署流程

模型开发流程



工具链支持



05

总结与展望

VLA技术发展面临的挑战

1. 缺少高质量的数据

- 真实数据采集成本高、效率低、难度大仿真数据和真实数据存在较大的gap

2. 模型结构尚不成熟

- 模型场景3D表达能力不足，导致模型泛化性差
- 缺少有效的时序融合能力，导致长程任务失败率高
- 模型缺少对物理世界的理解

3. 硬件本体精度和稳定性待提升

- 硬件构造和人类还有差距，影响机器人的灵活性、稳定性和灵敏度
- 尚不能完成复杂的操作，影响机器人的通用性

VLA模型架构及训练范式的迭代优化方向

多模态信息

增加其它传感器，例如触觉传感器从而提升任务的成功率和稳定性

3D空间表达/记忆信息/思维链

在VLM中注入3D信息让模型具备3D空间感知能力，从而提升模型性能



强化学习

让VLA模型学习更加灵活、鲁棒

World Model

结合世界模型，让VLA具备更好的理解能力

VLA的前沿方向

为VLA构建高效的数据闭环是智能化提升的关键

通过云端遥操进行安全兜底，并且回传数据进行性能迭代

随着智能驾驶/具身智能性能提升，云端接管次数变少，安全员可操作的车辆/机器人变多，成本降低

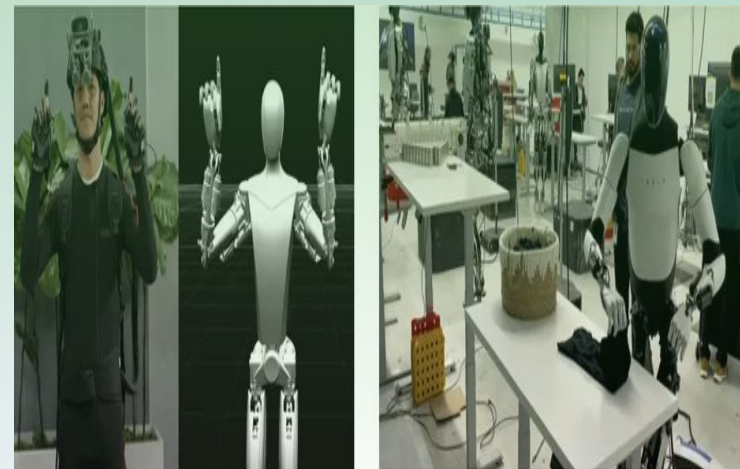
车端无人驾驶系统



云端遥操作



具身智能云端遥操作



📍 北京

QCon

全球软件开发大会

会议时间：4月10-12日

- 大模型赋能 AIOps
- 越挫越勇的大前端
- 云上业务架构演进
- 多模态大模型及应用
- 海外 AI 应用创新实践

📍 北京

AiCon

全球人工智能开发与应用大会

会议时间：6月27-28日

- 端侧智能
- AI Agent
- 多模态大模型
- 金融+大模型

📍 上海

QCon

全球软件开发大会

会议时间：10月23-25日

- AI Agent
- 大模型训练推理
- 端侧 AI
- 搜推深度融合
- 数智金融

4月

5月

6月

8月

10月

12月

📍 上海

AiCon

全球人工智能开发与应用大会

会议时间：5月23-24日

- 通用大模型
- AI Agent
- 垂直领域应用
- 模型可解释性

📍 深圳

AiCon

全球人工智能开发与应用大会

会议时间：8月22-23日

- 模型效率与部署
- 多模态大模型
- 大模型安全
- 智能硬件

📍 北京

AiCon

全球人工智能开发与应用大会

会议时间：12月19-20日

- 通用大模型
- 智能硬件
- LMOps
- 具身智能

极客邦科技 2025 年会议规划

促进软件开发及相关领域知识与创新的传播



参会咨询



查看会议

THANKS

大模型正在重新定义软件

Large Language Model Is Redefining The Software

