



智元大会具身智能与人形机器人分论坛交流

世界模型驱动的具身智能研究

孙富春

清华大学计算机科学与技术系
人工智能研究院智能机器人中心
中科第五纪智能技术有限公司

中国北京, 2025.06.07

一、空间智能与世界模型

让AI理解现实世界：

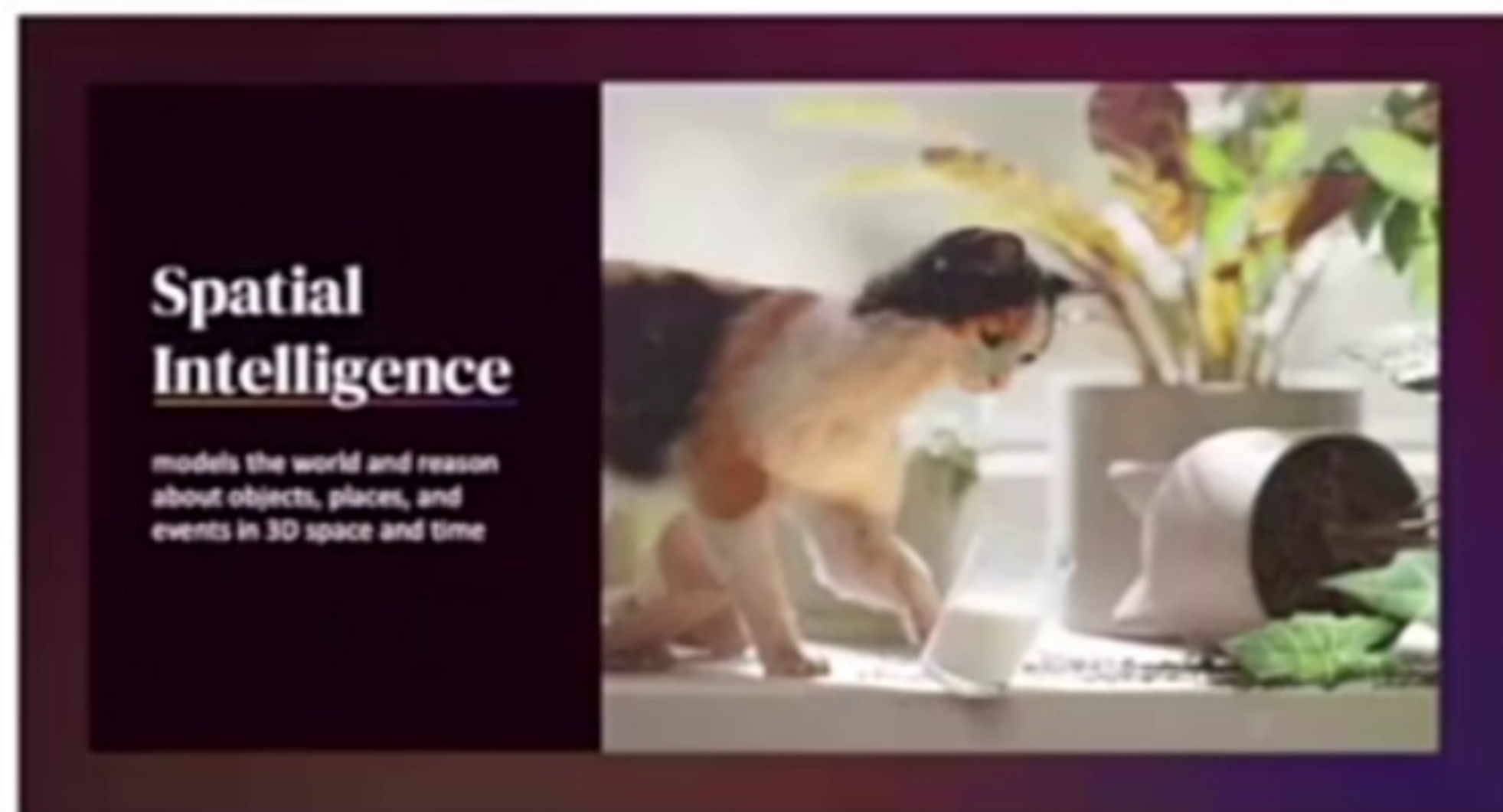


在温哥华举行的2024年TED大会上，发表了以《有了空间智能，AI将会理解现实世界》

■ “空间智能”是智能体在三维空间中感知、推理、和行动的能力，了解物体在三维空间中的定位，从而让机器对三维、甚至四维物理世界有更好的理解，以及基于这种理解做出合理预测和行动的能力。

■ 它是任何需要理解并与三维世界互动的具身智能系统的关键组成部分。

■ 空间智能开启AI发展新篇章，或成下一代AI的“原爆点”





一、空间智能与世界模型

空间智能

指智能系统感知、理解和操作空间信息的能力,包括对物体的识别、定位、物理规律的理解等。

- 💡 **视觉化为洞察**: 解码编码在图像中的空间关系
- 💡 **“看见”成为理解**: 理解三维空间中的位置、形态、运动状态等
- 💡 **理解导致行动**: 基于理解, 给出在三维世界中合理有效的操作或互动

VoxPoser: 利用大模型实现零样本操作



VoxPoser随着“眼睛”的发育而出现的能力涌现

- **评估物理特性**: 给定两个质量未知的方块, 机器人可进行物理实验, 确定哪个方块更重。
- **行为常识推理**: 在摆餐具任务中, 用户指定行为偏好, 比如“我是左撇子”, 机器人在任务的上下文中理解其含义。
- **细粒度语言校正**: 对于需要高精度的任务, 比如“给茶壶盖上盖子”, 用户可以给机器人准确的指令, 比如“你偏离了1厘米”。
- **多步的任务规划**: 给定一个任务“将抽屉精确地打开一半”, 机器人根据视觉反馈提出多步操作策略, 即首先完全打开抽屉同时记录手柄位移, 然后将其推回至中点。



一、空间智能与世界模型

世界模型 用于描述、理解和预测外部环境的状态及变化的抽象模型

花会随时间如何变化?



物理规律
理解



物理规律
应用



AI在脑海中想象整个世界，
预测未来情景
——就像人类做梦一样

世界模型从根本原理上理解世界，
并更高效地学习
——如同人类总结客观规律

世界模型基于对客观规律的理解与
动作预演，实现高效规划与决策
——如同人类三思而行

💡 构建内部表征以**理解世界的机制**

💡 预测未来状态以**模拟和指导决策**

核心作用：**反事实推理**
Counterfactual reasoning

一、空间智能与世界模型

空间智能与世界模型共同指导智能体与世界交互

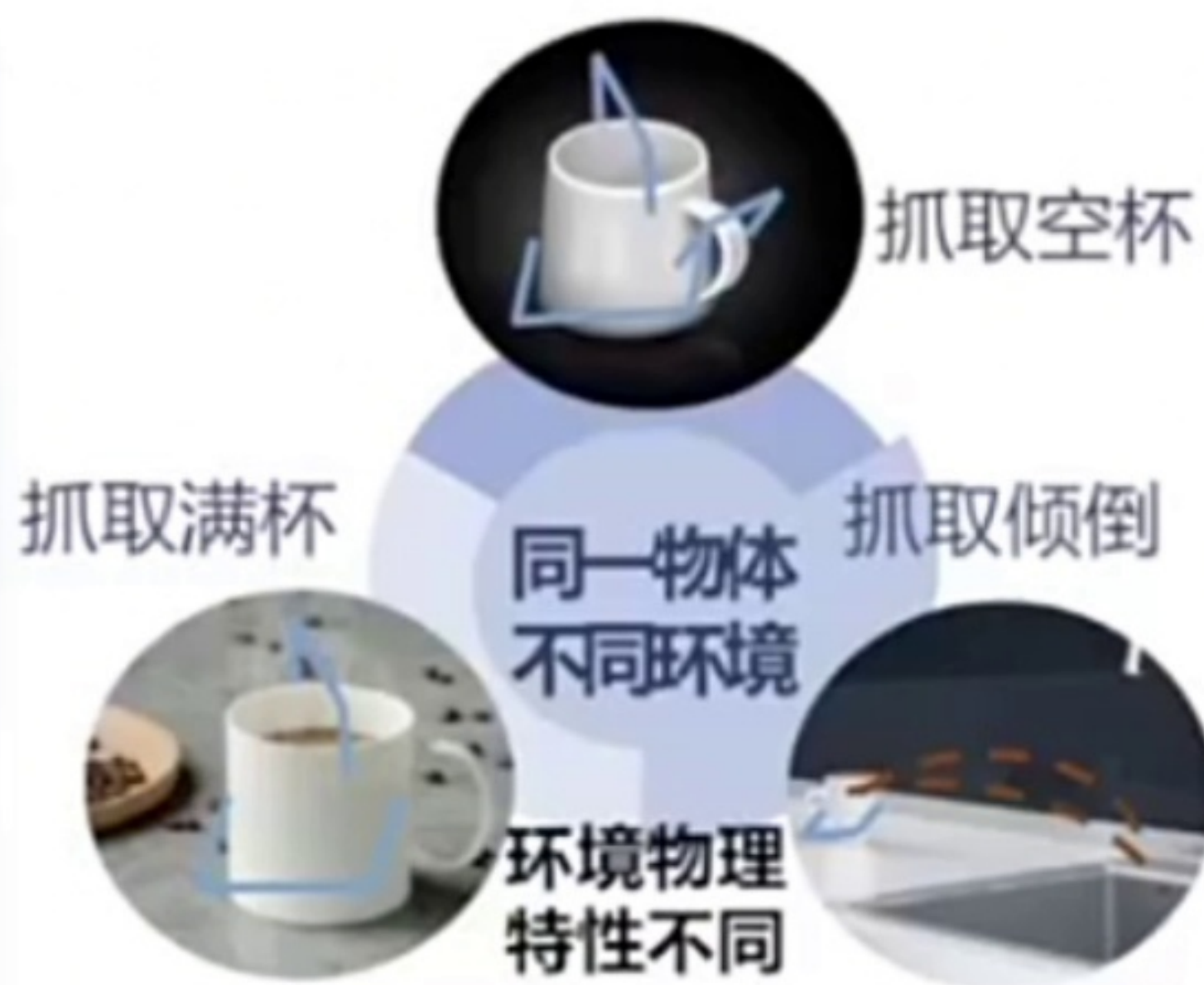


二、数字物理世界模型

为什么需要数字物理世界模型？

人类具身智能

- 深度感知环境物理属性（形状、质地、柔性、力反馈）
- 利用手部/工具目标导向操作
- 感知—预测—调整闭环
- 操作过程中能即时适应与修正



智能体关键挑战

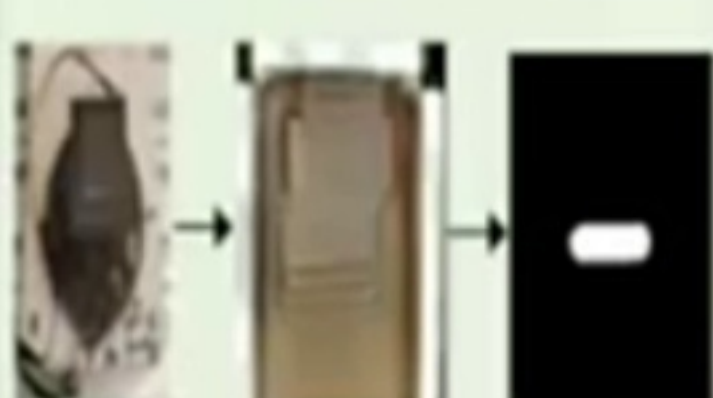
- 仅依靠视觉/语义模型无法获得物理约束信息
- 操作策略缺乏真实物理反馈
- 模型对真实世界变化缺乏鲁棒性

人类具备充分理解环境的物理特性
并以此使用人手或者功能工具进行
操作的能力

数字孪生建模



语义-物理耦合

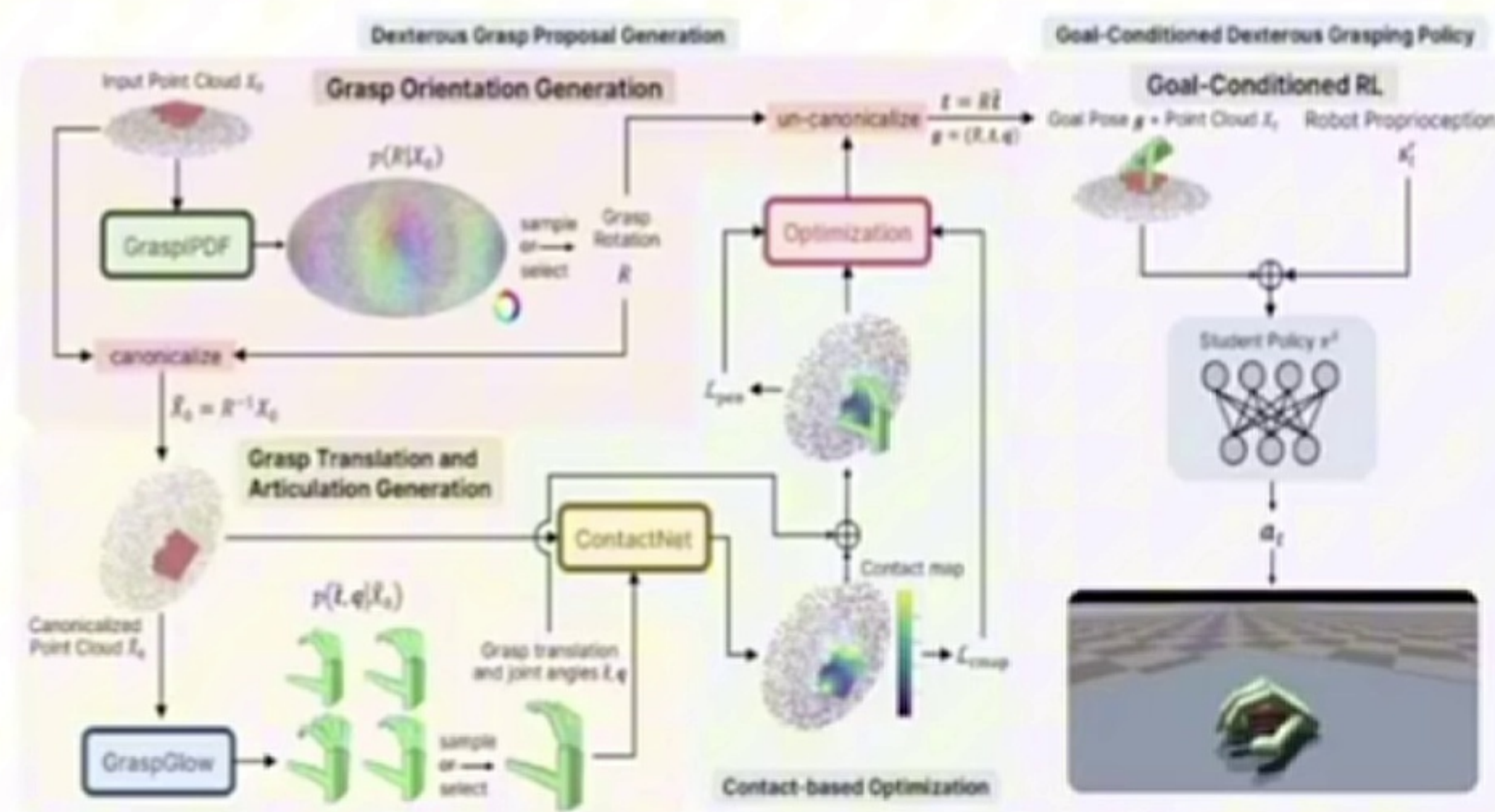


感知—建模—控制





二、数字物理世界模型

URDFormer:
具有语义和运动结构的环境模拟UniDexGrasp:
抓取手势的生成式三维信息重构



二、数字物理世界模型

空间智能体在三维世界中的能力体现：

01 3D感知与建模

3D感知技术可获取环境深度信息，为空间智能提供基础数据。NeRF和3D Gaussian建模将感知数据转化为精确的三维模型，支持后续的理解与推理。

02 空间推理与决策

空间推理技术处理和分析空间中几何透视，相对位置，物理运动属性等关系，为决策提供逻辑支持。决策模块根据推理结果规划决策序列，实现智能体的自主决策。

03 动态交互与操作

智能体能在多变环境中与人或其它智能体进行有效动作交互。行为智能体根据空间推理与决策结果执行复杂的物理任务，如抓取、移动物体等。

04 移动操作

耦合移动与操作，在空间中更高效的完成任务，结合动态感知、联合规划、稳定性控制、全身控制，多任务协调。

二、数字物理世界模型

目前数字孪生/世界模型：

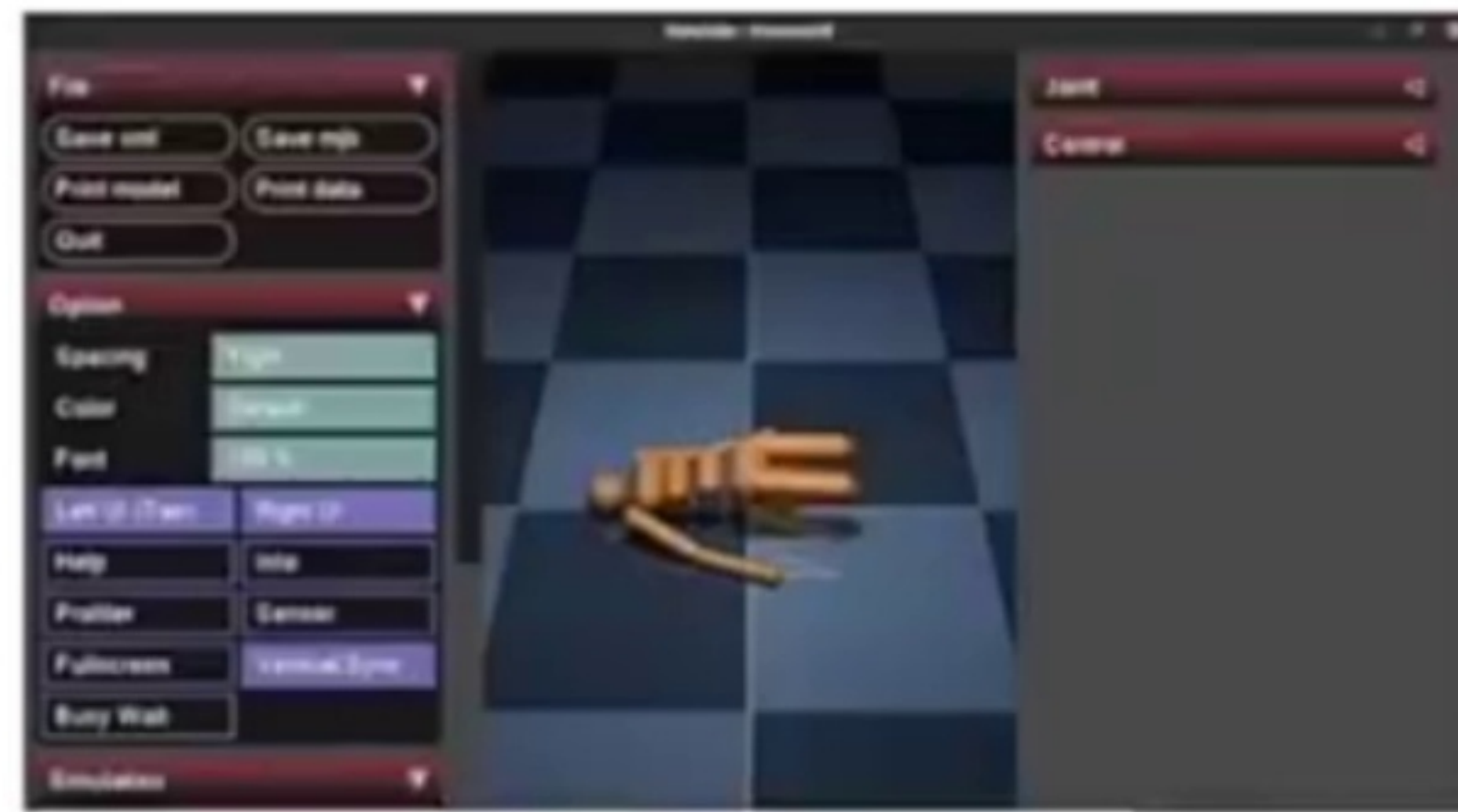
- 仿真环境逼真度低、缺乏多模态信息的感知（**力觉、触觉、滑觉**等），与真实场景存在较大差异；
- 无法复现真实环境中智能与环境的交互，基于仿真环境所习得的技能也难以应用到真实场景中
- 限制了机器人操作能力的发展，尤其是灵巧操作能力。



Gazebo



Pybullet

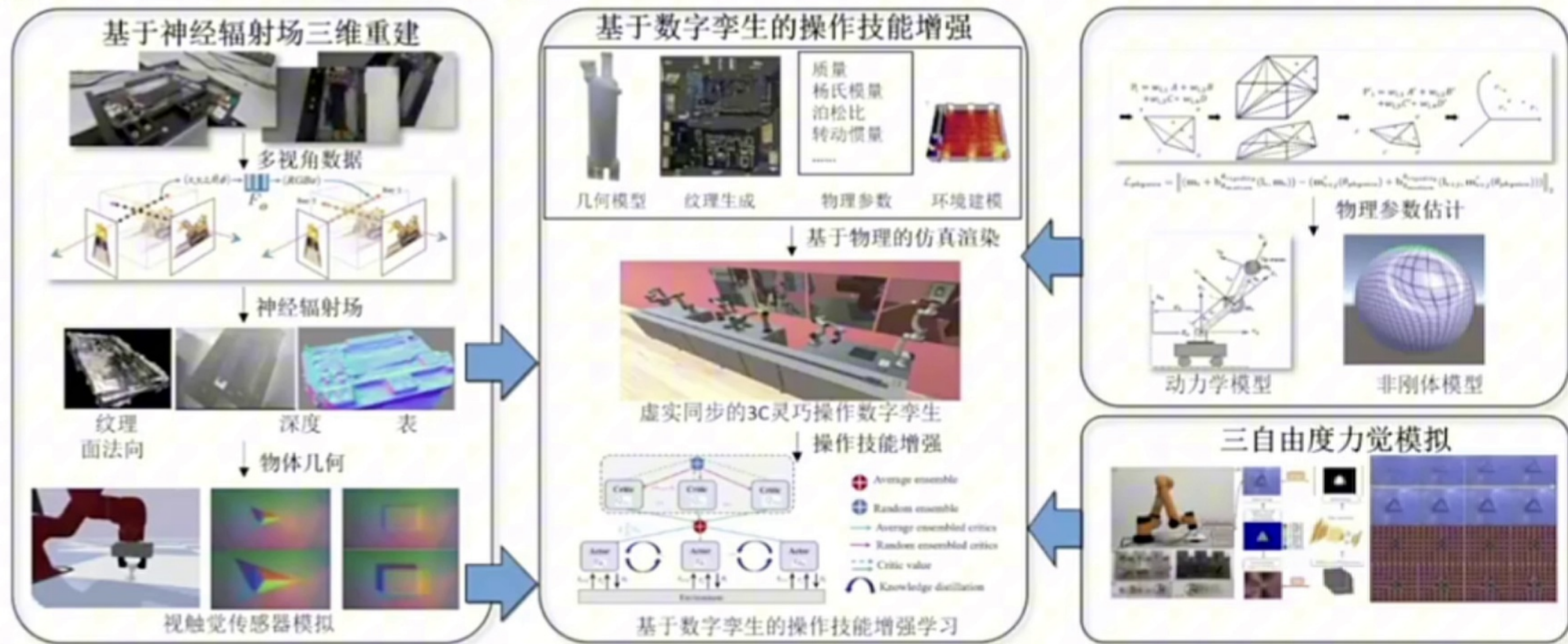


MuJoCo

构建数字物理世界模型，实现高鲁棒性、强可用性技能迁移具有重要意义

二、数字物理世界模型

具有物理交互特性的数字孪生环境构建



- 神经辐射场可见光重建，隐式场表征

- 数字孪生的虚实同步，操作技能虚实增强与迁移

- 几何、纹理和物理参数估计，虚拟触觉力模拟，



二、数字物理世界模型

刚性物体重建:

- 建立**SDF和Opacity之间关系**用于几何结构重建

$$\alpha_i = \max \left(\frac{\Phi_s(f(\mathbf{x}_i)) - \Phi_s(f(\mathbf{x}_{i+1}))}{\Phi_s(f(\mathbf{x}_i))}, 0 \right)$$

- 哈希编码**神经辐射场实现高精度重建

$$\mathbf{x}_{i,l} = \mathbf{x}_i \cdot V_l, \quad \gamma(\mathbf{x}_i) = (\gamma_1(\mathbf{x}_{i,1}), \dots, \gamma_L(\mathbf{x}_{i,L}))$$



真实值

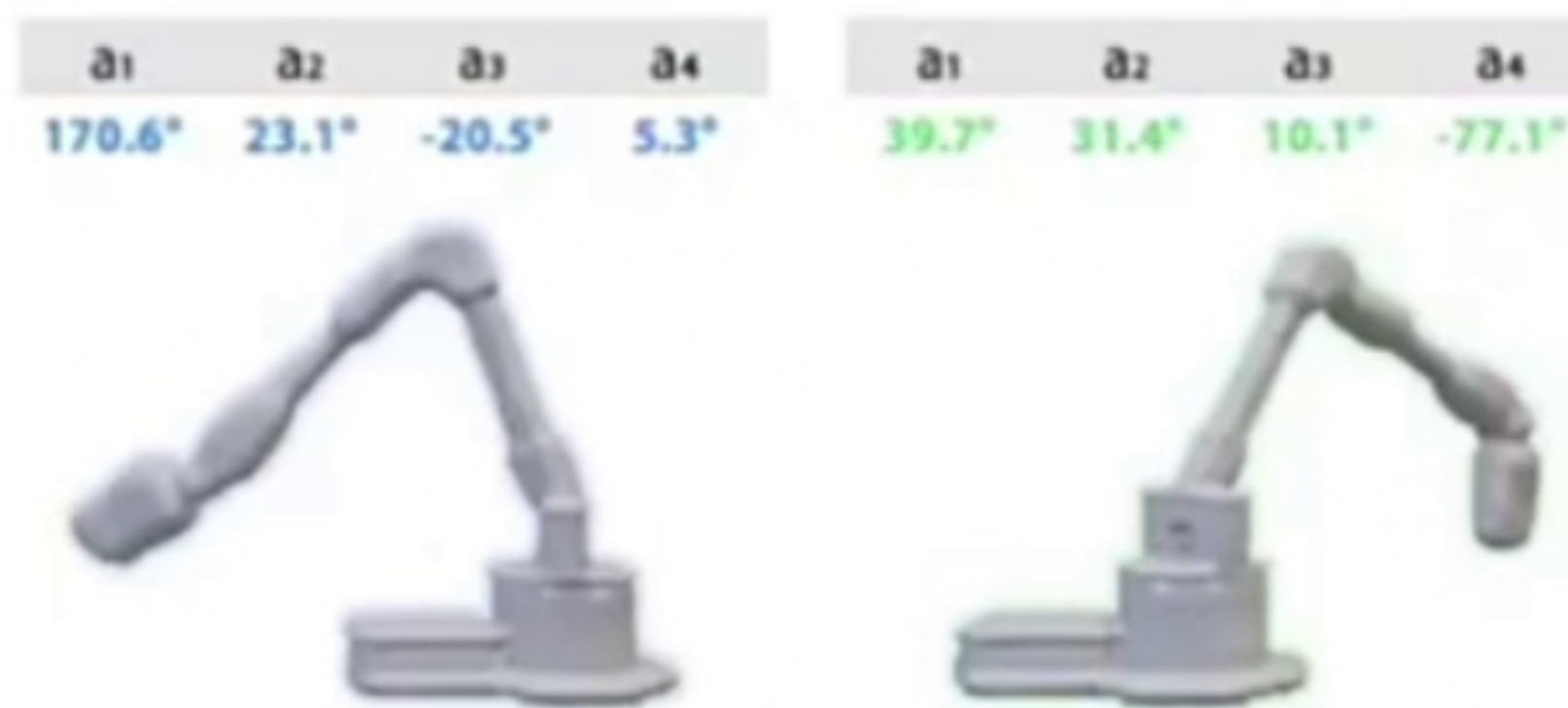
渲染结果

表面法向

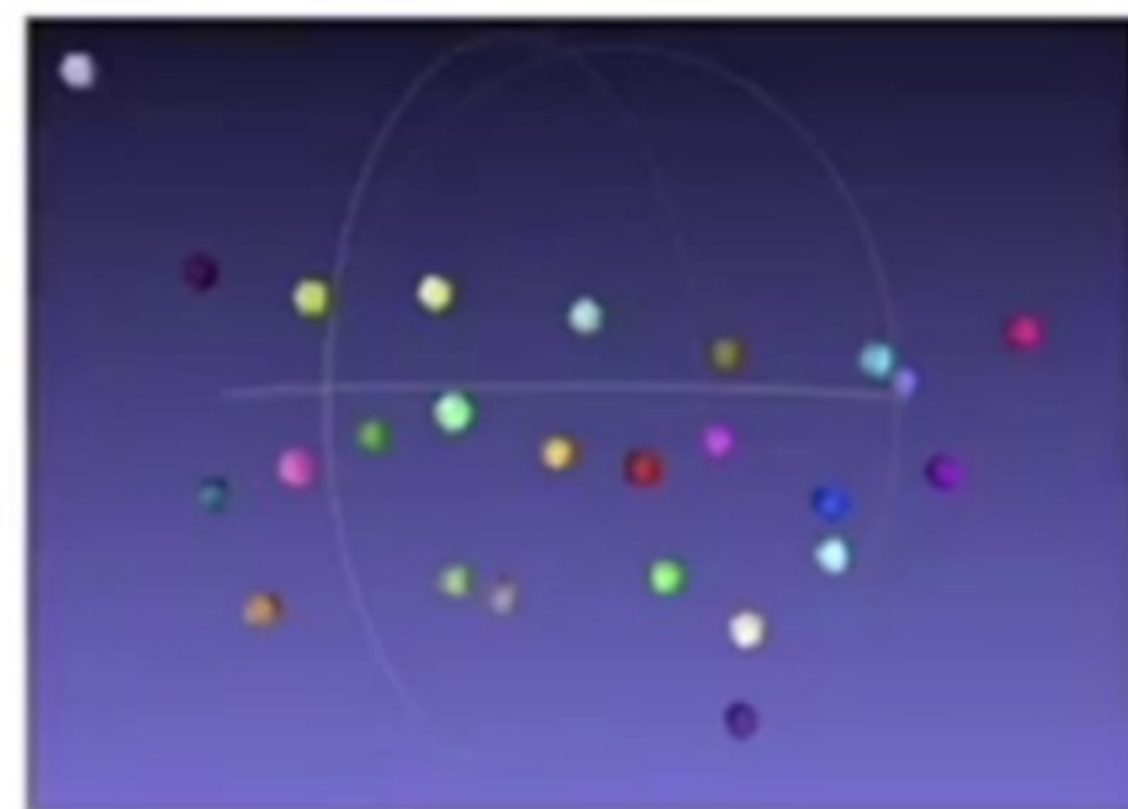
铰链物体重建:

- 将机械臂、同轴线等简化成铰接物体
- 神经辐射场作为正规空间模型
- 利用铰链骨骼和神经混合蒙皮将正规空间变换到相机空间

$$\mathbf{J}^{t,\rightarrow} = \sum_{b=1}^B \mathbf{W}_b^{t,\rightarrow} \Delta \mathbf{J}_b^t, \quad \mathbf{J}^{t,\leftarrow} = \sum_{b=1}^B \mathbf{W}_b^{t,\leftarrow} (\Delta \mathbf{J}_b^t)^{-1} \quad \begin{aligned} \mathbf{X}^t &= \mathbf{W}^{t,\rightarrow}(\mathbf{X}^*) = \mathbf{G}^t(\mathbf{J}^{t,\rightarrow} \mathbf{X}^*), \\ \mathbf{X}^* &= \mathbf{W}^{t,\leftarrow}(\mathbf{X}^t) = \mathbf{J}^{t,\leftarrow}((\mathbf{G}^t)^{-1} \mathbf{X}^t) \end{aligned}$$



相机空间渲染结果



铰链骨骼



基于生成式大模型蒸馏的单目铰链物体重建

规范空间隐式模型:

- 使用符号距离函数场的神经辐射场来表达**规范空间**当中的3D模型:

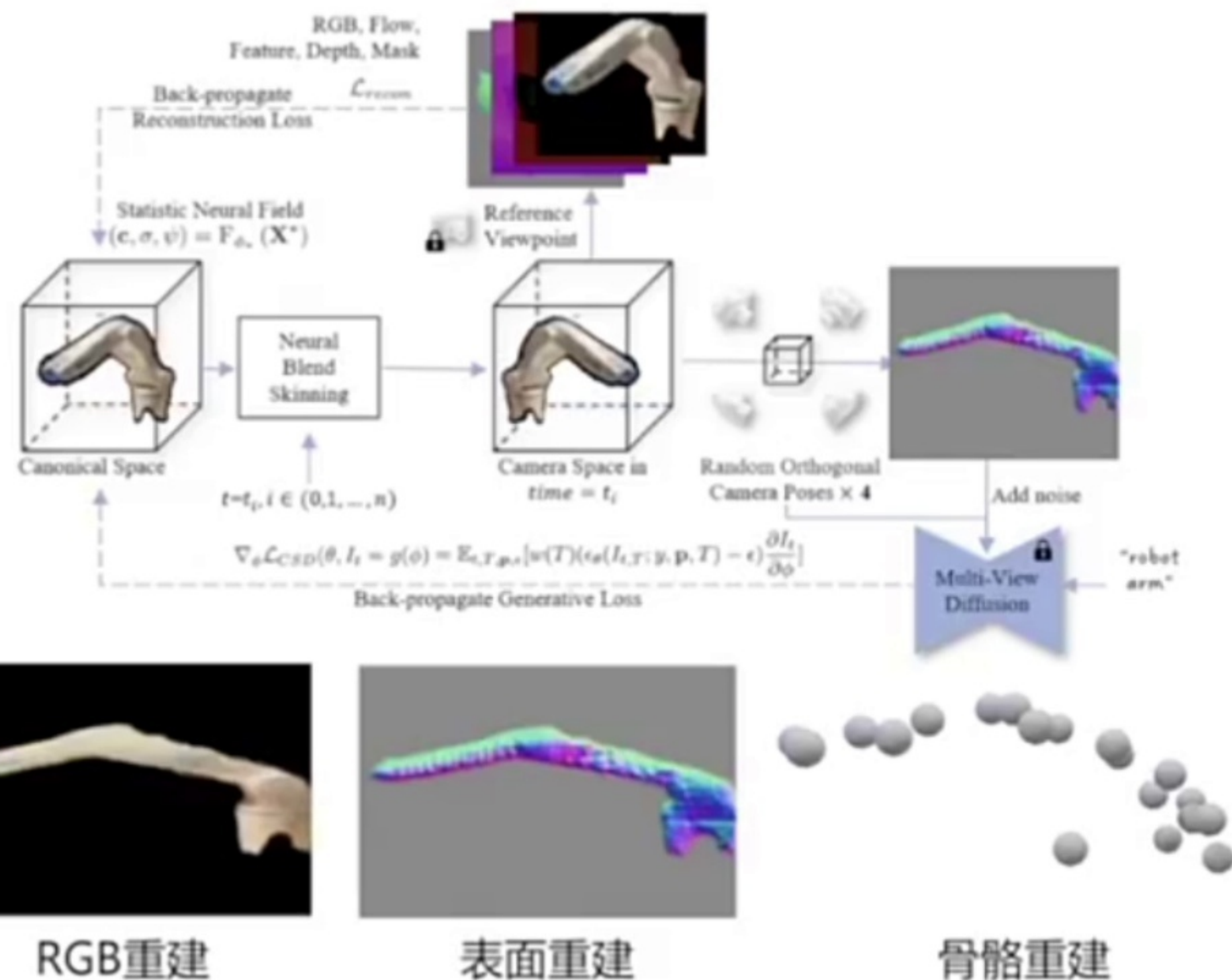
$$\begin{aligned} \mathbf{c} &= \text{MLP}_{\mathbf{c}}(\mathbf{X}^*, \mathbf{v}), \\ \sigma &= \Gamma_{\beta}(\text{MLP}_{\text{SDF}}(\mathbf{X}^*)), \\ \psi &= \text{MLP}_{\psi}(\mathbf{X}^*), \end{aligned}$$

- 渲染**自监督**DINO V2的特征, 用于规范空间与相机空间之间点的**配准**

隐式动画骨骼模型: 利用采样点变换场将规范空间与相机空间进行**可微分运动场**建模:

$$\begin{aligned} \mathbf{X}^t &= \mathcal{W}^{t, \rightarrow}(\mathbf{X}^*) = \mathbf{G}^t(\mathbf{J}^{t, \rightarrow} \mathbf{X}^*), \\ \mathbf{X}^* &= \mathcal{W}^{t, \leftarrow}(\mathbf{X}^t) = \mathbf{J}^{t, \leftarrow}((\mathbf{G}^t)^{-1} \mathbf{X}^t) \end{aligned}$$

$$\mathbf{J}^{t, \rightarrow} = \sum_{b=1}^B \mathbf{W}_b^{t, \rightarrow} \Delta \mathbf{J}_b^t, \quad \mathbf{J}^{t, \leftarrow} = \sum_{b=1}^B \mathbf{W}_b^{t, \leftarrow} (\Delta \mathbf{J}_b^t)^{-1}$$



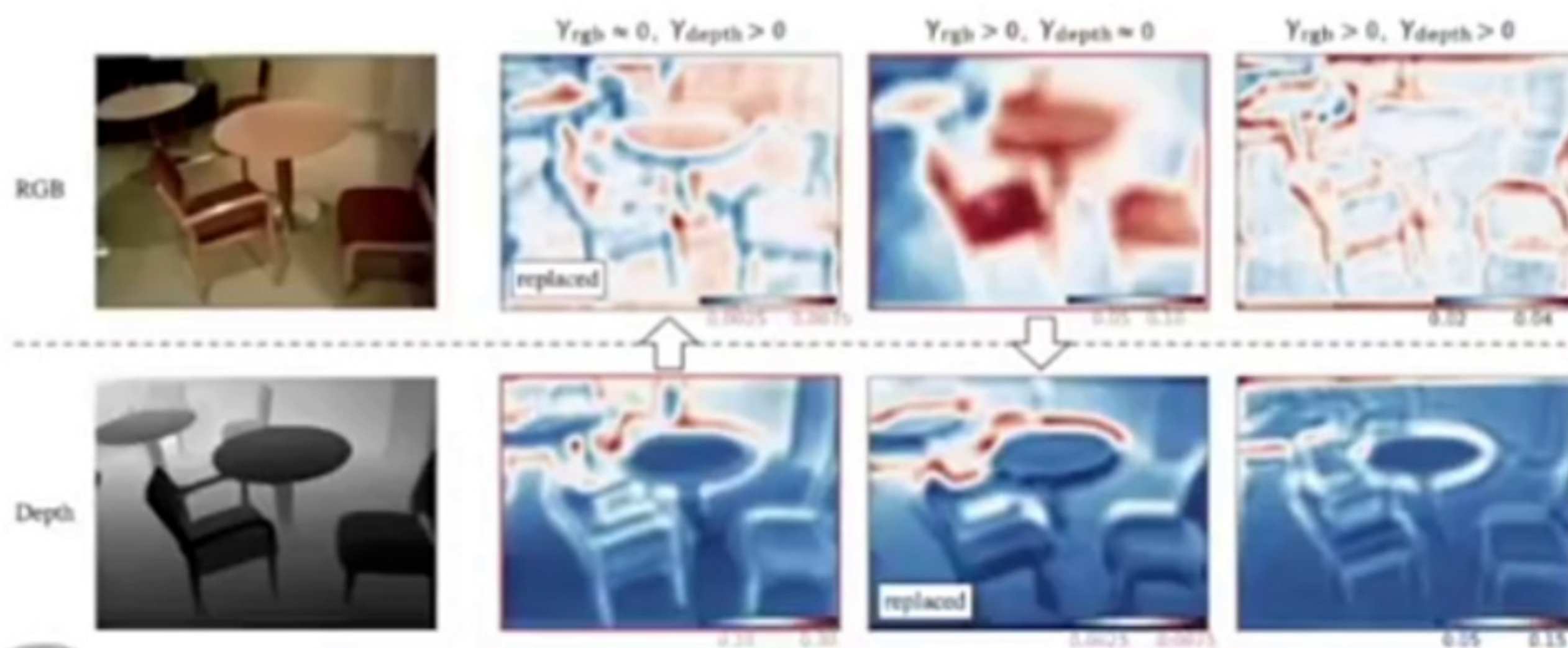
二、数字物理世界模型

空间智能体在三维世界中的推理能力体现：



空间关系推理（几何、拓扑、动力学）

空间关系推理技术需要处理复杂的几何和拓扑关系，为决策提供逻辑基础。



多模态融合推理（视觉+语言+触觉）

多模态融合推理技术能够整合视觉、语言、触觉等信息，提供更全面的环境理解。

融合推理通过弥补单一信息源的信息丢失，对于提高智能体的交互能力具有重要意义。

二、数字物理世界模型

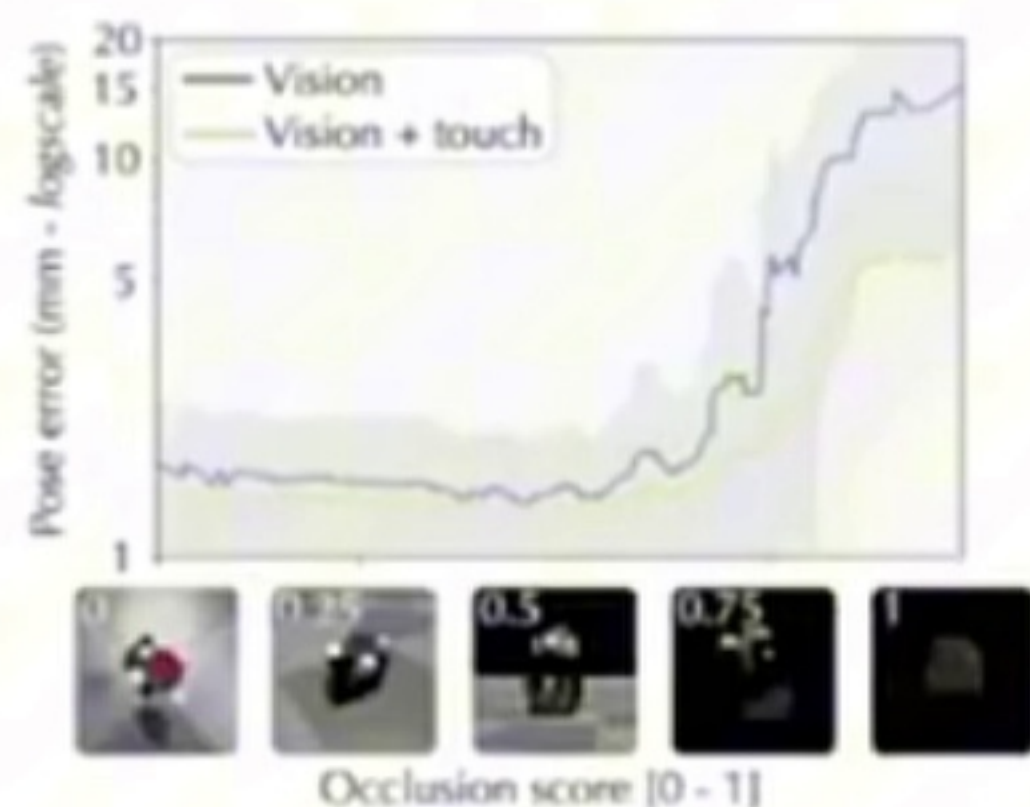
空间智能体在三维世界中的智能化与人性化提升：

结合物理属性的精细操作 01

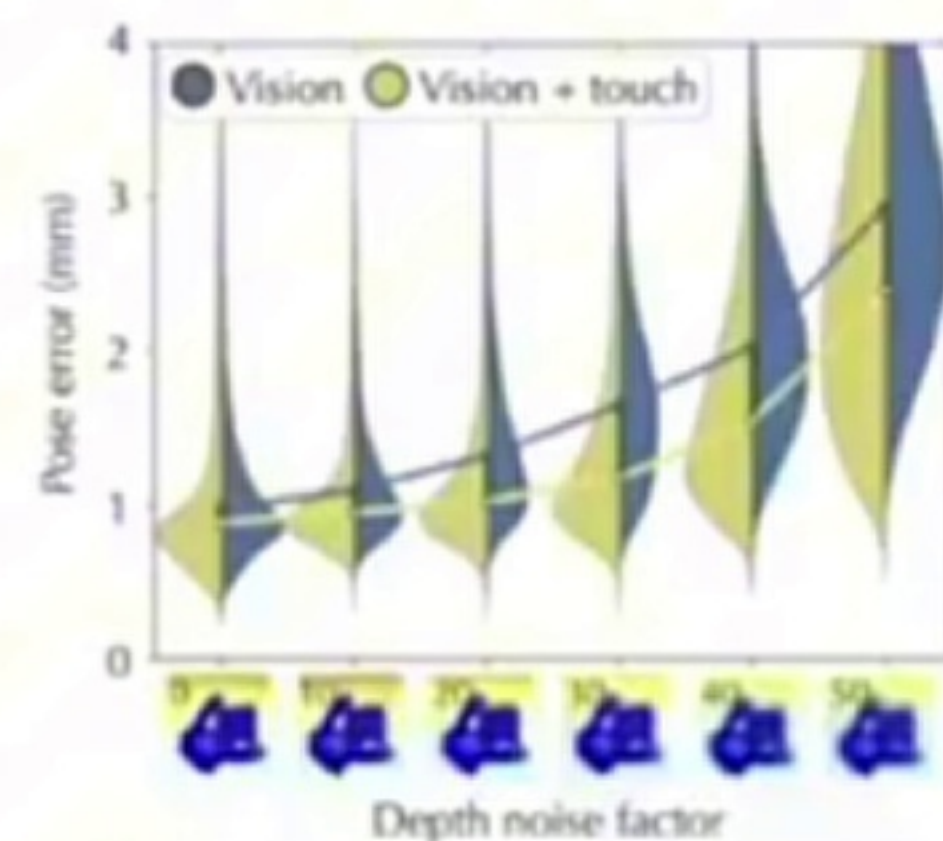
空间智能体根据物体的位置，几何属性，材质等物理属性进行动态操作与交互。

人机协作意图识别 02

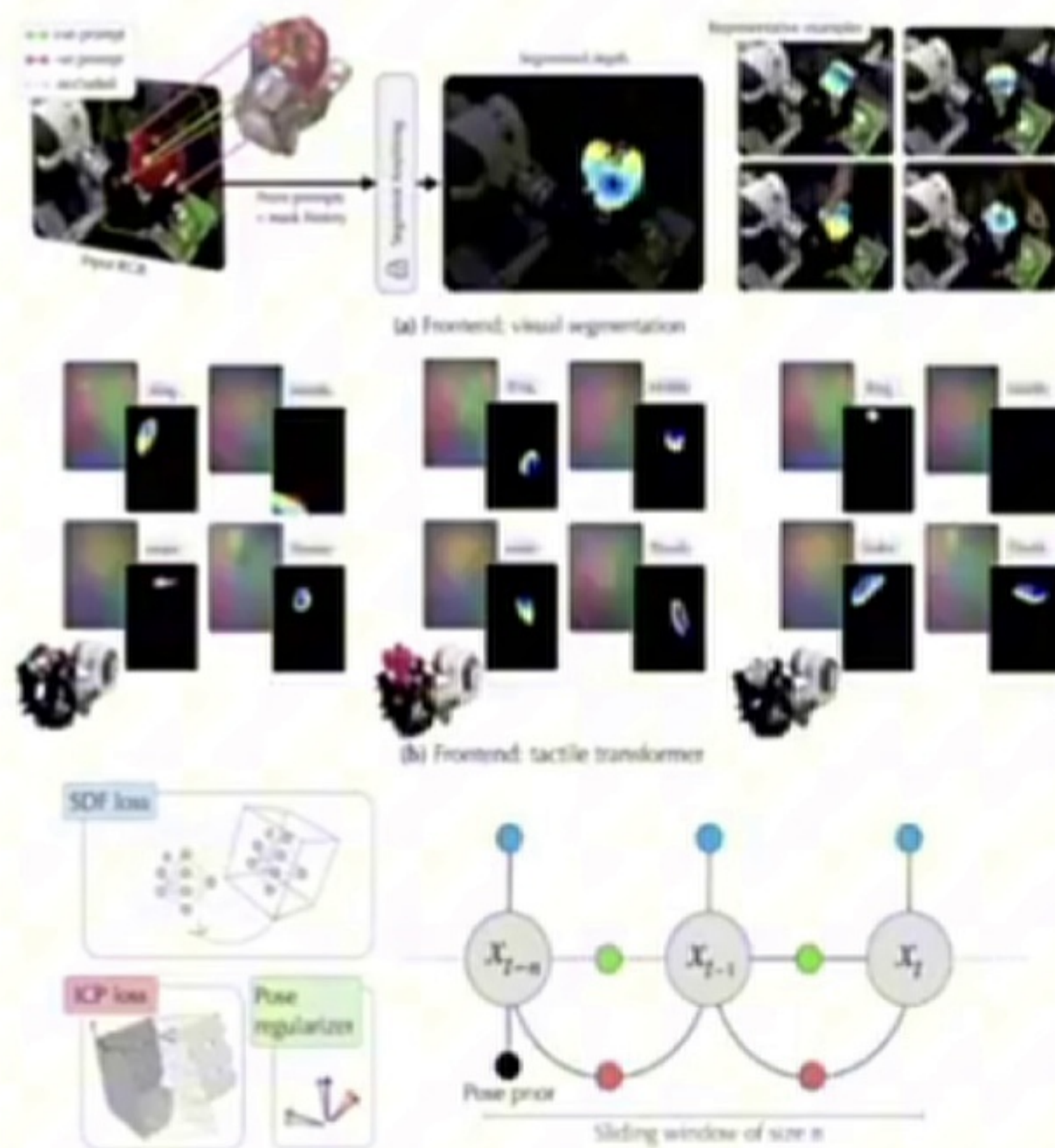
人机协作的意图识别技术在于提高人机协作的效率和交互的安全性，能够使智能体理解人的意图，实现更自然的交互。



(b) Pose error with object occlusion (↓)



(c) Pose error with depth noise (↓)

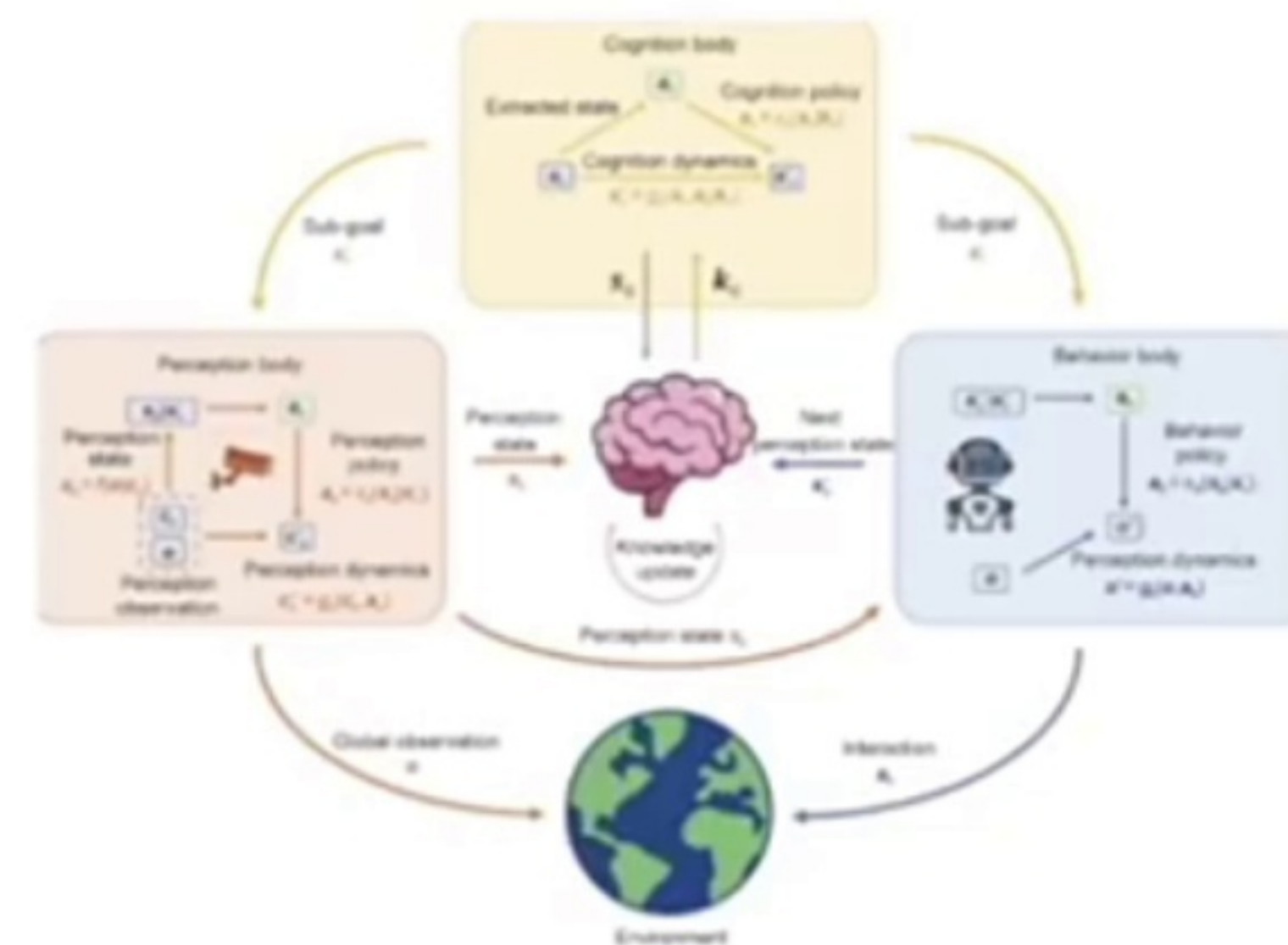
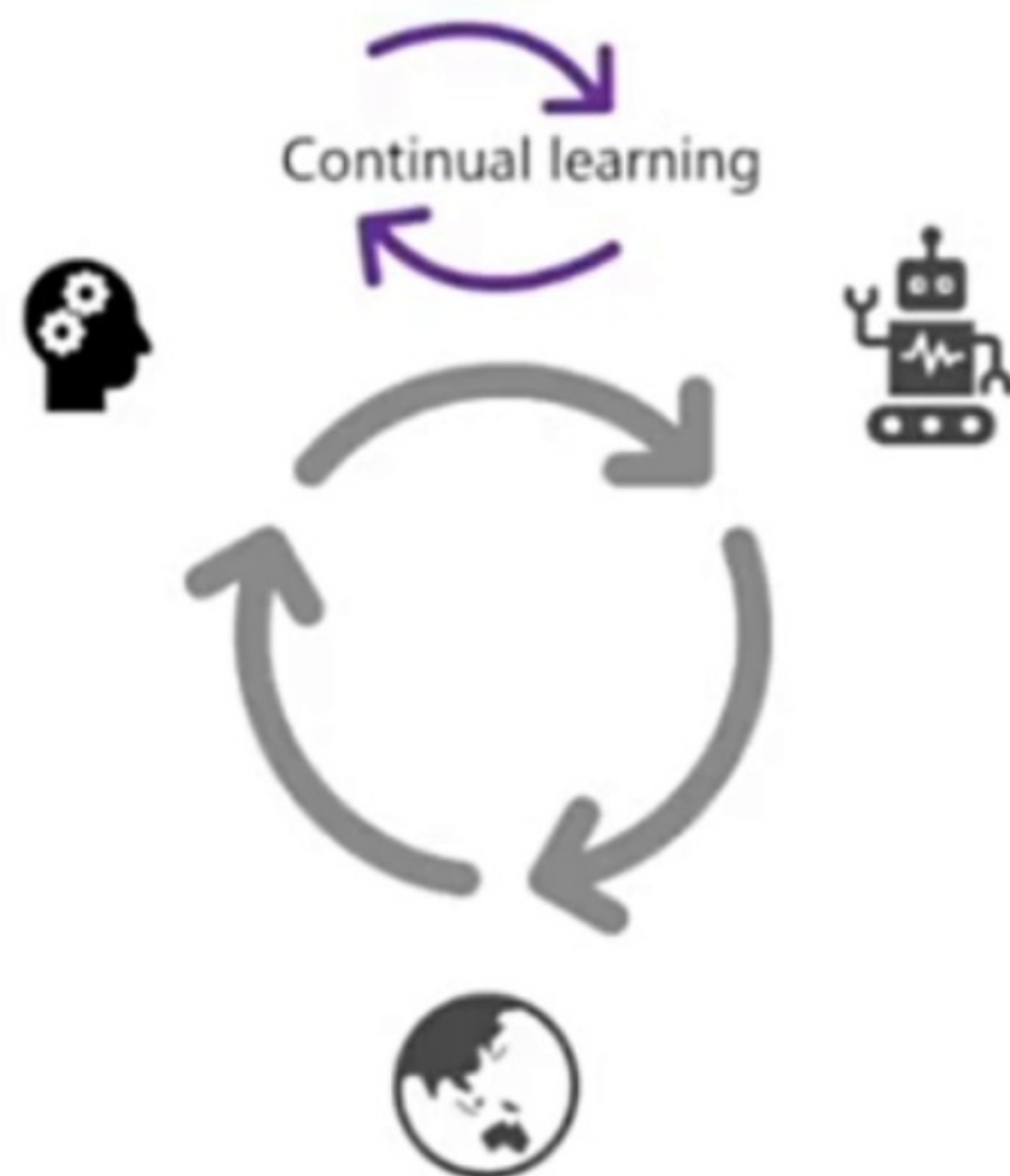
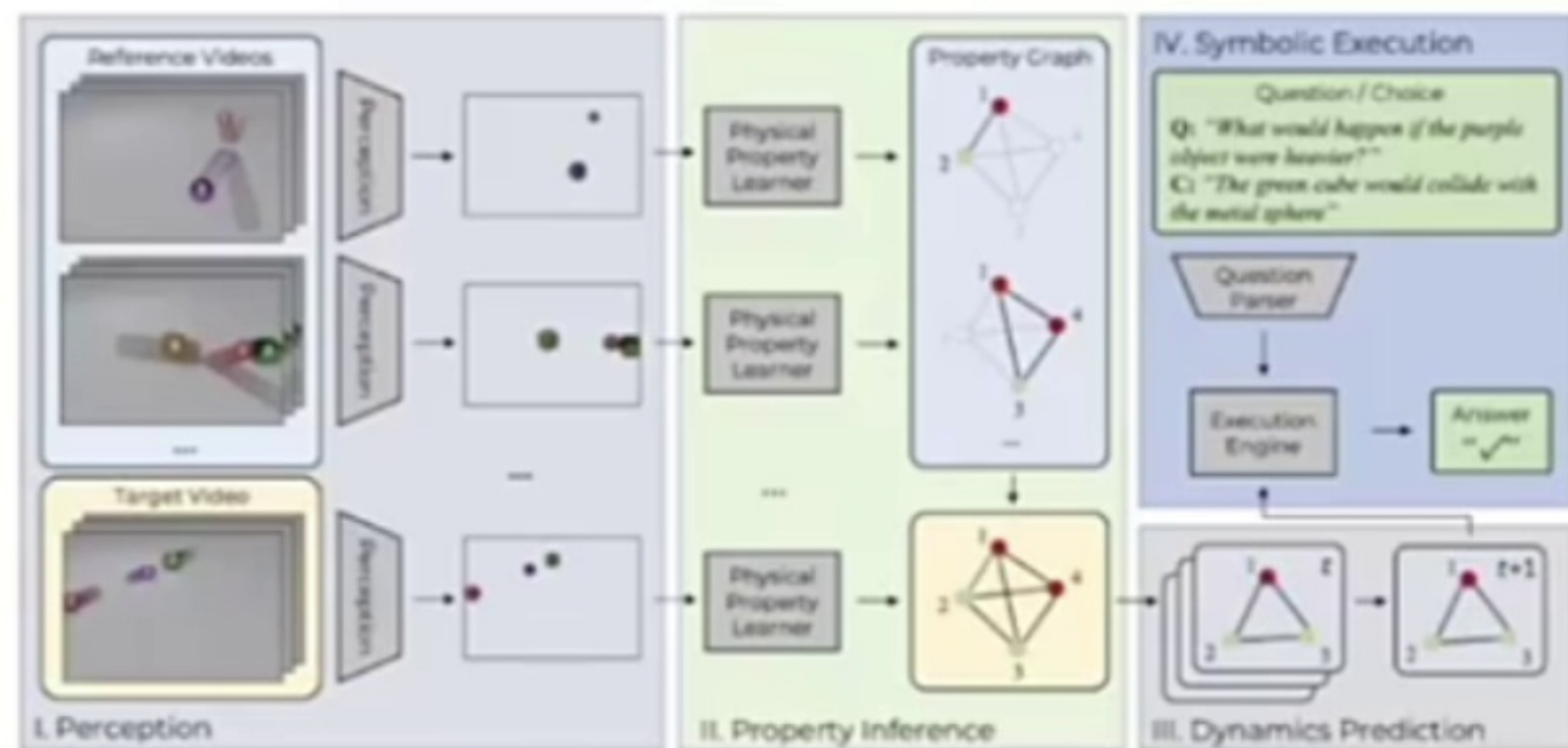


二、数字物理世界模型

时空联合推理与持续学习

从三维感知到四维感知，从单纯几何信息到时空联合的物理信息，从空间推理到物理推理

空间智能体在交互和主动感知过程中持续学习，弥补空间信息的不足



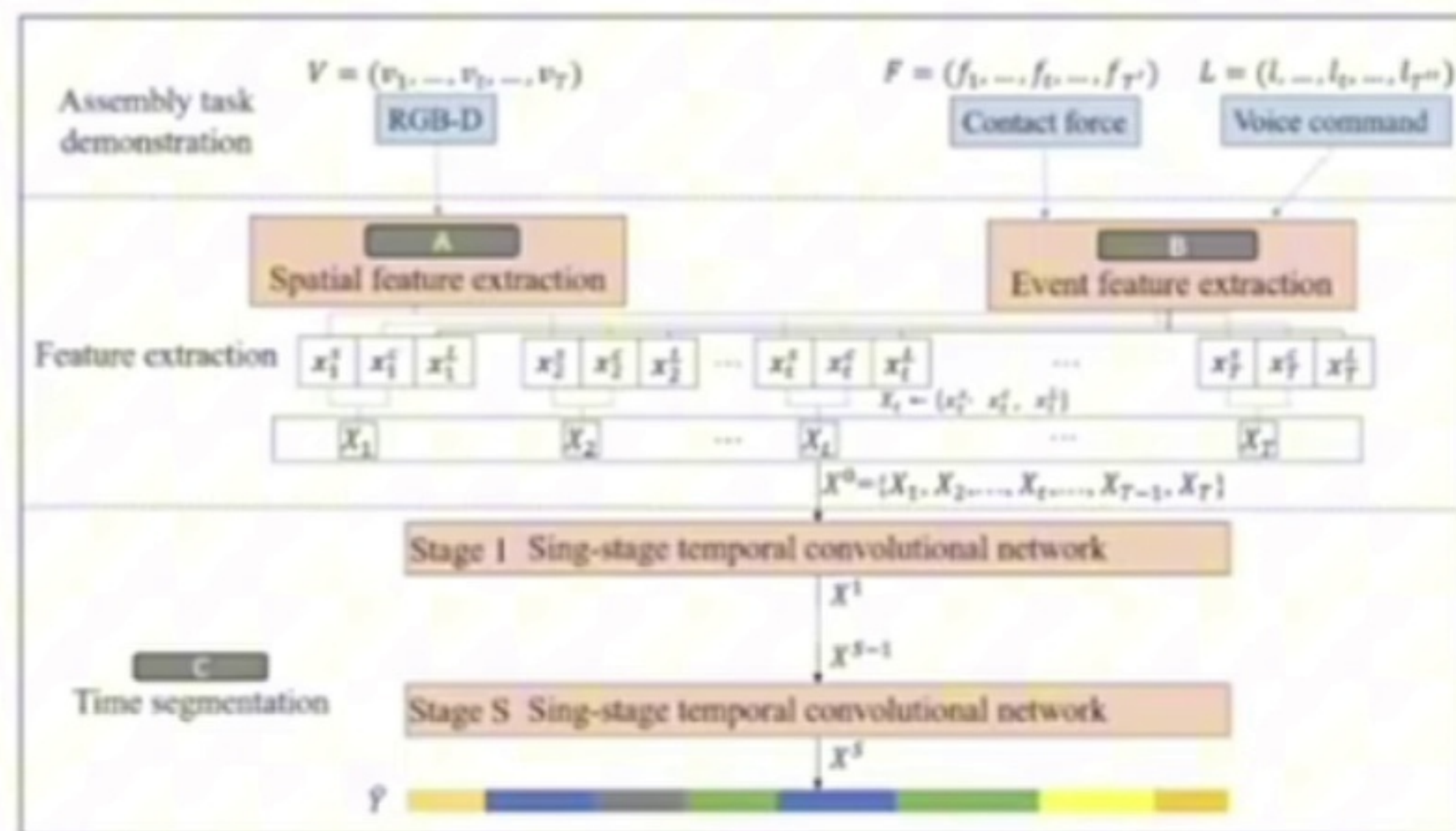
二、数字物理世界模型



更强的感知能力和推理能力



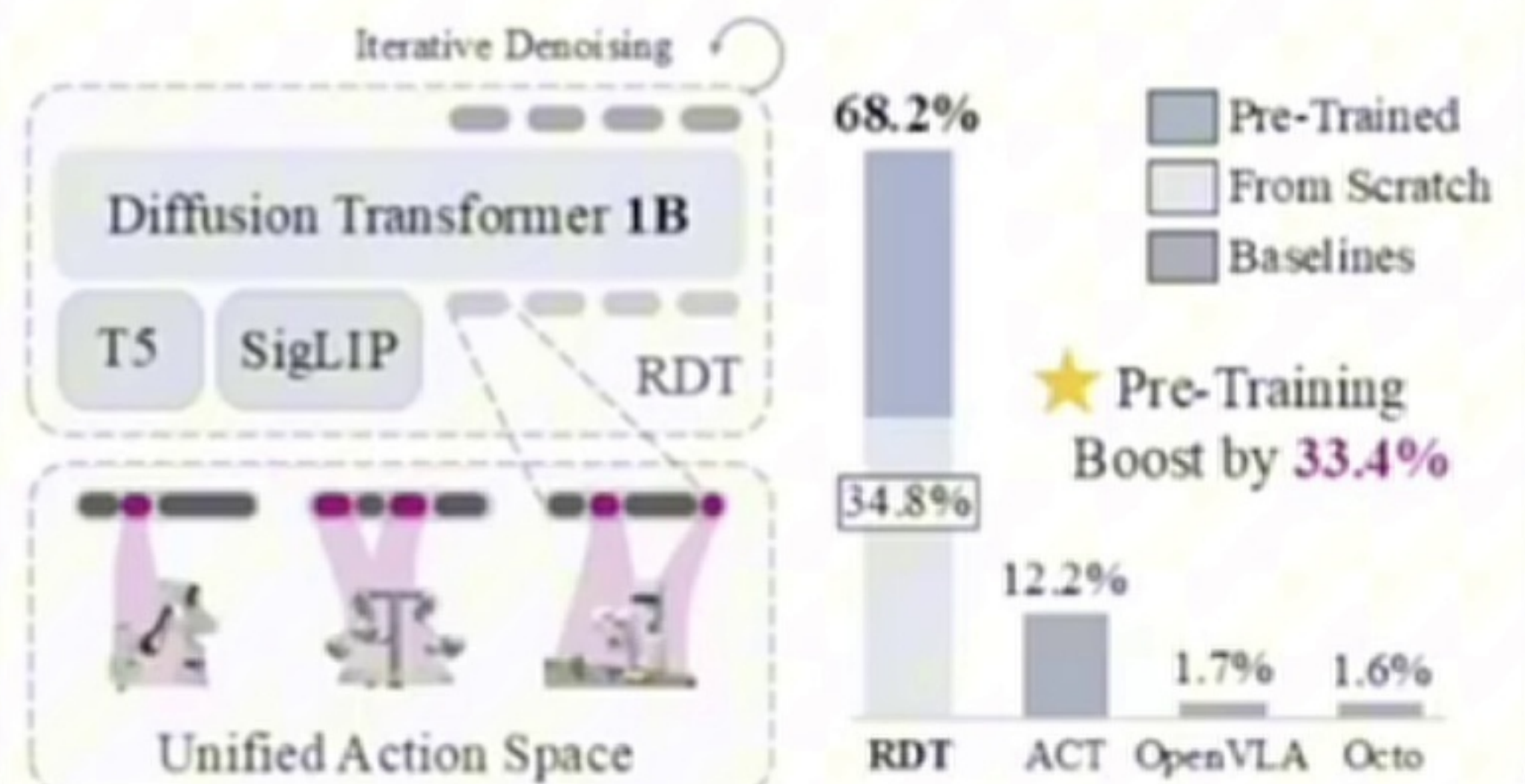
从被动感知到主动感知，
从单一模态到多模态，
从仅依赖模型到融合知识



表达能力更强的Policy head,参数集中在策略头:
——从感知大模型到决策大模型

RDT-1B

Robotics Diffusion Transformer as Language-Visuomotor Policy



孙富春, 刘乃军等 "Digital-twin-assisted Skill Learning for 3C Assembly Tasks" *IEEE Transactions on Cybernetics*, 2024, 54(7): 3852-3863.

Liu, S., Wu, L., Li, B., Tan, H., Chen, H., Wang, Z., Xu, K., Su, H., & Zhu, J. (2024). RDT-1B: A Diffusion Foundation Model for Bimanual Manipulation. *arXiv preprint arXiv:2410.07864*.

三、知识的迁移与增强



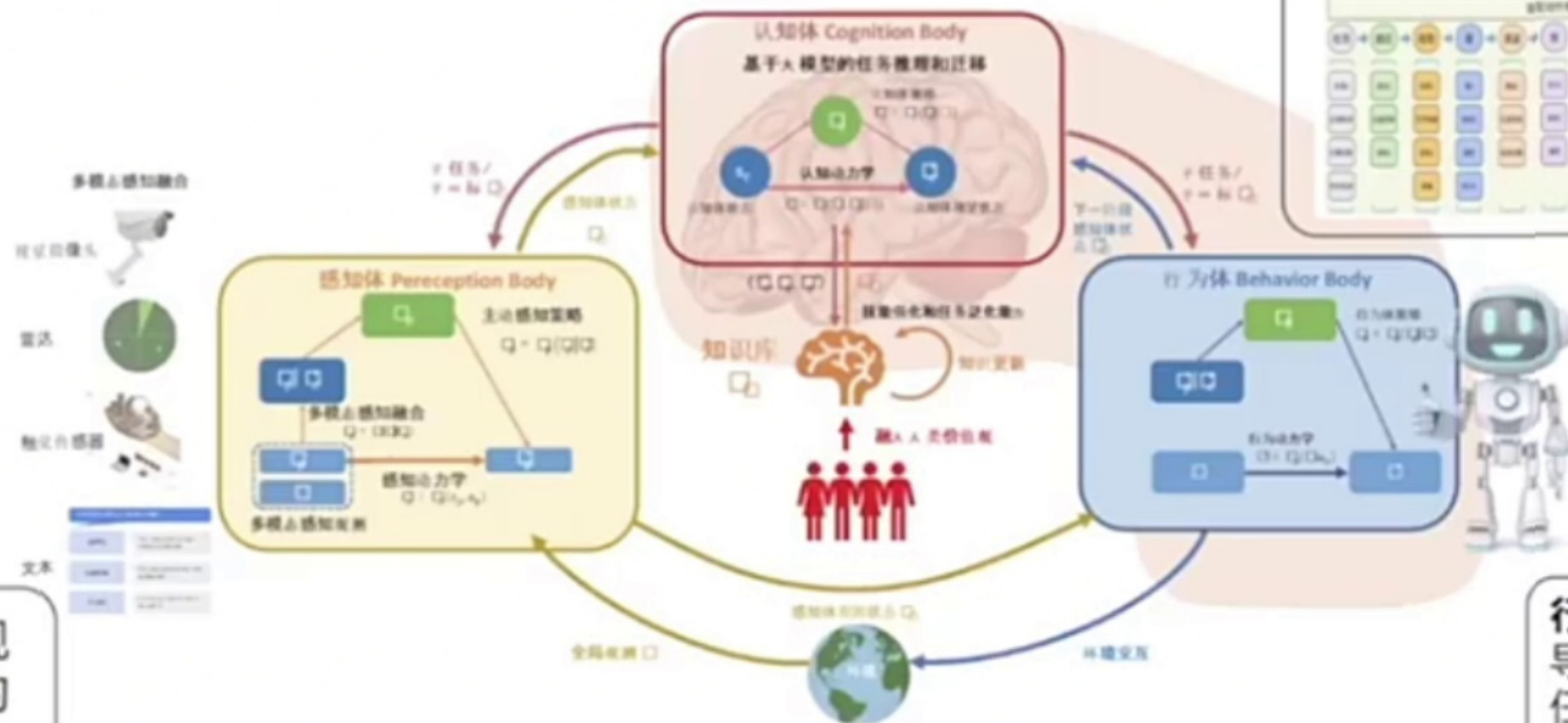
2021 ICRA大会报告

Behavior AI: Robot Cognitive Learning from Agent to Bcent

感知体：通过认知体的规划引导，采用主动感知的方式，获取多模态感知信息并解析



具身智能体Bcent范式



认知体：构建知识库，技能规划器和残差更新。通过残差更新知识库，实现知识的更新与生长，完成新任务下操作技能的规划



行为体：通过认知体的规划引导，实现小样本/零样本下多任务的操作技能迁移与泛化



具备**自上而下、自下而上**的交互，实现通用智能的**生长、更新和发育**

孙富春, 黄文炳, 刘华平等, "Robot cognitive learning by considering physical properties," *Engineering*, 47(2025):168-179

Nature 发现世界 (英), Science 理解世界 (美), Engineering 改变世界 (中)



迁移学习的误差界理论

$\mathcal{H}$ 为 VC-dim= d 的假设空间, $\mathcal{U}_S$ 和 $\mathcal{U}_T$ 为从 $\mathcal{D}_S$ 和 $\mathcal{D}_T$ 采样得到的 m' 个 unlabeled 的实例. S 是从 $\mathcal{D}_S$ 采样得到的 $(1 - \beta)m$ 个 labeled 的实例和 $\mathcal{D}_T$ 采样得到的 βm 个 labeled 的实例的集合, 其中使用分别的 ground truth functions f_S, f_T 进行 labeling. 如果 $\hat{h} = \arg \min_{h \in \mathcal{H}} \hat{\epsilon}_\alpha$, 其中训练集为 S , $h_T^* = \arg \min_{h \in \mathcal{H}} \epsilon_T$, 那么, 对于任何的 $\delta \in (0, 1)$, 下式至少有 $1 - \delta$ 的概率成立:

$$\epsilon_T(\hat{h}) \leq \underbrace{\epsilon_T(h_T^*)}_{\text{最优策略}} + 4 \underbrace{\sqrt{\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}}}_{\text{样本数量分布}} \underbrace{\sqrt{\frac{2d \log(2(m+1)) + 2 \log \frac{8}{\delta}}{m}}}_{\text{样本数量}} + 2(1-\alpha) \left(\underbrace{\frac{1}{2} \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_S, \mathcal{U}_T)}_{\text{无标签样本分布的 } H\Delta H\text{-Divergence}} + 4 \underbrace{\sqrt{\frac{2d \log(2m') + 2 \log \frac{8}{\delta}}{m'}}}_{\text{样本数量}} + \lambda' \right) \underbrace{\lambda'}_{\text{联合错误率}}$$

最优策略

样本数量分布

样本数量

无标签样本分布的
 $H\Delta H$ -Divergence

样本数量

联合错误率



增加仿真环境带标签样本数



减少样本的虚实误差

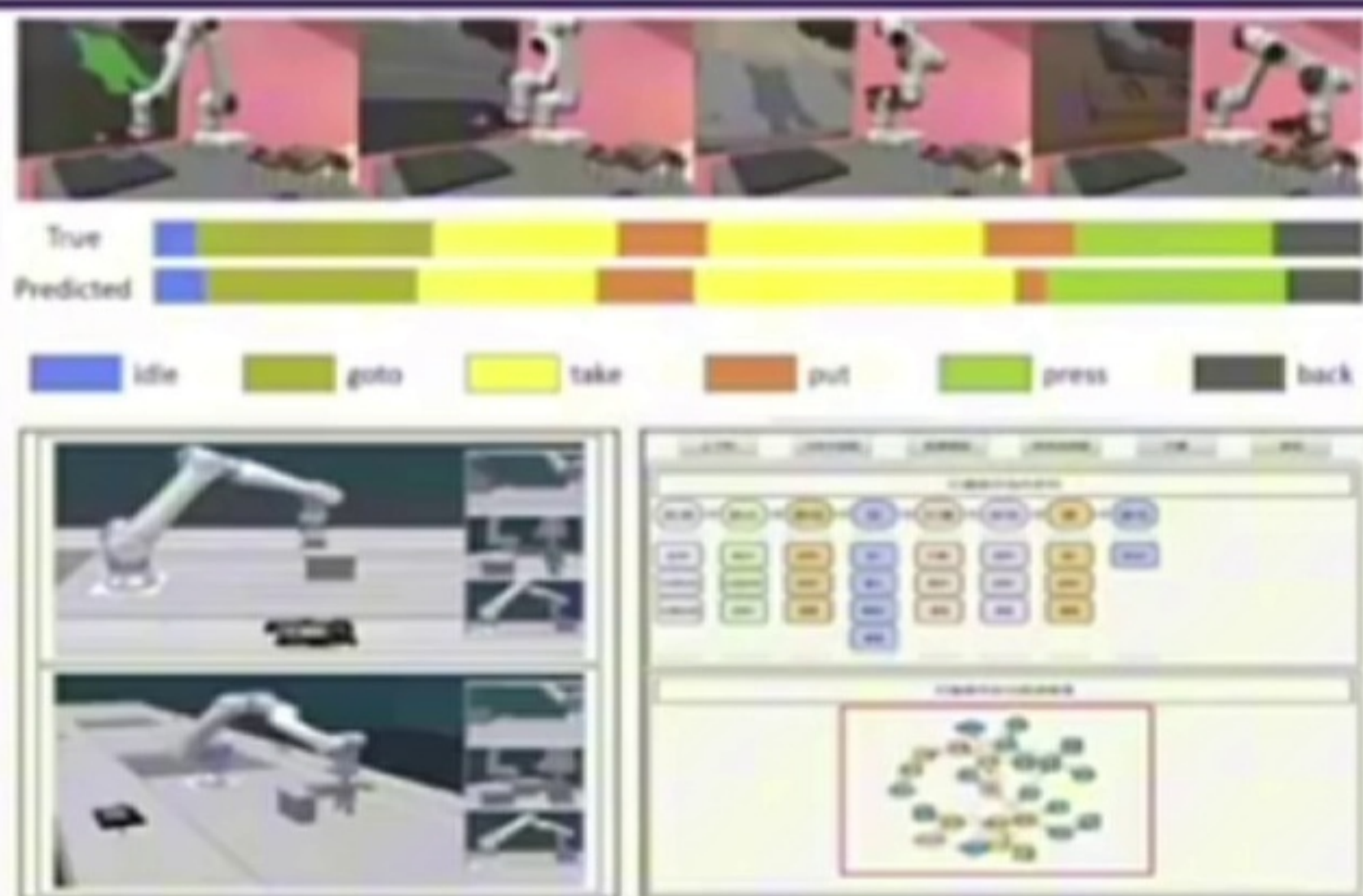
$$LPIPS(X, \hat{X}) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \|w_l \odot (X_{hw}^l - \hat{X}_{hw}^l)\|_2^2$$

$$MSE(X, \hat{X}) = \frac{1}{HW} \sum_{h,w} (X_{hw} - \hat{X}_{hw})^2$$

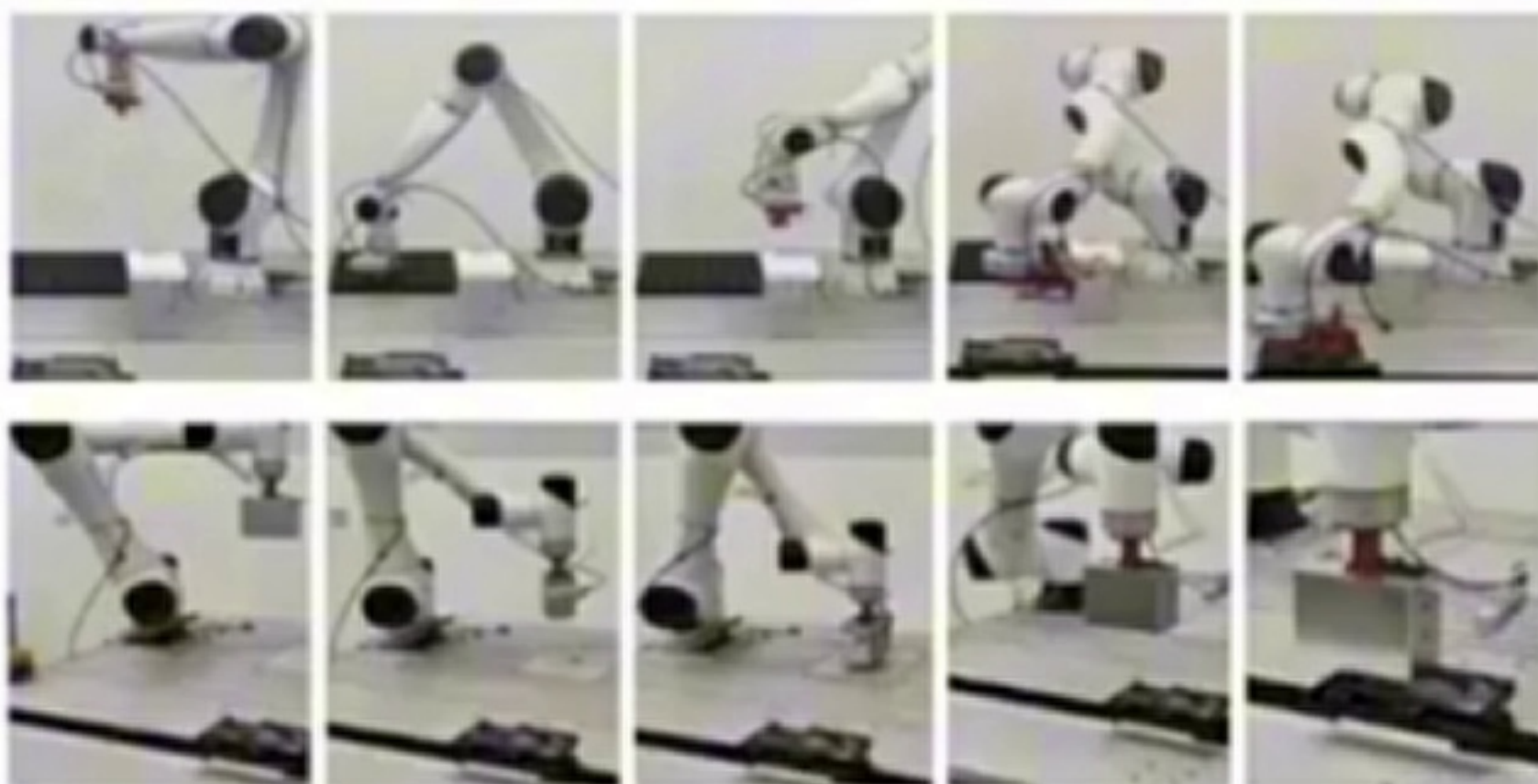
机器人虚实迁移学习



虚实同步的物理数字系统构建与技能迁移

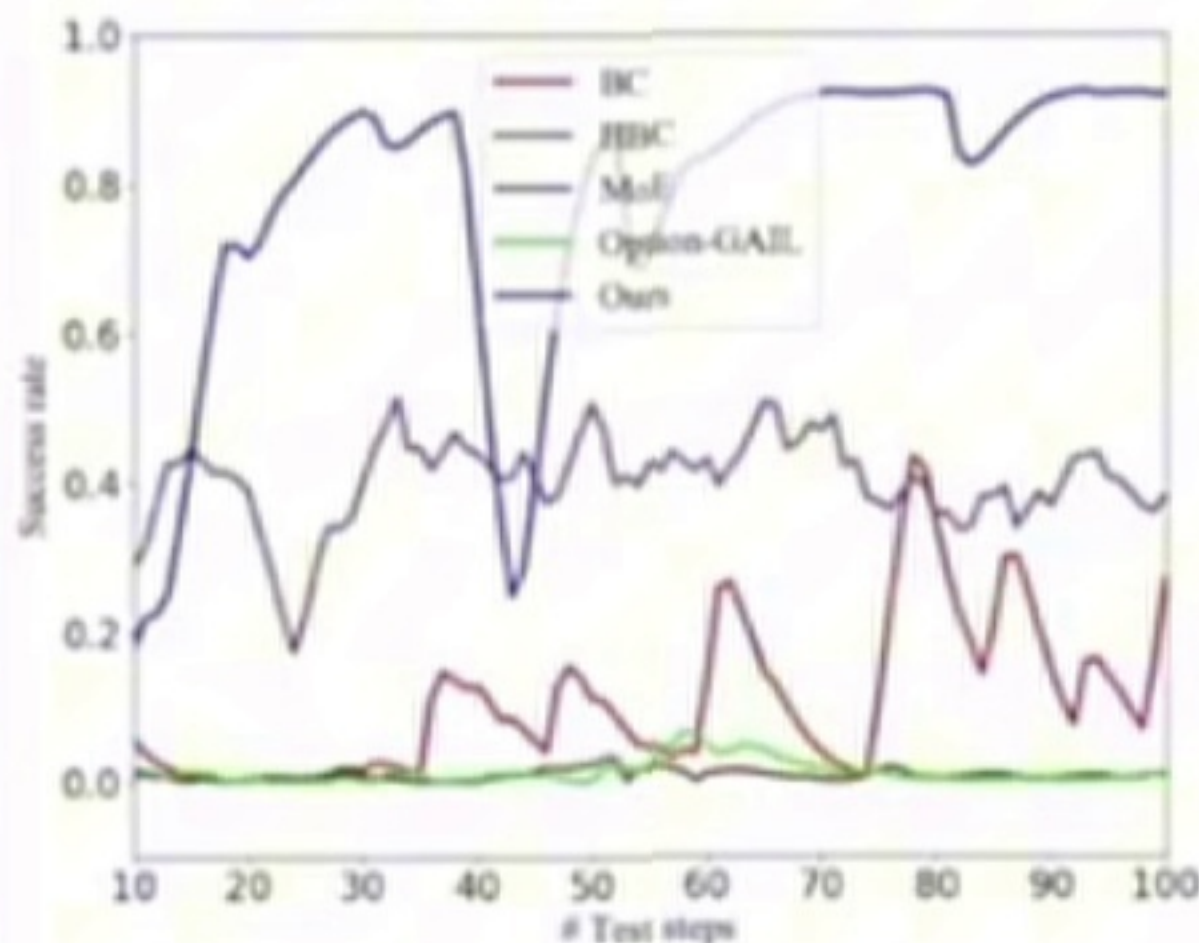


技能解析与知识库构建

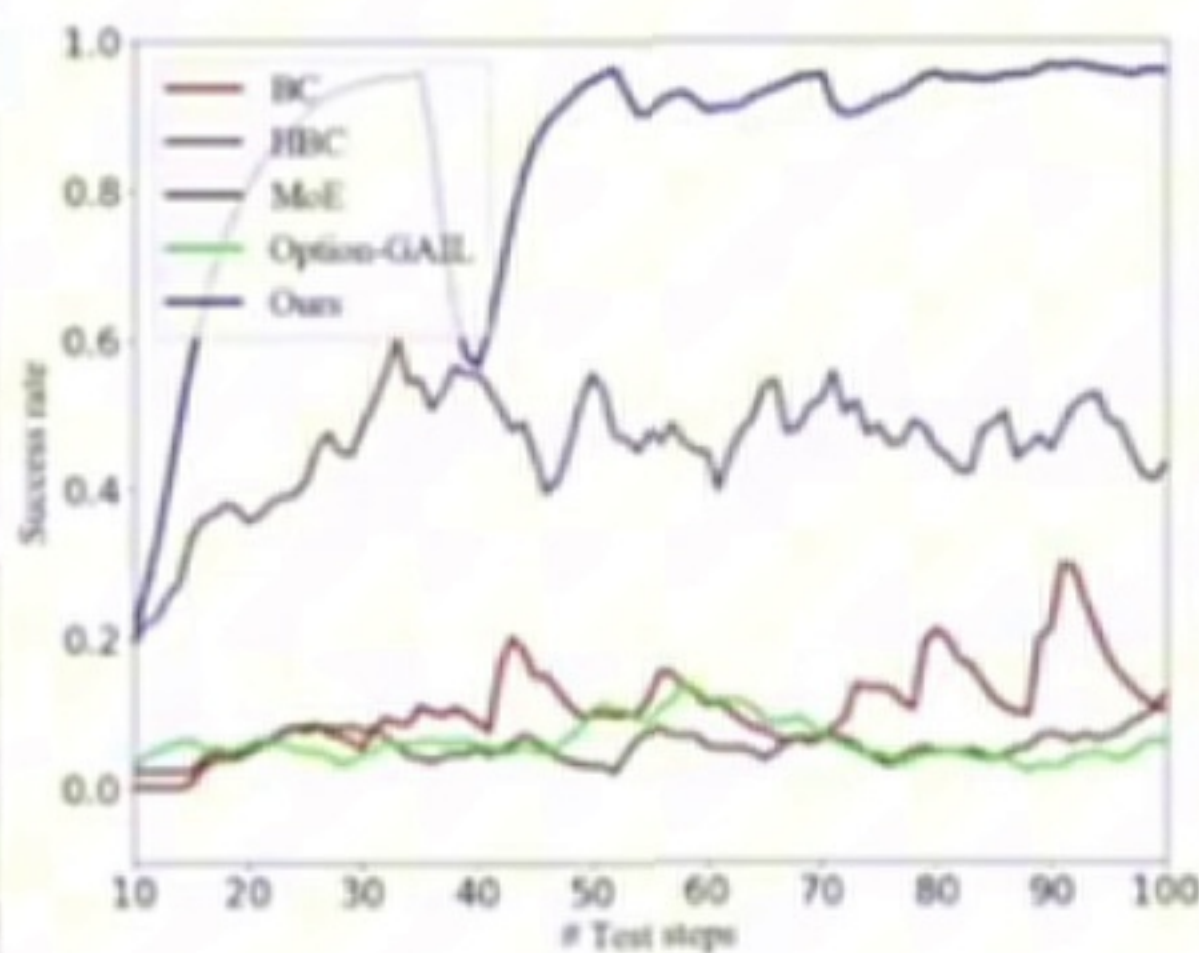


软排线与手机摄像头装配任务

学习性能曲线比较

F
P
C

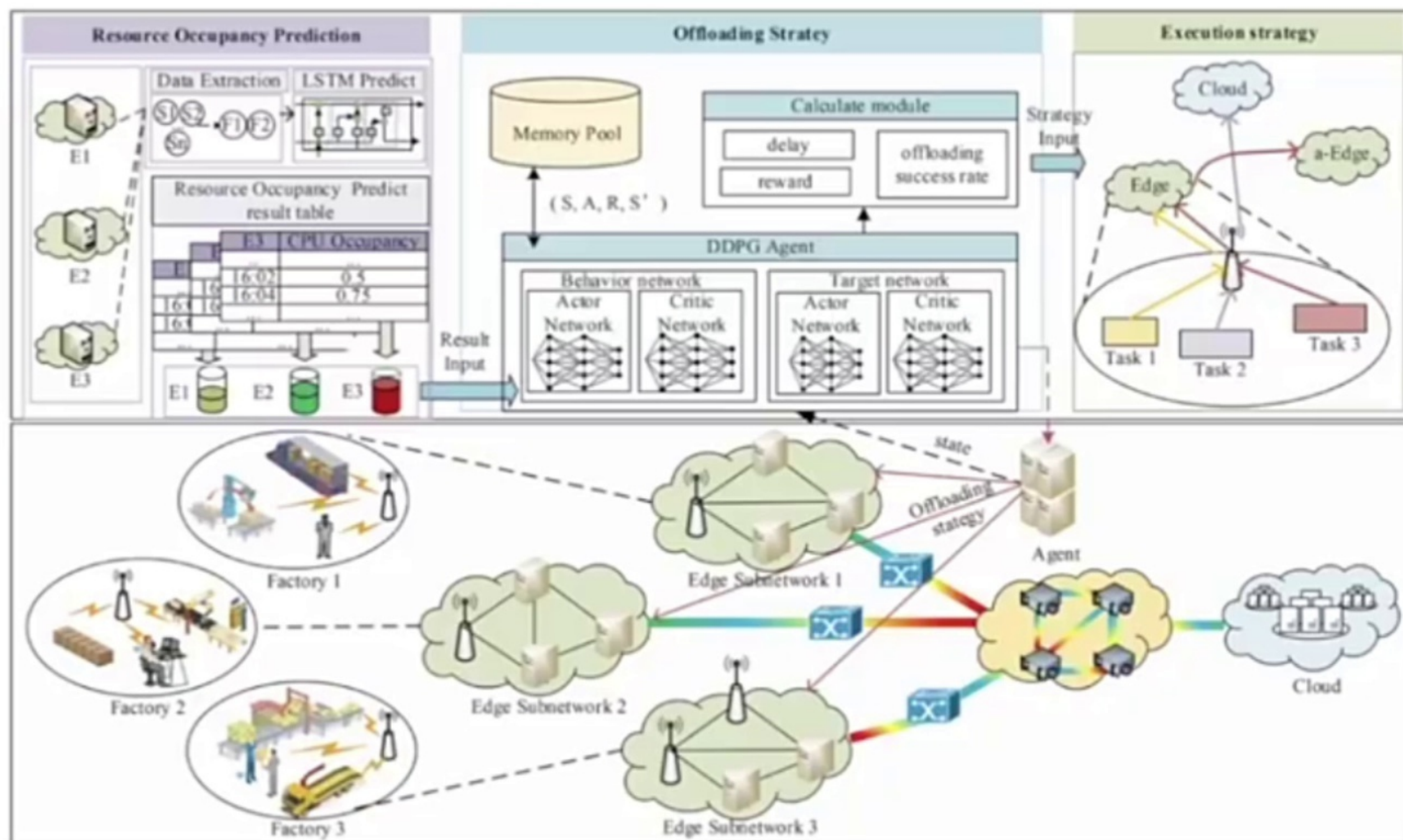
手机摄像头





四、智能制造行业的应用

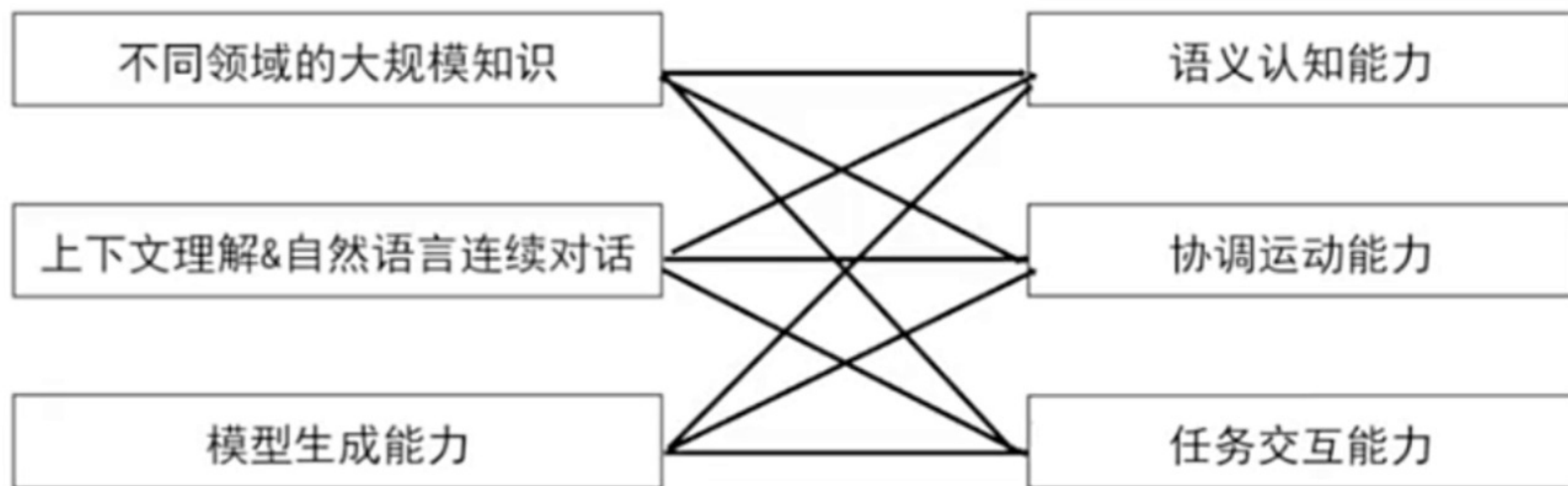
云边协同多任务学习框架



- 利用云边协同实现工业互
联的大规模数据资源共享;
- 大规模数据采集与共享;
- 多任务集群的规模化训练
与验证;
- 跨场景、跨智能体、跨任
务的通用操作验证



多模态大模型与具身智能的结合



体现：感知与交互

感知：通过多类型传感器的配合，机器人处理多个模态接收的信息，从而实现对文本、图像、视频、音频等模态信息的学习和理解。

交互：同时使用语音、手势、文本等与机器人交互，可以减轻用户认知负担（使交互更自然）、消除任务歧义（使交互更准确）并降低环境干扰（使交互更加鲁棒）。

具身大模型在通用操作中的作用

具身智能系统

感知层 (感官)

感知大模型:

负责将各种传感器采集信息（如视觉、听觉、触觉信息等）融合表征为高层语义表示，从而为后续决策和动作执行提供基础信息的大模型。例如DINOv2, CLIP, SigLIP, 3D-MVP, Voltron等。

认知层 (大脑)

推理、规划大模型: 基于感知层信息，分析场景环境和任务要求

1. 将高层次的任务指令分解为具体的原子动作任务步骤。
2. 分析当前环境状态的反馈，对后续任务进行修改，实现动态决策。

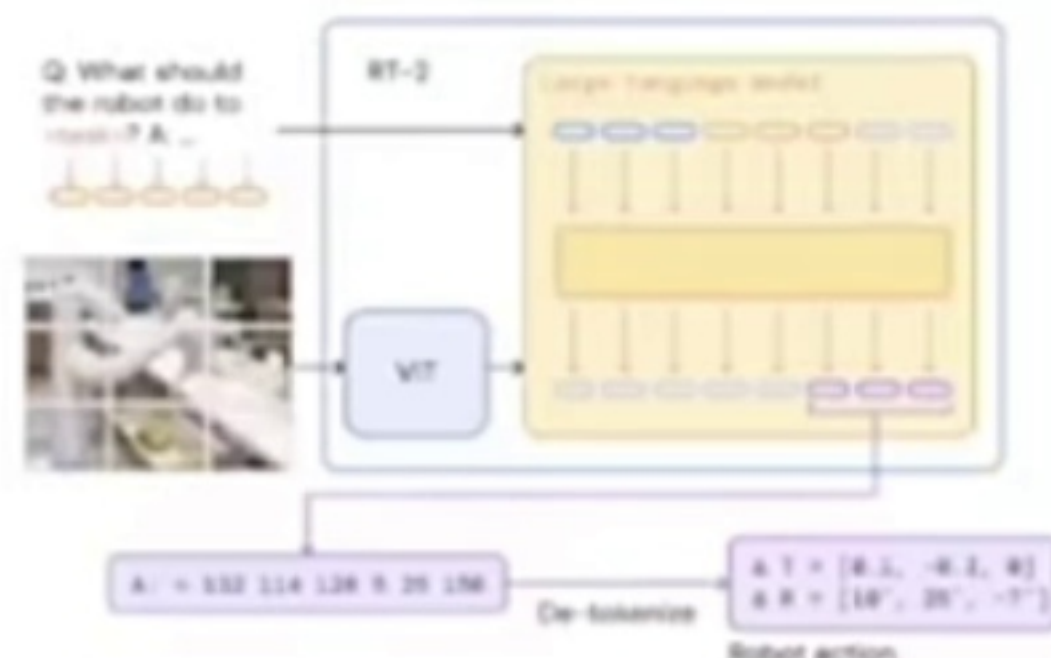
行为层 (小脑)

动作大模型:

控制逻辑，执行具体动作。例如：Diffuion Policy, Aloha ACT, 以及RT1, RT2等

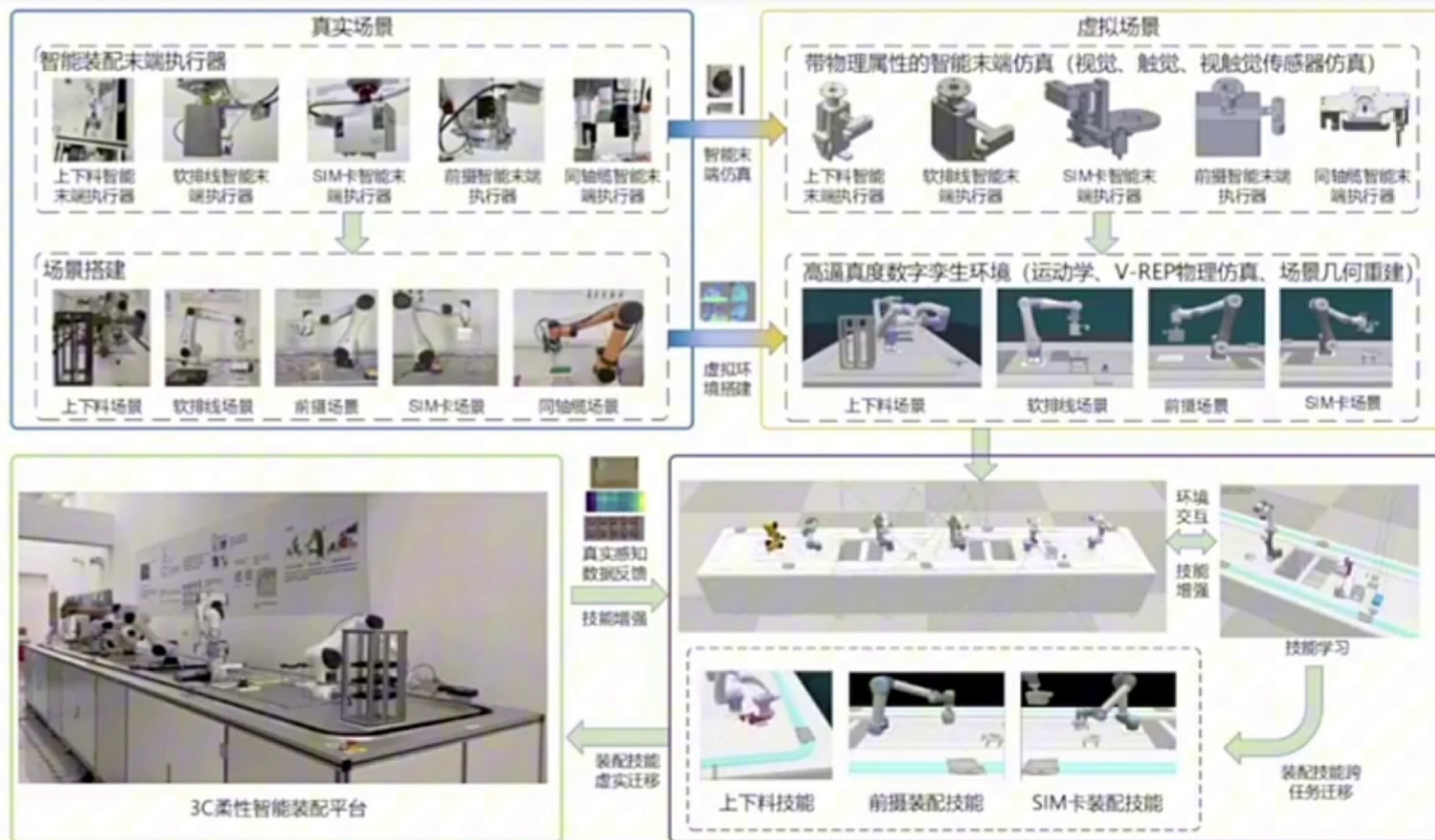
端到端具身大模型

集成感知、决策与执行，实现自主、智能的具身行为，能够直接从感知输入到动作输出完成复杂任务。





四、智能制造行业的应用



面向3C装配的上下料、软排线、前摄、SIM卡和同轴镜装配场景，我们构建了高逼真度的数字孪生环境。通过在数字孪生环境中使用基于知识的技能学习和基于知识蒸馏的技能增强，学习各个场景装配技能。

四、智能制造行业的应用

立讯苹果耳机产线基础及智能产线规划

上下料



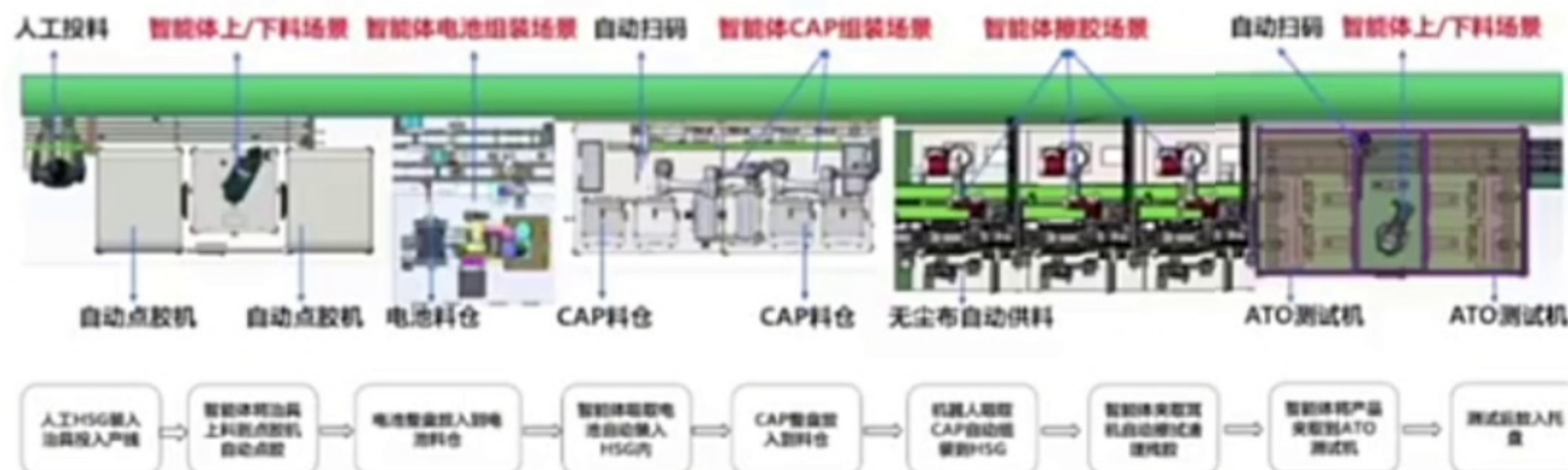
装配



擦胶

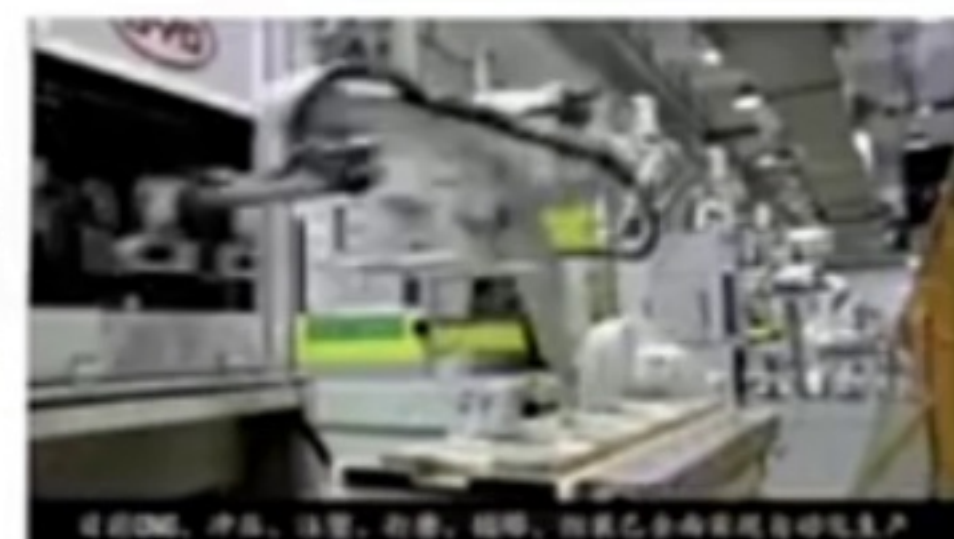


耳机智能产线应用推广



比亚迪华为手机产线基础及智能产线规划

上下料



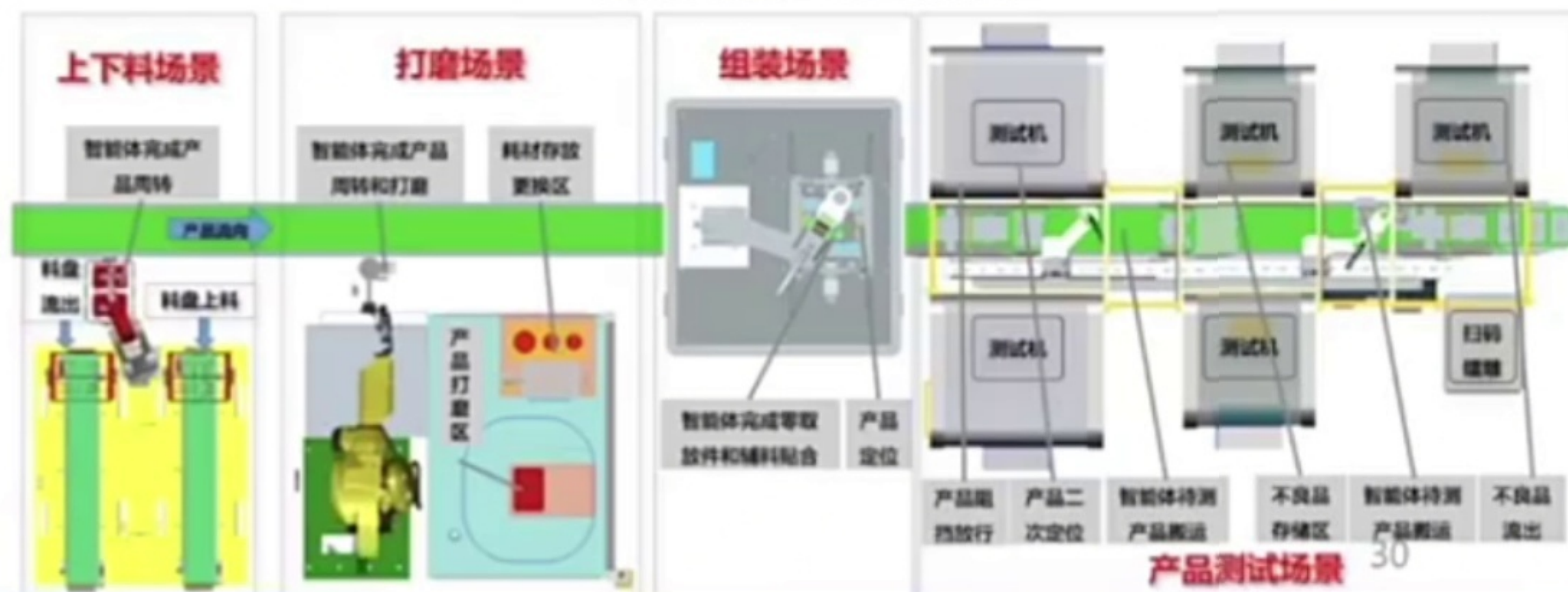
装配



打磨



手机智能产线应用推广





四、智能制造行业的应用





具身大模型在通用操作领域的挑战

应用层次

优势

瓶颈

感知层

数据优势：经过海量数据训练，天然具有优秀多模态语义感知和对齐能力

感知大模型参数规模庞大，主要针对图文两种模态。对比专业感知小模型，没有优势。

认知层

语言智能性：大模型自身具有强大推理分析能力，泛化性强。

推理规划输出如何有效转为具体的行为控制策略还需要针对性设计。

行为层

基于VLA实现**端到端**的具身智能。任务泛化性强。

缺乏充足控制轨迹数据。
机器本体多样，设置多样，难泛化。
推理速度慢。

多模态大模型具备同时处理图文输入能力，因此，基于多模态大模型可以实现感知与认知的统一。

机器动作位姿(action) 看作一种新模态，需要数据

关键问题

1. 当前大模型感知模态信息有限，缺乏对环境的全面建模。
2. 当前大模型缺乏对现实世界物理规律的认识，难以准确认知与精准行为。

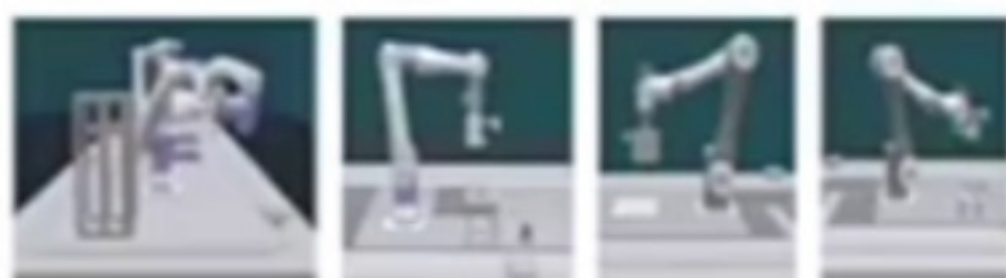
未来方向：复杂操作技能的高效主动思考与动态决策

- 人类大脑中对物理世界的建模，在执行某一操作前能够提前预测该动作的后果。
- 物理数字系统作为世界模型对物理世界进行实时建模，提供行为决策的预测模拟。

构建数字物理系统



仿真编程一体化平台



虚实多技能迁移

任务初步规划



思考尝试



虚拟世界感知交互



虚拟环境反馈



决策执行

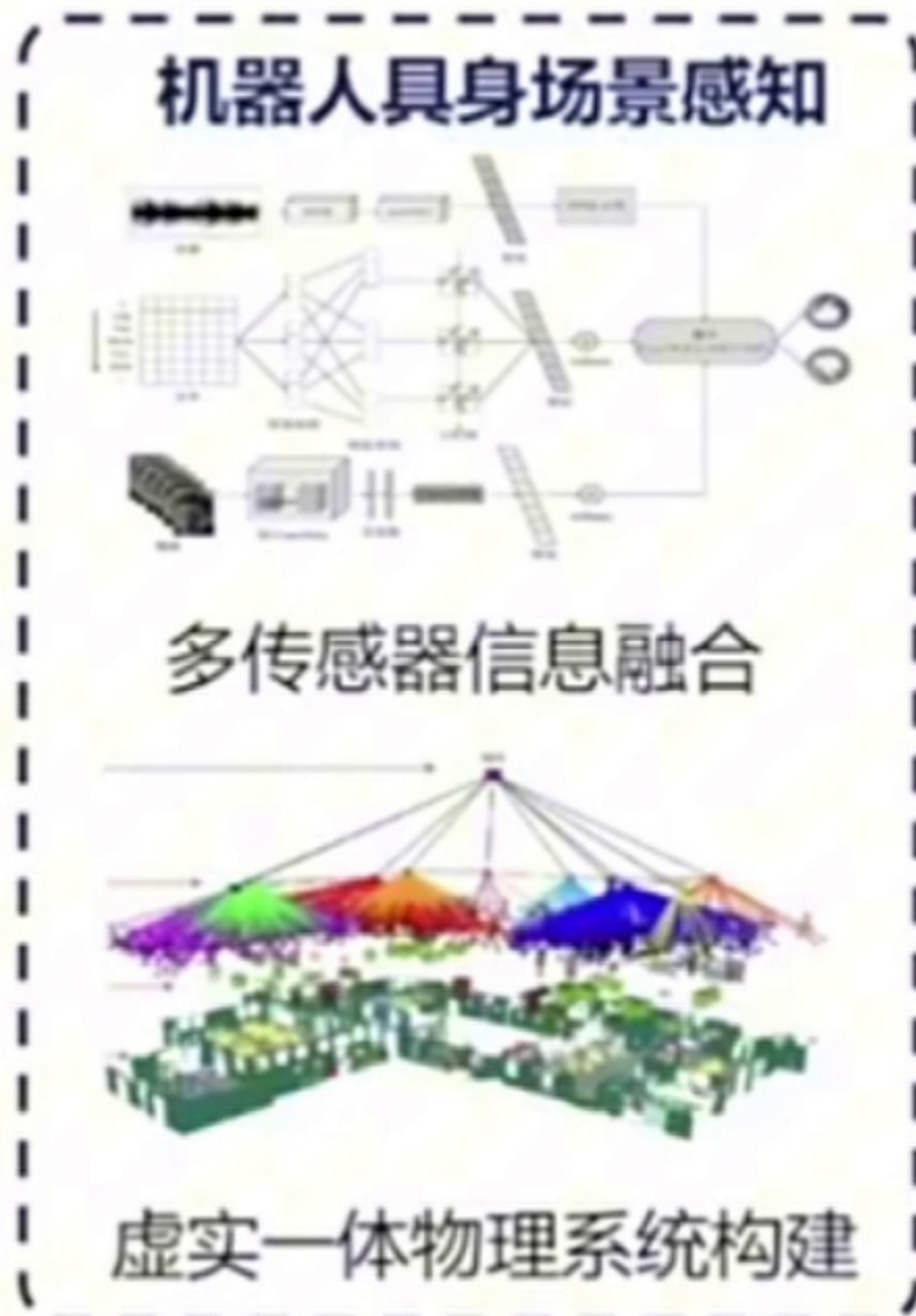


具身物理场景交互

五、未来展望

未来方向：泛在场景具身数据采集与预处理

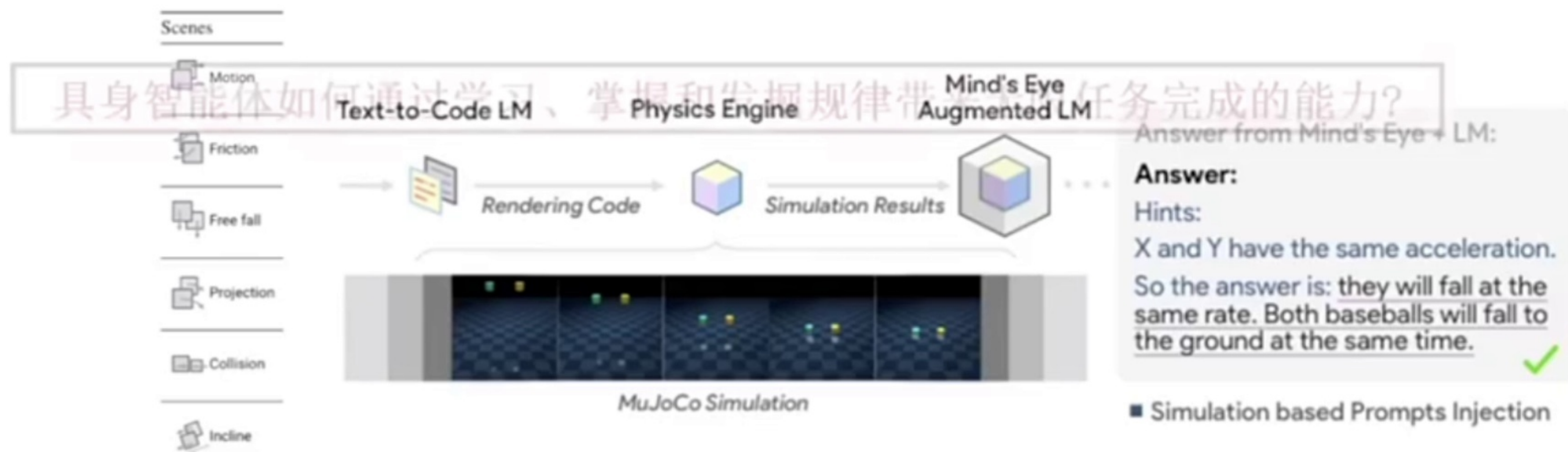
□ 研究机器人各类环境信息融合的**信息感知与多模态数据采集**，构建以数字物理系统为基础、全域数据为支撑的机器人**数据采集与预处理系统**，结合机器人实时场景反馈不断更新数据集。





未来方向：具身智能体如何通过学习、掌握和发掘规律

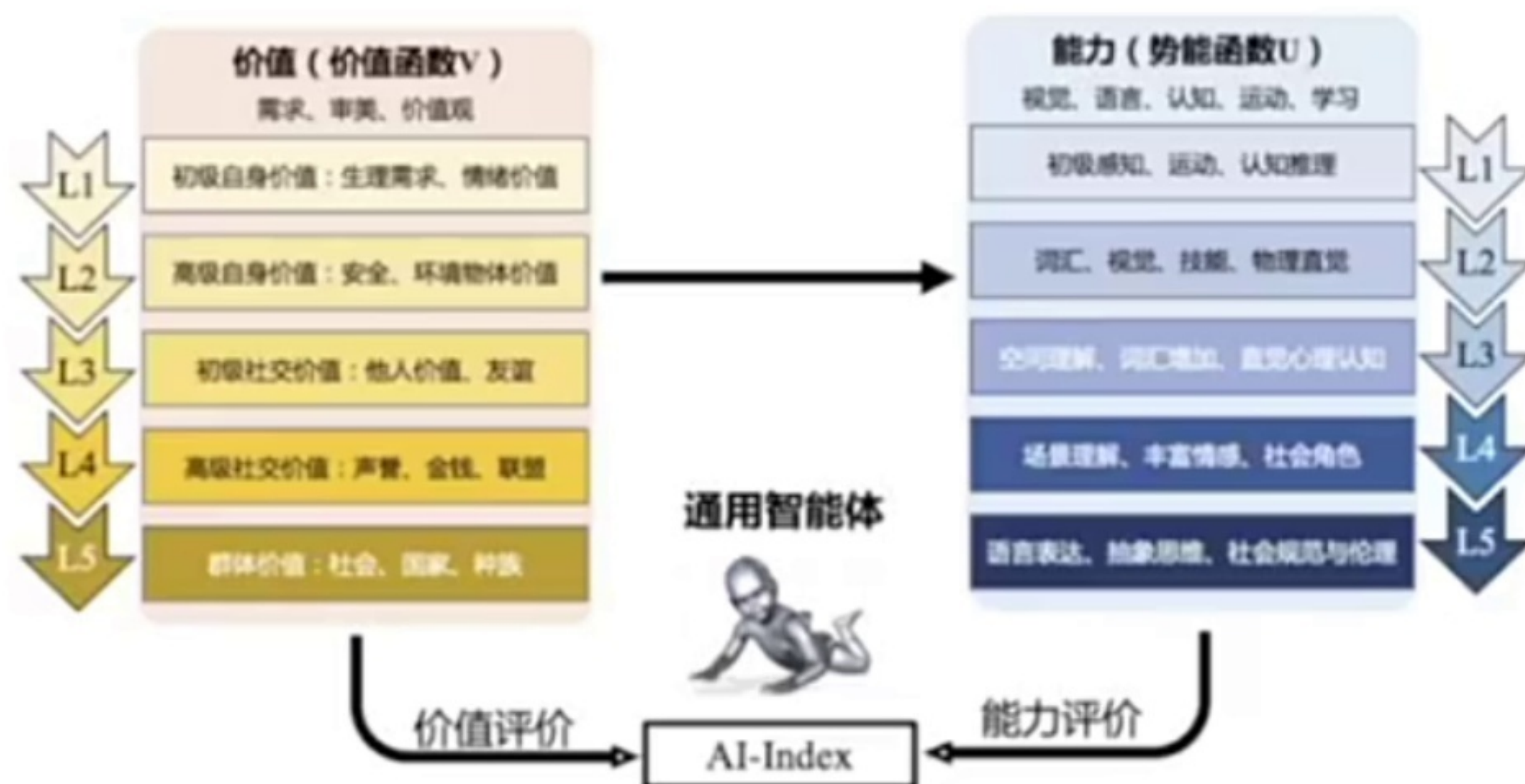
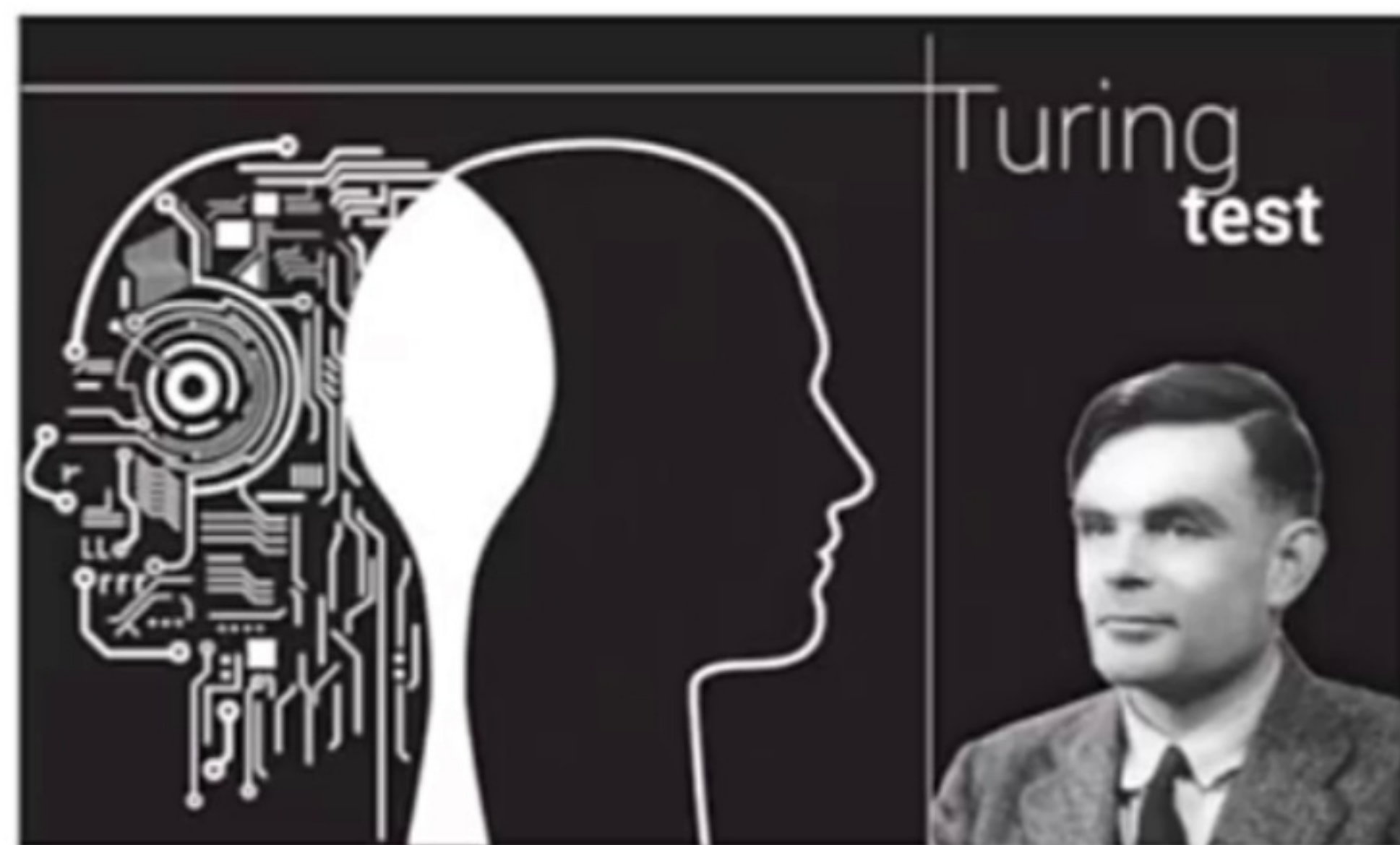
- 智能的第一个前提：存在规律。如果宇宙是没有规律、完全随机的，那么智能不会影响结果，也就无法做出有意义的行为。通用智能体发挥作用时利用了若干种规律，其中的规律包括各种物理常识、因果逻辑关系，是通用智能体能够有效达成目的的关键要素。





未来方向：如何设计和开发评价体系和测试系统来检验通用智能性？

- 智能的第二个前提：存在目的。如果没有目的，那么任何行为都无法被认为是或不是智能行为。目的不一定是智能自身意识到或者设定的，也可能是评估智能时假设的。为了衡量通用智能体的通用智能性，就需要设计和开发评价体系和测试系统，针对不同的目的（任务、场景）对通用智能体进行评价。



五、未来展望

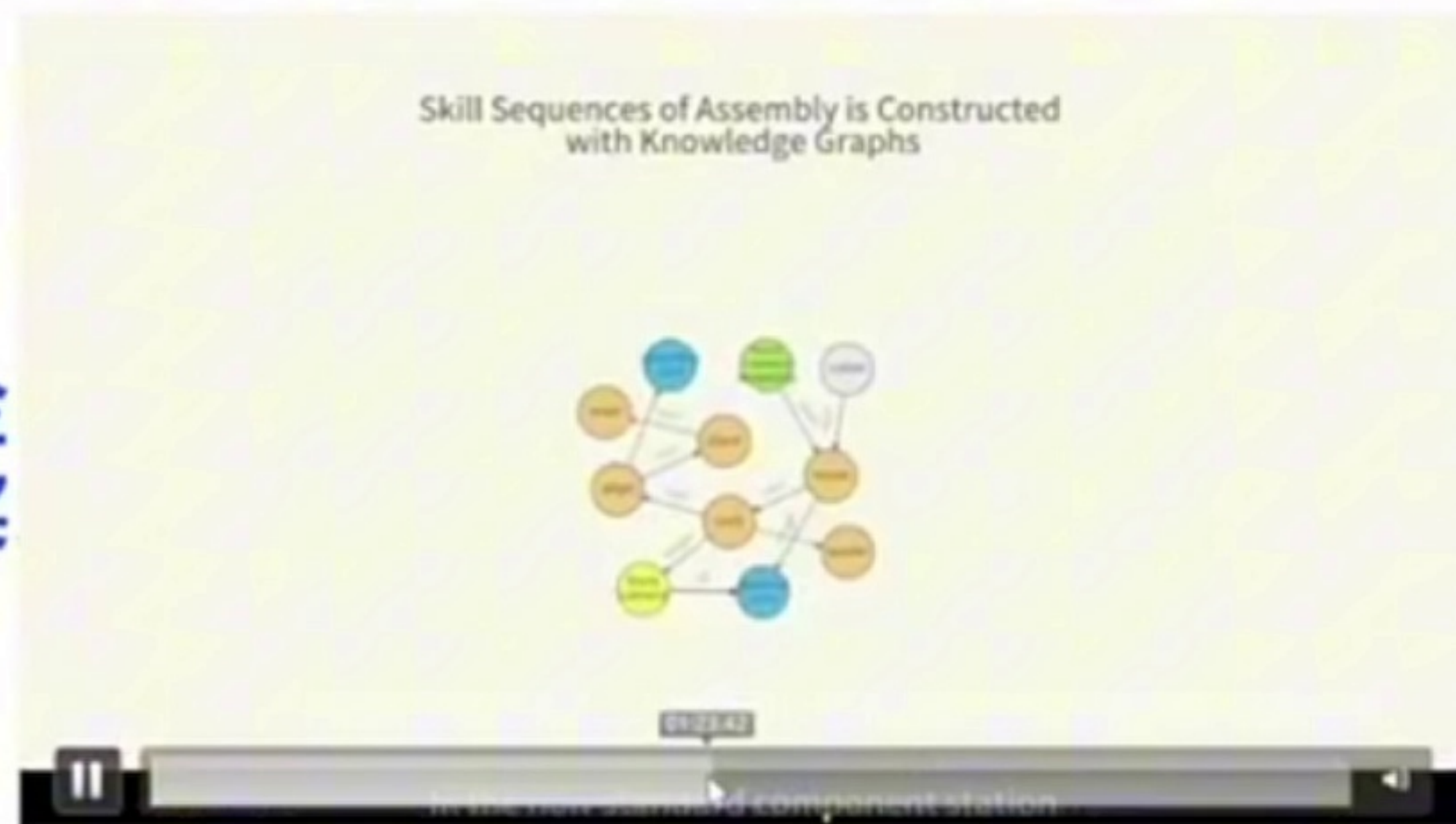
ICRA2024 抓取与操作比赛 装配(MT)赛道冠军

ICRA 2024机器人
抓取与操作比赛
MANUFACTURING TRACK

Task Analysis & Coding



第49届日内瓦国际发明展
金奖



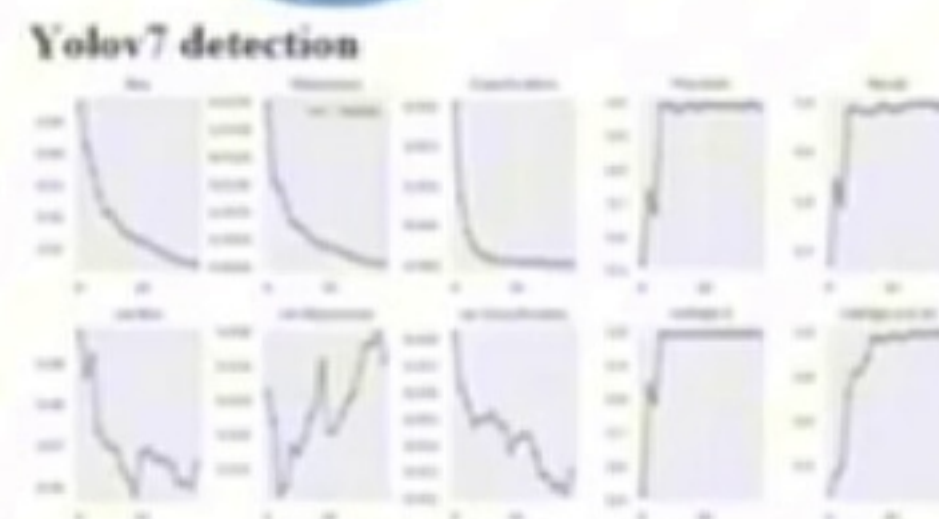
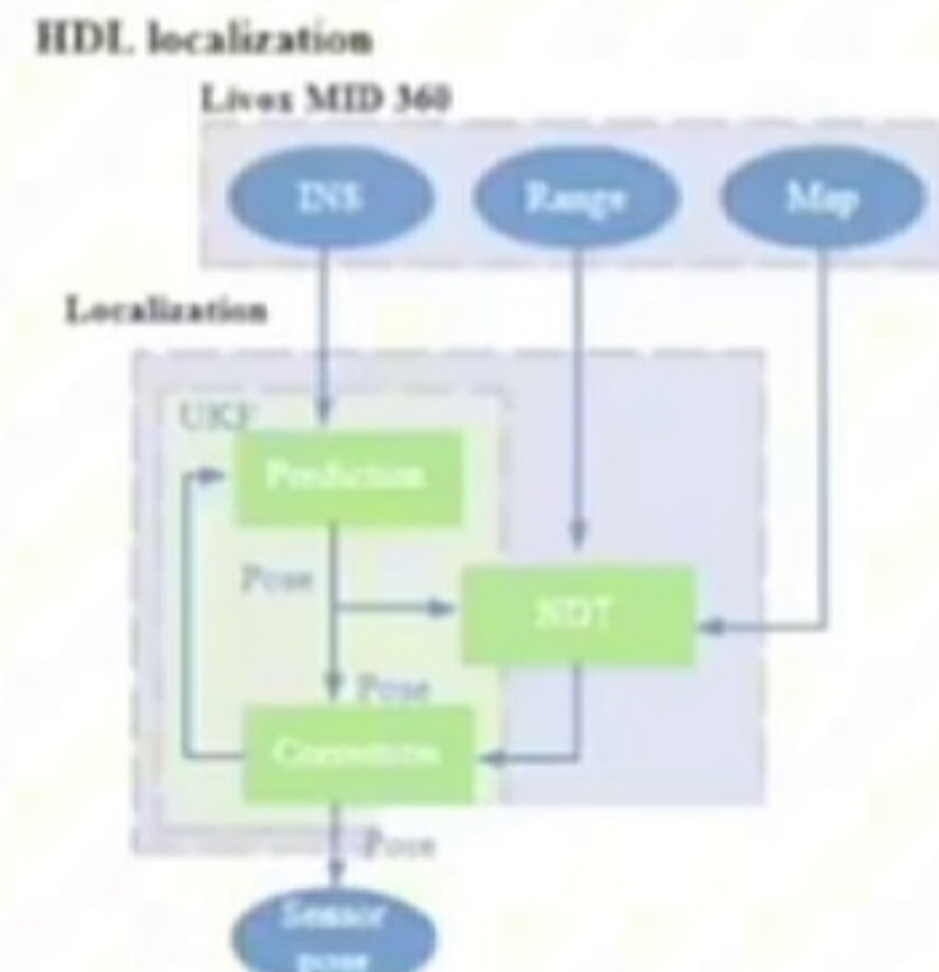
Knowledge-DrivenSim2real Transfer 3C intelligent Assembly System

ICRA2024 Sim2Real挑战赛 冠军

Robotic Sim2Real
Challenge



Method Baseline





第二代

五指

灵巧手







五、未来展望





五、未来展望





五、未来展望



2024年清华大学计算机系孙富春教授 团队获得深圳智能机器人灵巧手大赛冠军



ICRA2025 Sim2Real挑战赛冠军

经过与83支队伍初赛，美国亚特兰大远程线下比赛与8支决赛队伍的激烈比拼，清华大学团队三轮比赛（sim, sim2real和on-site）均为第一，总成绩第一，再次获得ICRA2025抓取与操作比赛 Sim2Real 赛道-类人型机器人操作比赛冠军。





清华大学 计算机科学与技术系
Department of Computer Science and Technology, Tsinghua University

五、未来展望



清华大学
Tsinghua University

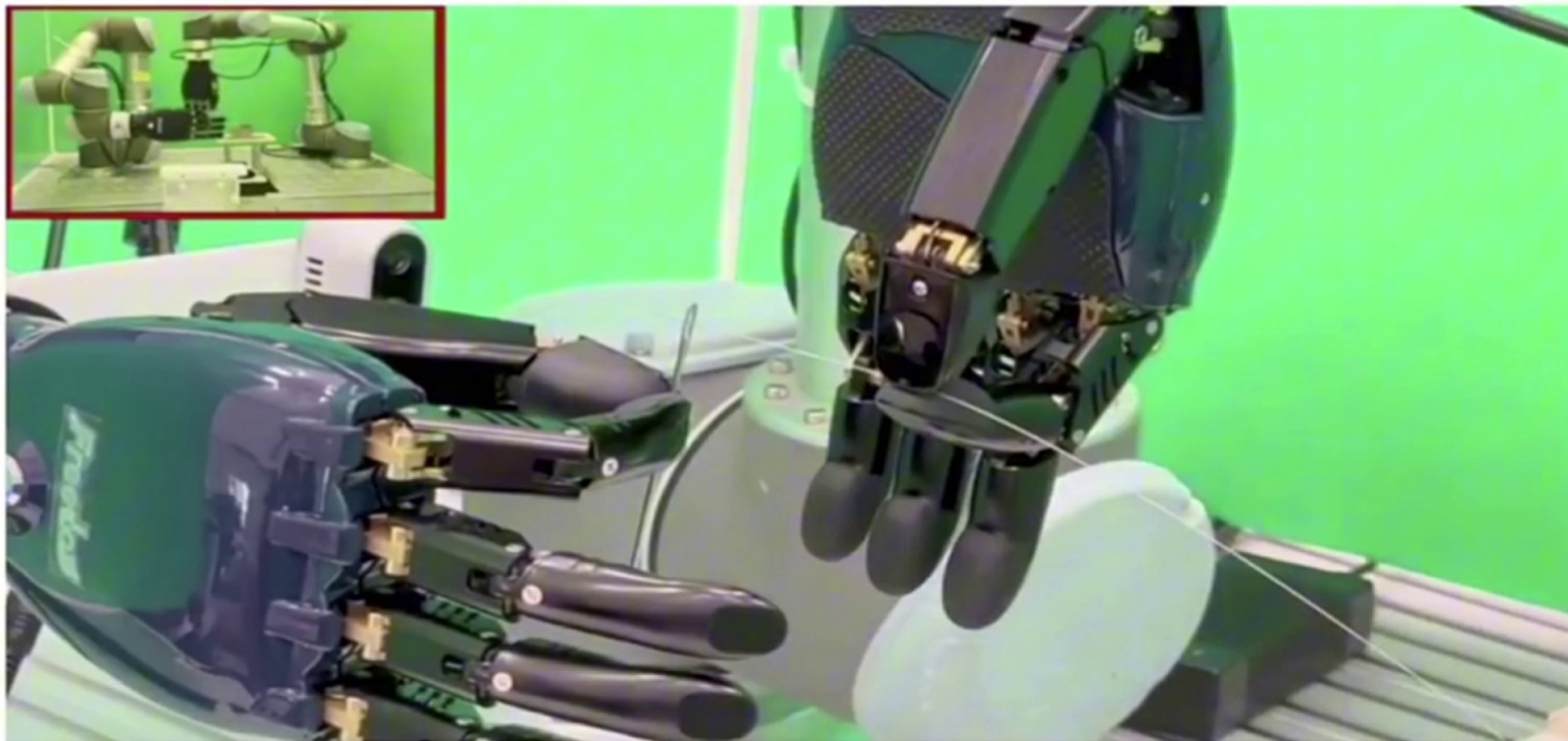


五指灵巧手 “穿针引线”

--清华大学孙富春教授团队

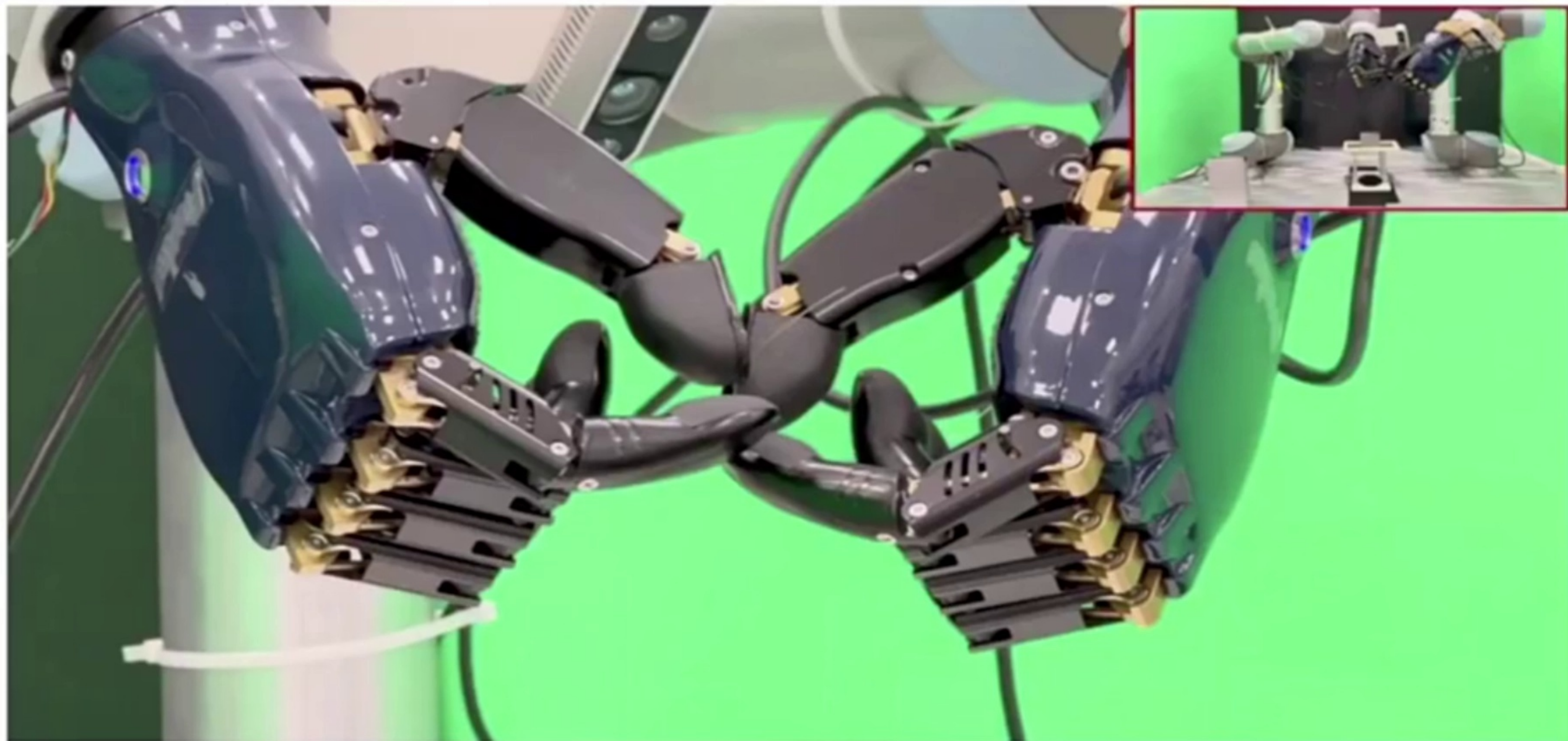


五、未来展望





五、未来展望





贺团队灵巧手穿针引线

人间大爱话真情，深嵌时空旷世惊
慈母密缝游子衣，华佗揆刺杏林名
仿生灵手新生力，引线穿针助远行
众志成城兴伟业，智能时代启征程

——孙富春

2024世界机器人大会



贺具身智能论坛 孙富春诗作

脑体相合创智能，知行发育竞鲲鹏。
感知交互呈佳绩，类脑认知启远征。
跨境迁移妆技艺，人形灵巧扮龙腾。
八方学者今欢聚，产业具身励志登。

