大模型时代的具身智能

什么是智能机器人?

公元前9世纪

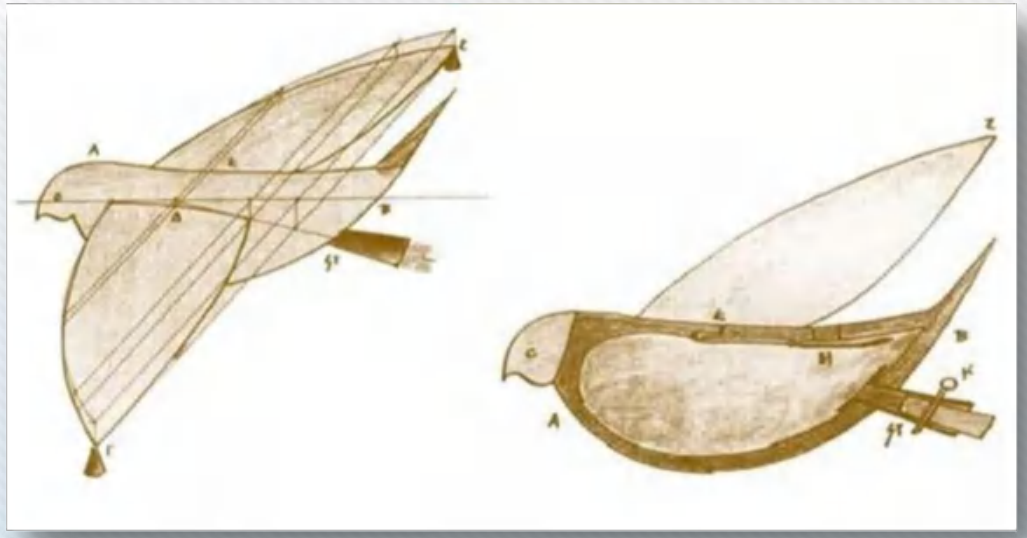
穆王惊视之，趋步俯仰，信人也。巧夫！领其颅，则歌合律；捧其手，则舞应节。千变万化，惟意所适。王以为实人也，与盛姬内御并观之。

——《列子·汤问》

周穆王西巡狩猎遇见了一个名叫偃师的奇人。偃师造出了一个机器人，与常人的外貌极为相似，达到了以假乱真的程度。那个机器人会做各种动作。掰动它的下巴，就会唱歌；挥动它的手臂，就会翩翩起舞。

古希腊数学家 **阿基塔斯** 研制出一种由机械蒸汽驱动的鸟状飞行器，并被命名为 **“鸽子”** 。

其腹部是一套用于产生蒸汽的密闭锅炉。



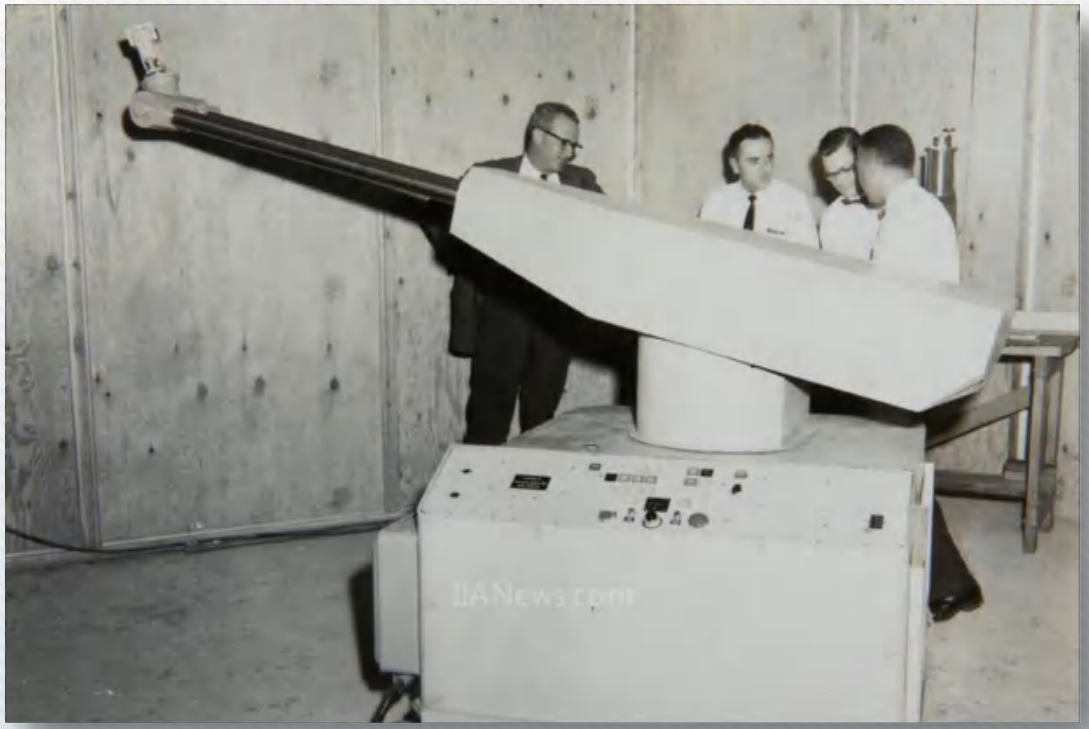
“鸽子”设计图

阿基塔斯

莱昂纳多·达·芬奇在 1495 年左右绘制了人形机器人的草图。现在被称为莱昂纳多的机器人，能够坐起、挥动手臂、移动头部和下巴。

莱昂纳多的机器人

机器人从“玩具”变成“工具”，并应用于工业领域



1961年，世界上第一台工业机器人**Unimate**，用于堆叠金属

1973 年,KUKA公司推出的世界第一台拥有六个机电驱动轴的工业机器人，**FAMULUS**

一定的自主性：编程后可自主运行，自主判断和决定接下来的操作

工业机器人已经相对成熟，人们开始探索更多场景、更智能的机器人

医疗微创机器人

物流运输机器人

展厅服务机器人

家庭清洁机器人

更好的**自主性**：应对的场景和任务更复杂，涉及多机器人协调

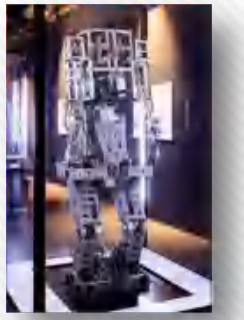
①**自主能力**：尽可能少的人类干预

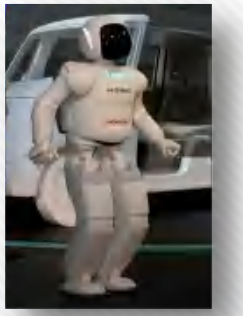
机器人 → **智能机器人 ≈ 人类**

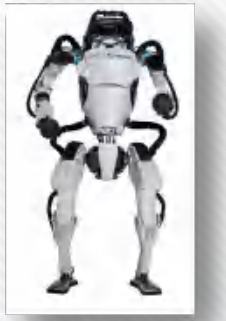
②**泛化能力（通用能力）**：具备强大的综合能力

最受关注的智能机器人——类人机器人

运动控制型机器人 → **智能机器人**







1972 → 2000 → 2008 → 2013

世界第一台全尺寸人形机器人

WABOT-1，日本早稻田大学加藤实验室，行走一步需要45秒，步伐也只有10公分

人形运动能力重大进步

ASIMO，日本本田制造，历经数次迭代，掌握双足奔跑、搬运托盘、上下楼梯等功能

人形机器人成功商业落地

法国 Aldebaran公司研发的小型教学陪伴用人形机器人NAO

人形动作能力迈入新纪元

Atlas机器人，美国波士顿动力公司研发，有很强的运动控制能力

重点关注机器人的运动能力 → **新的关注点：机器人智能**

①**自主能力**：尽可能少的人类干预

②**泛化能力（通用能力）**：具备强大的综合能力

机器人 → **智能机器人 ≈ 人类**

↓

一种算法　人工智能？

工业机器人已经相对成熟，人们开始探索更多场景、更智能的机器人

医疗微创机器人

视觉技术

物流运输机器人

视觉技术

展厅服务机器人

视觉技术
语音技术
自然语言处理

家庭清洁机器人

语音技术

人工智能真的让机器人智能了吗?

我们设想中的智能机器人是什么?

像人类一样工作的机器人?

各方面强于人类的机器人?

有意识和情感的机器人?

纵观人工智能发展

- 1956年—20世纪60年代初，使用人工智能做符号推理，进行**数学证明**
- 20世纪60年代—70年代初，启发式的搜索算法能力有限
- 20世纪70年代初—80年代中，构建**专家系统处理医疗、化学、地质等特定领域**应用
- 20世纪80年代中—90年代中，专家系统需要海量的专业知识，实用价值有限
- 20世纪90年代中—2010年，机器学习算法处理实际问题
- 2011年之后，深度学习算法用于**图像、文本、语音等信息处理**
- 2022年之后，可以处理通用任务的**大模型**
  - ✓ **一定的自主能力**
  - ✓ **一定的泛化能力（通用能力）**

但离我们设想的智能还有多远?

大模型与人形机器人结合形成智能机器人

- 上个世纪对未来人工智能的幻想，主要表现为智能人形机器人，但目前人工智能技术仍然停留在电脑屏幕，没有以实体的方式进入物理世界
- 目前智能程度最强的**大模型**，与目前最先进的**人形机器人**，能否结合形成**智能机器人**？

人工智能真的让机器人智能了吗?

先要说明的问题:

如何构建一个智能机器人?

构建智能机器人（以人形机器人为例）

硬件方面：

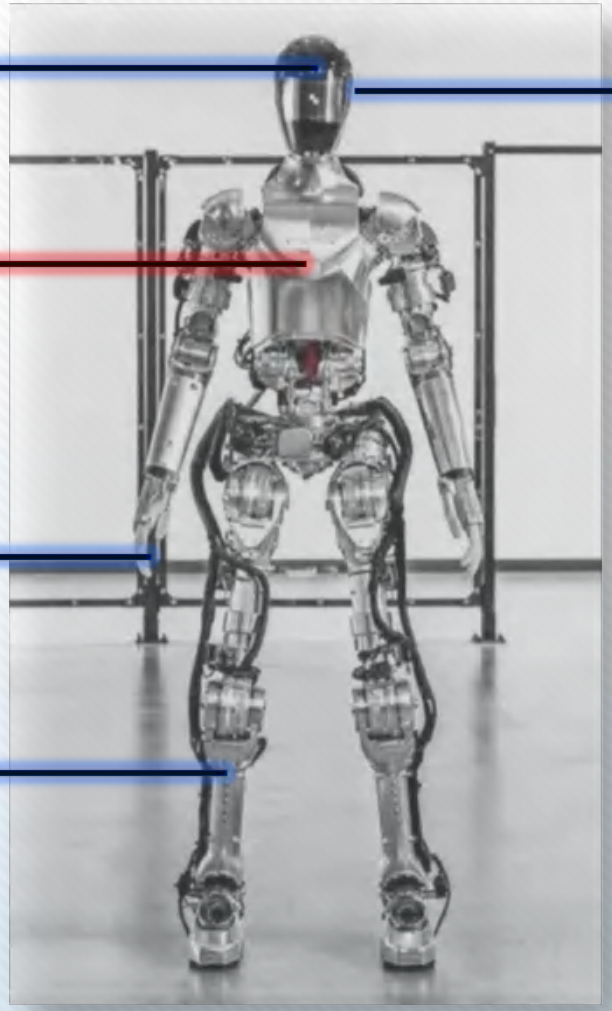
2D视觉信号或
3D点云信号

语音信号

机器人躯体的
所有硬件结构

触觉信号或
力反馈信号

位姿信号



软件及算法方面：

大脑

- 收集所有传感器采集的环境信息和自身状态。
- 并综合分析当前所有状态（**具身感知**）
- 根据当前状态，对自身下一步的运动做出决策和规划（**具身推理**）

小脑

- 向下位机下发送运动指令（**具身执行**）
  （形式包括代码、技能库API、关节旋转角度等）
- 下位机通过运控技术执行指令

举个不是特别形象的简单例子

机器人视觉传感器信号

- 收集所有传感器采集的环境信息和自身状态。并综合分析当前所有状态（**具身感知**）

机器人采集视觉信息，分析出应对咖啡进行清理

- 根据当前状态，对自身下一步的运动做出决策和规划（**具身推理**）

清理咖啡需要如下几步：
1. 扶正杯子并拿起杯盖
2. 找到抹布
3. 用抹布擦拭地面
4. 将抹布放回
5. 将杯子和杯盖扔掉

- 向下位机下发送运动指令（**具身执行**）

生成机器人的运动轨迹，包括手臂如何运动、手掌如何运动、腿部如何运动等

- 下位机通过运控技术执行指令

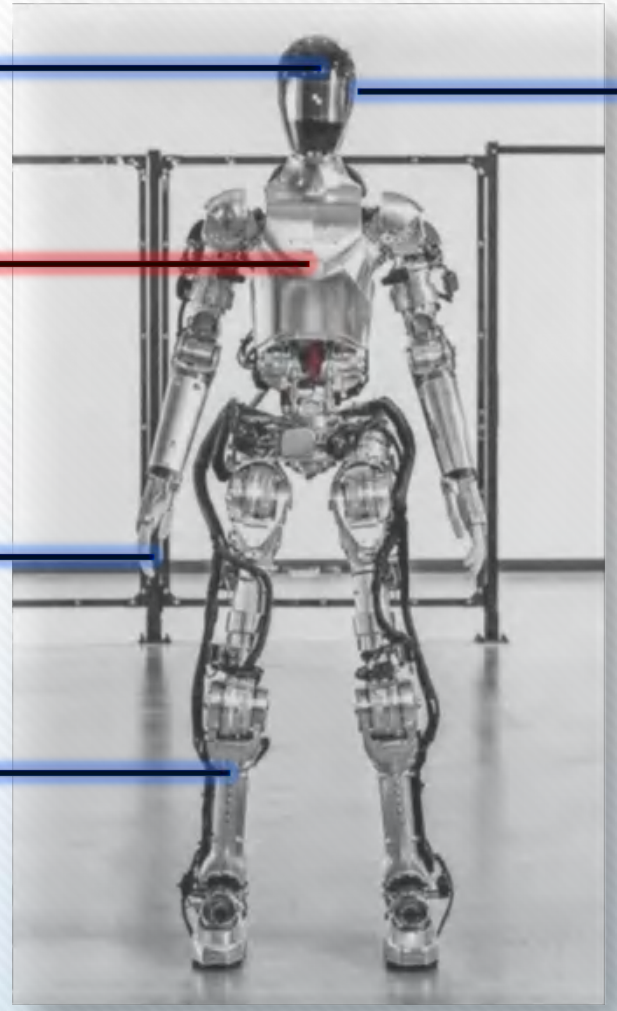
机器人执行

回到问题：

人工智能真的让机器人智能了吗?

构建智能机器人的技术，我们具备和不具备哪些?

硬件方面：



我们已经能造出具备基本性能的机器人硬件和高精度的传感器

构建智能机器人的技术，我们具备和不具备哪些?

软件及算法方面：

还存在诸多问题

大脑

- 收集所有传感器采集的环境信息和自身状态。并综合分析当前所有状态（**具身感知**）
- 根据当前状态，对自身下一步的运动做出决策和规划（**具身推理**）

小脑

- 向下位机下发送运动指令（**具身执行**）
  （形式包括代码、技能库API、关节旋转角度等）

- 下位机通过运控技术执行指令

运控技术相对来说已经较为成熟

当前人工智能这几个方面存在哪些问题?

- 收集所有传感器采集的环境信息和自身状态。并综合分析当前所有状态（**具身感知**）

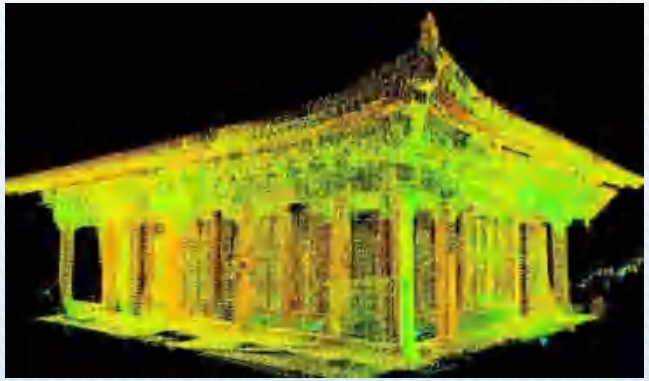
多模态大模型LLaVA已能做到：

请标记出抓握图中插着花的花瓶的位置

但实际场景远比此复杂

这是什么？如何打开它?

人的一些手势是什么意思?



3D点云图如何理解?

当前人工智能这几个方面存在哪些问题?

- 根据当前状态，对自身下一步的运动做出决策和规划**（具身推理）**

来看目前大模型在一组数据集中的表现：

问题：根据视频中显示的当前状态，请问接下来我应当进行那个动作来制作拿铁？

Question: Considering the progress shown in the video and my current observation in the last frame, what action should I take next in order to *make coffee with milk (Task Goal)*?

Candidate Action Plans:

A. pick milk bottle B. open fridge

C. put down mug D. close milk bottle **Answer:** D

候选：A. 拿起牛奶瓶 B. 打开冰箱 C. 放下马克杯 D. 盖上牛奶瓶 答案：D

EgoPlan比赛任务样例

当前人工智能这几个方面存在哪些问题?

- 根据当前状态，对自身下一步的运动做出决策和规划**（具身推理）**

主流大模型在该数据集上的表现：

Table 1: Performance of MLLMs on EgoPlan-Val.

Model	LLM	Acc%
BLIP-2	Flan-T5-XL	26.71
InstructBLIP	Flan-T5-XL	28.09
InstructBLIP Vicuna	Vicuna-7B	26.53
LLaVA	LLaMA-7B	27.00
MiniGPT-4	Vicuna-7B	28.11
VPGTrans	LLaMA-7B	27.38
MultiModal-GPT	Vicuna-7B	27.81
Otter	LLaMA-7B	28.08
OpenFlamingo	LLaMA-7B	27.67
DeepSeek-VL-Chat	DeepSeek-LLM-7B	27.57
mPLUG-Owl-2	LLaMA2-7B	27.84
Yi-VL	Yi-6B	28.67
Gemini-Pro-Vision	-	30.46
SEED-X	LLaMA2-Chat-13B	31.07
XComposer	InternLM-7B	37.17
GPT-4V	-	**37.98**

当前人工智能这几个方面存在哪些问题?

➢ 向下位机下发送运动指令（**具身执行**） （形式包括代码、技能库API、关节旋转角度等）

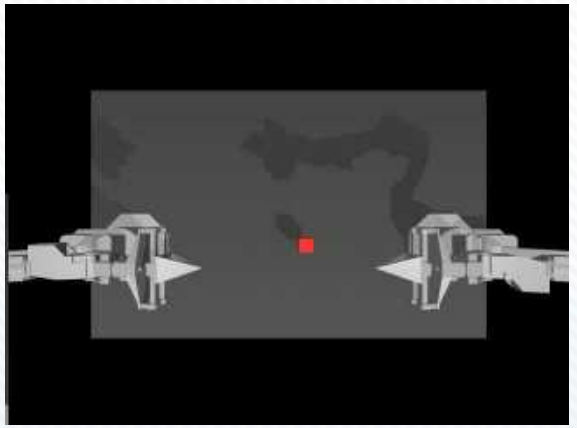
- 对于生成关节旋转角度形式的运动指令：

多模态大模型

拿起可乐

关上抽屉

扩散小模型



转移红色方块

	执行的成功率	执行的流畅度	泛化能力
多模态大模型	较低（60%~70%）	不够流畅	物品泛化
扩散小模型	较高（90%以上）	流畅	位置泛化或无泛化

技能泛化
场景泛化
物品泛化
位置泛化
无泛化

泛化能力

- 对于生成技能库API或代码API形式的运动指令：**现实世界场景过于复杂，构建完整的技能库几乎不可能**

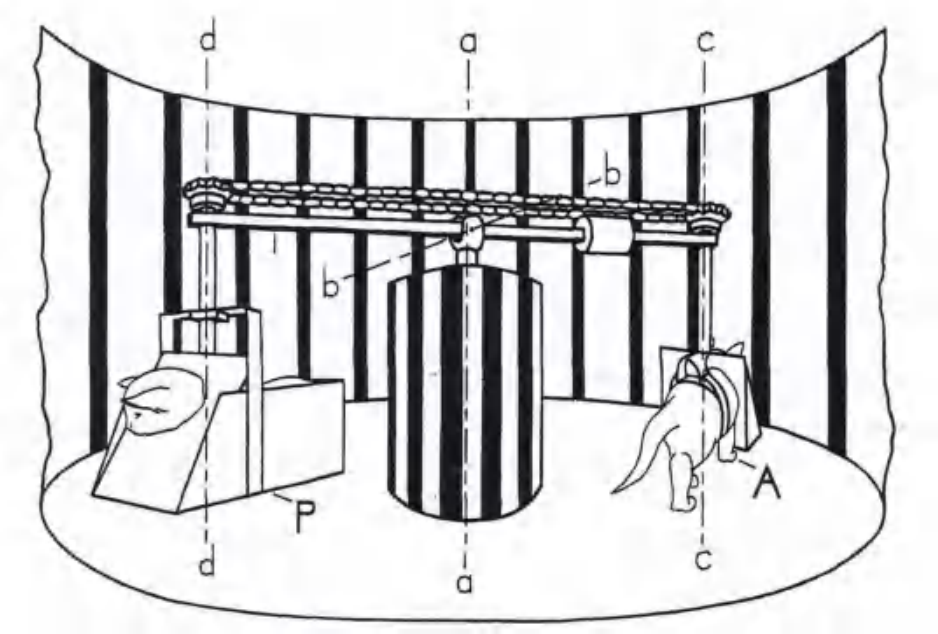
因此，当前人工智能还不足以让机器人更智能，需要**具身智能**

什么是具身智能?

机器人能学习文本图像，能学会走路吗?

1963年进行了一场心理学实验，两只猫自出生起便在黑暗封闭的环境中生活。

- 被动移动位置
- 只能注意到眼中的物体在变大、缩小
- 没有学会走路，甚至不能意识到眼中物体逐渐变大就是在靠近自己



- 可以自由的移动
- 随着腿部动作，眼中物体的大小有相应的变化
- 最终学会走路

有行走条件才能学会走路：有物理身体，可以进行交互

[1] Richard Held, Alan Hein. Movement-produced stimulation in the development of visually guided behavior. 1963 Jouranal of Comparative and Physiological Psychology

具身智能的定义

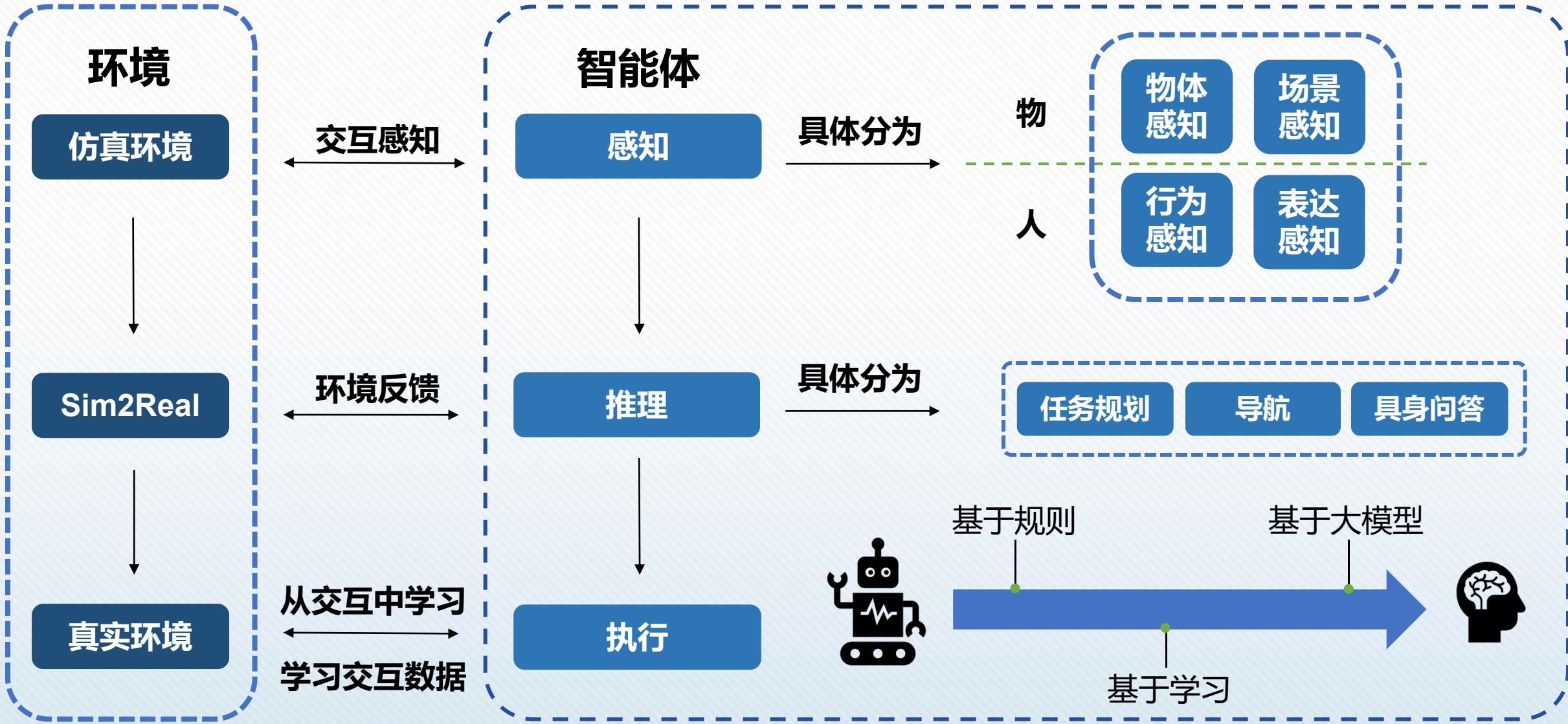
- **定义**：一种基于物理身体进行感知和行动的智能系统，其通过智能体与环境的交互获取信息、理解问题、做出决策并实现行动，从而产生智能行为和适应性。

- **实质**：强调有**物理身体**的智能体通过与**物理环境**进行交互而获得智能的人工智能研究范式。

抽象的智能（围棋、文本处理、图像识别）	VS	有物理身体、与环境进行交互的具身智能
学习“有遮挡的物体识别”		学习“移开遮挡后的物体识别”

具身智能 | CCF专家谈术语，卢策吾，王鹤

具身智能划分：感知、推理、执行



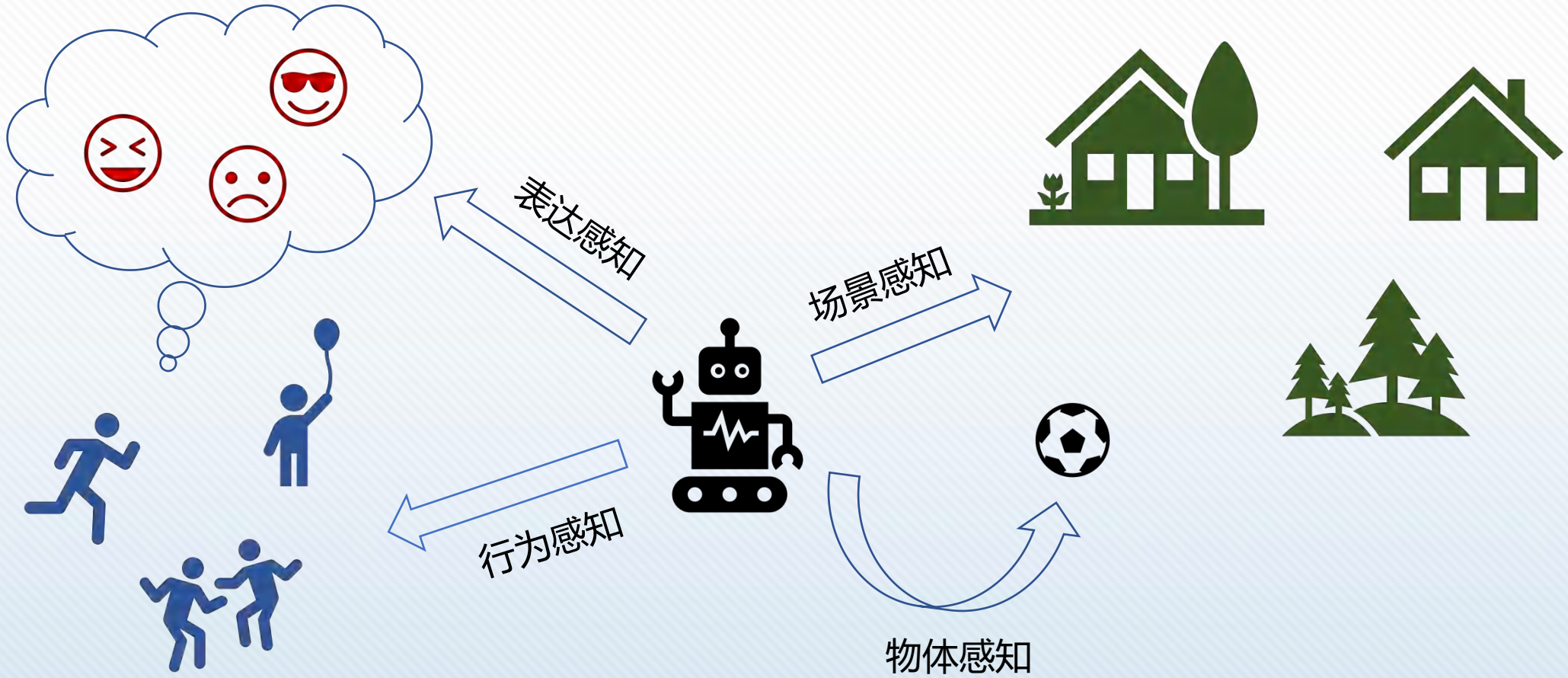
目 录

CONTENTS

1 具身感知

具身感知

- 机器人需要具备环境感知能力，依据感知对象的不同，可以分为四类：



具身感知

- 机器人需要具备环境感知能力，依据感知对象的不同，可以分为四类：
  - 物体感知
    - 几何形状、铰接结构、物理属性
  - 场景感知
    - 场景重建 & 场景理解
  - 行为感知
    - 手势检测、人体姿态检测、人类行为理解
  - 表达感知
    - 情感检测、意图检测
- 重点需要感知能力的机器人：服务机器人、人机协作场景下机器人、社交导航机器人、环境探索机器人

具身感知

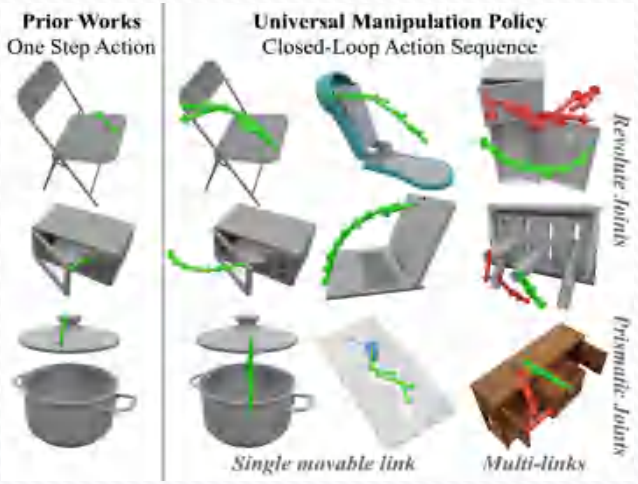
- 具身感知的过程主要包括以下几步：

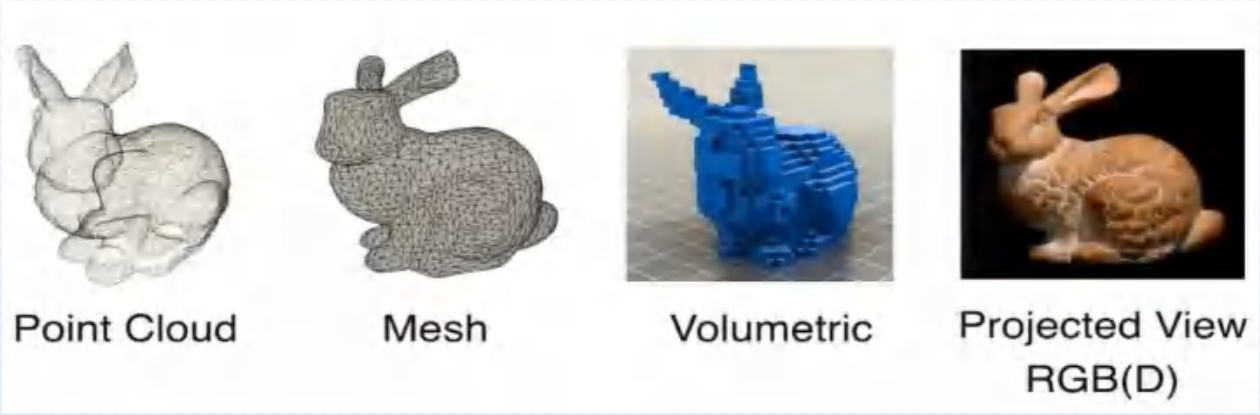
	感受信息	信息的表示/建模/重建	信息理解	下游任务
物体感知	RGB相机 3D相机 麦克风 触觉传感器 热传感器 …	获取3D物体的几何形状、铰接结构、物理属性信息	信息编码方法	位姿估计、物体抓取、交互感知、可供性预测
场景感知		场景重建（主动探索、主动定位、场景的表示）	基于交互的物体检测、空间关系理解、时序变化检测	
行为感知		主要基于2D图片和深度信息	手势识别、姿态检测、人类行为理解	社交导航、自动驾驶、人机协作
表达感知		主要基于视觉表情和声音	情感识别、意图推断（指代表示）	

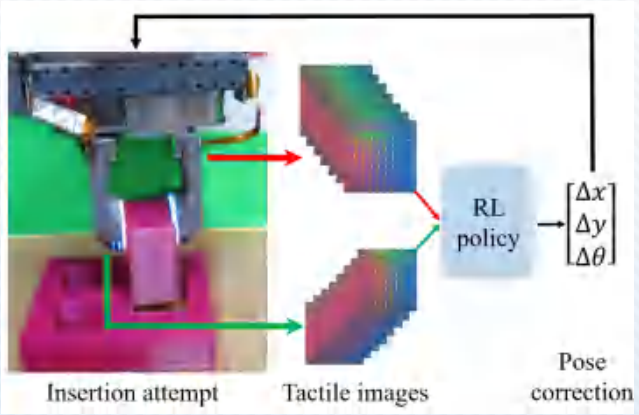
1-1 物体感知

物体感知范畴

- 对于3D空间中的物体，有必要感知其：
  - 几何形状
  - 铰接结构
  - 物理属性







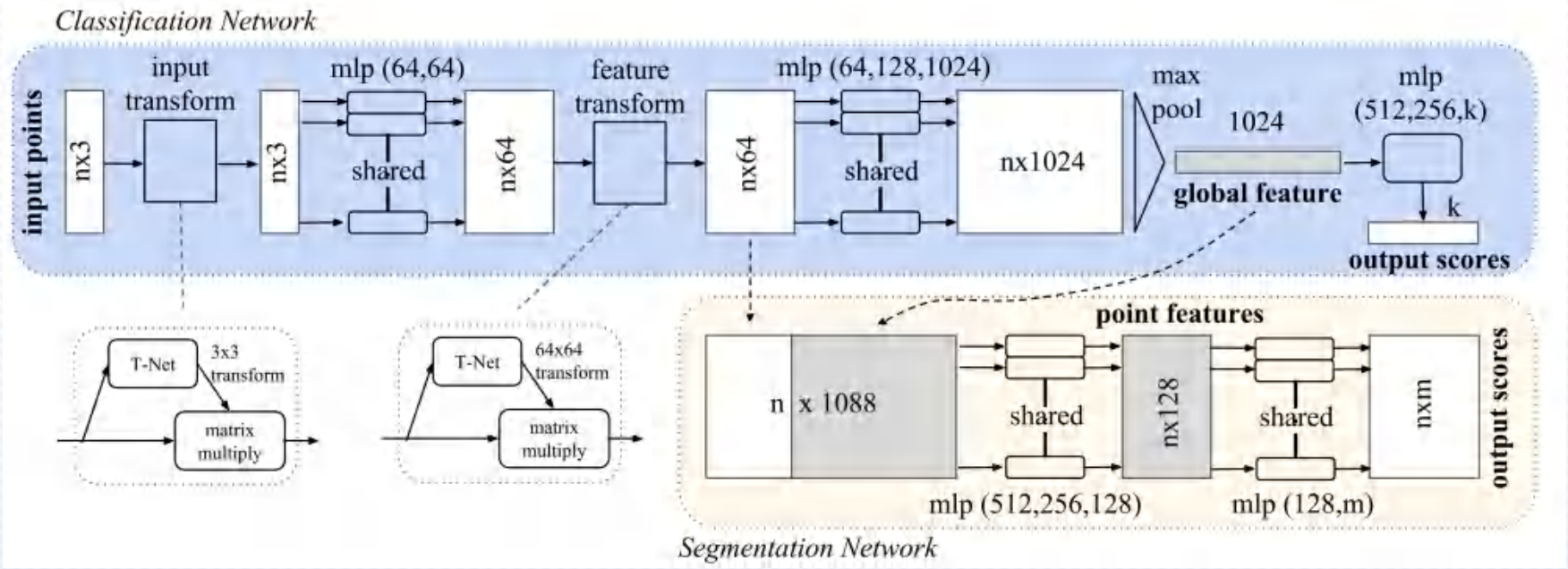
[1] https://adioshun.gitbooks.io/deep_drive/content/intro3d-cloudpoint.html
[2] Xu et al. UMPNet: Universal Manipulation Policy Network for Articulated Objects. 2022 RA-L
[3] Dong et al. Tactile-RL for Insertion: Generalization to Objects of Unknown Geometry

物体几何形状感知

数据格式	描述	来源	编码方法
点云	一组点，每个点包括3D坐标和特征	LiDAR	PointNet，PointNet++
网格	基于点、线、面（三角形）表示物体表面	CAD模型、点云转换	MeshNet
体素	一组立方体，每个立方体包括坐标、体积和特征	点云转换	VoxelNet、DeepSDF、Occupancy Network
深度图	为2D图片每个像素匹配一个深度	双目立体相机、结构光相机、ToF相机	GVCNN

PointNet

- 基于多层感知机，编码点云数据，可以获得点云整体的表示、每个点的表示

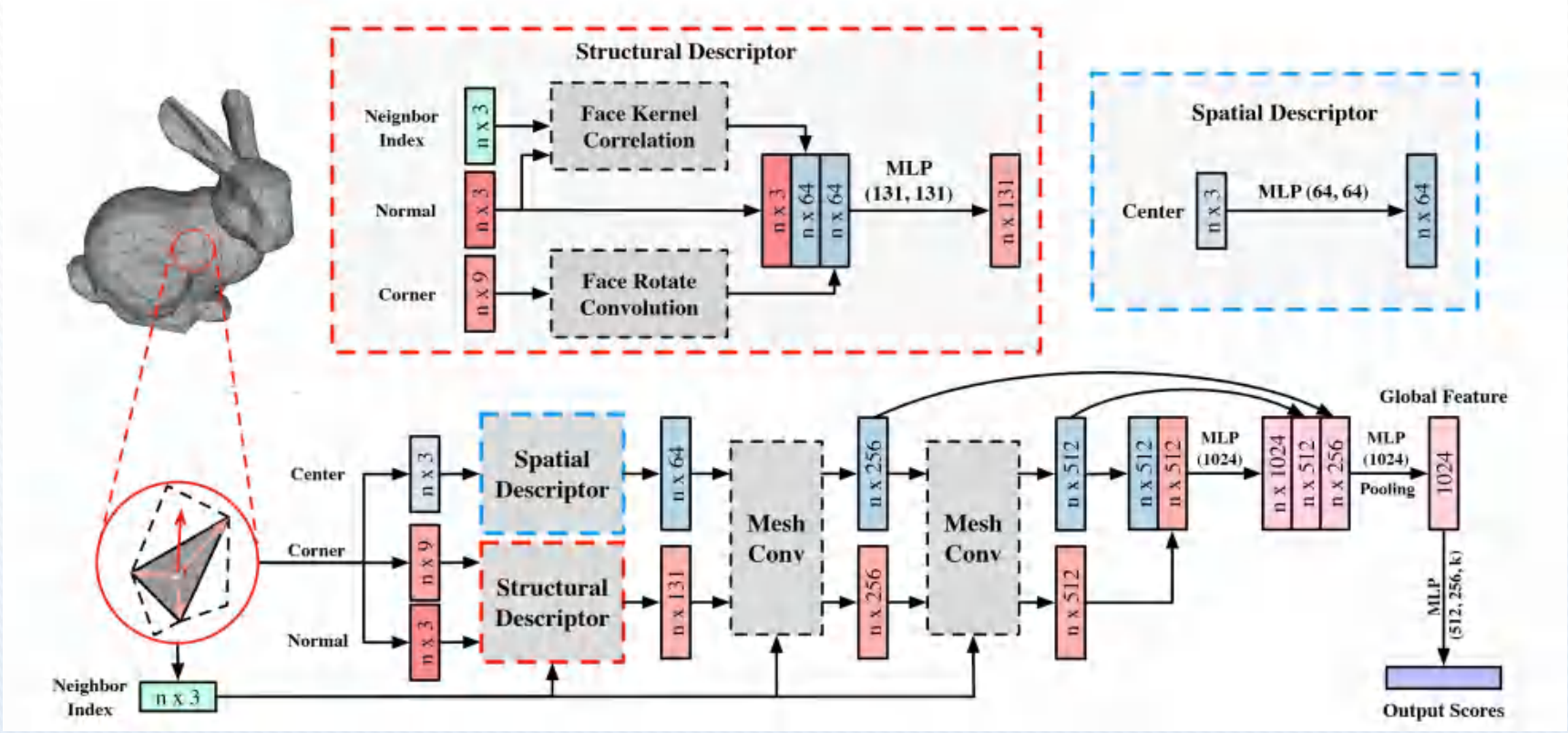


- PointNet为点云数据编码的经典方法，针对其难以捕捉局部特征的缺点又提出了改进版本PointNet++

[1] Qi et al. Pointnet: Deep learning on point sets for 3d classification and segmentation. 2017 CVPR
[2] Qi et al. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. 2017 NIPS

MeshNet

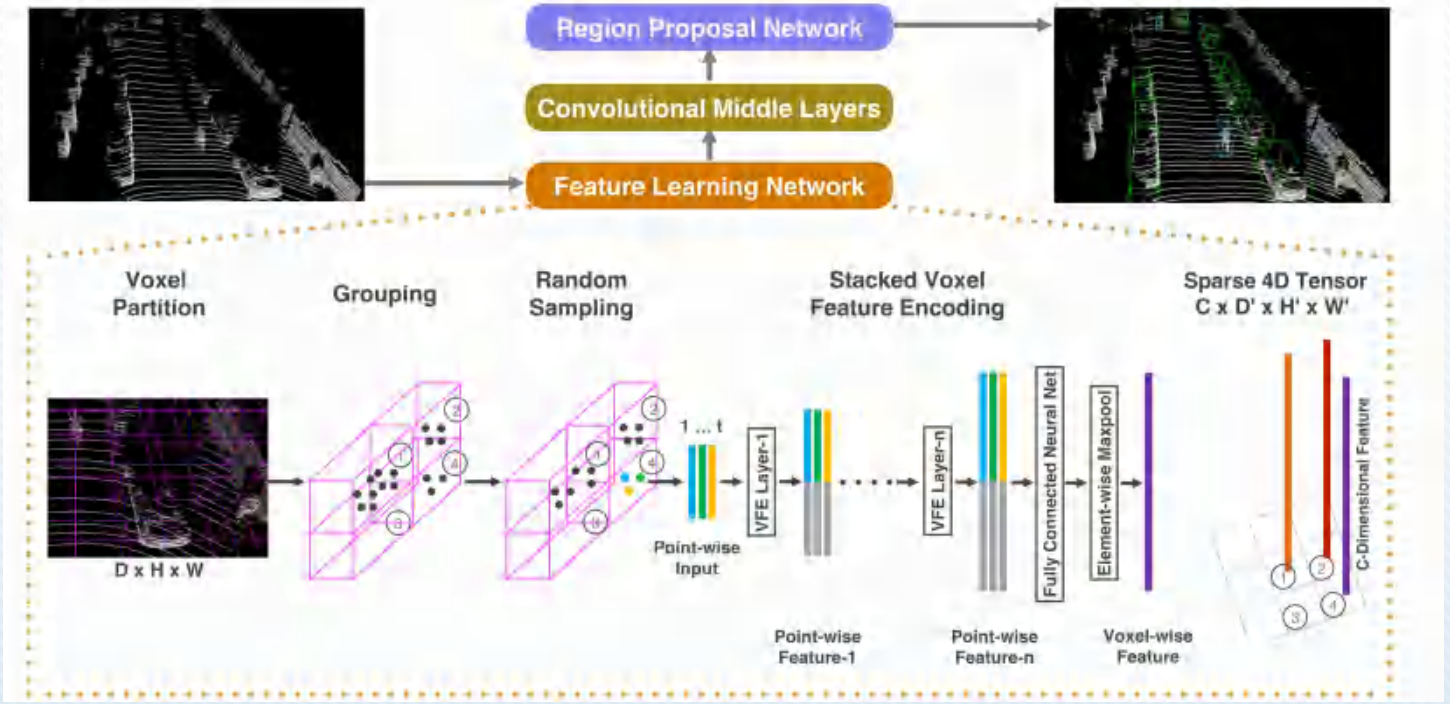
- 基于MLP和CNN，编码每个面的空间特征和结构特征，最后获得整体的物体外形表示



[1] Feng et al. Meshnet: Mesh neural network for 3d shape representation. 2019 AAAI

VoxelNet

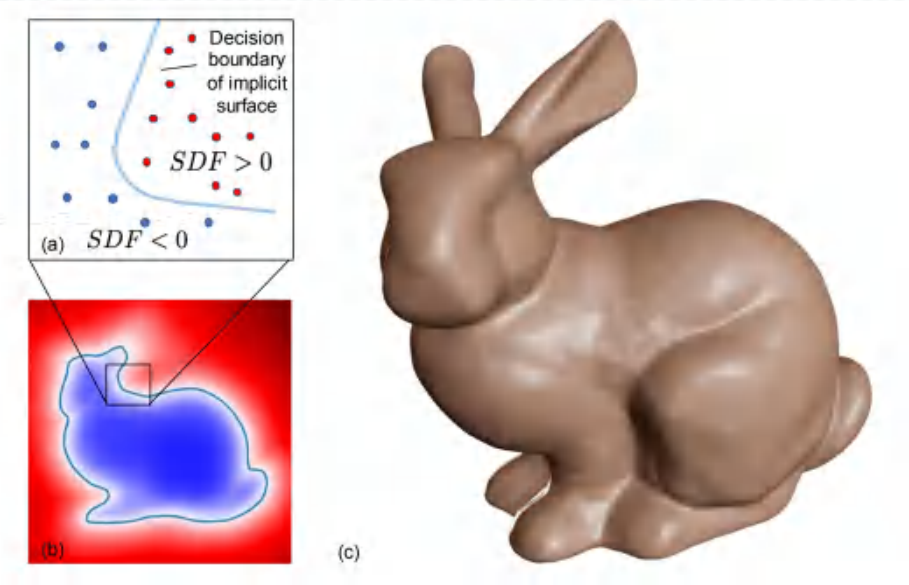
- 首先将点云体素化，然后使用基于MLP和CNN的网络编码体素
- PointNet、MeshNet、VoxelNet对3D数据的卷积编码方式，类似于CV中对2D图片的编码

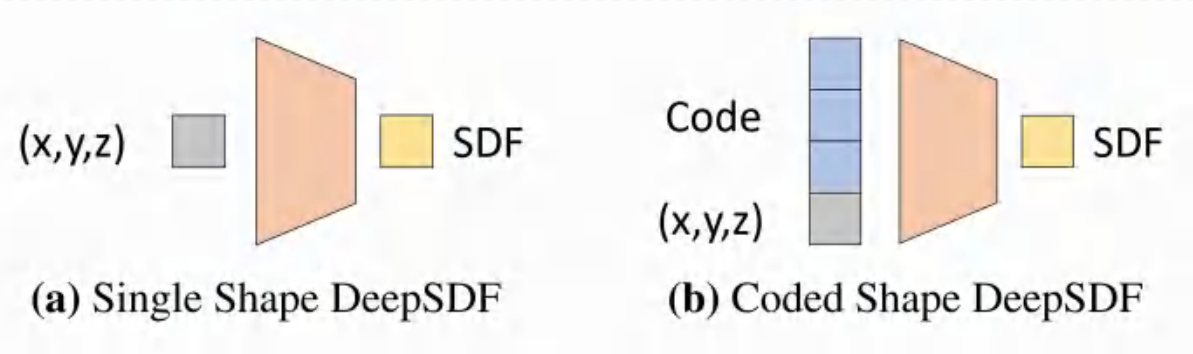


[1] Zhou et al. VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection. 2018 CVPR

DeepSDF （Signed Distance Function）

- 之前PointNet、MeshNet和VoxelNet将3D数据视为离散的单元进行卷积编码
- DeepSDF训练神经网络，拟合一个连续函数：以体素坐标为输入，输出其离最近物体表面的距离。这个连续函数同样蕴涵物体的几何形状信息。



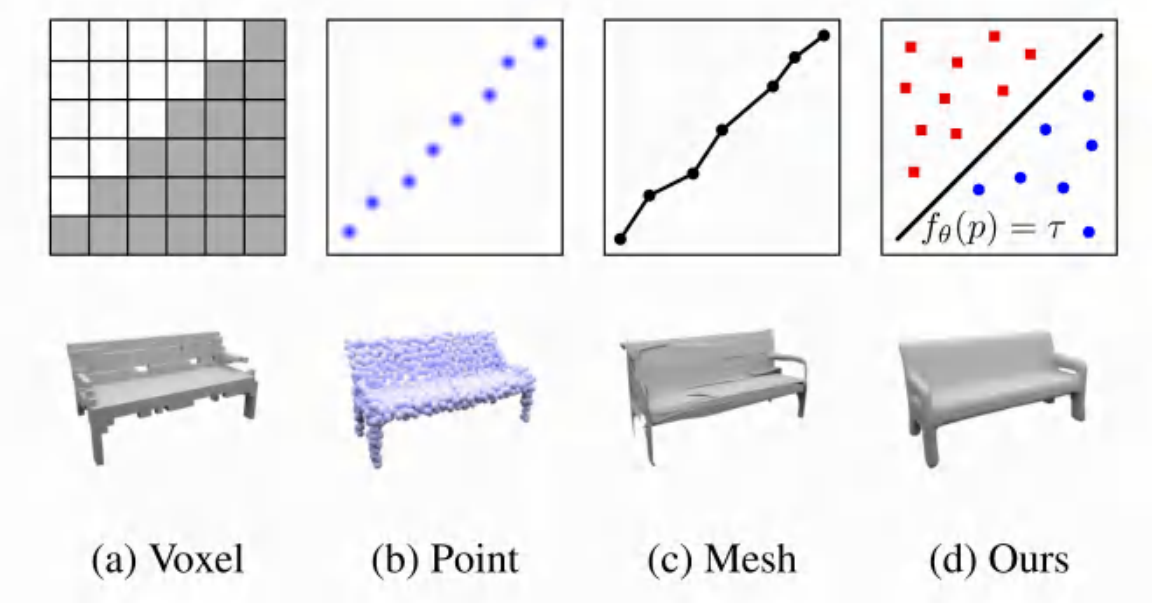


为使训练的SDF不局限于一个物体，引入Code作为物体形状标签

[1] Park et al. DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation. 2019 CVPR

Occupancy Networks 占用概率网络

- 类似于DeepSDF使用一个连续的函数来表示整个空间的体素分布情况，Occupancy Network同样使用神经网络来拟合一个连续的函数，该函数以体素坐标为输入，输出该坐标处体素出现的概率

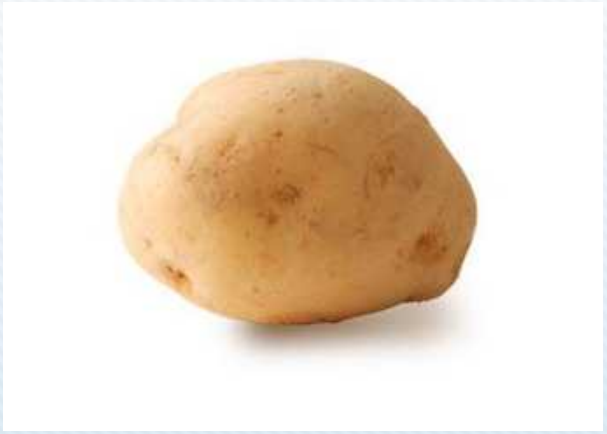


[1] Mescheder et al. Occupancy Networks: Learning 3D Reconstruction in Function Space. 2019 CVPR

几何形状感知的下游任务：物体位姿估计

- 位姿估计任务是预测一个物体在3D空间中的位姿，包括三自由度的平移，与三自由度的旋转，或者可视为物体的位置与朝向
- 根据是否物体的CAD模型是否已知，位姿估计可以分为：
  - 实例级别的位姿估计：需要物体CAD模型，从而获取平移的中心和旋转的初始朝向
  - 类别级别的位姿估计：不需要物体CAD模型

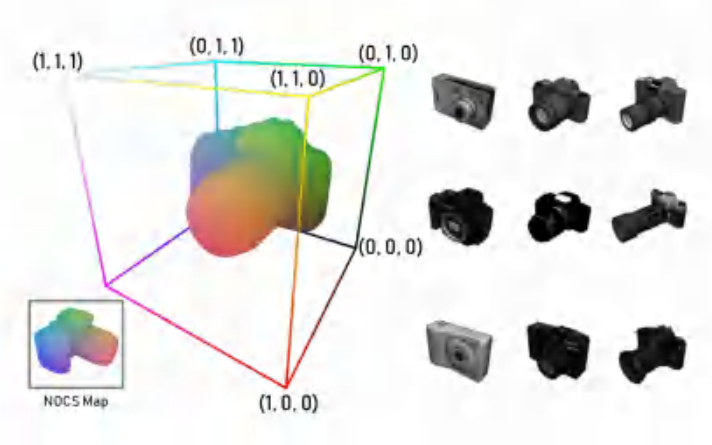
中点是哪里？正面（初始朝向）是哪？没有这些信息如何知道平移和旋转的情况？

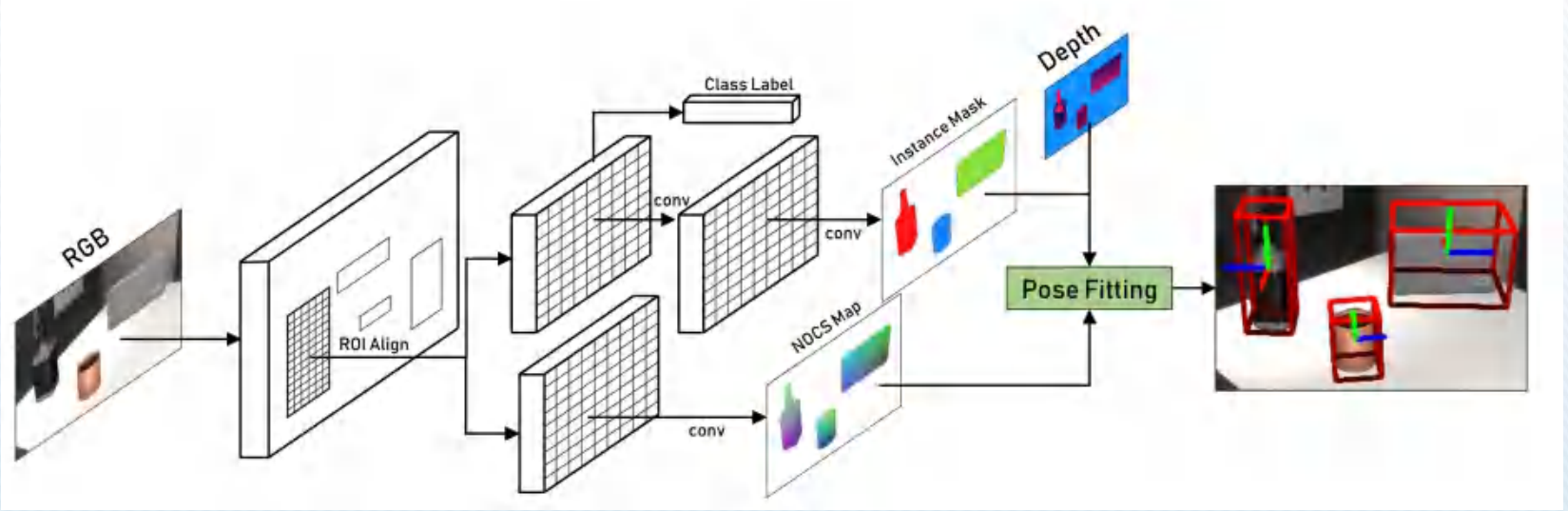


通过“见过”训练集中一个类别下很多物体的中心点和初始朝向，从而可以在测试时对未见过的物体“预设”一个中心点和朝向，然后估计位姿

构建同类别物体统一坐标空间

- 物体上每一个点对应一个（x，y，z），代表该点在标准空间中的位置。给定任意一个图片，分割其中物体，然后在每个像素上预测（x，y，z）。mask上的（x，y，z）就代表这个物体在标准空间中的朝向，结合深度可得位移
- CNN预测：类别、分割Mask、标准空间Map





[1] Wang et al. Normalized Object Coordinate Space for Category-Level 6D Object Pose and Size Estimation. 2019 CVPR

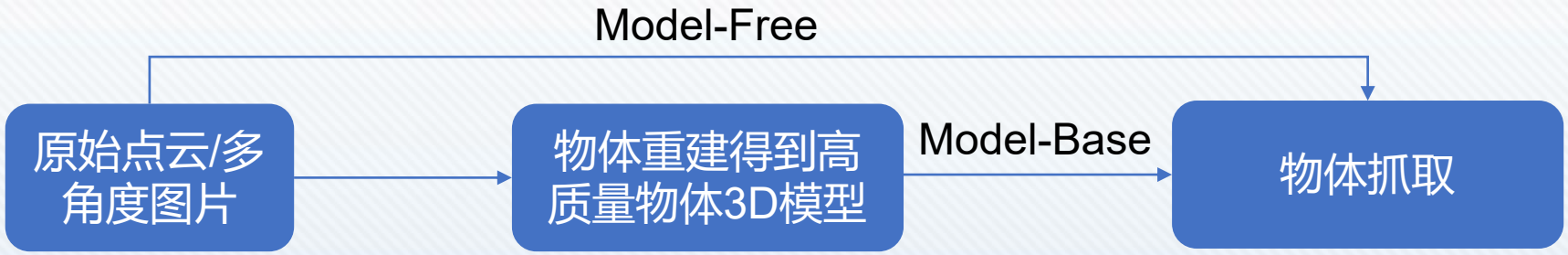
具身感知小结一（提前放在这里，应对可能的疑惑）

- 以上物体外形的研究，与智能机器人根据人类指令执行特定动作的关联在哪里？
- 上述研究与大模型有什么关联？

- 在我们能很好的端到端解决具身智能任务前，以感知物体作为中间任务，助力下游的推理、执行任务，满足实际应用的需要，是很有意义的。
  - 正如句法分析、词性标注之于早期的NLP领域，以及T5模型统一自然语言理解与生成
  - 有观点认为，一个显式的世界模型是人工智能的后续方向，该观点下感知具有更重要的意义
- 在深度学习范畴内，3D数据的处理方式与对2D图片的处理方式非常相似，或许不久之后就会出现很多3D领域的大模型

物体抓取

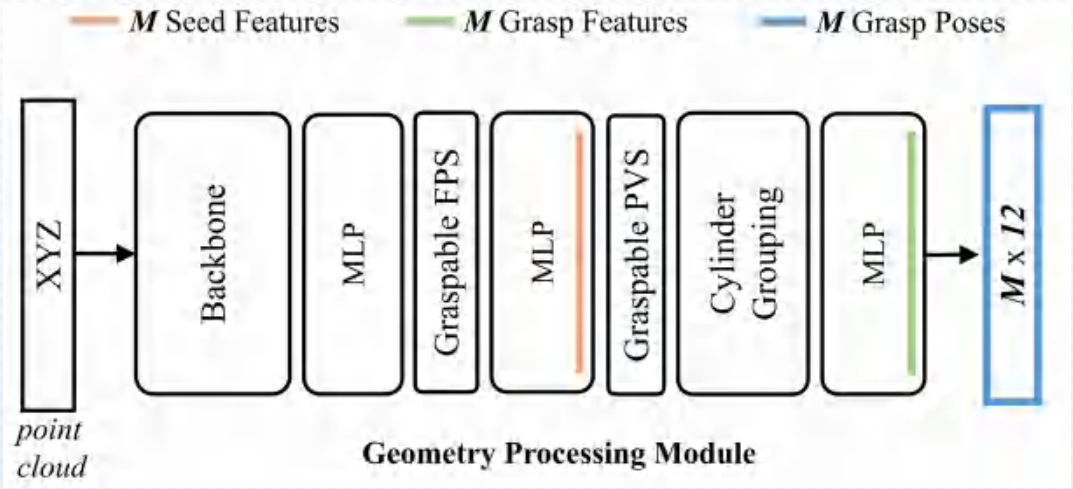
- 传统的物体抓取：
  - 需要已知物体的3D模型，然后使用分析的方法通过数学建模求解抓取点位
- 基于深度学习的物体抓取：
  - 依赖3D相机获取初步点云，不进行显式的物体重建，直接基于点云通过神经网络求解抓取位姿

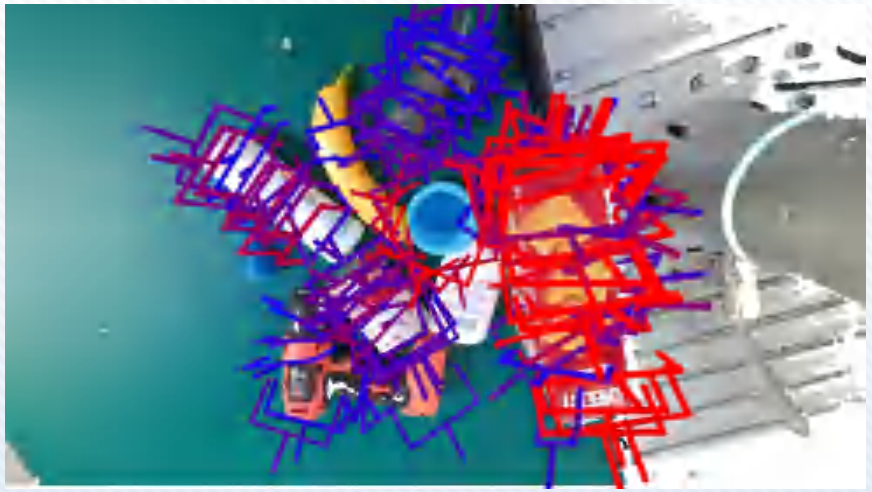


- 感知3D物体的几何形状，与计算机图形学（CG）中的物体重建有密切联系，即使不进行显式的物体重建，一个好的物体重建方法往往也是很好的3D物体和场景的表示方法，例如有研究将CG中3DGS方法用于机器人任务

Anygrasp

- 经典的物体抓取方法，基于物体几何外形信息，并支持动态物体抓取和碰撞检查
- 基于单张RGBD图片，即可生成多个7自由度抓取位姿

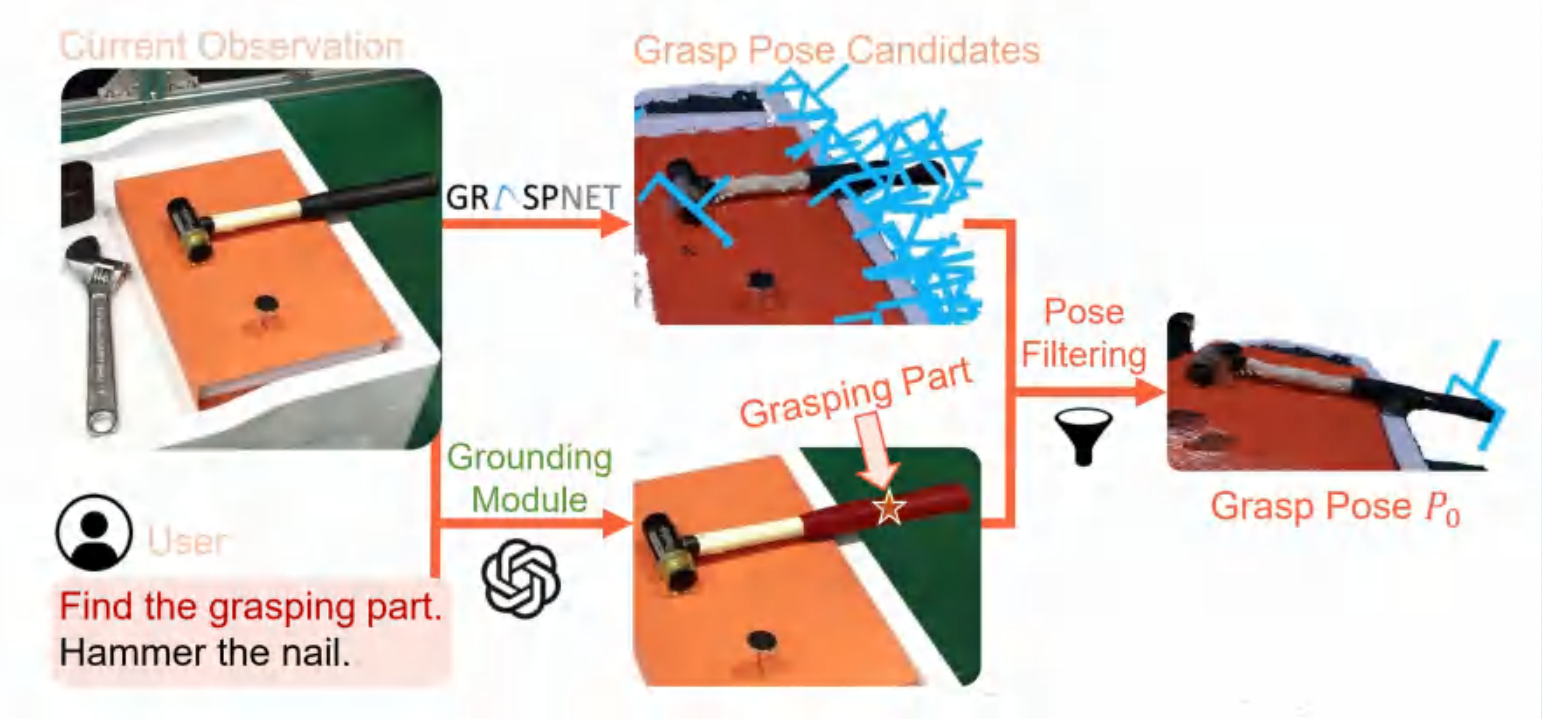




[1] Fang et al. AnyGrasp: Robust and Efficient Grasp Perception in Spatial and Temporal Domains. 2022 T-RO

CoPA：结合大模型物体感知的抓取

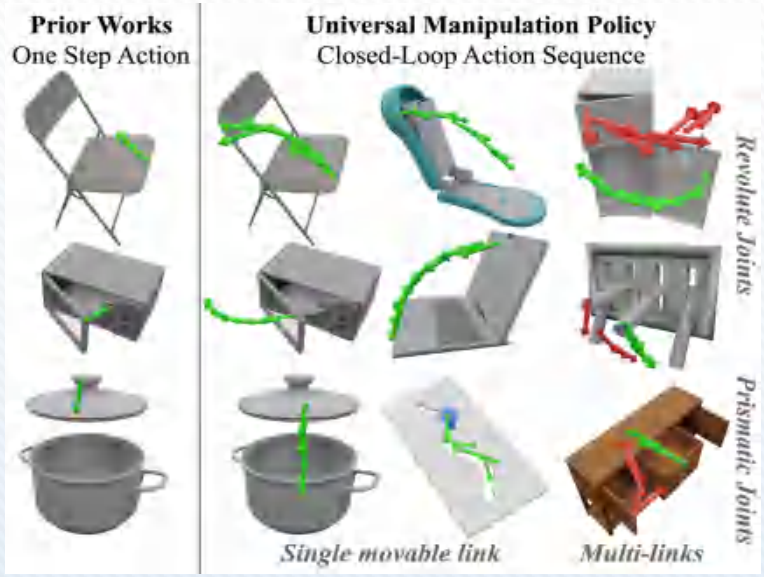
- 多模态大模型结合物体分割模型由粗到细确定抓取点位（物体部件级别）
- 抓取小模型GraspNet生成多个抓取位姿，与大模型给出的抓取点位接近的分数更高

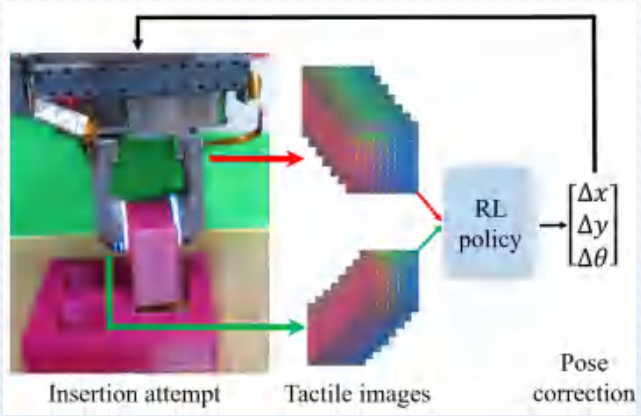


[1] Huang et al. CoPa: General Robotic Manipulation through Spatial Constraints of Parts with Foundation Models. 2024 ICRA

物体感知范畴

- 对于3D空间中的物体，有必要感知其：
  - 几何形状：点云、体素、网格、深度图的编码表示，以及位姿估计，物体抓取下游任务
  - **铰接结构**
  - 物理属性

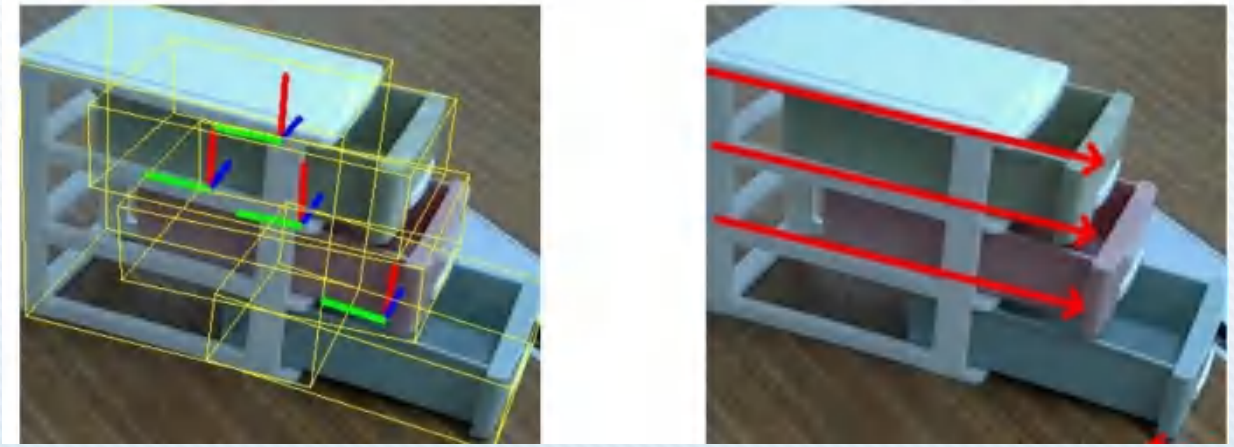




[1] https://adioshun.gitbooks.io/deep_drive/content/intro3d-cloudpoint.html
[2] Xu et al. UMPNet: Universal Manipulation Policy Network for Articulated Objects. 2022 RA-L
[3] Dong et al. Tactile-RL for Insertion: Generalization to Objects of Unknown Geometry

铰接结构

- 铰接物体与刚性物体：
  - 刚性物体内部构件刚性连接，无法变形
  - 铰接物体内部构件由关节或其他铰接结构连接，部件可以旋转、平移
- 刚性物体关注几何形状，对其的操作主要为抓取、放置，即位姿估计和物体抓取任务
- 铰接物体除几何形状外，还关注对其铰接结构。铰接物体支持复杂的操作，例如开关柜门，拧瓶盖



[1] Liu et al. Toward Real-World Category-Level Articulation Pose Estimation. 2022 TIP

铰接结构数据来源

- 铰接物体数据格式主要为URDF，通过定义物体的边、关节属性来定义物体铰接结构
- 铰接结构数据来源主要包括
  - 手工收集，e.g. AKB-48
  - 在已有3D数据集上标注铰接信息
  - 合成数据

Figure 3. The task-specific model acquisition equipment. (a) 1 is a Rotating turntable for objects with multiple scales. 2 is a tracking marker. 3 is a light-absorbing item. 4 is a lift bracket. 5 is the Shining 3D scanner. 6-8 are the realsense L515 cameras for capturing multiviews of objects.

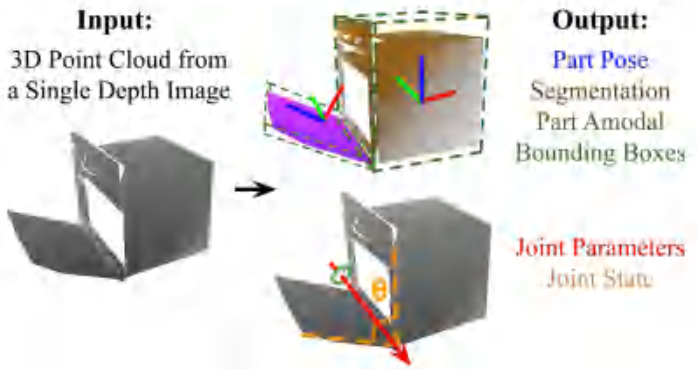
[1] Liu et al. AKB-48: A Real-World Articulated Object Knowledge Base. 2022 CVPR
[2] Cage et al. CAGE: Controllable Articulation GEneration. 2024 CVPR

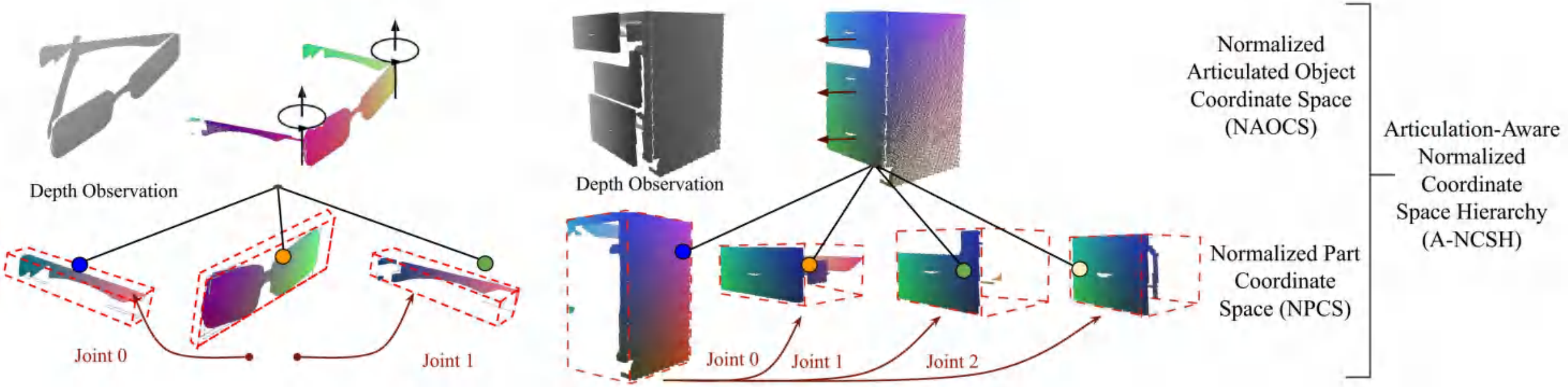
铰接物体表示方法

- 铰接物体的表示，应该主要包括以下信息：
  - 每个组件的几何形状信息
  - 每个组件的运动学信息，包括：位移类型（平移、旋转）、位移参数（平移方向、旋转轴）、位移限制（最大移动距离、最大旋转角度）
- 一个好的铰接表示有助于机器人理解铰接物体
- 两种铰接结构表示方法
  - 直接建模关节参数
  - 建模位移变化情况

建模关节参数表示铰接物体

- 通过分别建模物体部件和整体两个层次的信息来表示铰接物体，实现基于RGBD图片预测物体铰接结构。
- 物体层次信息主要为关节参数和状态，部件层次信息为部件的位姿和规模

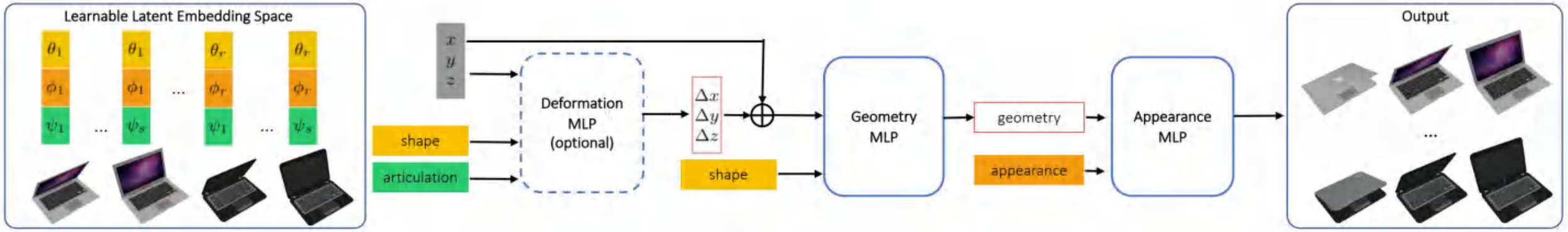




[1] Li et al. Category-Level Articulated Object Pose Estimation. 2020 CVPR

建模位移情况表示铰接物体

- 该论文同样希望通过多视角图片得到物体的形状、外观、铰接结构信息。
- 其认为物体状态可以由形状、外观、铰接状态来表示，并使用不同的code来表示，通过一个变形网络分离物体铰接状态（位移情况）得到新的物体位置，然后分别得到几何形状和物体外观
- 变形网络使用有监督训练的方式，以形状和铰接code为输入，预测物体每个点的位移



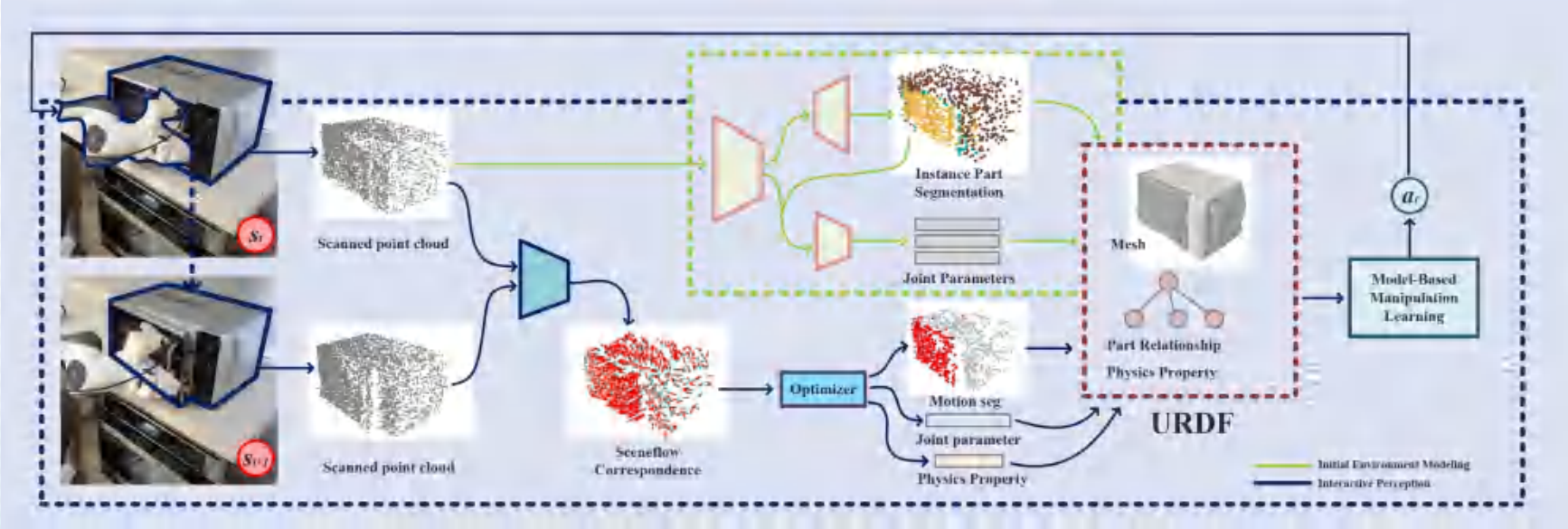
[1] Wei et al. Self-supervised Neural Articulated Shape and Appearance Models. 2022 CVPR

铰接结构的下游任务

- 几何结构部分与主流计算机视觉领域相比，其特点在于主要基于3D信息
- 但对3D信息的处理并非具身智能的核心，具身智能的核心在于其是一种行为智能，在感知领域具体体现为：可以通过与环境的主动交互，增强对环境的感知效果
- 铰接物体支持机器人进行丰富的操作任务，并提供相应的反馈。与之相关的下游任务有**交互感知**、**物体可供性预测**两类
  - 交互感知：机器人通过与物体交互获取更多信息
  - 物体可供性预测：预测物体能否支持机器人进行某种操作

交互感知获取物体铰接结构

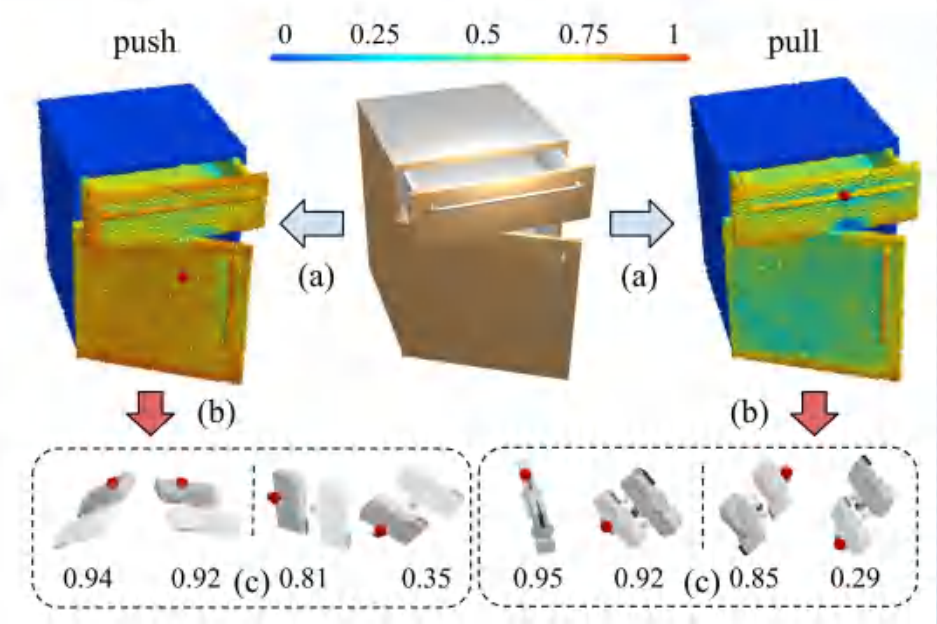
- 之前介绍的工作基于静态数据集预测物体铰接结构，该工作通过实际物理交互行为获取物体铰接结构
- 首先以原始物体点云作为输入，基于物体组件级分割，得到物体初始URDF文件
- 机器人操作物体，基于当前URDF文件可以预测操作后的物体状态，与实际观察到的物体状态进行对比，该监督信号对于物体模型参数（URDF文件）是可微的，从而进行参数更新

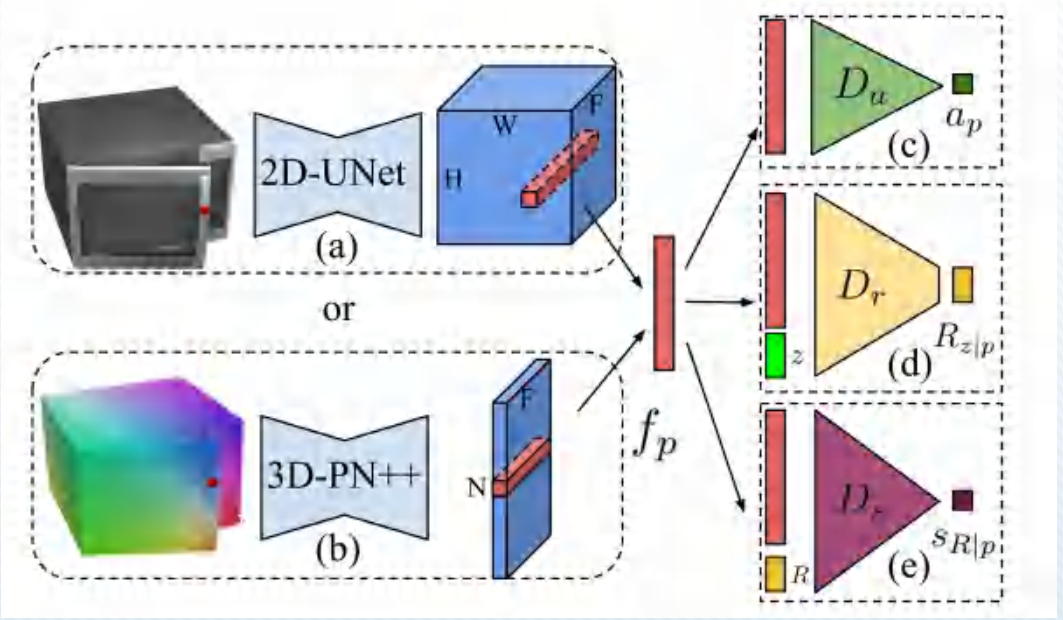


[1] Lv et al. SAGCI-System: Towards Sample-Efficient, Generalizable, Compositional and Incremental Robot Learning. 2022 ICRA

物体可供性预测

- 对于任务规划和导航任务，知道一个物体可以施加哪些动作是很重要的，也可以用于指导物体操作
- Where2act训练一个预测网络，给定一个原子动作（推、拉），对于图片或点云中每一个像素预测1）可行性分数；2）动作轨迹；3）成功概率
- 基于此，机器人可以知道每一个原子动作在物体上的最佳操作点位与轨迹

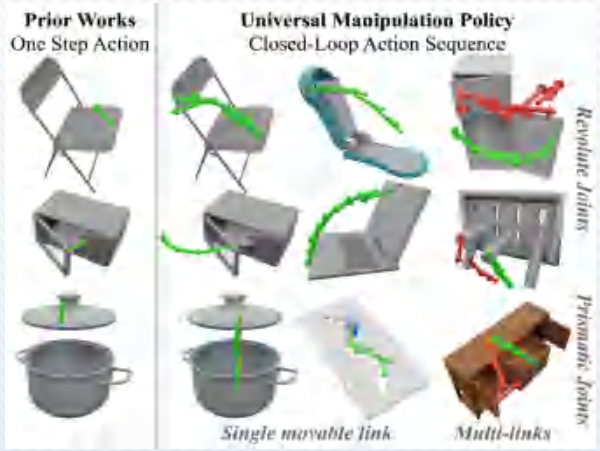


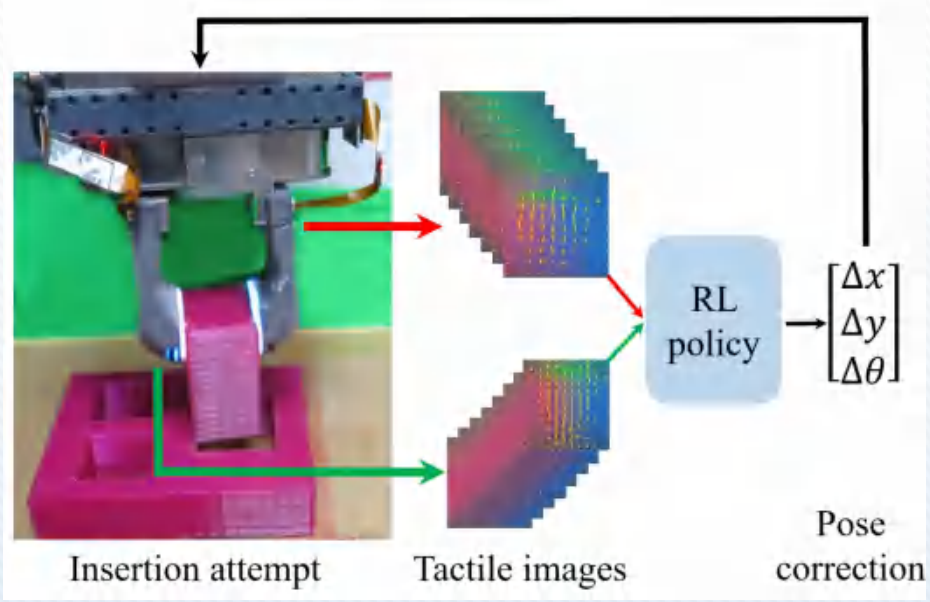


[1] Mo et al. Where2Act: From Pixels to Actions for Articulated 3D Objects. 2024 ICCV

物体感知范畴

- 对于3D空间中的物体，有必要感知其：
  - 几何形状：点云、体素、网格、深度图的编码表示，以及位姿估计，物体抓取下游任务
  - 铰接结构
  - **物理属性**





[1] https://adioshun.gitbooks.io/deep_drive/content/intro3d-cloudpoint.html
[2] Xu et al. UMPNet: Universal Manipulation Policy Network for Articulated Objects. 2022 RA-L
[3] Dong et al. Tactile-RL for Insertion: Generalization to Objects of Unknown Geometry

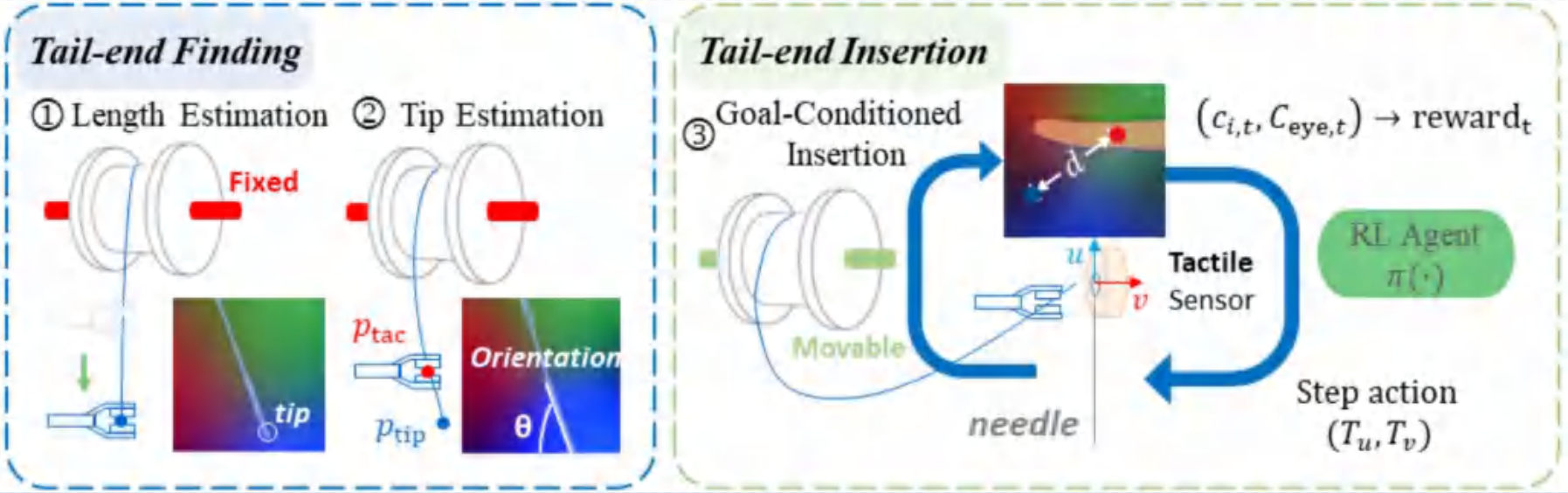
物理属性

- 物体的物理属性种类及来源包括：
  - 触觉：触觉传感器
  - 力矩：六轴力矩传感器，3自由度力，3自由度扭矩，
  - 温度：温度传感器
  - 材质、硬度…
- 物理属性的表示
  - 与其他模态融合，如图像和点云：IMAGEBIND、LANGBIND
  - 单独使用物理信息：强化学习端到端的方式利用触觉信息

[1] Girdhar et al. Imagebind: One embedding space to bind them all. 2023 CVPR
[2] Zhu et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. 2024 ICLR
[3] Dong et al. Tactile-rl for insertion: Generalization to objects of unknown geometry. 2024 ICRA

物理属性辅助操作解决视觉遮挡问题

- 利用触觉传感器理解物理属性：T-NT
- 根据视觉和触觉反馈，用强化学习训练机器人将线穿过针孔
- 使用触觉传感器查找线的末端，以及判断针是否穿过针孔



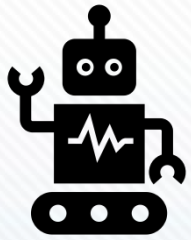
[1] Yu et al. Precise Robotic Needle-Threading with Tactile Perception and Reinforcement Learning. 2023 CoRL

1-2 场景感知

场景感知简述

- 定义：场景感知是通过实现与场景的**交互**来理解现实世界场景
- 意义：赋予机器人理解周围环境并与之交互的能力
- 内核：
  - 对空间布局的**几何理解**
  - 对场景中物体的**语义理解**
- 组成：
  - 粗粒度：场景中物体的组成、物体的语义、物体的空间关系
  - 细粒度：场景中每个点的精确空间坐标和语义
- 具体形式：点云、地标、拓扑图、场景图、隐表示

场景感知的研究内容



环境

原始信息 → 场景表示 → 场景认知

抽取 ↓　构建 ↑　形成 ↑

场景信息获取 → 场景重建 → 场景理解

易利用：视觉信息　激光信息　雷达信息

难利用：听觉　触觉　化学　红外　超声　…

传统：SLAM

新方法：深度学习

新任务：主动映射　主动定位

新表示：拓扑结构　场景图

对象：单个物体　多个物体　场景图像

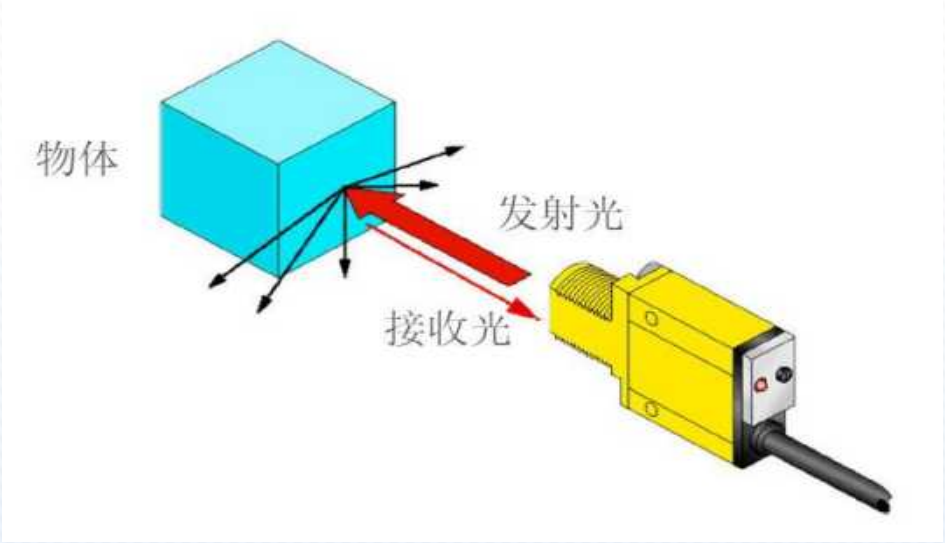
实现：物体识别　空间关系推理　时序变化检测

场景信息构成

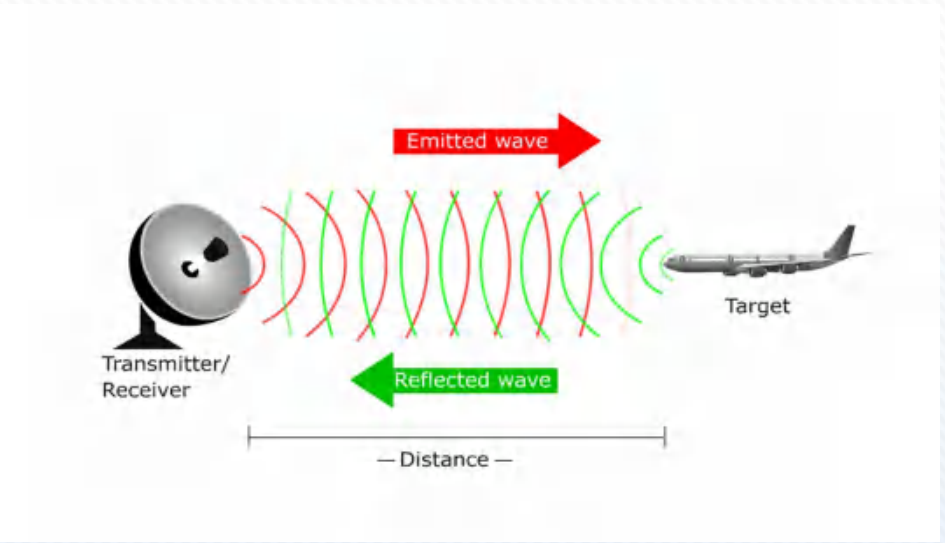
- 场景信息组成
  - 粗粒度
    - 场景中物体的组成
    - 场景中物体的语义
    - 场景中物体的空间关系
  - 细粒度
    - 场景中每个点的精确空间坐标和语义
- 场景信息提取方式
  - 构建场景表示
    - 点云、地标、拓扑图、场景图及隐式表示

易利用的场景信息

- 视觉：符合人类的先验知识，相关研究工作多
- 激光/雷达：可以直接获取准确的场景表示，无需视觉重建



激光传感器工作原理



雷达传感器工作原理

[1] Sun, et al. A quality improvement method for 3D laser slam point clouds based on geometric primitives of the scan scene. 2021 IJRS
[2] Kong, et al. Multi-modal data-efficient 3d scene understanding for autonomous driving. 2024 arXiv
[3] Zheng, et al. Scene-aware learning network for radar object detection. 2021 PCMR
[4] Yang, et al. An ego-motion estimation method using millimeter-wave radar in 3D scene reconstruction. 2022 IHMSC

难利用的场景信息

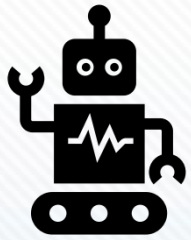
- 听觉：可用于视听导航任务
- 触觉：可用于感知物体表面
- 化学：可用于特殊任务，如识别气味来源
- 红外：可用于特殊场景，如烟雾场景下
- 超声：可用于深度测量

应用范围狭窄

并非场景感知任务焦点

[1] Singh, et al. Sporadic Audio-Visual Embodied Assistive Robot Navigation For Human Tracking. 2023 PETRA
[2] Gan, et al. Look, listen, and act: Towards audio-visual embodied navigation. 2020 ICRA
[3] Roberge, et al. StereoTac: A novel visuotactile sensor that combines tactile sensing with 3D vision. 2023 RAL
[4] Padmanabha, et al. Omnitact: A multi-directional high-resolution touch sensor. 2020 ICRA
[5] Armada, et al. Co-operative smell-based navigation for mobile robots. 2004 CLAWAR
[6] Ciui, et al. Chemical sensing at the robot fingertips: Toward automated taste discrimination in food samples. 2018 ACS sensors
[7] Sinai, et al. Scene recognition with infra-red, low-light, and sensor fused imagery.1999 IRIS
[8] Kim, et al. Firefighting robot stereo infrared vision and radar sensor fusion for imaging through smoke. 2015 Fire Technology
[9] Shimoyama, et al. Seeing Nearby 3D Scenes using Ultrasonic Sensors. 2022 IV
[10] Mulindwa, et al. Indoor 3D reconstruction using camera, IMU and ultrasonic sensors. 2020 JST

场景重建



环境

原始信息 → 场景表示 → 场景认知

抽取 ↓

构建 ↑

形成 ↑

场景信息获取 → 场景重建 → 场景理解

易利用：视觉信息、激光信息、雷达信息

难利用：听觉、触觉、化学、红外、超声、…

传统：SLAM

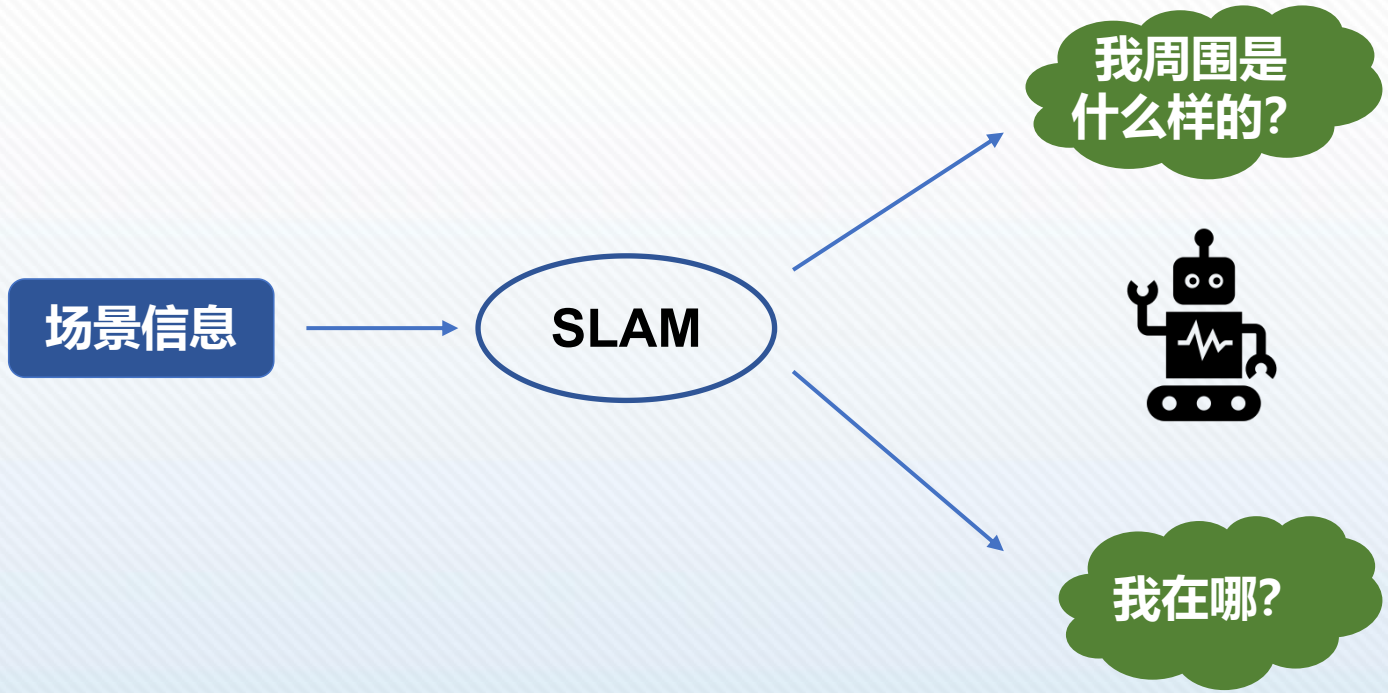
- 新方法：深度学习
- 新任务：主动映射、主动定位
- 新表示：拓扑结构、场景图

对象：单个物体、多个物体、场景图像

实现：物体识别、空间关系推理、时序变化检测

场景重建

- 场景重建的核心技术是SLAM（同步定位与映射）
- SLAM是机器人在未知环境下移动，逐步构建周围环境的连续地图，并同时估计其在地图中位置的技术
- 传统的SLAM技术：
  - 滤波算法
  - 非线性优化技术
- 引入深度学习后的SLAM：
  - 新方法
  - 新任务
  - 新表示



[1] Durrant et al. Simultaneous localization and map: part I. 2006 RAM
[2] Taketomi e al. Visual SLAM algorithms: A survey from 2010 to 2016. 2017 IPSJ

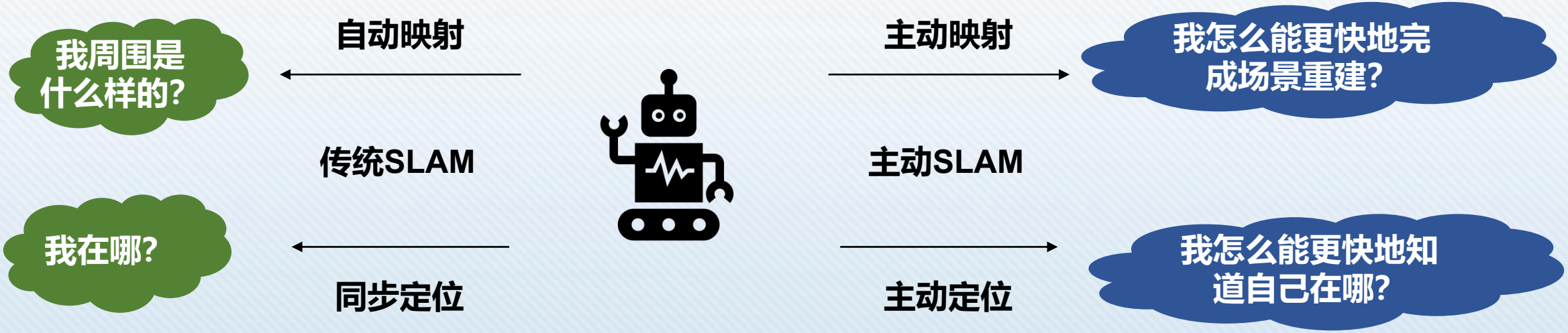
场景重建-新方法

- 将深度学习集成到SLAM
- 用深度学习方法替换传统的SLAM模块
  - 特征提取
  - 深度估计
- 在传统SLAM上加入语义信息
  - 图像语义分割
  - 语义地图构建
- 基于深度学习的新方法主要为SLAM领域的**自我优化**或**迭代**，很少有方法从具身智能的角度出发

[1] DeTone, et al. Toward geometric deep slam. 2017 arXiv
[2] Tateno, et al. Cnn-slam: Real-time dense monocular slam with learned depth prediction. 2017 CVPR
[3] Li, et al. Undeepvo: Monocular visual odometry through unsupervised deep learning. 2018 ICRA

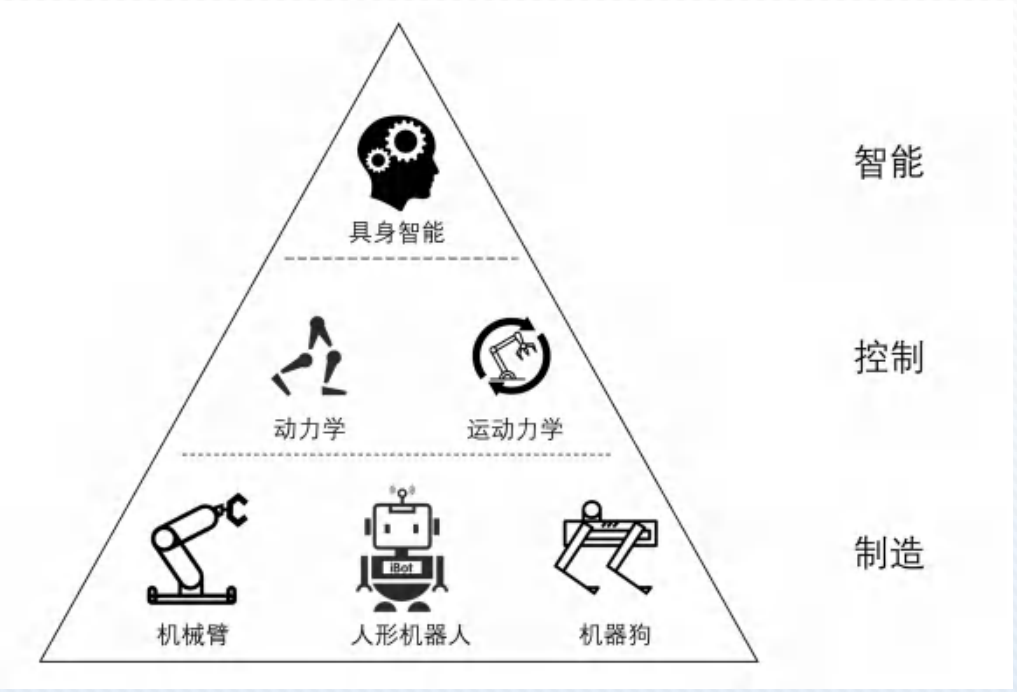
场景重建-新任务

- 传统SLAM
  - 机器人由人类控制，或使用预定义的航点，或基于给定的路径规划算法进行导航
- 主动SLAM
  - 机器人可以自主行动，以实现更好的场景重建和定位
  - **主动映射**：机器人自主选择下一步视点，以获得更好的观察，进行环境探索
  - **主动定位**：机器人自主规划路径，旨在解决模糊位置定位，而不仅仅是导航



具身感知小结二

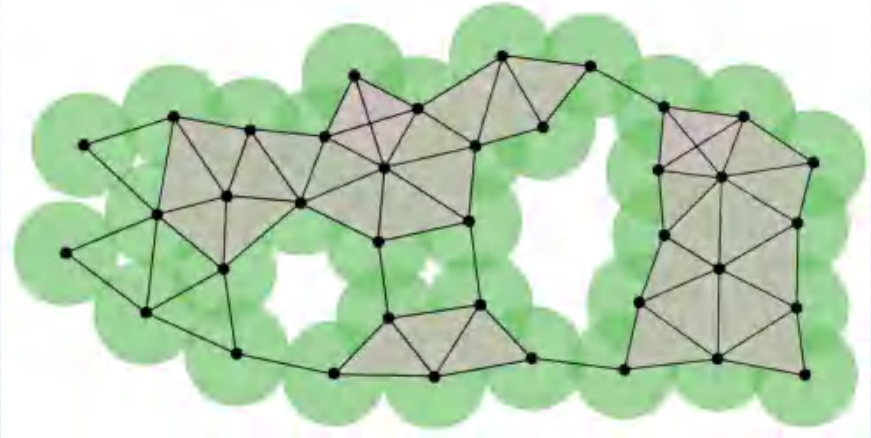
- 具身智能与非具身的智能，乃至其他领域，例如CV、NLP、CG（计算机图形学）、Robotics、Control，核心区别在哪里？
- 3D数据？机器人任务中的深度学习技术？



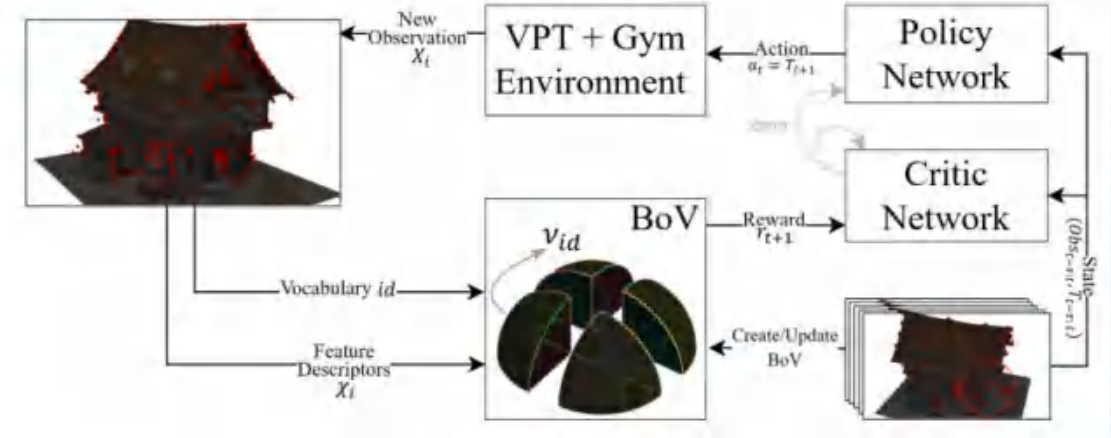
- 在于行为智能，在于交互，在于告诉机器人怎么动
- 此处的交互具体指 空间中一条7自由度的轨迹
- 操作铰接物体、主动探索、主动定位
- 多模态大模型和文本大模型没见过轨迹数据，如果将轨迹数据压缩为大模型，或许有更智能的交互效果

主动映射

- 主动映射任务，即下一个最佳视图（Nest Best View）任务，旨在找到更好的观测视点或更有效的观测策略
- 视图的评估标准：信息增益、机器人运动成本和场景重建的质量



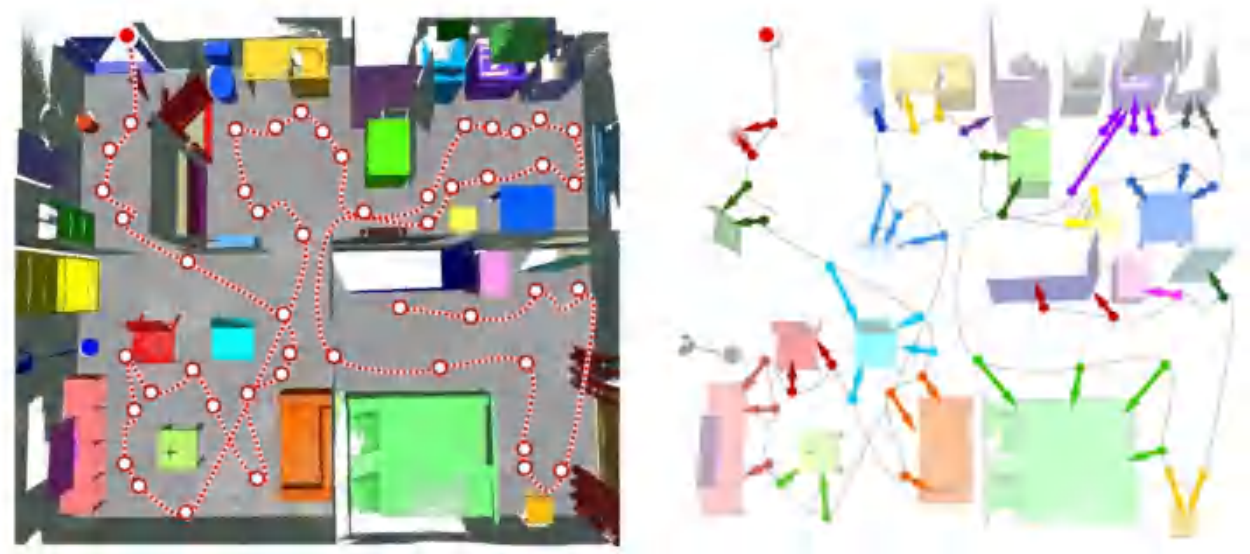
基于拓扑的信息增益度量确定下一个最佳视图



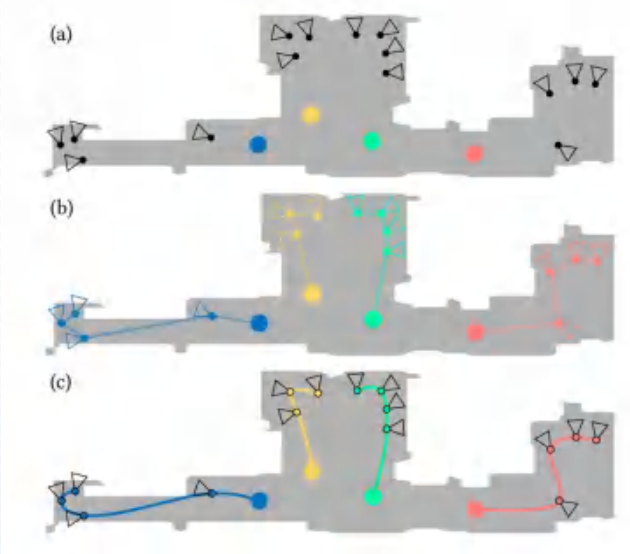
RL方法，目的是识别最大化其场景记忆变化的视图。核心思想是帮助智能体记住尽可能多的不可见的视觉特征

[1] Collander, et al. Learning the next best view for 3d point clouds via topological features. 2021 ICRA
[2] Gazani, et al. Bag of views: An appearance-based approach to next-best-view planning for 3d reconstruction. 2023 RAL

主动映射



将 NBV 任务与次优对象 （NBO） 任务集成，选择感兴趣的对象，确定重建它们的最佳视角

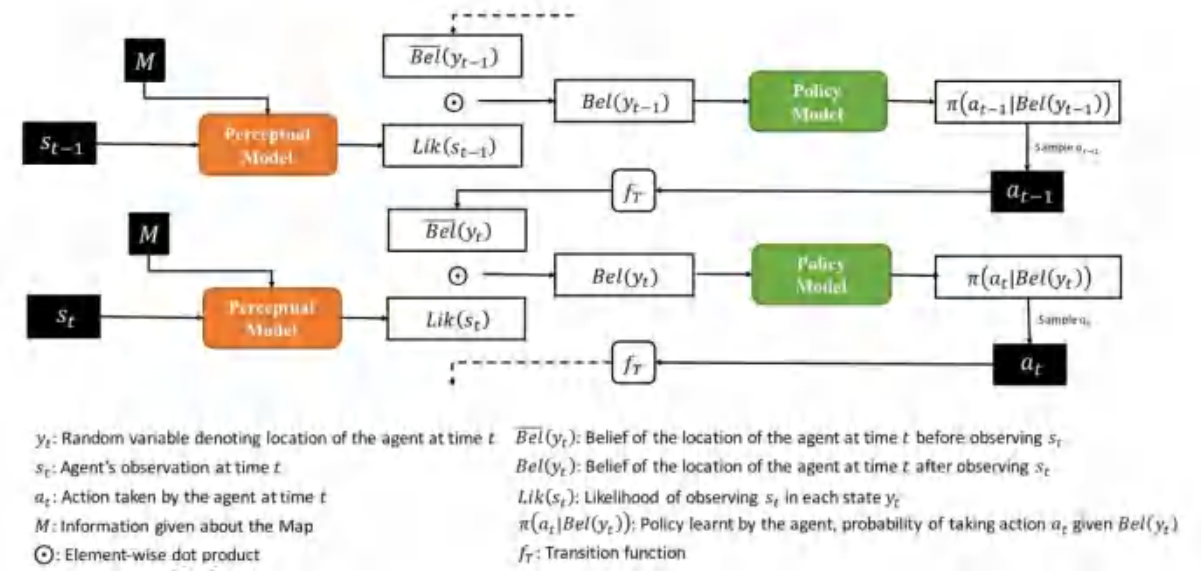


多智能体协作的主动映射

[1] Liu, et al. Object-aware guidance for autonomous scene reconstruction. 2018 TOG
[2] Dong, et al. Multi-robot collaborative dense scene reconstruction. 2019 TOG

主动定位

- 主动定位涉及在参考图中规划后续运动路径，以尽量地减轻机器人空间方向的模糊性
- 传统的定位算法与动作选择无关
- ANL（Active neural localization）通过端到端强化学习（包括感知模块和策略模块）最大化移动后的“后验概率”（可理解为位置的置信度），从而最小化定位所需的步骤数量

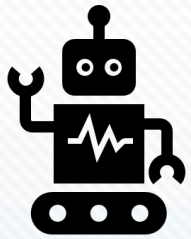


[1] Chaplot, et al. Active neural localization. 2018 arXiv

场景重建-新表示

- SLAM领域亦在探索几何外观等经典属性之外的环境表示，旨在对层次结构、功能、动态和语义等属性进行建模
- 主要的表示形式：
  - 拓扑模型
    - 描述环境连通性的拓扑图
  - 场景图
    - 将环境建模为有向图，其中节点表示对象或位置等实体，边缘表示这些实体之间的关系

场景理解



环境 → 原始信息 → 场景表示 → 场景认知

抽取 ↓ 构建 ↑ 形成 ↑

场景信息获取 → 场景重建 → 场景理解

易利用：视觉信息 / 激光信息 / 雷达信息

难利用：听觉 / 触觉 / 化学 / 红外 / 超声 / …

场景重建：传统 SLAM

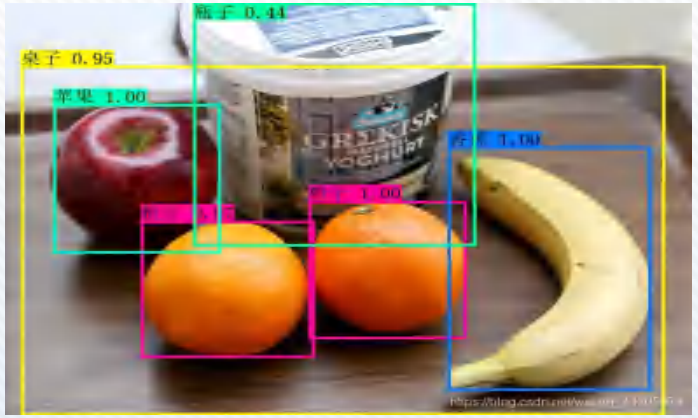
- 新方法：深度学习
- 新任务：主动映射、主动定位
- 新表示：拓扑结构、场景图

场景理解：对象 → 实现

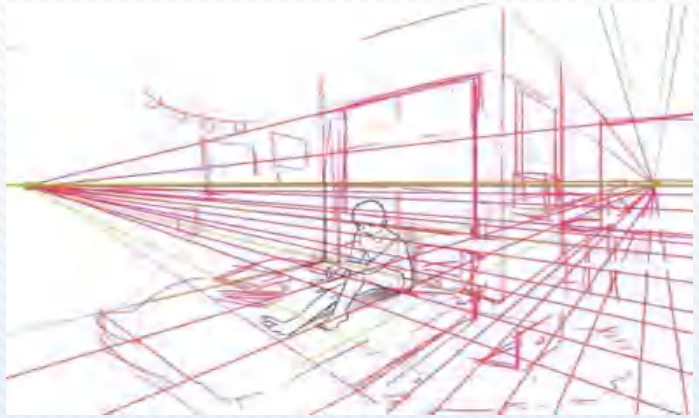
- 单个物体 → 物体识别
- 多个物体 → 空间关系推理
- 场景图像 → 时序变化检测

场景理解

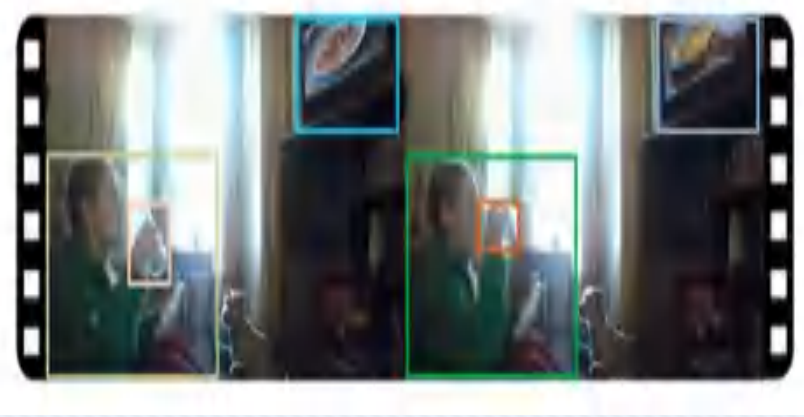
- 理解场景信息是场景感知的重要组成部分
- 高效的理解过程（例如分割、识别和检测）为智能体理解复杂环境
- 场景理解不仅包括物体的识别，还包括物体之间的空间关系和场景帧之间的时间变化



物体识别



空间关系推理



时序变化检测

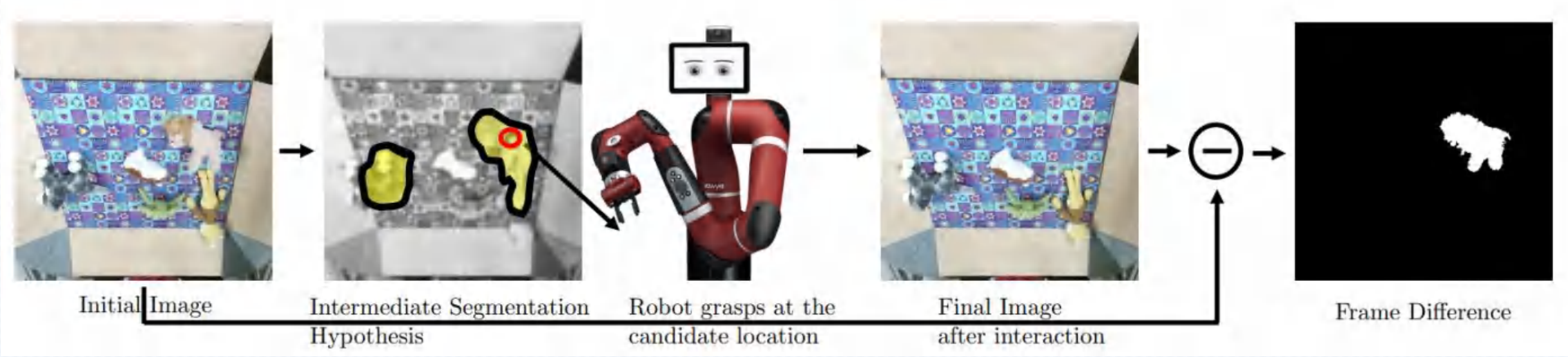
物体识别

- 常规的、主流的物体识别方法：
  - YOLO
  - MASK RCNN
  - ResNet
- 这些方法的局限性：难以利用机器人与环境的交互能力
- 具身智能的物体识别：
  - 物理交互：通过移动（触碰）物体实现更好的物体识别
  - 更改视点：通过移动改变自身在场景中的位置，结合多视角信息实现更好的物体识别

[1] Redmon, et al. You only look once: Unified, real-time object detection. 2016 CVPR
[2] He, et al. Mask r-cnn. 2017 ICCV
[3] He, et al. Deep residual learning for image recognition. 2016 CVPR

物体识别-物理交互

- Pathak et al. 利用简单的对象操作来协助实例分割和对象识别

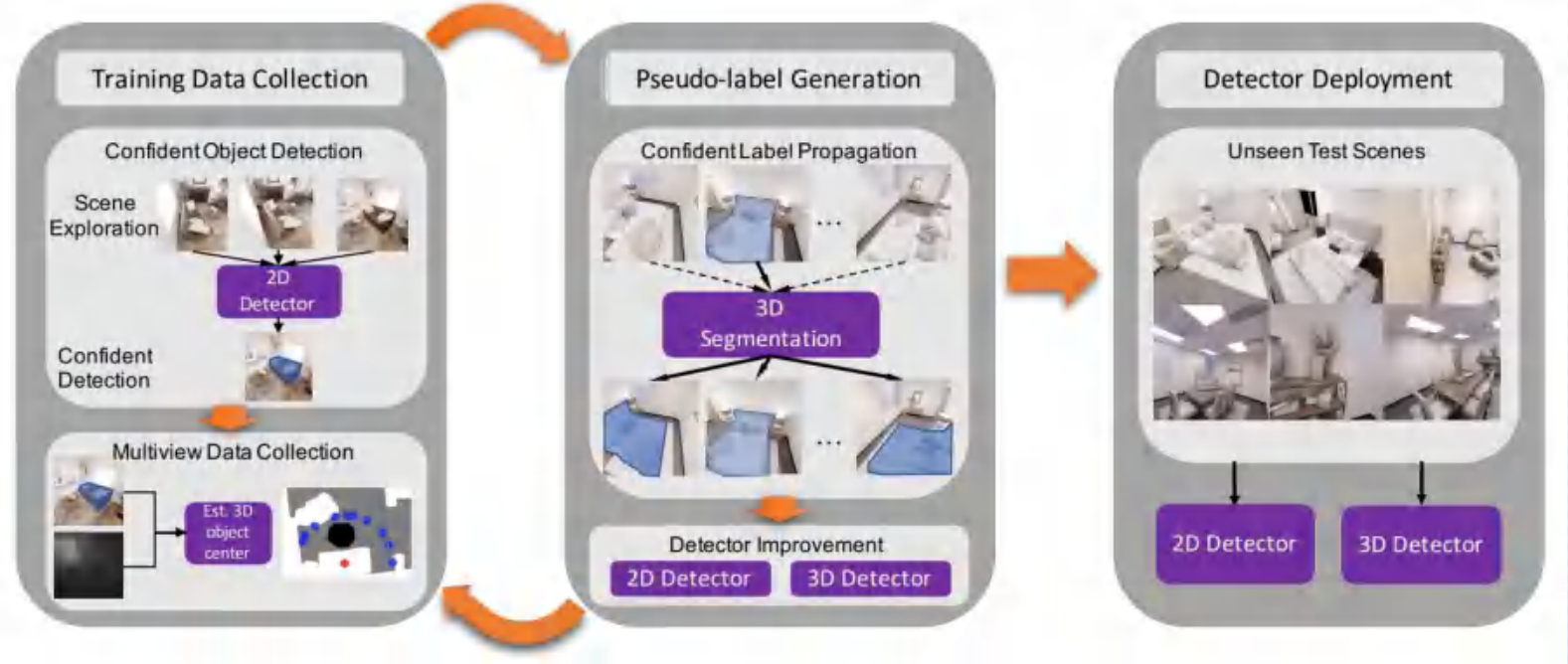


通过对象操作实现实例分割的流程

[1] Pathak, et al. Learning instance segmentation by interaction. 2018 CVPR

物体识别-更改视点

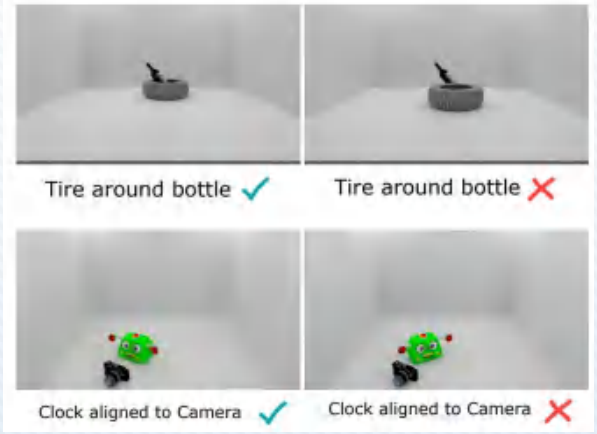
- Seeing by Moving模仿人类“通过绕着同一物体走动来获取多个观察视角”的策略，使机器人能够通过自主运动获取单个物体的多视图数据
- 该方法从人类的演示中学习移动策略，而其他方法则依靠强化学习来学习行为策略



[1] Fang, et al. Move to see better: Self-improving embodied object detection. 2020 arXiv

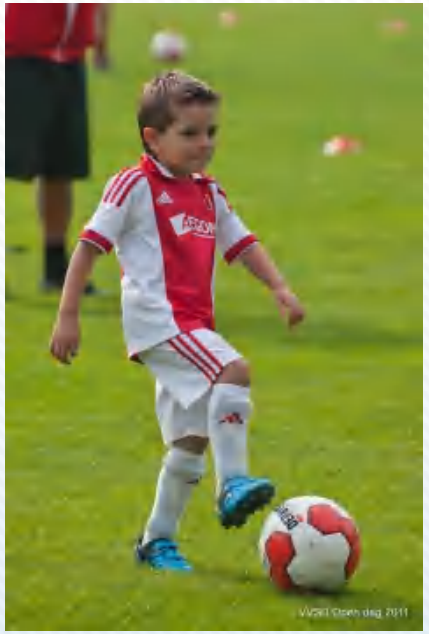
空间关系推理

- 空间关系主要涉及视觉检测和关系推理
- 相关的数据集以及空间关系推理的基准benchmark：
  - Rel3d
  - Spatialsense
  - open images



Rel3d

Spatialsense

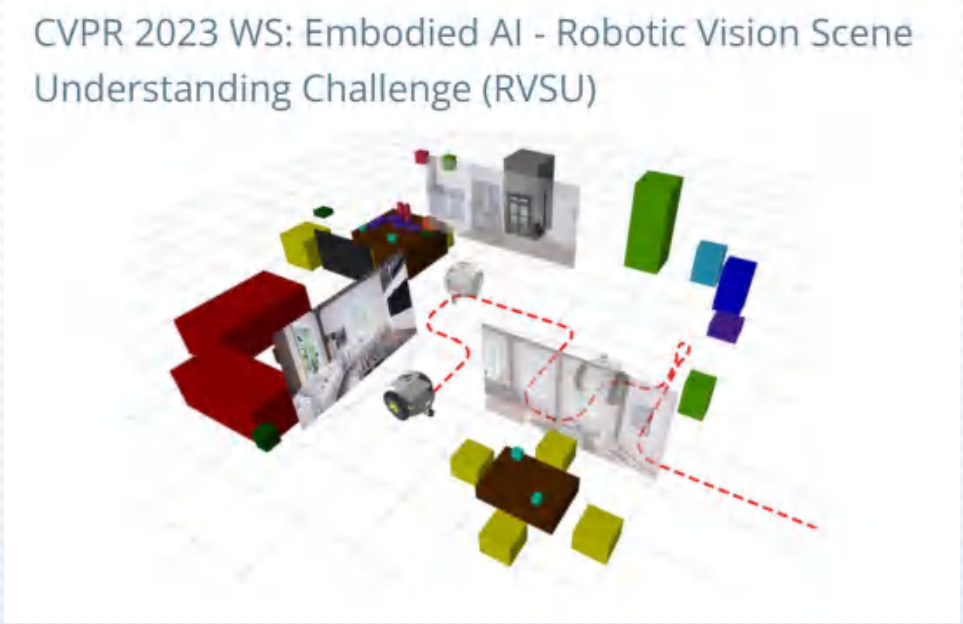


open images

[1] Goyal, et al. Rel3d: A minimally contrastive benchmark for grounding spatial relations in 3d. 2020 NIPS
[2] Yang, et al. Spatialsense: An adversarially crowdsourced benchmark for spatial relation recognition. 2019 ICCV
[3] Kuznetsova, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. 2020 IJCV

时序变化检测

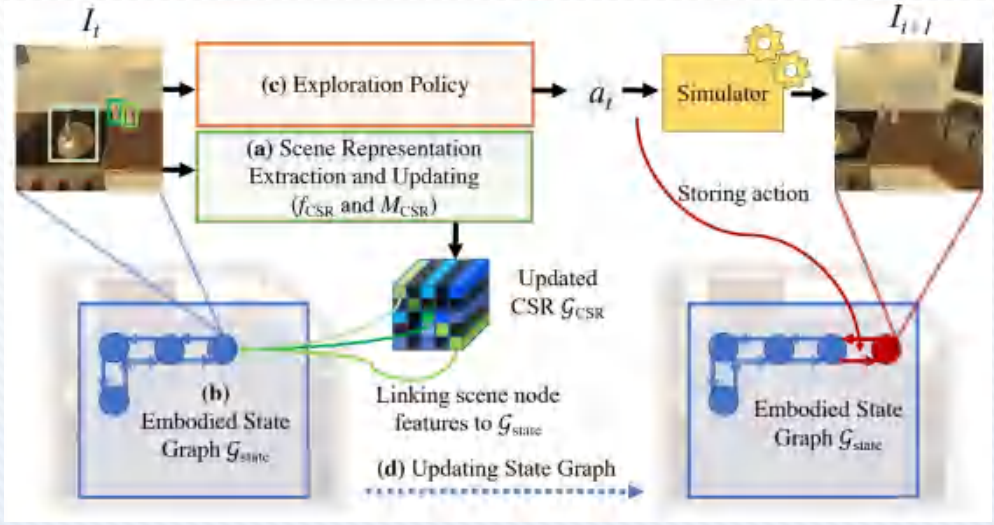
- 场景变化检测：一个机器人在两个不同的时间探索环境，并识别它们之间的任何物体变化。物体变化包括环境中添加和移除的物体
- 常用数据集：
  - robotic vision scene understanding challenge
  - ChangeSim
  - VL-CMU-CD
  - PCD



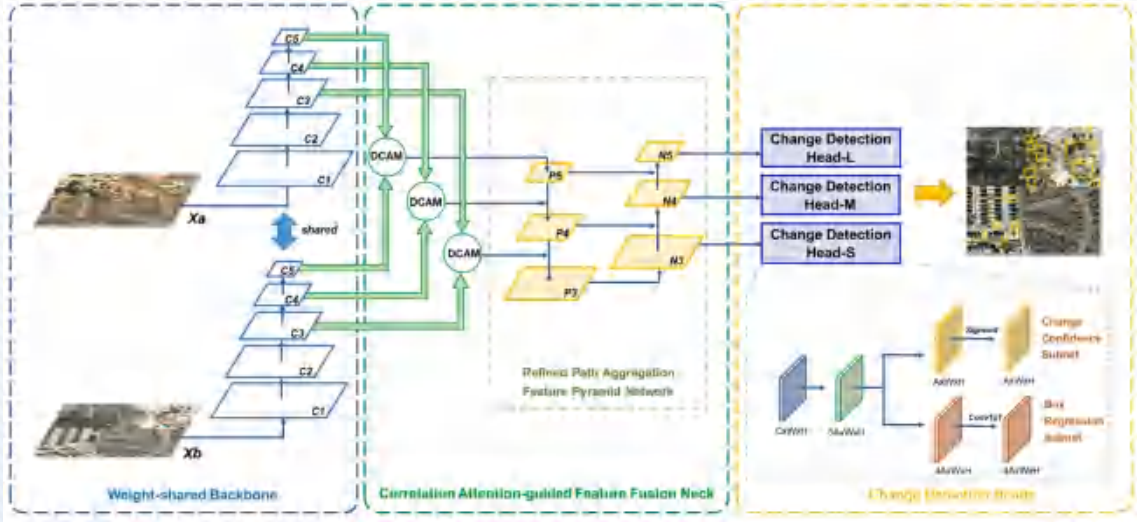
[1] Hall, et al. The robotic vision scene understanding challenge. 2020 arXiv
[2] Park, et al. Changesim: Towards end-to-end online scene change detection in industrial indoor environments. 2021 IROS
[3] Prabhakar, et al. Cdnet++: Improved change detection with deep neural network feature correlation. 2020 IJCNN
[4] Sakurada, et al. Weakly supervised silhouette-based semantic scene change detection. 2020 ICRA

场景变化检测

- CSR主要针对具身导航任务，智能体在移动穿越场景时跟踪物体，相应地更新表示，并检测房间配置的变化
- DCA-Det实现面向物体级别的变化检测



CSR框架图

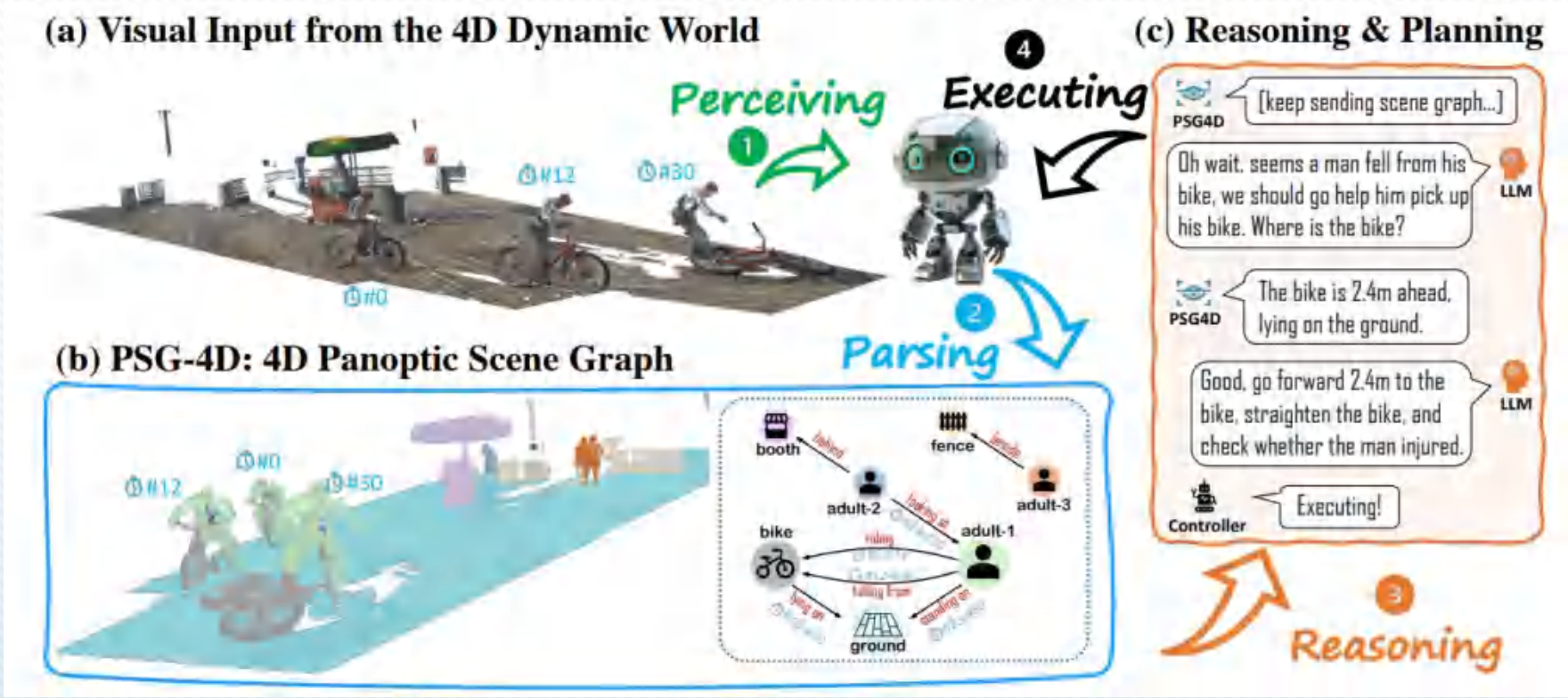


DCA-Det框架图

[1] Gadre, et al. Continuous scene representations for embodied ai. 2022 CVPR
[2] Zhang, et al. Object-level change detection with a dual correlation attention-guided detector. 2021 ISPRS

场景动态感知

- 4D全景场景图（PSG-4D），放弃了“动态”的概念，将时间视为场景表示中的变量，作为第四维度纳入现有的3D场景图中。这种新的表现形态适用于场景预测和动态场景理解



[1] Yang, et al. 4d panoptic scene graph generation. 2024 NIPS

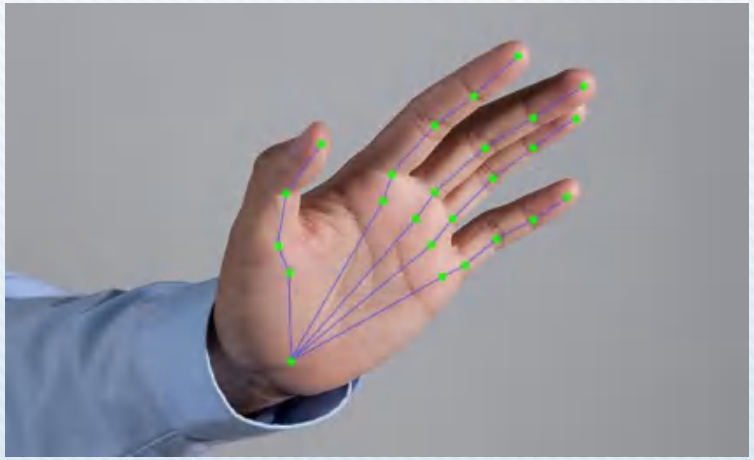
1-3 行为感知

行为感知

- 不同于对物体、场景的感知，对人的感知需要人的行为，包括：
  - 手势识别
  - 身体位姿识别
  - 人类行为理解
- 机器人对人的行为感知有助于人机交互应用：
  - 社交导航
  - 自动驾驶
  - 人机协作装配

手势识别

- 手势识别是识别图片中人体手势的类别，一般以分类任务的形式出现
- 手势识别的一般流程：
  - 使用RGB相机或RGBD相机获取图片
  - 手势的分割与检测：基于肤色、轮廓、深度信息等信息检测图中手势区域和手的关节点
  - 手势识别：在分割检测结果的基础上进行手势分类



人体姿态检测

- 人体姿态检测需要预测2D图像或3D数据中人体的关节点
- 单人的姿态检测，可以使用回归的方法或基于热图的方法
  - 回归：直接基于图片预测关节点位置
  - 热图：预测每个像素点属于某个关节的概率，进而基于概率决定关节位置
- 多人的位姿检测，可以分为自顶向下和自底向上
  - 自顶向下：识别图中人体后分别进行姿态估计
  - 自底向上：首先检测图中所有关节点，然后进行组合

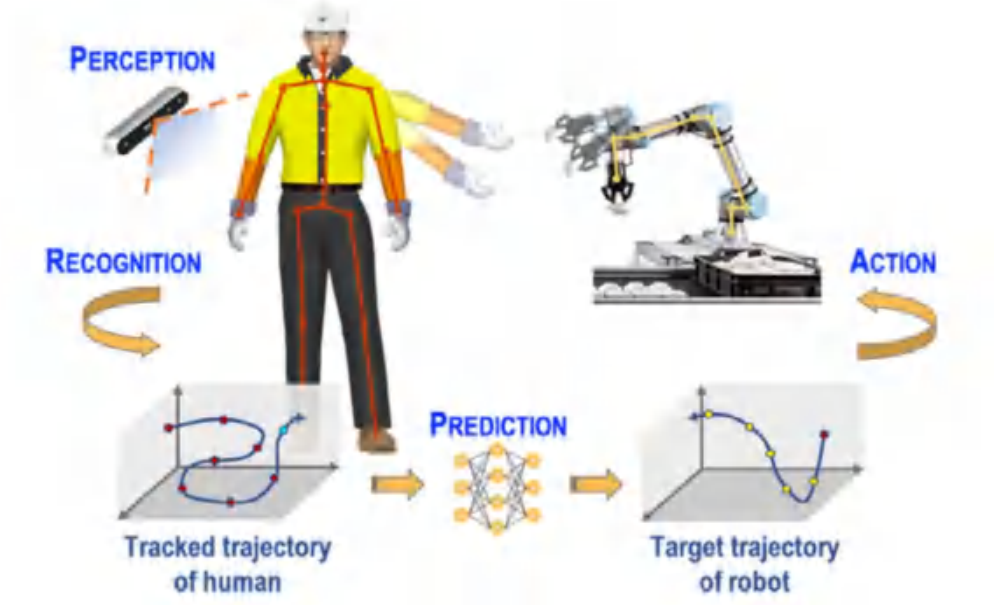
社交导航机器人 & 自动驾驶

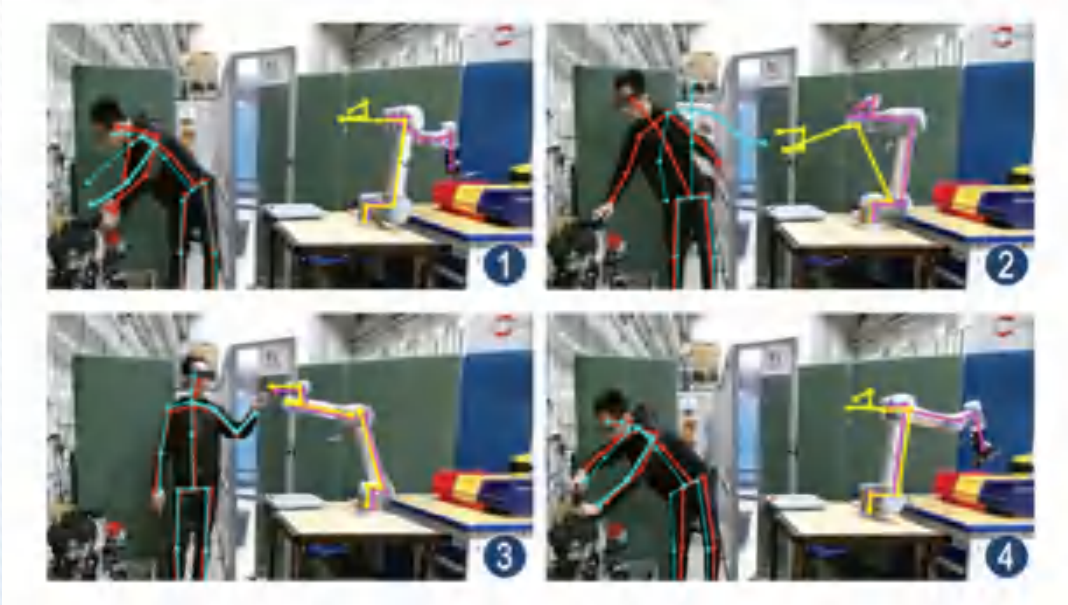
- 人体姿态估计的结果可以用于预测人类下一步动作，这有助于机器人进行决策
- 社交导航机器人基于人体位姿预测人类下一步方向，从而选择移动方向
- 自动驾驶决策时同样需要预测人类移动轨迹

[1] Narayanan et al. ProxEmo: Gait-based Emotion Learning and Multi-view Proxemic Fusion for Socially-Aware Robot Navigation. 2020 IROS

人机协作装配场景下防碰撞

- 除预测人类移动轨迹用于社交导航场景和机器人场景外，在工业场景中人机协作进行装配任务同样需要预测人类未来行为轨迹，以免机器人和人发生碰撞





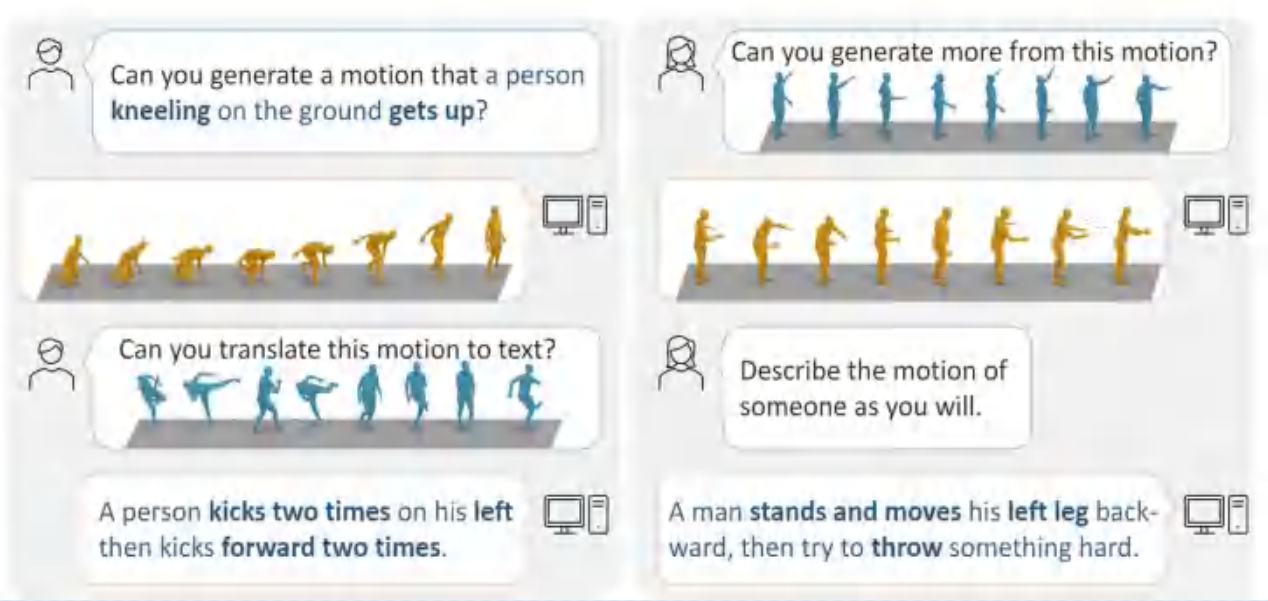
[1] Zhang et al. Recurrent neural network for motion trajectory prediction in human-robot collaborative assembly. 2020 CIRP.

人类行为理解

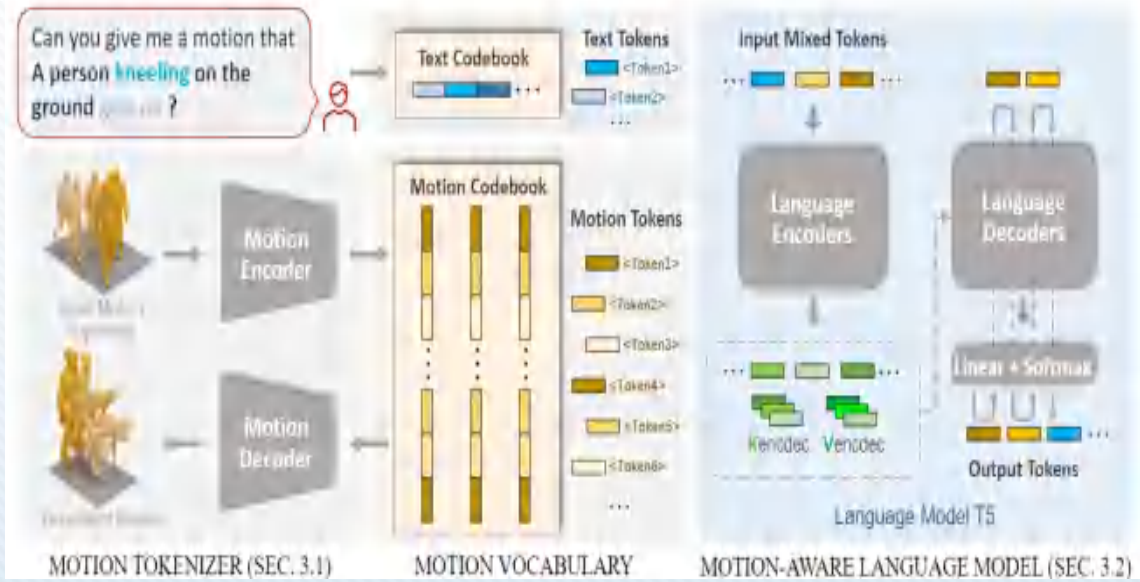
- 人类行为理解即通过检测姿势、运动和环境线索来推断其正在进行的行为
- 该领域超越了对基本动作的识别，还包括对复杂行为的分析
  - 人物交互
  - 多人协作
  - 动态环境中的自适应行为
- 最近的进展侧重于通过更深入的语义理解来建模这些行为

人类行为理解：统一的动作-语言生成预训练模型

- 统一的动作-语言生成预训练模型MotionGPT
  - 将人类动作视为一种外语，引入自然语言模型进行动作相关生成
  - 功能包括：给定文本生成动作，给定动作生成文本，动作扩增，文本动作描述生成



MotionGPT的演示

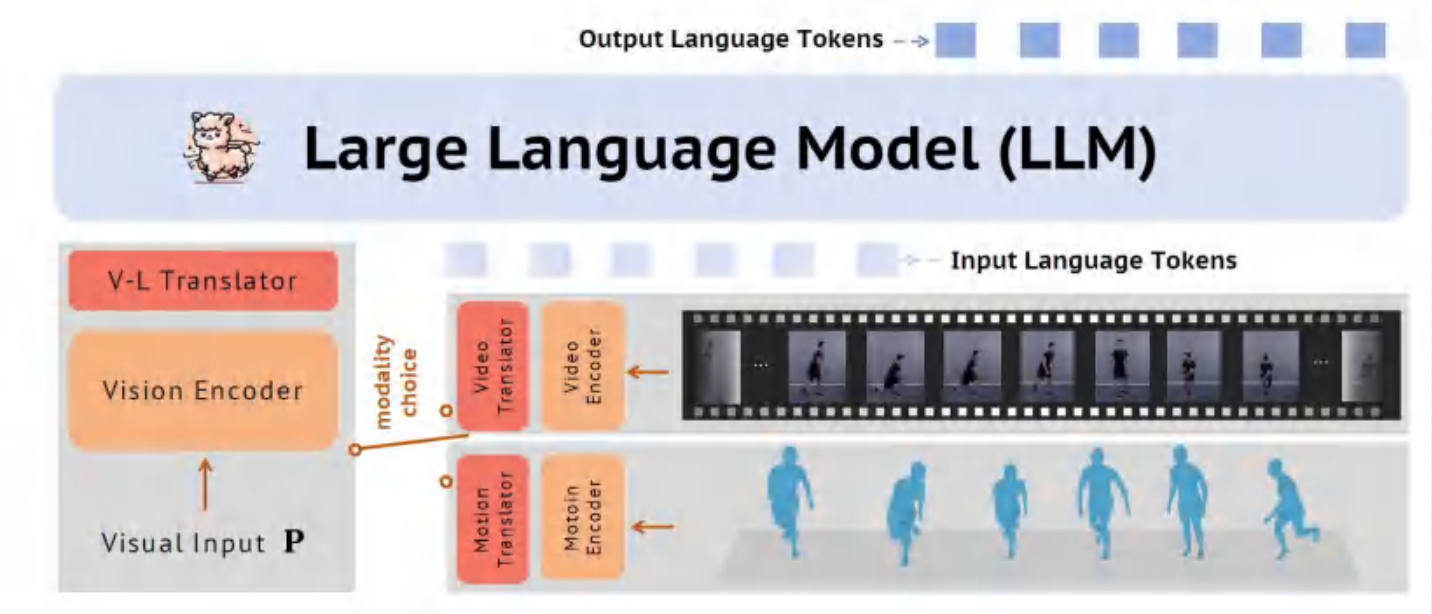


MotionGPT的方法总览

[1] Jiang et al. Motiongpt: Human motion as a foreign language. 2024 NIPS

人类行为理解：可理解人类动作和视频的大语言模型

- 可理解人类动作和视频的大语言模型MotionLLM
  - 收集并构建了一个名为MoVid的大规模数据集和MoVid-Bench的基准测试
  - 提出了一个结合视频和动作数据的统一框架，通过大语言模型来理解人类行为



MotionLLM的基本架构

MotionLLM的两阶段训练

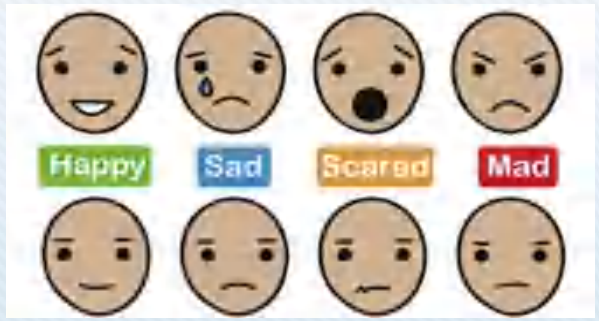
[1] Chen L H et al. MotionLLM: Understanding Human Behaviors from Human Motions and Videos. 2024 arXiv preprint arXiv:2405.20340

1-4 表达感知

表达感知概述

- 机器人想获取人类的情感和意图，可以通过人的：
  - 面部表情
  - 语音
  - 上述两种模态信号的结合

面部表情、语音 → 情感、意图



表达感知研究意义

- 表达感知的研究意义：
  - 增强任务协作能力，从而提升机器人在**人机交互**中的**自然性和有效性**
  - 更准确地感知用户的情感变化与意图，从而显著提高用户**体验和满意度**
  - 可能应用的实际场景：陪伴老年人、智慧家居、工业协作等

陪伴机器人

智慧家居

工业机器人

面部情感感知

- 面部表情数据采集一般是通过摄像头设备进行采集
- 特征提取
  - 如几何特征（关键点坐标）、纹理特征（局部二值模式，LBP）和动作单元（Action Units, AU）等
- 面部情感识别的主要挑战
  - **复杂环境**下的面部情感感知
  - 可能包括光照变化、姿态变化、遮挡和不同的背景场景等，对准确性和鲁棒性要求更高

[1] Ma F et al. Facial expression recognition with visual transformers and attentional selective fusion. 2021 IEEE Transactions on Affective Computing

面部情感感知：复杂环境下的面部情感感知

- **Visual Transformers与特征融合**
  - 针对在**野外**（即非实验室控制环境）中的FER任务，能够处理遮挡、不同的头部姿势、面部变形和运动模糊等复杂情况
- **区域注意力网络RAN**
  - 旨在解决现实世界中FER的**遮挡鲁棒性和姿态不变性**问题
  - 构建了若干具有姿态和遮挡属性的**野外FER数据集**，解决了对应领域数据集缺乏的情况
- **边缘AI驱动（Edge-AI-driven）的FER框架**
  - 该框架可以在**低功耗设备**上实现实时的面部表情识别，确保在**有限的计算资源和能源消耗**下，仍能保持高精度
  - 这对于**智能穿戴设备、智能手机和远程医疗**等应用场景尤为重要

[1] Ma F et al. Facial expression recognition with visual transformers and attentional selective fusion. 2021 IEEE Transactions on Affective Computing
[2] Wang K et al. Region attention networks for pose and occlusion robust facial expression recognition. 2020 IEEE Transactions on Image Processing
[3] Wu Y et al. Edge-AI-driven framework with efficient mobile network design for facial expression recognition. 2023 ACM Transactions on Embedded Computing Systems

语音情感感知&多模态情感感知

- 语音情感感知：
  - 从人类的语音信号中提取音高、音调、节奏、音色等特征作为输入
  - 表示声音频率内容的图像形式：梅尔频谱图（Mel-spectrogram）及其梅尔频率倒谱系数（MFCC）
  - 通过理解说话者的情感状态，系统能够做出更加人性化和智能化的响应
  - 在客服机器人、智能助理、心理健康监测等领域有广泛的应用
- 多模态情感感知：
  - 通过结合多种不同类型的数据源，如语音、面部表情、身体语言、文本等，来识别人类的情感状态
  - 在本节特指结合人类的**面部表情和语音**来进行情感感知
  - 相比单一模态的情感识别，多模态方法能够从不同维度捕捉情感特征，提高识别的**准确性和鲁棒性**

意图推断

- 表达感知不仅可以帮助机器人获知用户的情感变化，还可以辅助机器人进行对人类**意图的推断**
- 意图推测的精确度对于提升机器人在人机交互中的表现、提高用户体验和满意度具有重要意义
- 未来的机器人一定是能够“理解”人的想法的机器人

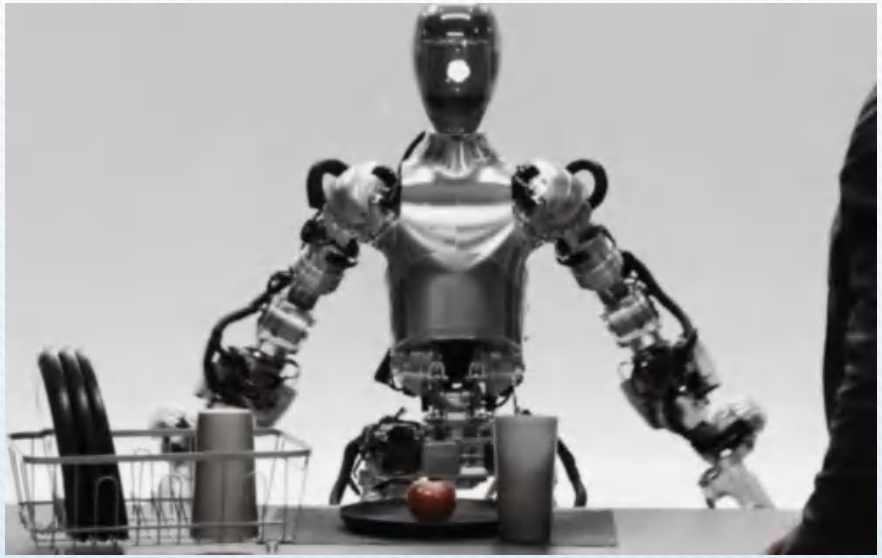


Figure 01机器人（基于OpenAI大模型）听到人类说“我饿了”之后，准确领悟到了人类的意图，选择了苹果放到盘子中

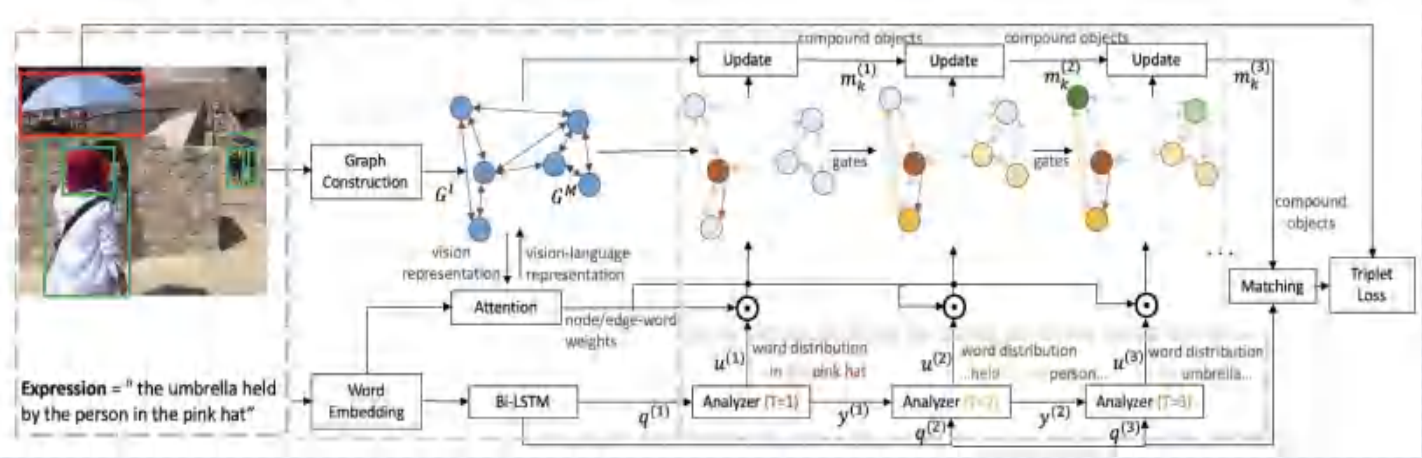
内嵌谷歌PaLM-SayCan模型的机器人正在厨房内帮人类拿零食，该模型能够帮助机器人更好地理解自然语言并执行复杂任务

指代表达

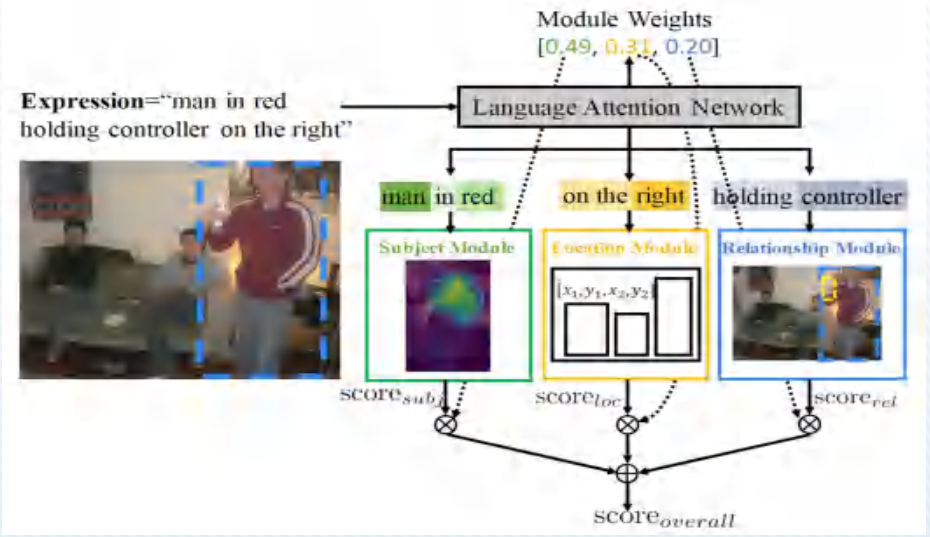
- 指代表达是指在特定上下文中生成**描述性语言或表达**，以便清晰地指代某个特定对象或实体
- 意图推断与指代表达之间的关系：
  - 指代表达的理解是意图推测的重要应用形式之一
  - 在理解指代表达的过程中，机器人需要推测用户的意图，以确定用户所指代的具体对象或位置
- 指代表达的研究方向
  - 指代表达的生成（从人类的角度出发）
  - 指代表达的理解（从机器人的角度出发）
- 最常见的指代表达形式为接收人类的**语言指令**来完成人类想要其完成的操作

指代表达的理解

- 动态图注意力网络DGA
  - 首次从语言驱动视觉推理的角度探索了引用表达式理解问题，显著提高了识别复杂引用表达式的能力
- 模块化注意力网络MAttNet
  - 通过模块化框架关注相关单词和视觉区域，并动态计算整体匹配分数，显著提高了机器人的理解性能。



指代表达理解的动态图注意网络 (DGA) 的整体架构



模块化注意网络(MAttNet)的示意图

[1] Yu L et al. Mattnet: Modular attention network for referring expression comprehension. 2018 Proceedings of the IEEE conference on computer vision and pattern recognition

[2] Yang S et al. Dynamic graph attention for referring expression comprehension. 2019 Proceedings of the IEEE/CVF International Conference on Computer Vision

1-5 具身感知小结

具身感知

- 具身感知的过程主要包括以下几步：

	感受信息	信息的表示/建模/重建	信息理解	下游任务
物体感知	RGB相机 3D相机 麦克风 触觉传感器 热传感器 …	获取3D物体的几何形状、铰接结构、物理属性信息	信息编码方法	位姿估计、物体抓取、**交互感知**、可供性预测
场景感知		场景重建（**主动探索、主动定位**、场景的表示）	**基于交互的物体检测**、空间关系理解、时序变化检测	
行为感知		主要基于2D图片和深度信息	手势识别、姿态检测、人类行为理解	社交导航、自动驾驶、人机协作
表达感知		主要基于视觉表情和声音	情感识别、意图推断（指代表示）	

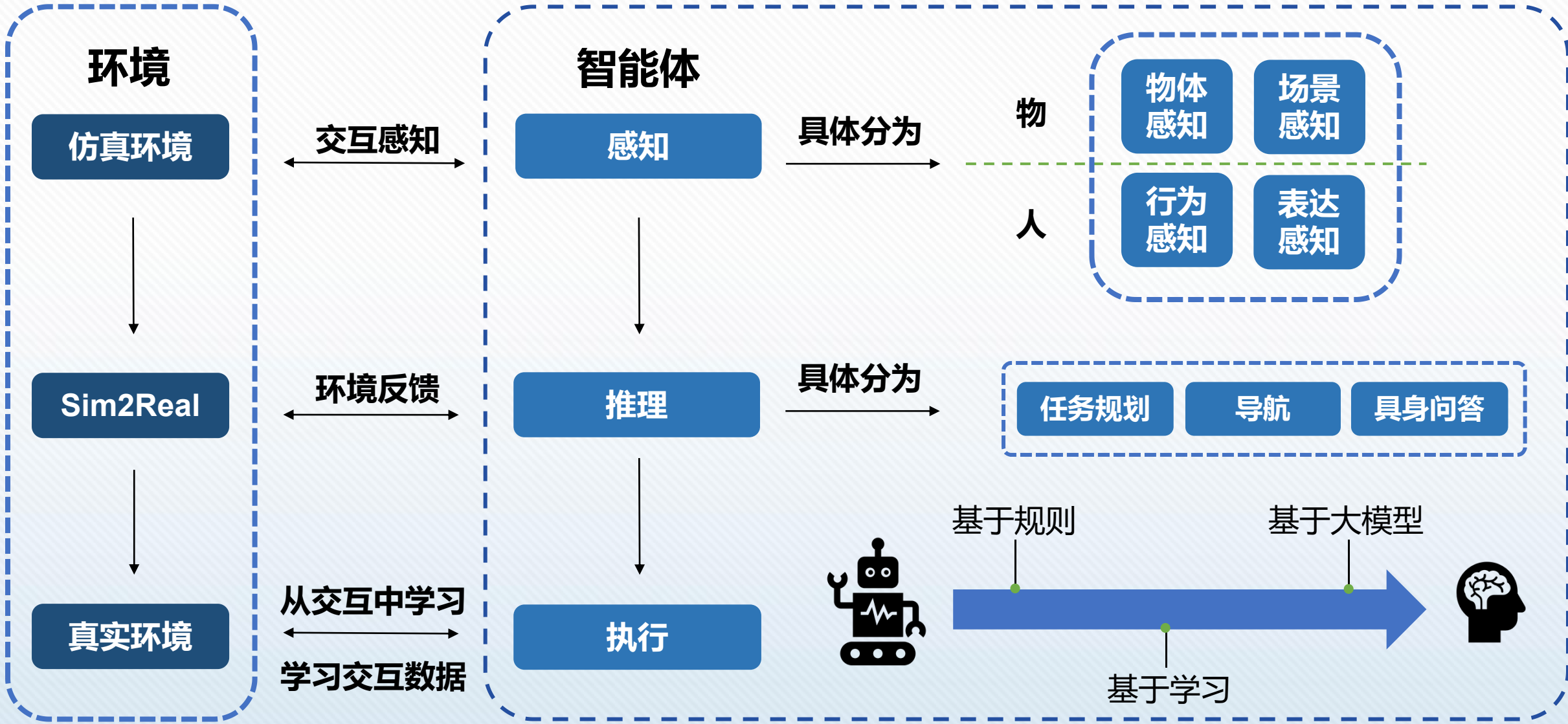
具身感知

- 感知能力强 AND 有一定的推理能力，就可以成为一个很好的机器人落地产品
  - 服务机器人、人机协作场景下机器人、社交导航机器人、环境探索机器人
- 感知能力也可以为抓取、操作等执行任务提供帮助，在端到端执行模型性能达标前，抓取等任务更多依赖感知能力

- 多模态大模型处理语言、2D图片、3D数据都没有超出我们的想象。但能处理交互数据的大模型还没有出现在地平线上
  - 在交互感知、主动探索等任务中，模型能否zero-shot的给出行为轨迹
- 基于已有大模型，依赖人类先验设计模型结构或训练算法来弥补这个缺陷？人类先验或许不那么有效

2 具身推理

具身智能划分：感知、推理、执行

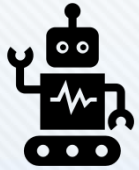


2-1 任务规划

任务规划简介

- 任务规划（Task Planning）是具身智能的核心任务之一（另一个核心是技能学习），将抽象的非可执行人类指令转换为具体的可执行技能
- 完成人类指令只需要两步：人类指令分解为机器人可执行的技能，执行技能

我渴了，可以帮我拿杯水放桌上吗



移动到水瓶附近（MoveTo， bottle）

拿起水瓶（PickUp，bottle）

移动到桌子附近（MoveTo，table）

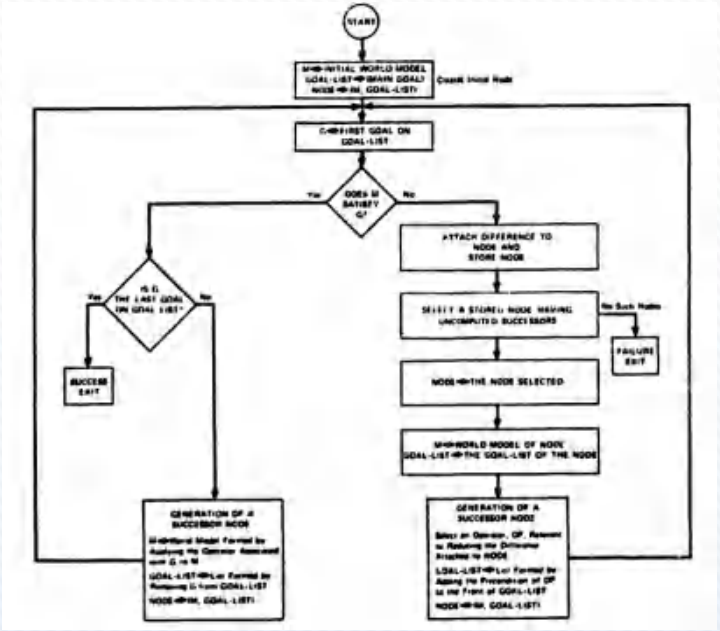
把水放到桌上（Put，bottle，table）

任务规划简介

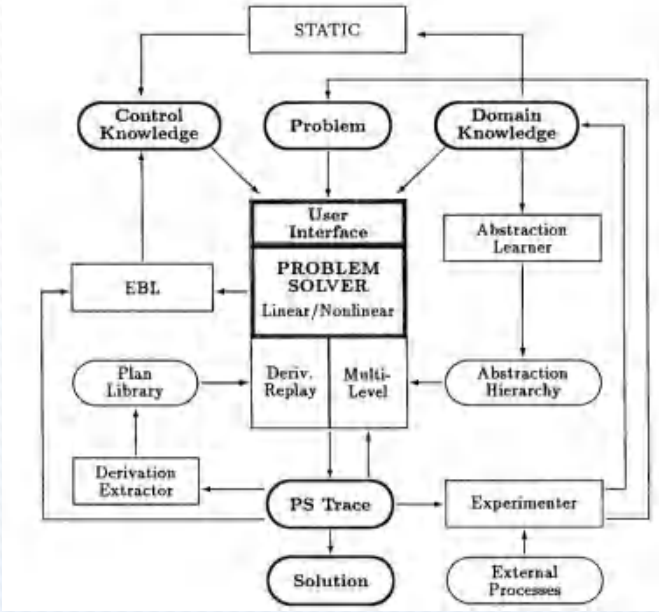
- 任务规划的假设：机器人有一组可执行技能集
  - 潜在含义一：机器人并非万能，技能集之外的不能执行
  - 潜在含义二：需要显式的指定技能，“你先拿捏住”这种自然语言要解析才能被执行
- 与轨迹规划的区别（Motion Planning）
  - 任务规划：将人类指令分解为给定技能集的离散技能序列
  - 轨迹规划：为机器人执行操作技能生成连续7-DOF轨迹
- 难点：
  - 需要理解人类指令
  - 需要理解周围环境
  - 需要理解技能集合

任务规划早期方法：专家系统

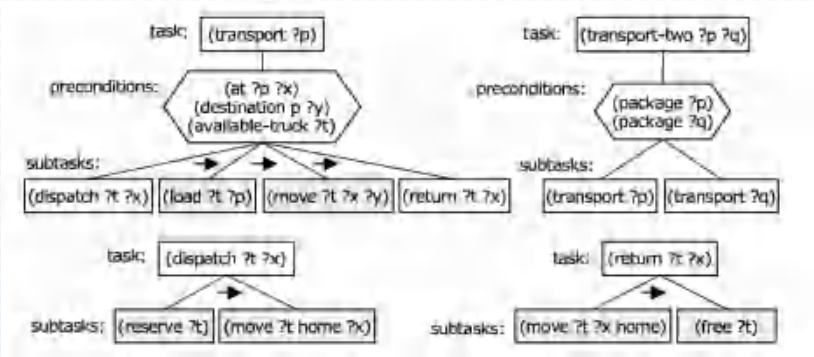
- 专家系统，如STRIPS，PRODIGYAI，SHOP2等，使用不同的形式化建模和搜索策略



STRIPS



PRODIGYAI



SHOP2

[1] Fikes et al. STRIPS: A New Approach to the Application of .Theorem Proving to Problem Solving. 1971 IJCAI.
[2] Carbonell et al. PRODIGY: an integrated architecture for planning and learning. 1991 SIGART Bull.
[3] Au et al. SHOP2: An HTN Planning System. 2003 arXiv.

任务规划早期方法：统一建模语言

- 建模语言，PDDL和ASP：统一规划建模语言，简化问题求解器的开发。

```
(define (domain vehicle)
 (:requirements :strips :typing)
 (:types vehicle location fuel-level)
 (:predicates (at ?v - vehicle ?p - location)
              (fuel ?v - vehicle ?f - fuel-level)
              (accessible ?v - vehicle ?p1 ?p2 - location)
              (next ?f1 ?f2 - fuel-level))

 (:action drive
   :parameters (?v - vehicle ?from ?to - location
                ?fbefore ?fafter - fuel-level)
   :precondition (and (at ?v ?from)
                      (accessible ?v ?from ?to)
                      (fuel ?v ?fbefore)
                      (next ?fbefore ?fafter))
   :effect (and (not (at ?v ?from))
                (at ?v ?to)
                (not (fuel ?v ?fbefore))
                (fuel ?v ?fafter))
 )
)
```

PDDL示例

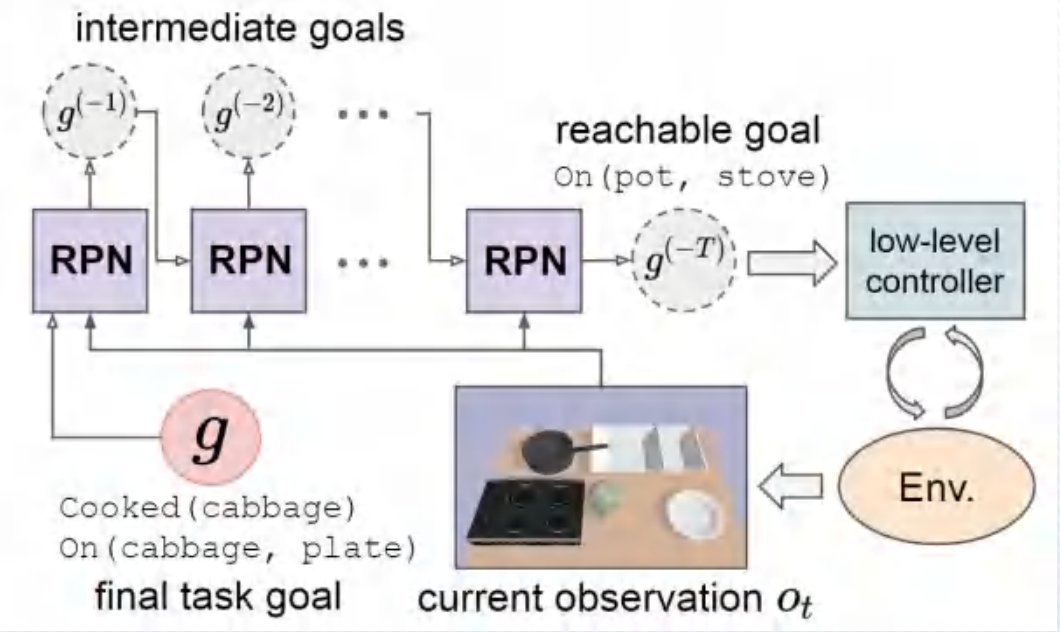
```
{in(X,Y)} :- edge(X,Y).

:- 2 {in(X,Y) : edge(X,Y)}, vertex(X).
:- 2 {in(X,Y) : edge(X,Y)}, vertex(Y).

r(X) :- in(0,X), vertex(X).
r(Y) :- r(X), in(X,Y), edge(X,Y).

:- not r(X), vertex(X).
```

ASP示例

[1] Howe et al. PDDL-the planning domain definition language. 1998 ICAPS
[2] Lifschitz et al. What Is Answer Set Programming?. 2008. AAAI

基于深度学习技术的任务规划：RPN网络

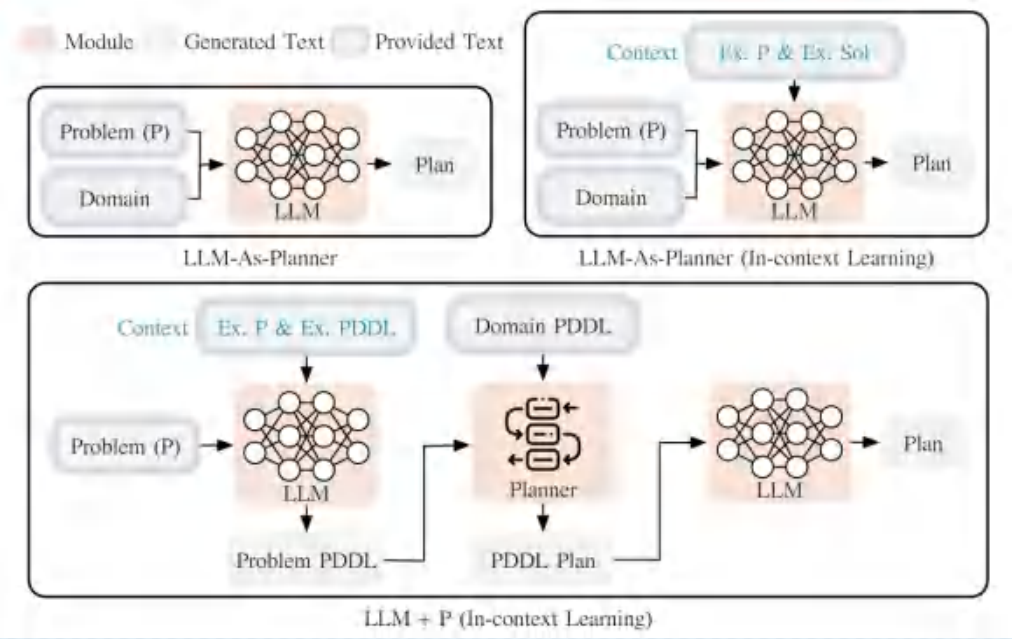
- RPN网络在符号空间进行回归规划，给定最终目标和当前观测，从后向前预测中间目标，直到中间目标对当前状态是可达的
- 使用神经网络进行搜索，而非传统的启发式搜索、深度优先或广度优先搜索（搜索空间随物体数量指数增加）



[1] Xu et al. Regression Planning Networks. 2019. NeurIPS

结合大模型的任务规划：大模型作为转换器

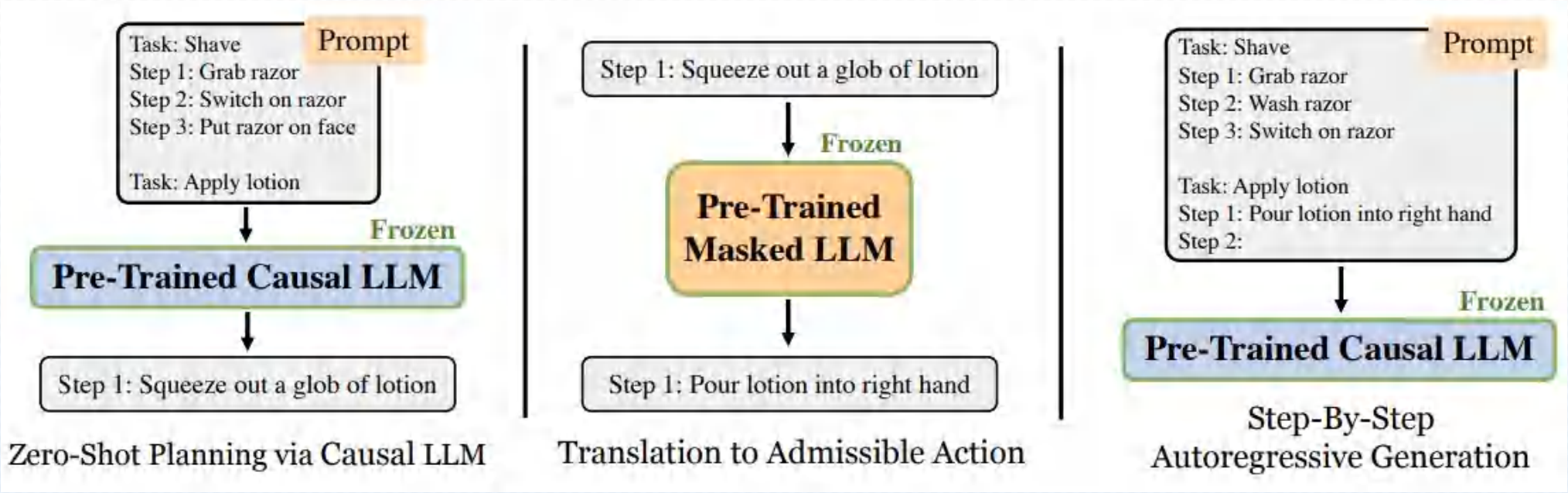
- 大模型作为转换器
- LLM+P，用LLM将状态信息描述成PDDL语言再进行规划，取代以往需要人工针对实际问题书写PDDL语言对任务进行建模



[1] Liu et al. LLM+P: Empowering Large Language Models with Optimal Planning Proficiency. 2023 arXiv.

结合大模型的任务规划：大模型作为规划器

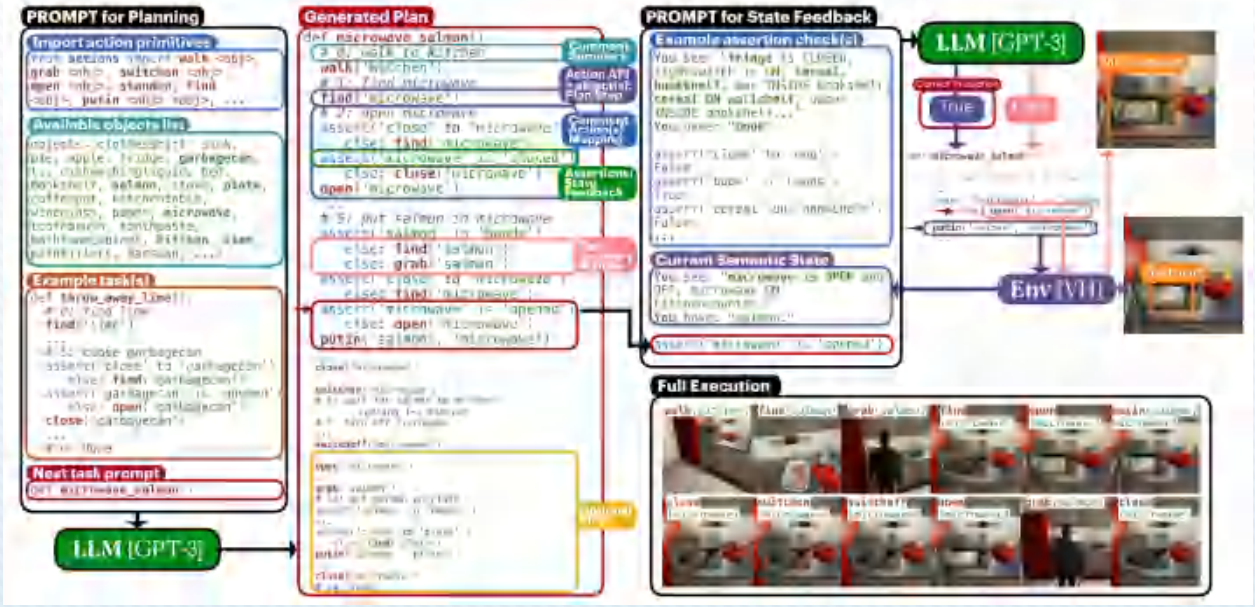
- 大模型作为规划器
- 可以zero-shot进行任务规划



[1] Huang et al. Language Models as Zero-Shot Planners: Extracting Actionable Knowledge for Embodied Agents. 2022. arXiv.

构建用于任务规划的具身智能体

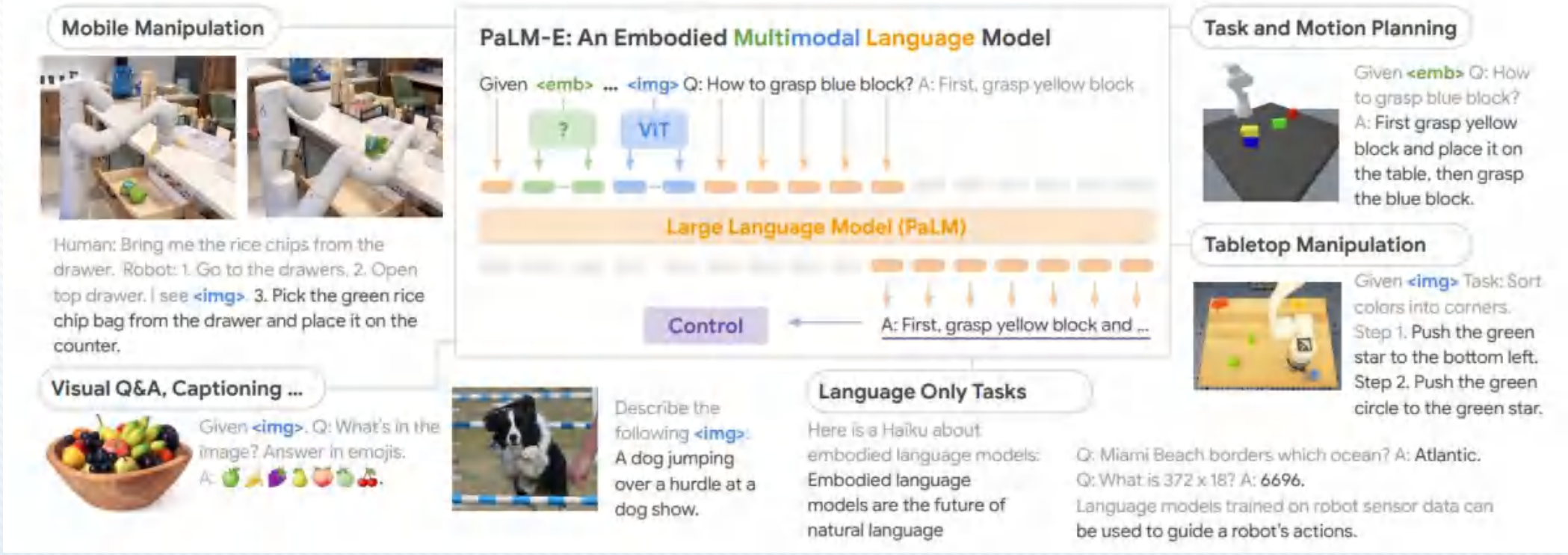
- 结合大模型（多模态大模型）、小模型，以图片、自然语言、代码构建Prompt，设计Pipeline框架，作为具身智能体
- Prompt中还可以加入记忆、反馈等信息。充分利用大模型的图文理解能力、推理能力



[1] Singh et al. ProgPrompt: Generating Situated Robot Task Plans using Large Language Models. 2023 Autonomous Robots.

微调大模型用于任务规划

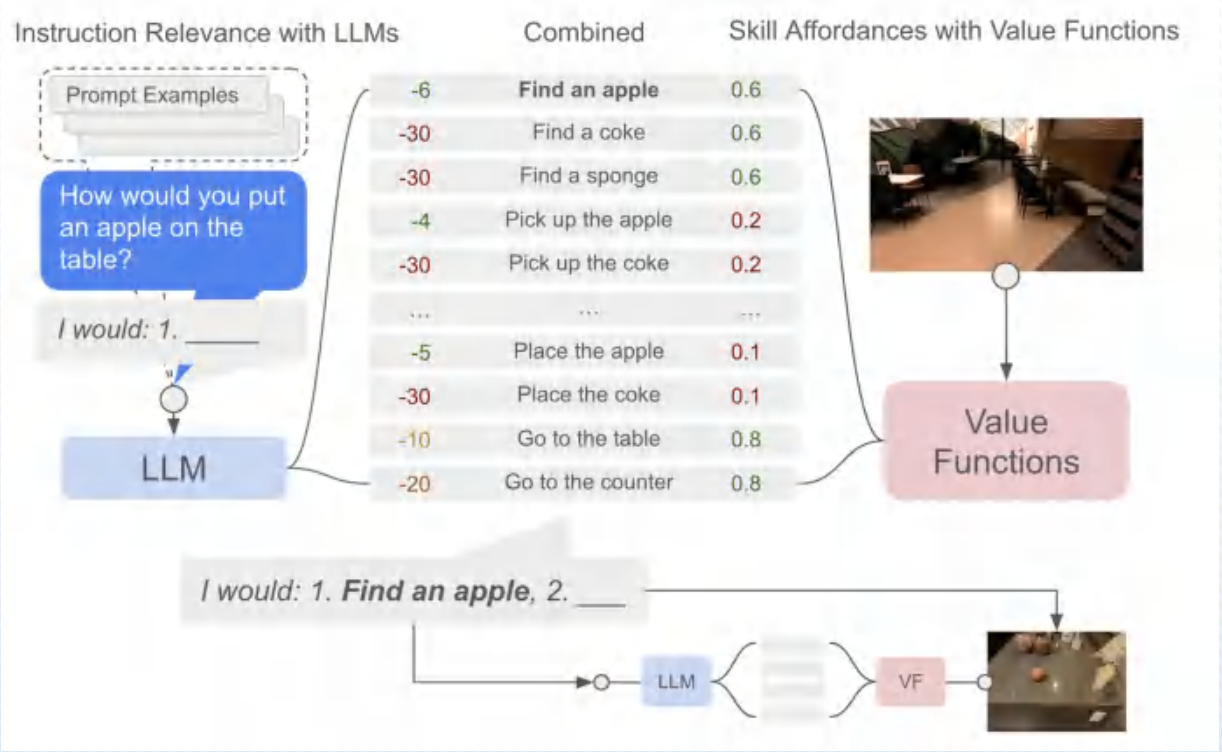
- Palm-E搜集了大量具身智能数据对LLM进行训练，并支持多模态输入



[1] Driess et al. PaLM-E: An Embodied Multimodal Language Model. 2023. arXiv.

训练小模型检测可行性，与大模型结合

- 技能小模型为技能提供可行性评分，为大模型任务规划提供参考
- 该工作经典之处在于，大模型难以判断一个技能的可执行性，没有grounding到物理世界



[1] Ahn et al. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. 2022. arXiv.

任务规划的评估

- 基准集和相关指标

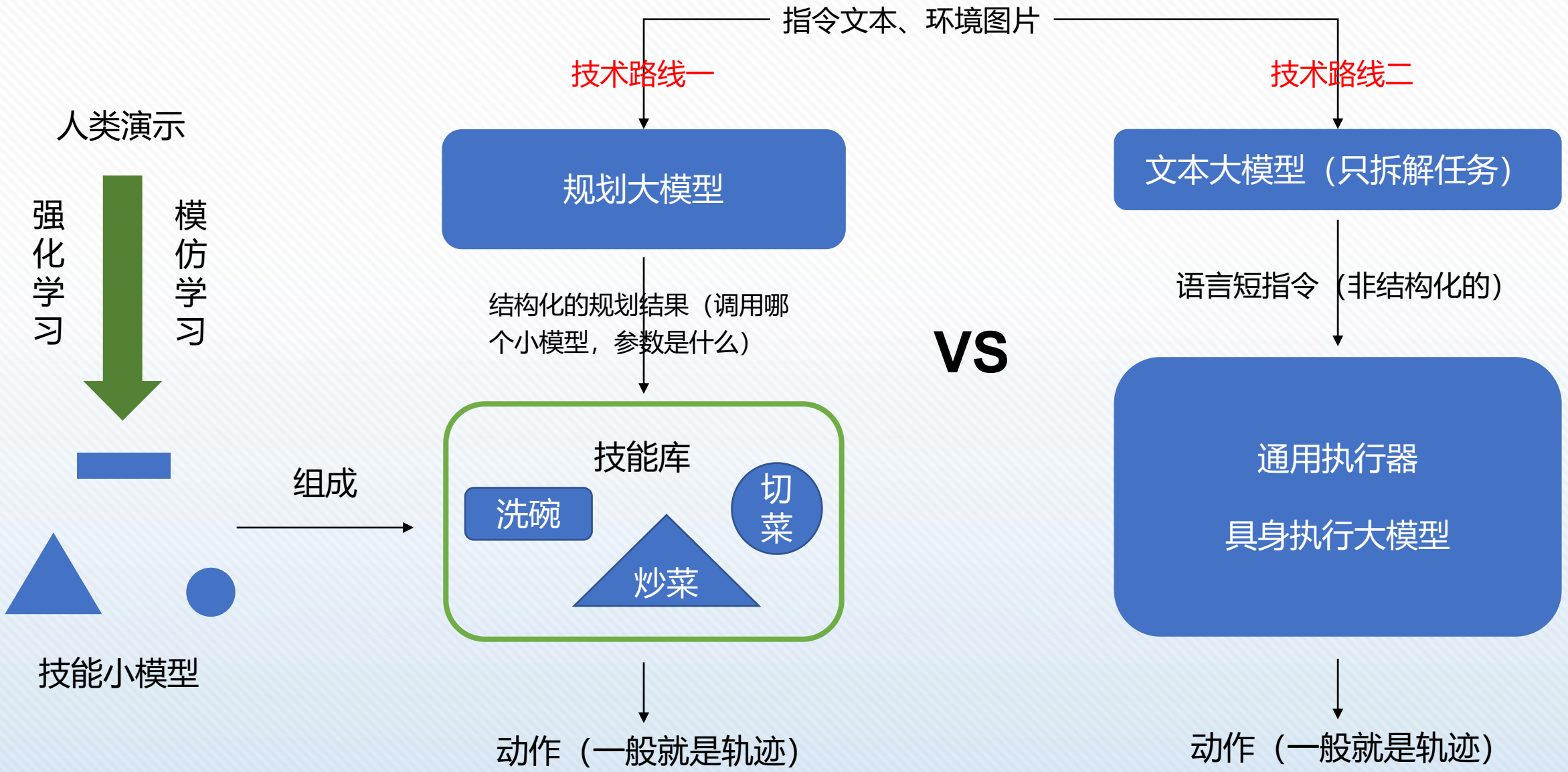
Benchmark	Size	Environment	Metrics
EgoPlan-Bench	57,656	Real World	*Accuracy* of multiple choice questions
BEHAVIOR	500	iGibson 2.0	*Success score* (Satisfied goal literals per step)
			Efficiency (Total time cost)
WAH	1,211	VirtualHome	*Success rate* (Task completion)
			Speedup (Reduced episode length)
			Cumulative reward under designed reward function
ALFRED	25,726	AI2-THOR	*Task Success* (Task completion)
			Goal-Condition Success (Sub-task completion)
			Path Weighted Metrics (Sequence redundancy)

[1] Yi et al. EgoPlan-Bench: Benchmarking Multimodal Large Language Models for Human-Level Planning. 2023. arXiv.
[2] Srivastava et al. BEHAVIOR: Benchmark for Everyday Household Activities in Virtual, Interactive, and Ecological Environments. 2021. arXiv.
[3] Puig et al. Watch-and-help: A challenge for social perception and human-ai collaboration. 2020. arXiv.
[4] Shridhar et al. ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. 2020. CVPR2020.

任务规划的关键问题、关键信息

- 任务规划的关键问题
  - 结构化
    - 大模型输出自然语言，需要解析后才能被程序使用，可能出现未精准匹配，无法映射为单技能的情况
  - 可行性
    - 结构化的动作不具备执行的可行性，例如操作不存在的物体，物体离机器人太远，物体不支持操作
  - 有用性
    - 即使结构化的动作、被操作物体可被执行，它们还应该是有助于完成任务的
- 任务规划的两种关键信息
  - 观测信息
    - 场景描述、物体列表、图片。不同类型的观察信息有不同的特点
  - 反馈信息
    - 即使通过可行性检测，机器人仍有可能执行技能失败，此时反馈信息很重要

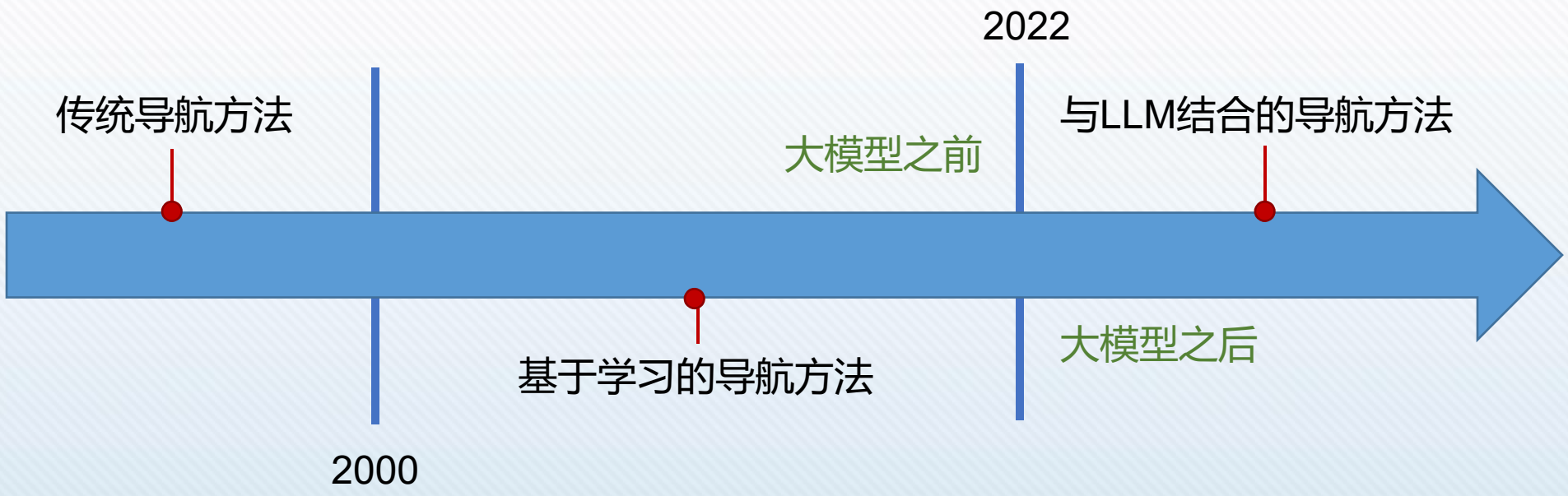
通用执行模型出现后的任务规划



2-2 导航

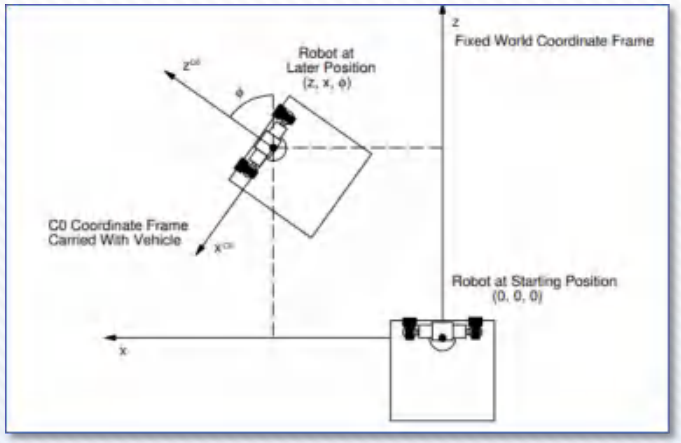
训练小模型检测可行性，与大模型结合

- 具身导航（Embodied Navigation）：智能体在3D环境中移动完成导航目标
- 目标的形式可以是点、物体、图像、区域；目标可以结合声音、自然语言指令、人类先验

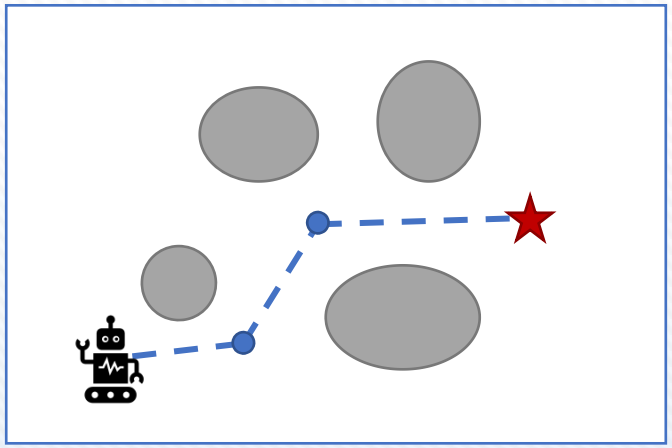


基于规则的导航

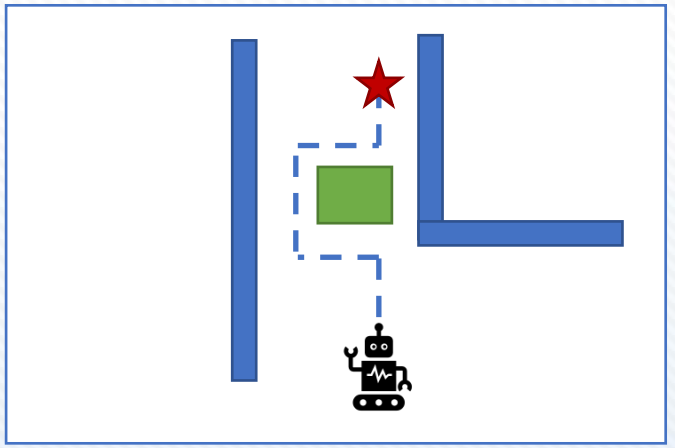
- 早期的具身导航，通过构建一系列基于规则的组件和算法，实现有效的环境感知、定位、路径规划和避障
- 关键技术包括：



SLAM建图技术



路径规划算法



避障算法

- 优点：鲁棒性强、计算效率高、实现方法简单、确定性高
- 缺点：适应性差、依赖地图，建图成本高、缺乏学习能力

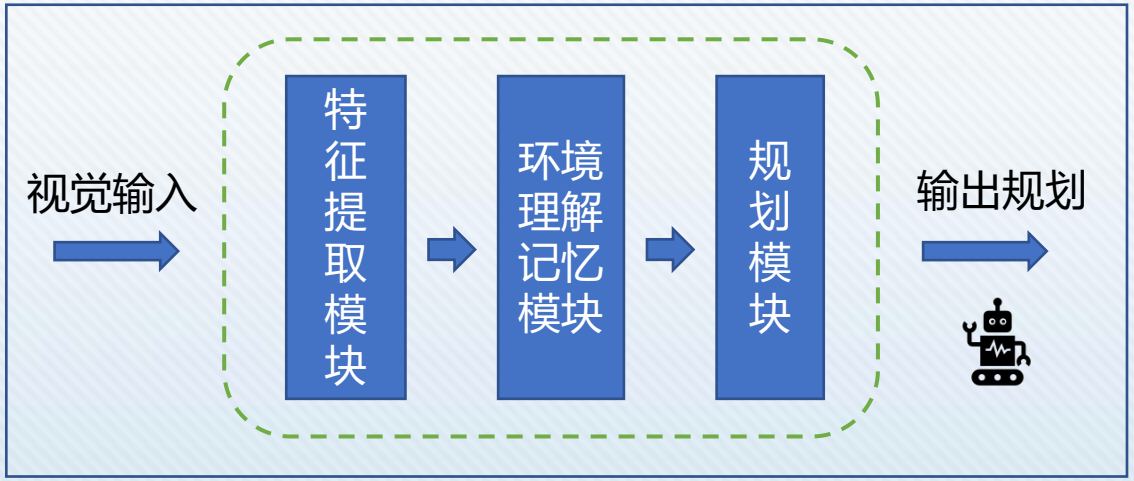
基于学习的导航

- 基于学习的导航利用深度学习与强化学习技术，提高模型对复杂环境和新场景的泛化能力
- 不同于传统算法依赖预定义的规则和手工设计的特征，基于学习的导航算法从大量数据中学习环境特征和导航策略，实现强自适应性和高灵活性

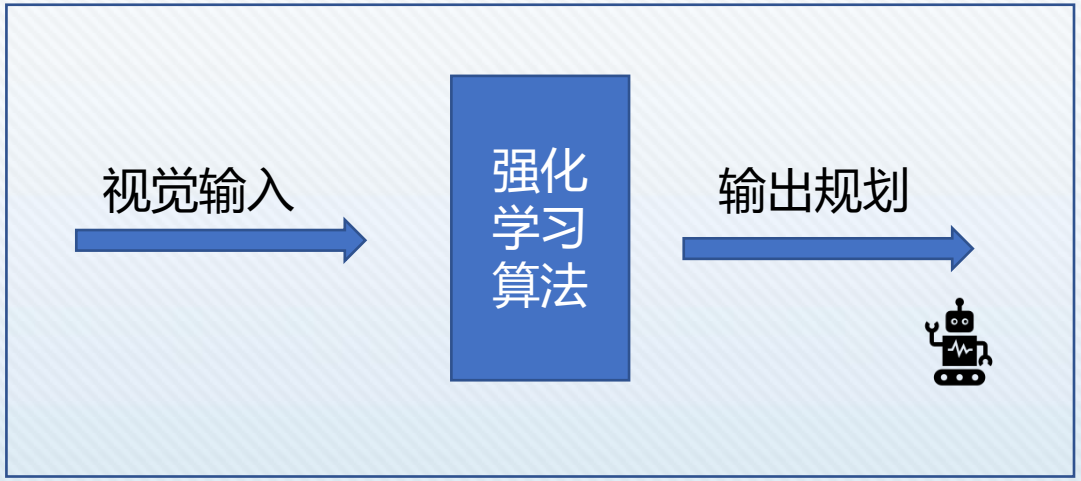
- 按照输入模态可以分为：
  - 视觉导航（Visual Navigation）
  - 视觉语言导航（Vision-Language Navigation）

视觉导航

- 视觉导航是基于学习的导航的一个重要分支，它依靠计算机视觉来理解环境信息并做出导航决策
- 视觉导航面临的主要挑战包括：
  - 如何从视觉输入中提取有用的信息
  - 如何理解和记忆环境的布局
  - 如何规划路径



非端到端的方法



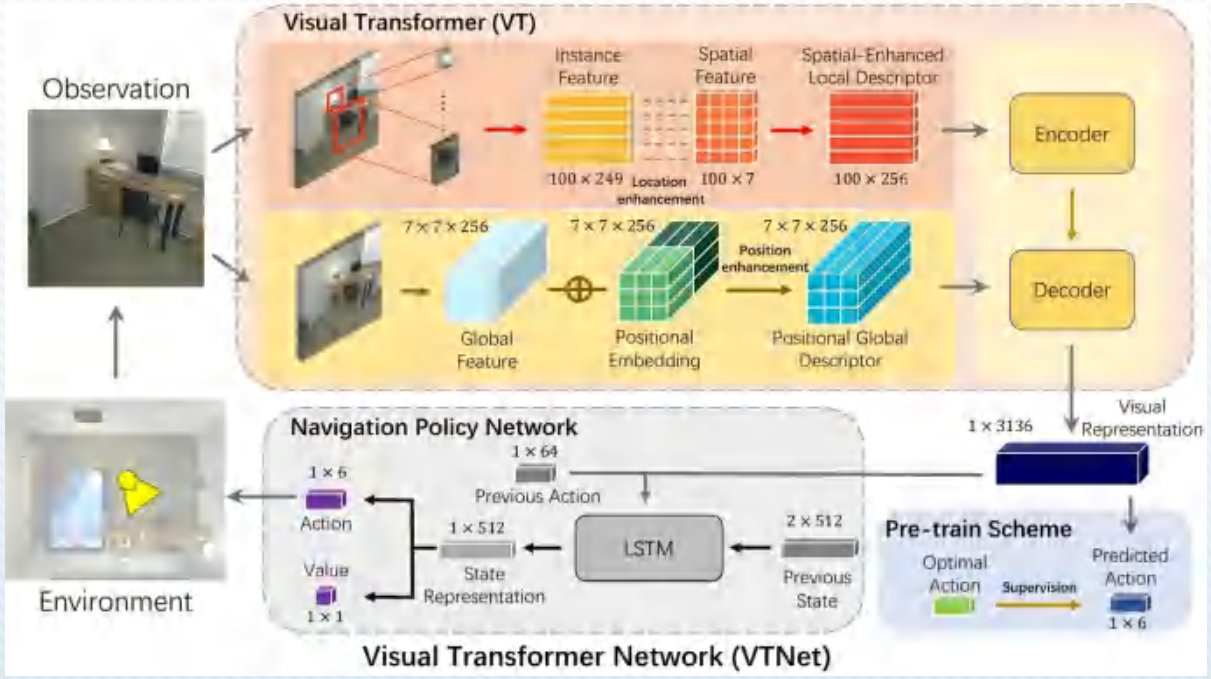
端到端的方法

提取视觉信息

- 从环境中提取信息有多种方法：
  - 卷积神经网络（CNN）在捕捉图像空间层次结构方面表现优异
  - 多种预训练模型的应用
    - ResNet、VGG、Inception等预训练模型大幅提升了目标识别的效率和成功率
  - 注意力机制在视觉感知中的应用
    - VTNet 通过学习场景中所有对象实例的关系和空间位置，生成方向性导航信号

提取视觉信息

- VTNet包含一个视觉 Transformer (VT) 和一个导航策略网络。
- 视觉策略网络首先分割出场景中物体实例。实例特征与空间特征融合，获取对局部的重点表示
- 同时全局特征也添加位置嵌入，获得对场景整体的表示，两个表示一起解码得到视觉表示



[1] Du et al. VTNet: Visual Transformer Network for Object Goal Navigation. ICLR 2021

建模环境信息

- 模型构建方法
  - 空间布局建模：使用地图构建和路径规划的联合架构，如Gupta et al提出的认知图构建
  - 拓扑图建模：基于图神经网络进行的定位与导航行为分解
- 场景占用状态推断：
  - 使用RGB-D观测进行超出可见区域的场景占用状态推断，提升空间感知能力
- 知识图谱与深度强化学习结合：
  - 将知识图谱与强化学习结合，推断目标对象可能位置，生成导航策略
- 贝叶斯关系记忆（BRM）：
  - 捕捉训练环境中的布局先验，在测试中更新记忆并高效规划导航路径

[1] Gupta et al. Cognitive mapping and planning for visual navigation. 2020 IJCV
[2] Chen et al. A behavioral approach to visual navigation with graph localizationnetworks. 2019 RSS
[3] Ramakrishnan et al. Occupancy anticipation for efficient exploration and navigation. 2020 ECCV
[4] Yang et al. Visual semantic navigation using scene priors. 2018 arxiv
[5] Yi Wu. 2019. Bayesian relational memory for semantic visual navigation. 2019 ICCV

路径规划

- 近年来，对于未知环境路径规划任务，基于强化学习的端到端系统取得了显著的成功
- DFP（直接未来预测）：基于未来行为预测的学习方法，解决感知运动控制问题
  - BDFP：引入中间地图表示，使黑盒更具解释性
- DD-PPO方法
  - 三种辅助任务（动作条件对比预测编码、逆动力学学和时间距离估计），提升样本和时间效率
- SAVN方法
  - 鼓励模型在动态环境中进行自适应学习，提高了路径规划的灵活性和效率

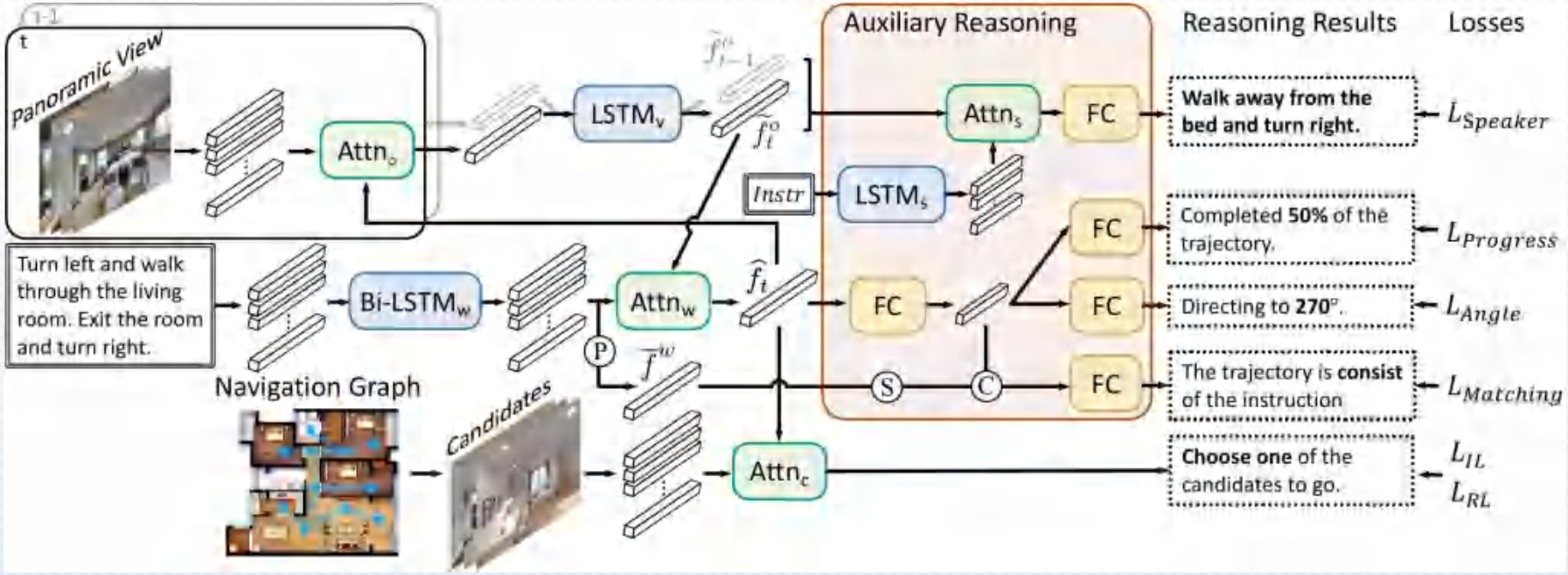
[1] Dosovitskiy et al. Learning to act by predicting the future. 2017 ICLR
[2] Mishkin et al. Benchmarking classic and learned navigation in complex 3d environments. 2019 arxiv
[3] Wijmans et al. DD-PPO: Learning Near-Perfect PointGoal Navigators from 2.5 Billion Frames. 2019 arxiv
[4] Ye et al. Auxiliary tasks speed up learning point goal navigation. 2021 CoRL
[5] Wortsman et al. Learning to learn how to learn: Self-adaptive visual navigation using meta-learning. 2019 CVPR

视觉语言导航

- 视觉语言导航是通过通过自然语言指令和视觉图像进行导航的任务，其目标是开发一种能够与人类进行自然语言交流并在现实3D环境中导航的具身智能体
- 视觉语言导航可以根据时间节点分为
  - 大模型之前的视觉语言导航
    - 主要通过RNN , LSTM, Transformer等网络来提取命令中的语义信息
  - 结合LLM的具身导航
    - 利用大模型作为辅助来帮助规划器输出规划或者大模型直接作为规划器来输出规划

自监督的辅助推理任务提高VLN效果

- 该论文针对视觉语言导航任务，提出四个辅助推理任务用于预训练：轨迹复述、进度评估、角度预测、跨模态匹配，基于LSTM和注意力机制，提高视觉信息和语义信息的融合效果，进而生成更好的导航动作



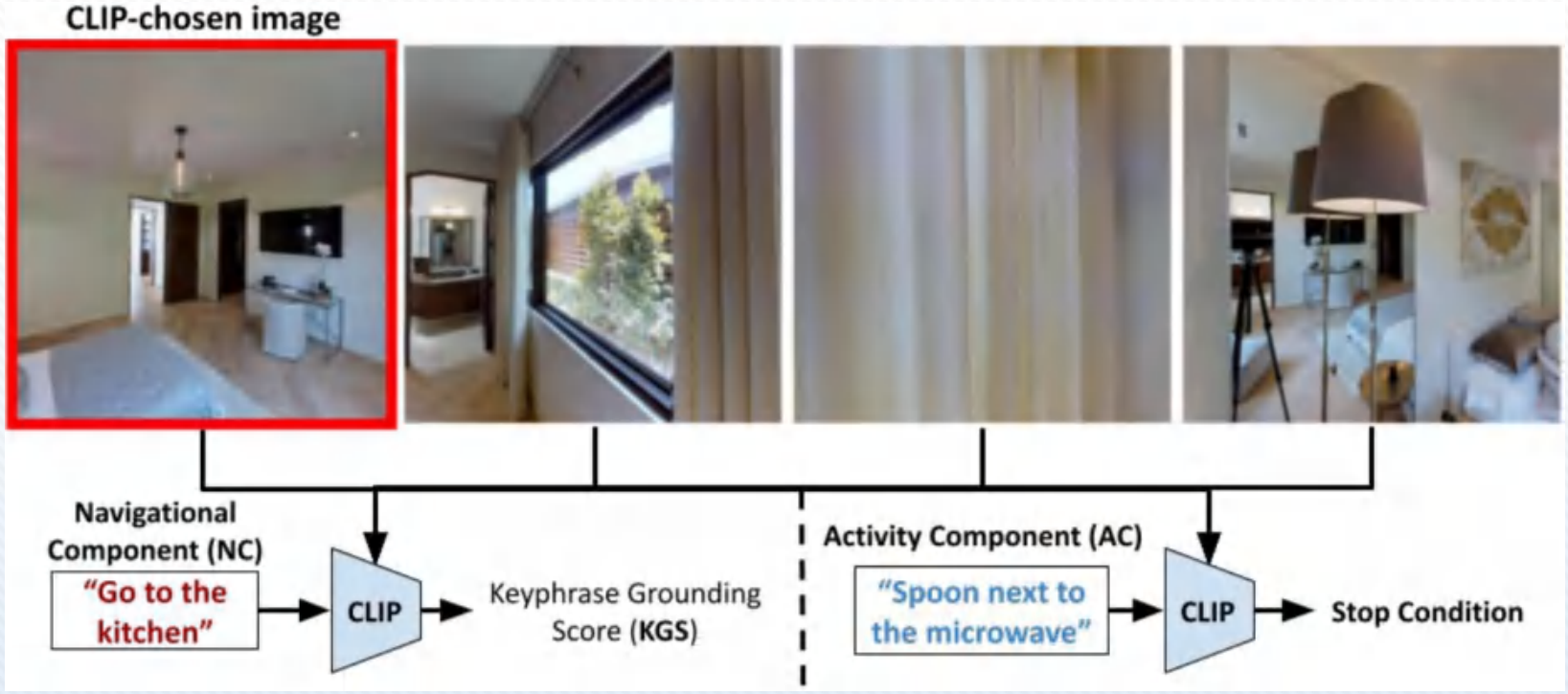
[1] Zhu et al Vision-Language Navigation with Self-Supervised Auxiliary Reasoning Tasks 2020,arxiv

结合LLM的具身导航

- 大模型的出现显著改变了视觉语言导航领域的发展
- 大模型，或者视觉语言联合预训练模型如CLIP等，为该领域带来了新的方法和思路，使得视觉语言导航系统变得更加智能和鲁棒
- 根据大模型的作用不同，这里我们将这些工作分为：
  - 视觉语言联合预训练模型的应用
  - 大模型基于构建的地图输出规划
  - 大模型基于图片转换的文本描述输出规划

视觉语言联合模型的应用

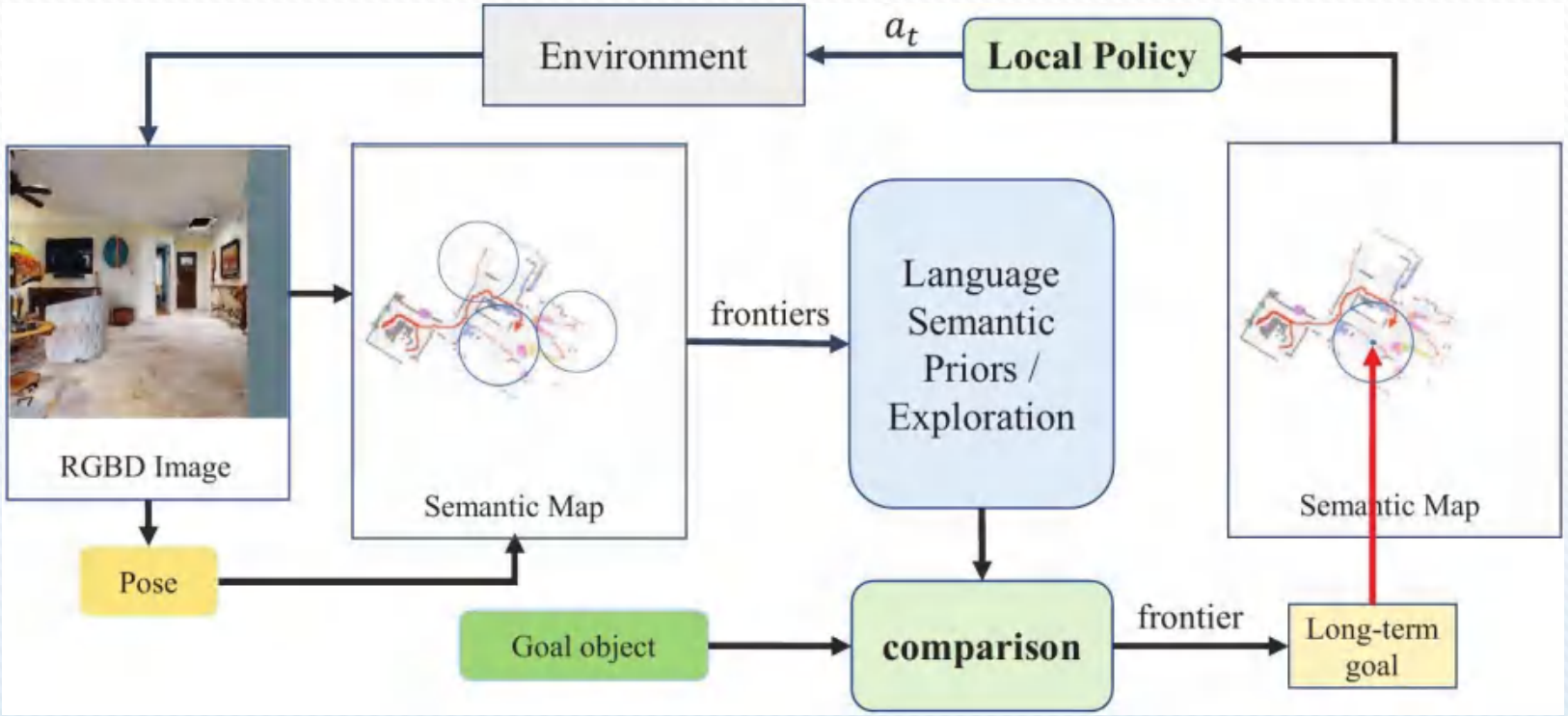
- 首先将粗粒度指令分解为关键词短语，然后使用 CLIP 将短语与当前输入图片计算相似度，选择最恰当的图片作为下一步方向



[1] Dorbala et al. CLIP-Nav: Using CLIP for Zero-Shot Vision-and-Language Navigation. 2022 arxiv

大模型基于构建的地图

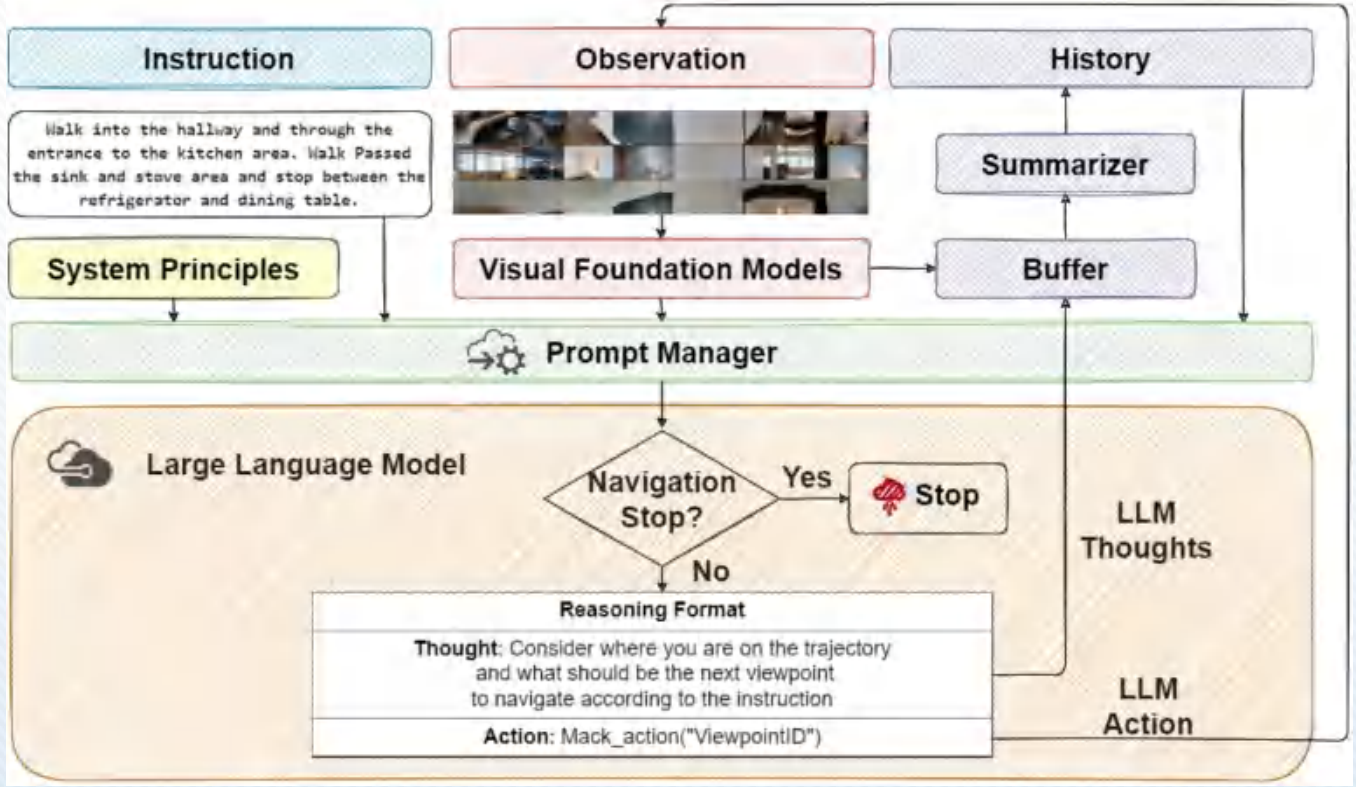
- 构建语义地图，从已探索地图和障碍物地图中提取边界地图，每个边界都是下一步候选搜索窗口
- LLM根据目标物体和候选搜索窗口的观测信息之间的相关性打分，选择下一步方向



[1] Yu et al L3MVN: Leveraging Large Language Models for Visual Target Navigation 2022,arxiv

大模型基于图片转换成的文本描述输出规划

- NavGPT 将视觉观察的文本描述、导航历史和未来可探索方向作为输入来推理代理的当前状态，并做出接近目标的决定

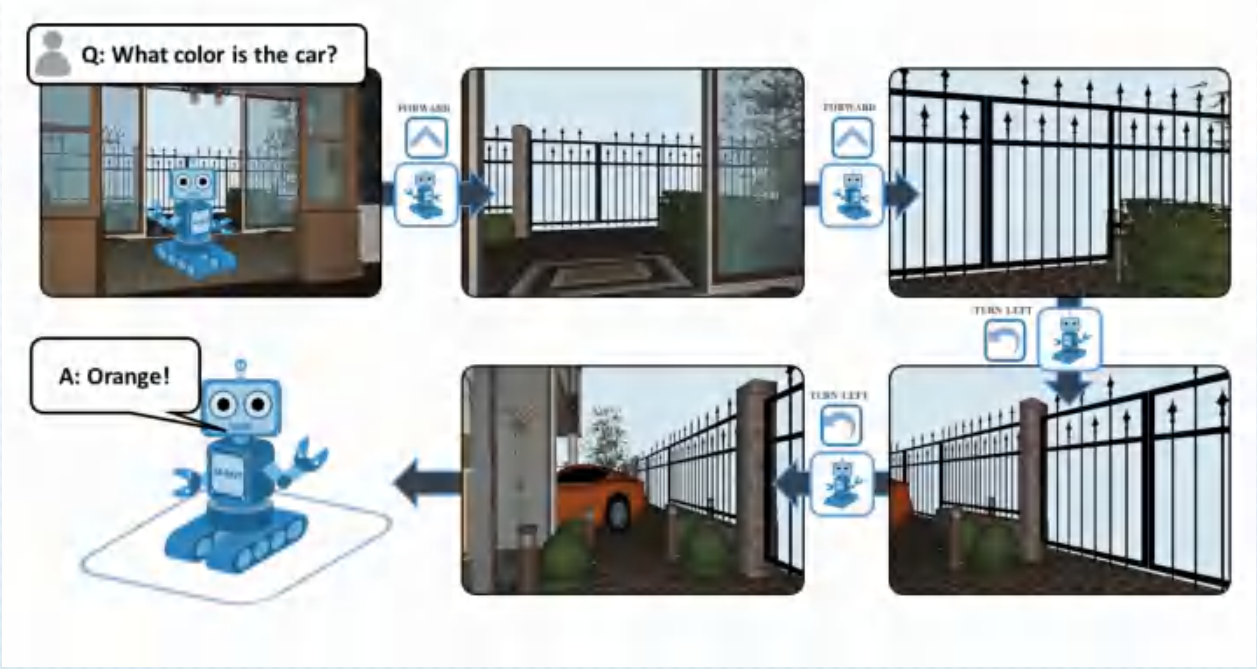


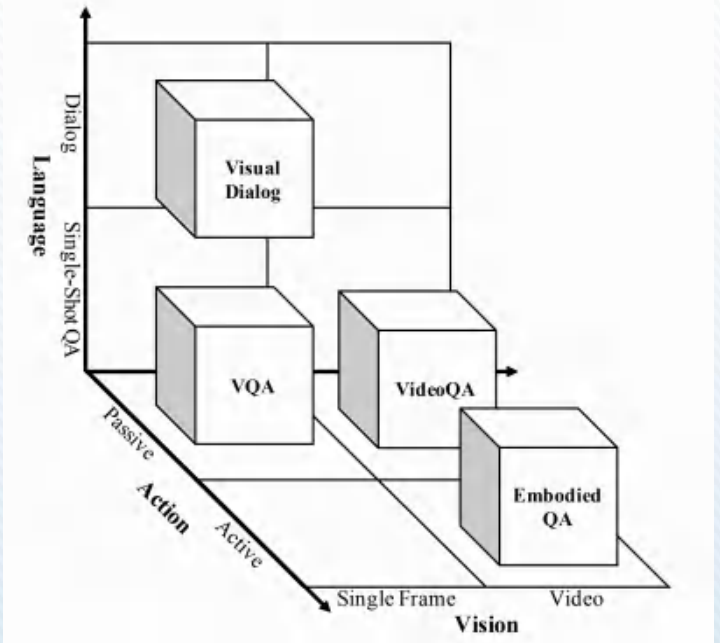
[1] Zhou et al NavGPT: Explicit Reasoning in Vision-and-Language Navigation with Large Language Models. 2023 arxiv

2-3 具身问答

具身问答

- 具身问答任务最早由Das et al.提出，该任务下机器人需要主动探索环境，定位目标物体或位置，获取环境中的信息，然后基于获取的信息回答问题。该任务可视为导航、VQA任务的结合
- 相比于VQA等已有问答任务，具身问答的特点在于机器人具有主动行动能力





[1] Das et al. Embodied Question Answering. 2018 CVPR

具身问答方法创新

- 最早的具身问答论文：一个模块处理一个子任务
  - 路径规划模块：导航
  - 视觉识别模块：目标识别
  - 问答模块：生成自然语言回复
- 方法优化：
  - Luo et al. 针对嘈杂环境设计鲁棒性更强的导航、问答模块，并提出两阶段鲁棒学习算法
  - Li et al. 提出Model-based RL的EQA算法，通过“想象”下一个子目标环境图片，提高探索效率，并使得智能行为更加具有可解释性
  - Chaplot et al. 多任务联合学习，将文本信息与视觉特征融合

[1] Das et al. Embodied Question Answering. 2018 CVPR
[2] Luo et al. Robust-EQA: Robust Learning for Embodied Question Answering With Noisy Labels. 2023 ITNNLS
[3] Li et al . Walking with MIND: Mental Imagery eNhanceD Embodied QA. 2019 ACM MM
[4] Chaplot et al. Embodied Multimodal Multitask Learning 2019 IJCAI

具身问答任务创新

- 多目标EQA：在单个EQA样例中，智能体需要找多个目标才能准确回答问题，更强调规划、推理、记忆能力
- 多智能体EQA：多智能体协作一起完成任务，不同的智能体需要有一个策略分配任务，避免重复搜索，还需要交换、整合信息，提高完成任务的效率
- 知识增强EQA：外源知识增强的EQA，外源知识可以帮助智能体回答更复杂，范围更全面的问题

[1] Yu et al. Multi-Target Embodied Question Answering. 2019 CVPR
[2] Tan et al. Multi-agent Embodied Question Answering in Interactive Environments. 2020 ECCV
[3] Tan et al . Knowledge-Based Embodied Question Answering. 2021 TPAMI

具身问答数据集与盲人Benchmark

- EQA数据集：750个场景，45个不同物体，7种房间类别，5000个问题
- MT-EQA数据集：6种复合问题类型，包括多目标比较。588个场景，19287个问题
- Blind Benchmark：在EQA v1数据集上，不基于环境观察回答问题，2018年提出时成为当时的sota，仅在智能体初始点离物体特别近的情况下效果略有落后

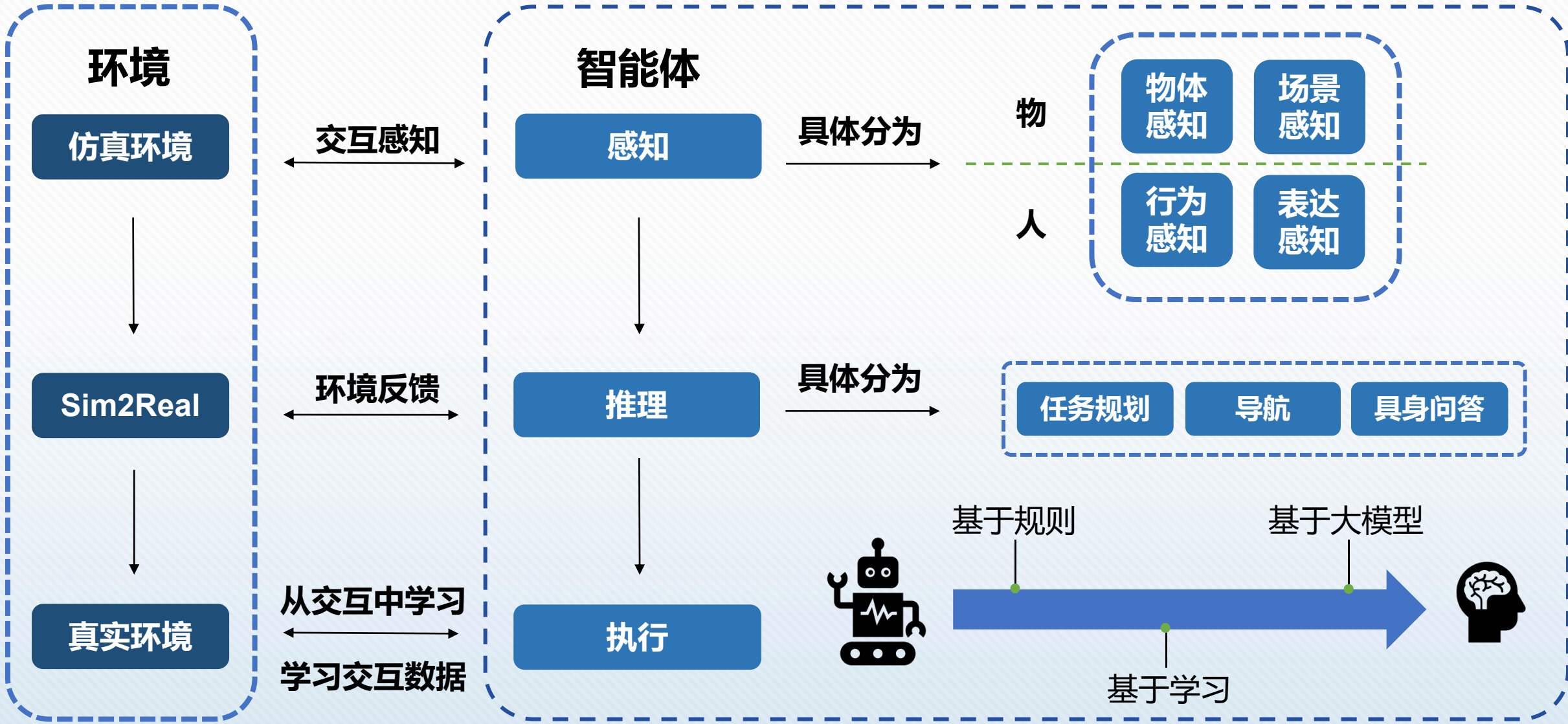
2-4 具身推理小结

具身推理

- 在推理任务上，大模型具有非常显著的优势，其相比于小模型推理能力有了显著的提升
- 基于大模型构建具身智能体是一个非常自然的选择，但也存在许多问题：
  - 推理速度慢，推理开销大
  - 生成结果不够稳定
  - 复杂的Agent结构难以维护，且时间开销会更大

3 具身执行

具身智能划分：感知、推理、执行



具身执行：技能学习

- 在具身感知中我们介绍了很多任务，并根据感知对象的不同分为四大类
  - 对非人的感知：物体感知、场景感知
  - 对人的感知：行为感知、表达感知
- 在具身推理中我们介绍了三个重点任务：任务规划、导航、具身问答
- 在具身执行中我们仅介绍一个任务：技能学习

- 技能学习：以技能描述、环境观察为输入，输出完成技能所需的7Dof轨迹
- 7Dof轨迹主要指：人手腕或者机械臂末端执行器的位置、朝向、末端状态
- 主要为手部操作，虽然不足以表达人或机器人全部动作空间，但足以覆盖生活中绝大多数技能

技能学习的两类方法

- 模仿学习：收集专家演示数据，用神经网络拟合
- 强化学习：设计奖励函数，机器人通过交互学习行为策略
- 行为策略：给定技能描述，在当前观察下，选择动作执行

- 其本质差别在于
  - 模仿学习从样例中学习；机器人学习过程中**不与**环境进行交互
    - 样例一般也是提前收集的交互样例数据，也可以算广义的交互学习
  - 强化学习从交互中学习；机器人学习过程中**与**环境进行交互

3-1 模仿学习

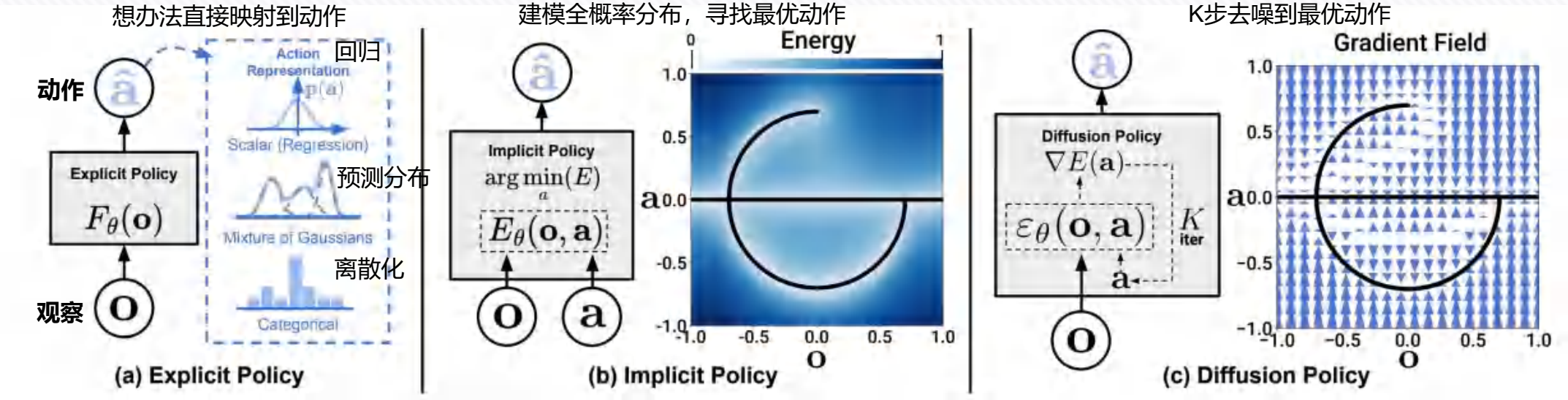
模仿学习

- 模仿学习主要理解并复制人类或机器人行为
- 数据采集方法
  - 基于通过摄像头捕捉的专家演示
  - 摄像头通常安置在执行主体的手腕、头部或侧面
- 演示数据来源
  - 动觉教学（手动接触）、VR、手柄、GUI界面控制等
- 演示数据组成
  - 收集到的专家演示包括一系列观察和行动
  - 示范数据展示了末端执行器或关节的位置或者速度
- 学习策略与目标
  - 核心是学习行为策略，将观察映射到连续动作空间

模仿学习总览

- 模仿学习可以分为两部分：**对图像的编码，图像表示映射到动作**
- 一般而言，图像的编码器使用预训练的视觉编码器更好，如果只使用样例数据集训练编码器会导致实际应用中缺乏泛化性
- 机器人的动作空间一般是连续的。对于连续动作值的预测一般有以下几类：

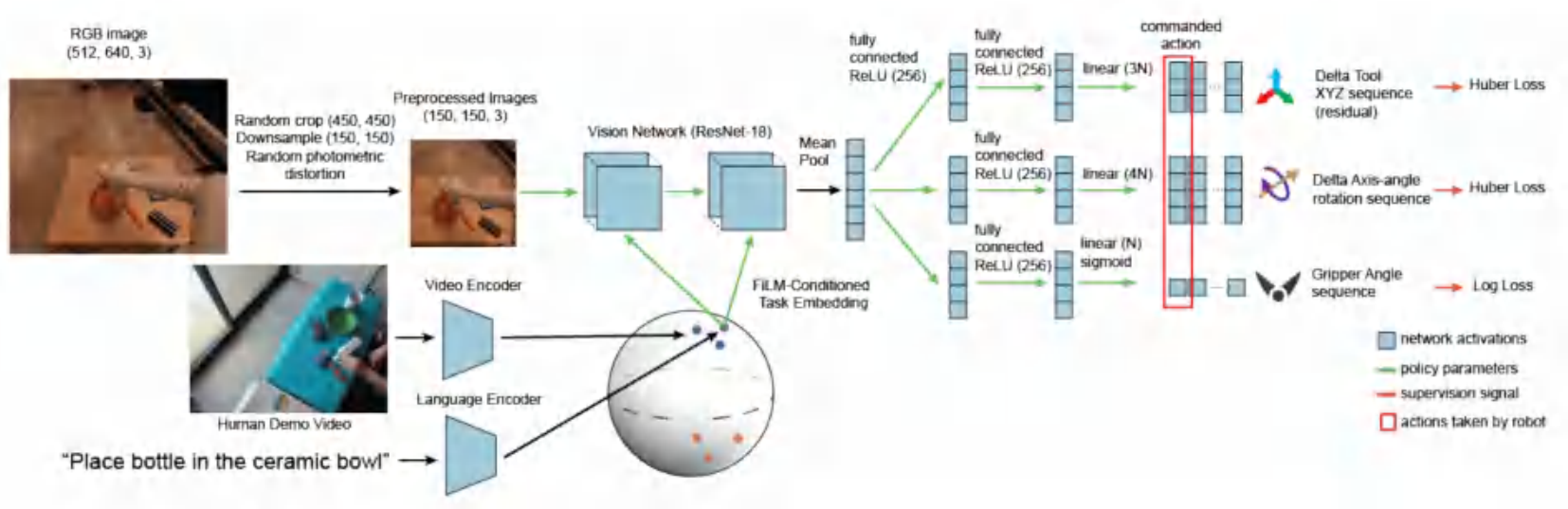
想办法直接映射到动作　　建模全概率分布，寻找最优动作　　K步去噪到最优动作



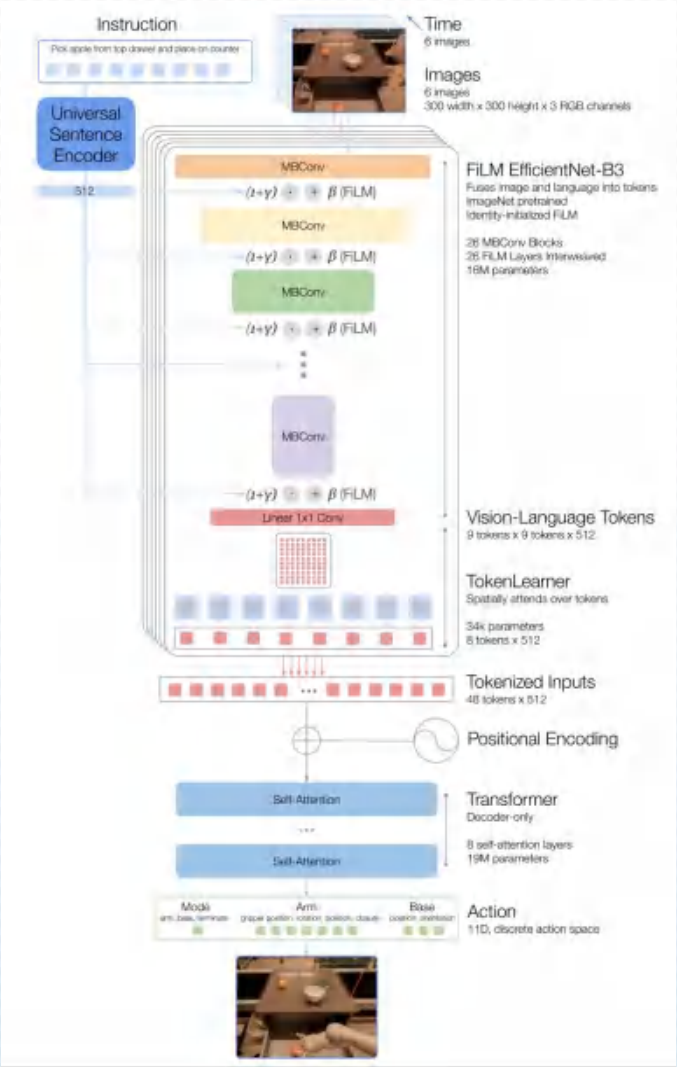
直接策略　　隐式策略　　扩散策略

直接策略：行为克隆

- 最经典、使用最广泛的策略学习算法，是将图像编码后**直接映射**到动作
- 唯一的区别在于损失函数的设计，即预测的动作值与真实动作值之间的损失
- 包括RT-1、RT-2在内的具身多模态大模型均采用该方法



BC-Z模型，使用回归的方式设计Loss

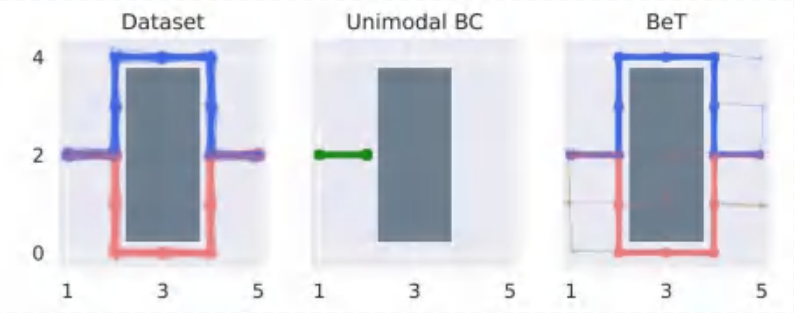


RT-1模型，离散成256个bin

[1] Jang et al. BC-Z: Zero-Shot Task Generalization with Robotic Imitation Learning. 2021 CoRL
[2] Brohan et al. RT-1: Robotics Transformer for Real-World Control at Scale. 2022 Arxiv

行为克隆可能出现的问题 以及 动作聚类方法

- 模仿学习数据集往往假设：专家完成任务只使用一种方式

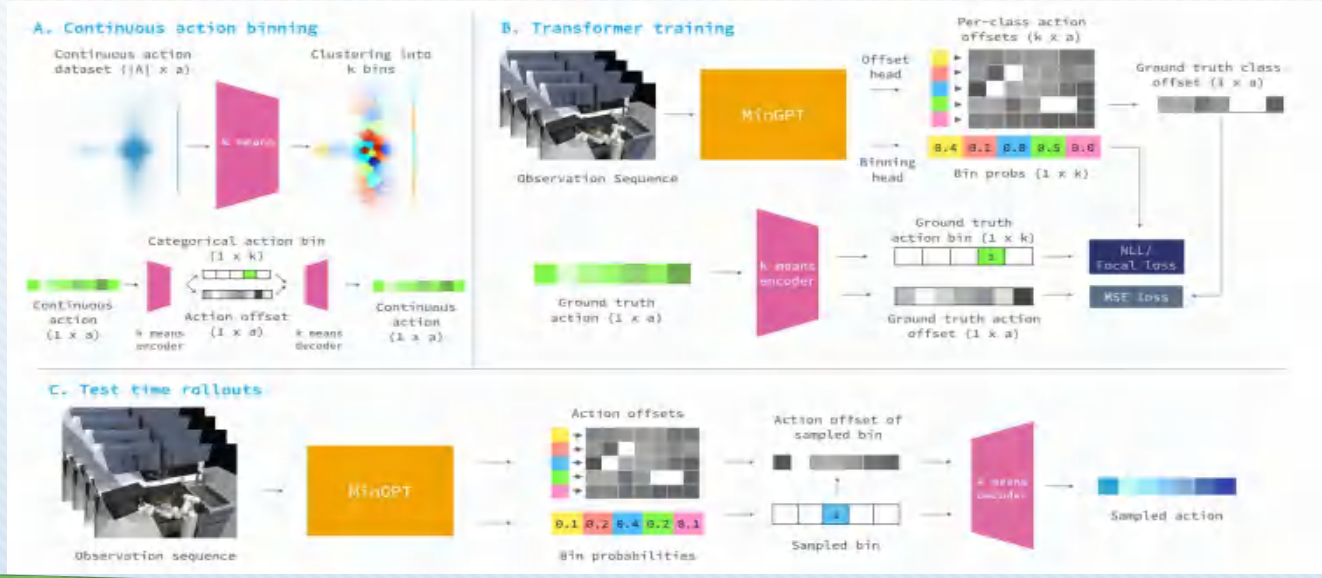


中间是障碍物，可以向上绕过，可以向下绕过，但不能走中间

可以向左，可以向右，但是不能中间

真实轨迹有多条的情况下，使用回归的方式就会有问题

- 作者认为上述假设错误地认为，样例数据集的动作来自同一个分布。真实情况是多个分布。

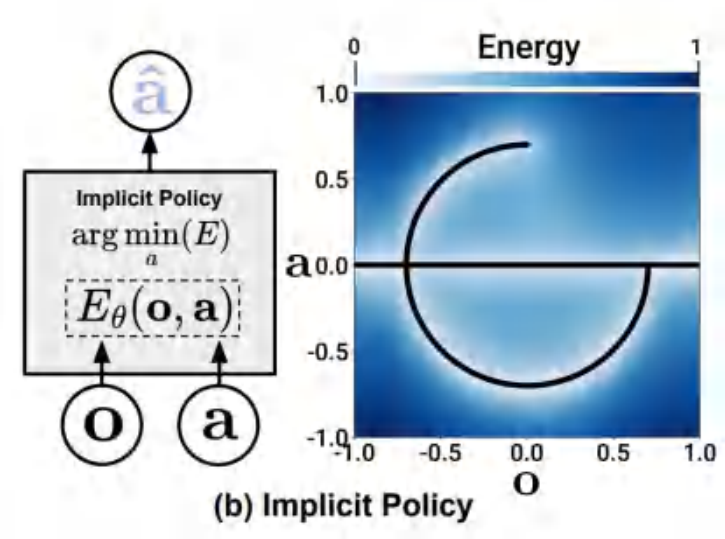


此方法将动作进行**聚类**，然后分别预测动作的类别和偏移量

[1] Shafiullah et al. Behavior Transformers: Cloning k modes with one stone. NIPS 2022

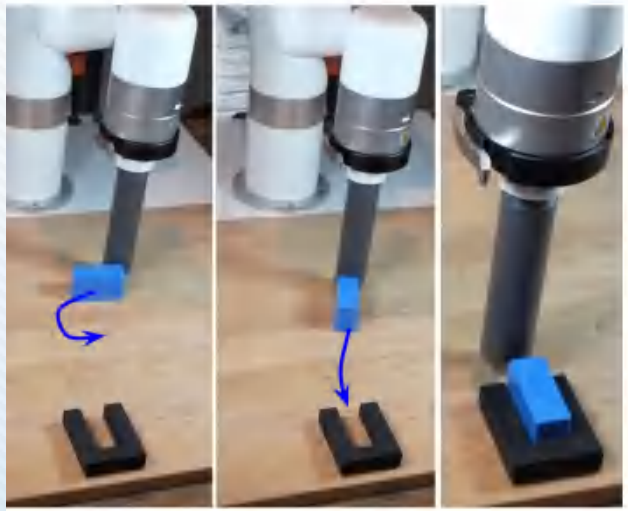
隐式策略：隐行为克隆

- 直接的动作映射存在一些问题，包括：
  - 轨迹不连续
  - 多种模式（存在多种轨迹）
- 作者提出隐式策略，不直接建模条件概率分布，而是建模观察与动作的联合概率分布。在实际推理中，需要基于联合概率分布，基于优化的方法寻找最优动作。

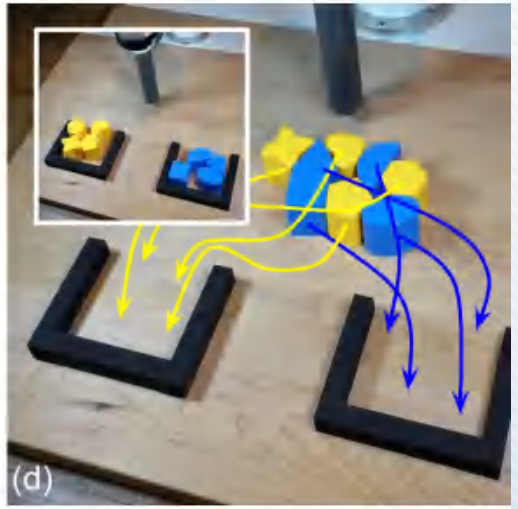


(b) Implicit Policy

横轴是观察空间，纵轴是动作空间，图中每个点表示（观察、动作）对，颜色的深浅表示（观察、动作）对的好坏，黑色轨迹就是最好的那些（观察、动作）对



两次离散操作轨迹

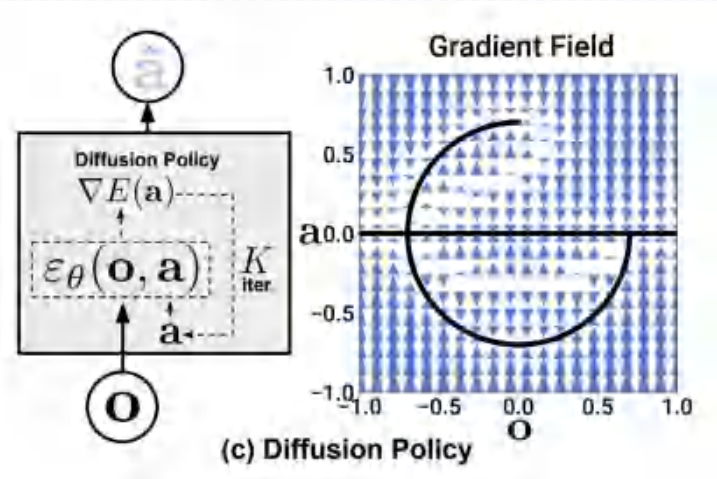


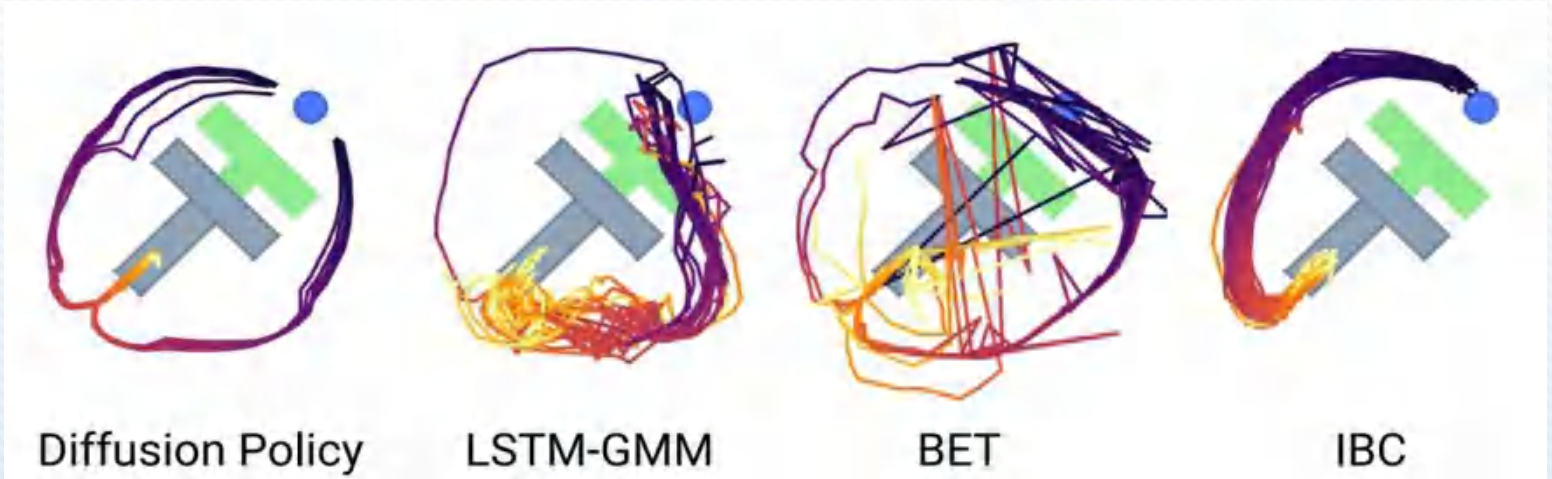
多种模式（操作轨迹）

[1] Florence et al. Implicit Behavioral Cloning. CoRL 2021

扩散策略

- 为机器人生成动作其实是生成轨迹，因为生成动作一般不是预测一个动作，而是生成动作序列。
- 轨迹生成需要考虑序列的连贯性，因此有研究者将其他领域中的生成方法引入策略学习中：
  - 联合概率建模 → 隐行为克隆
  - 扩散模型生成图片 → 扩散策略生成轨迹序列





图中蓝色箭头表示梯度，是某个观察下最好的动作取值点的方向，可以看到梯度都指向轨迹上的点，也就是最好的（观察、动作）对

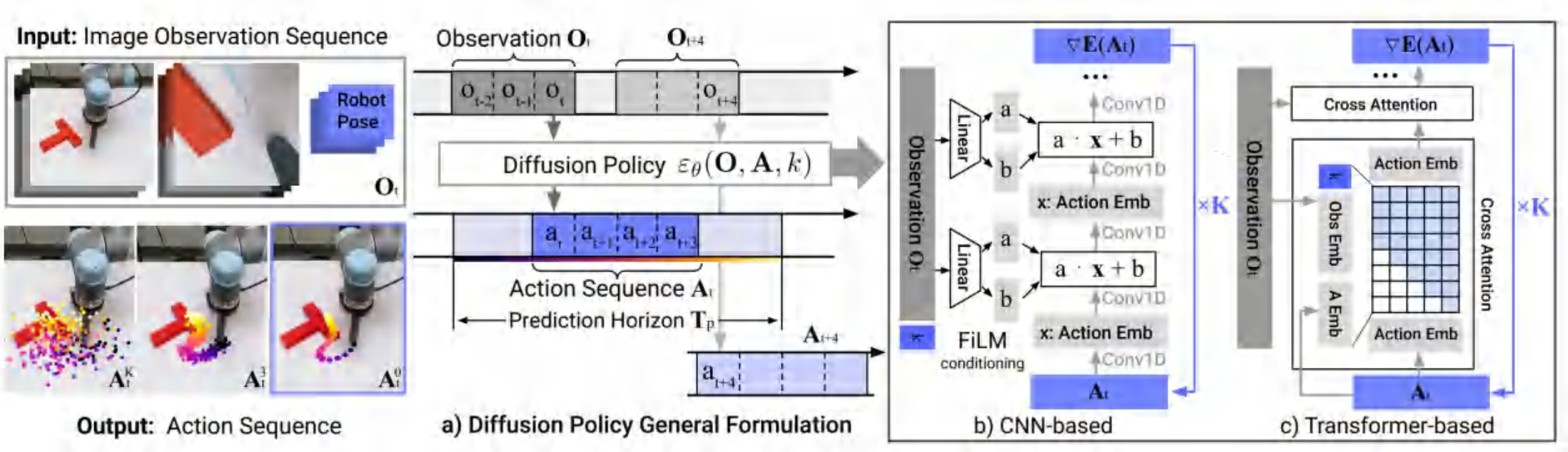
既能建模多种模式的轨迹，又带有随机性可以采样出多种可行的轨迹

建模多种模式的轨迹失败

能建模多种模式的轨迹，但整个过程是确定性的，只能产生一种模式的轨迹

扩散策略

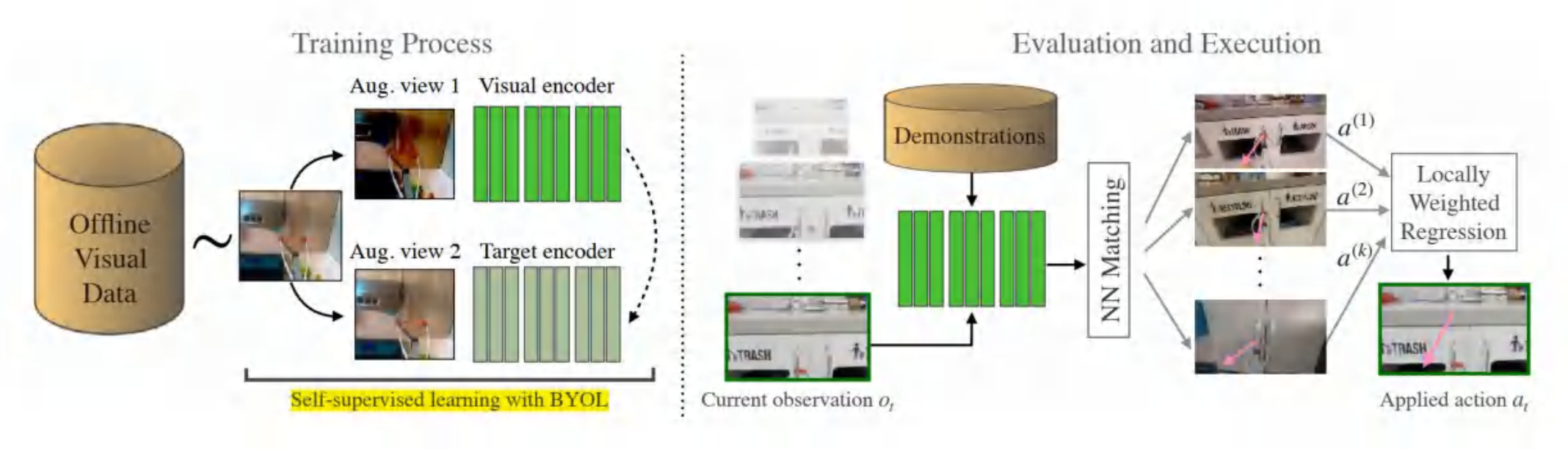
- 训练过程：
  - 加噪：从专家演示中随机采样一条轨迹，然后不停的向其中加入噪声；
  - 去噪：取加噪后的轨迹，基于观察预测加入的噪声，进行轨迹的还原；
- 推理：
  - 初始从高斯分布中采样一小段轨迹，基于这段轨迹和观察预测噪声，然后减去噪声，去噪持续K步生成最后的轨迹



[1] Chi et al. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. RSS 2023

策略学习：最近邻居算法

- **无参数**的策略学习算法：对于给定图片，寻找和它最相似的一组图片，取这组图片对应的动作进行加权平均，即为输出动作
- 证明了视觉编码的重要性。使用好的视觉编码，结合简单的策略学习算法，效果可以达到与行为克隆同等水平



[1] Pari et al. The Surprising Effectiveness of Representation Learning for Visual Imitation. RSS 2022

3-2 强化学习

强化学习：奖励函数指导的交互学习框架

- **强化学习（Reinforcement learning, RL）**
  - 强化学习专注于如何让智能体在环境中采取行动以**最大化某种累积奖励**
  - 基于**与环境的交互**，智能体学习选择最佳行动，逐步改善其行为策略
  - 应用广泛，包括自动驾驶、游戏、机器人控制等
- **无模型强化学习（Model-Free RL）**

  直接从环境交互学习最优策略或价值函数，依赖尝试和错误经验提升策略性能

  优点：能学习复杂行为，无需构建显式的环境模型

  挑战：可能需大量环境交互，样本效率较低
- **基于模型的强化学习（Model-Based RL）**

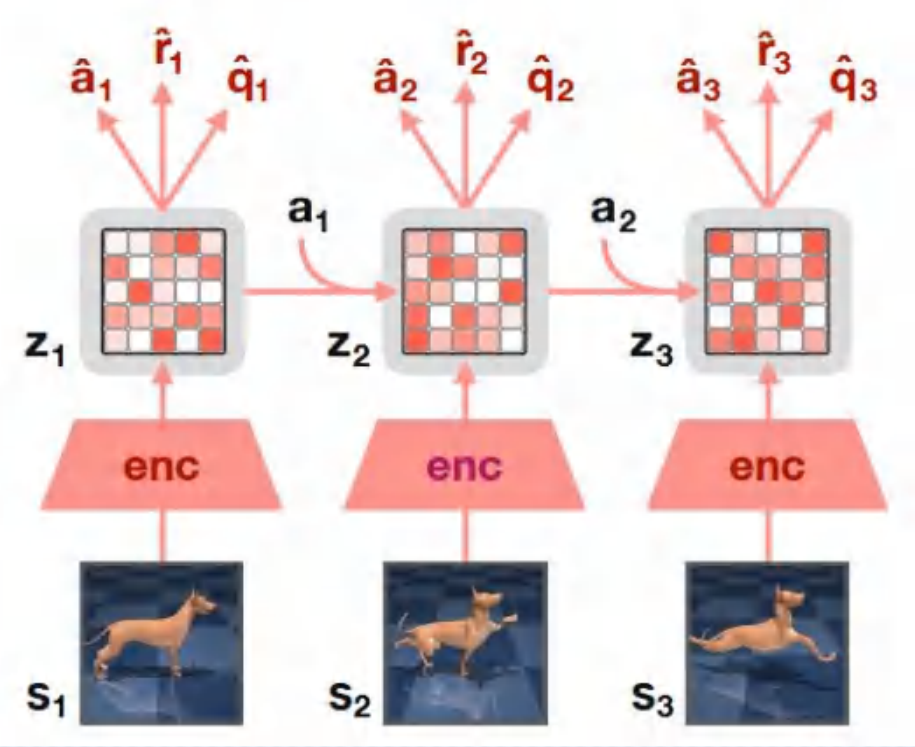
  学习环境的动态模型，用于规划或预测未来状态。通过“想象”未来的环境状态，提高样本效率，减少实际交互需求

  优点：更高的效率和规划能力

  挑战：环境模型的准确性对性能起到至关重要的作用

TD-MPC2 Model-based

- Dynamic model
  - 奖励函数 $\hat{r} = R(z, a, e)$
  - 总回报 $\hat{q} = Q(z, a, e)$
    - 多任务训练中不同任务之间奖励大小差别可能太大，将奖励预测R和回报预测Q看作离散回归(分类)问题。用CE损失函数
  - 状态预测模型 $z' = d(z, a)$
- 策略 $\hat{a} = p(z, e)$
- 编码器 $z = h(s, e)$



以上五个组件均为MLP(LayerNorm，Mish激活函数)

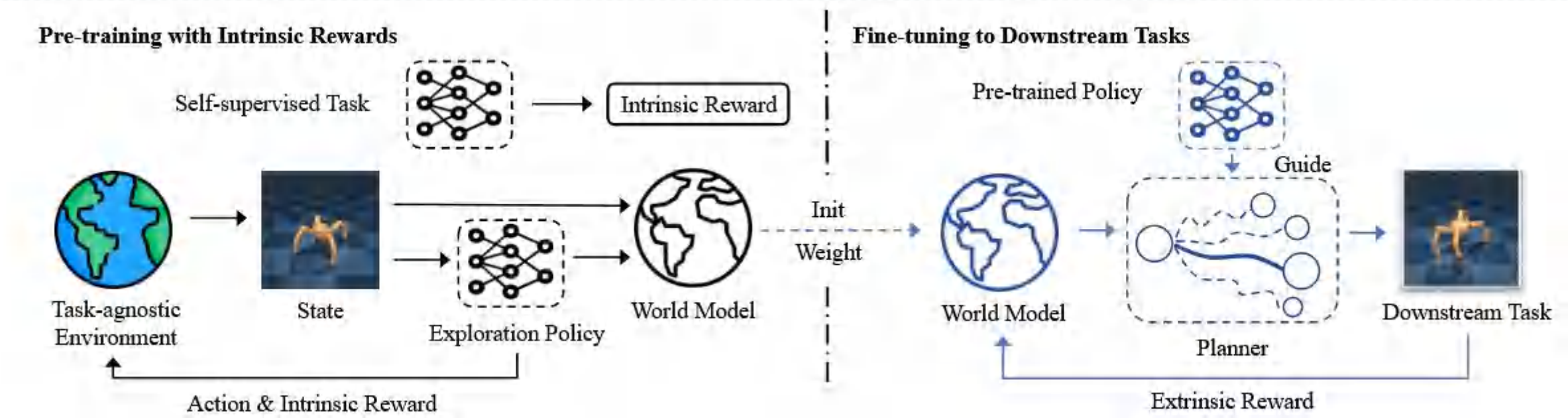
e(任务表示)　　s(环境观测)
a(动作)　　z(环境编码)

在TD-MPC基础上添加了多任务e的支持

[1] Hansen et al TD-MPC2: Scalable, Robust World Models for Continuous Control.2024 ICLR

EUCLID Model-based

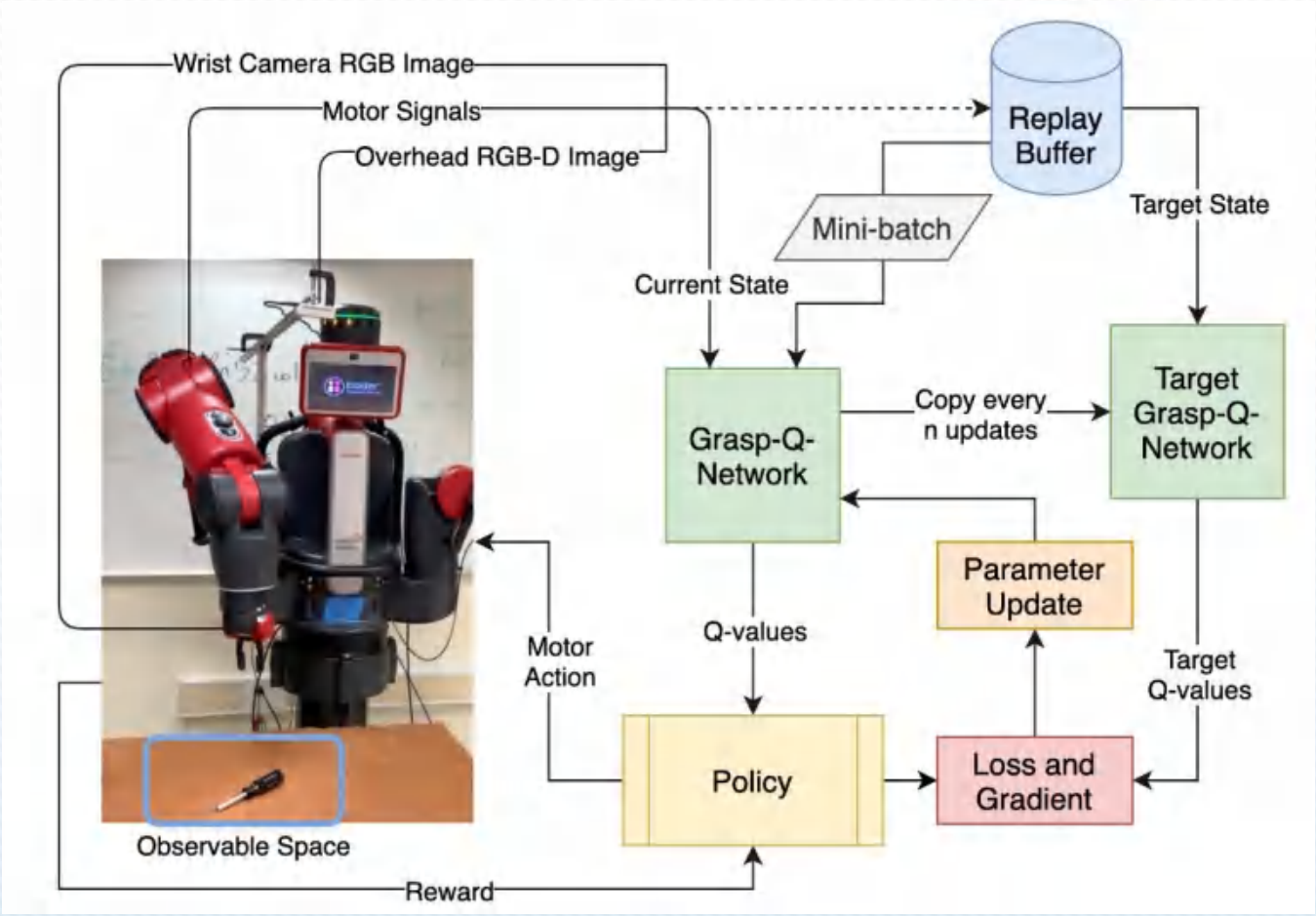
- 预训练
  - 交互 $\pi_{exploration}$
  - 更新 （基于 TD-MPC 的）dynamic model
  - 更新策略 $\pi_{exploration}$得到$\pi_{\mathrm{pretrained}}$
- 微调（下游任务）
  - 交互 $\pi_{\mathrm{pretrained}}$
  - 更新 dynamic model
  - 更新策略$\pi_{\mathrm{pretrained}}$得到$\pi_{task}$



[1] Yuan et al EUCLID: Towards Efficient Unsupervised Reinforcement Learning with Multi-choice Dynamics Model.2023 ICLR

Grasp-Q-Network Model-free

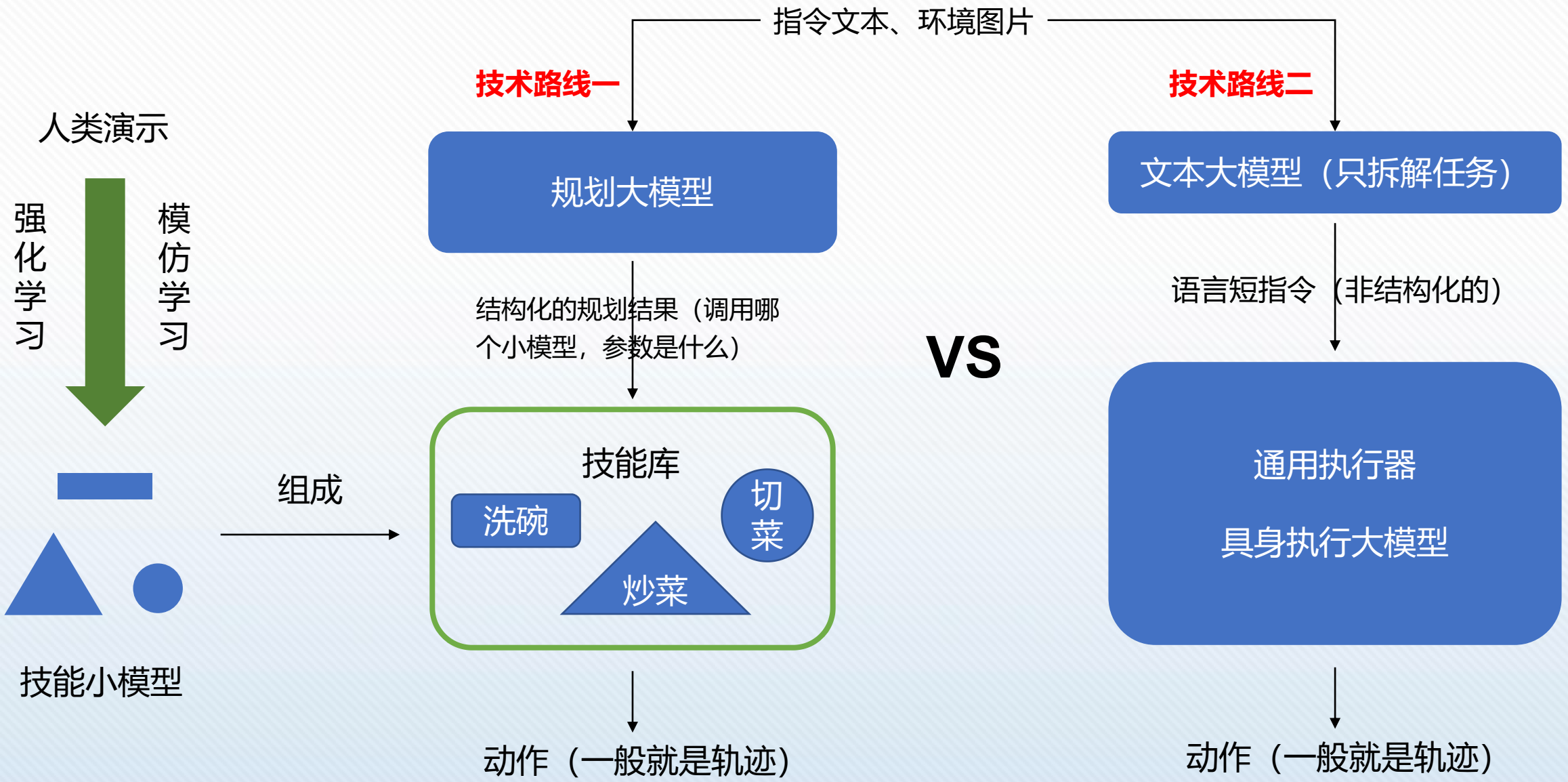
- DQN在抓取中的应用
  - Q网络
  - 目标网络与当前训练网络分离
  - 经验重放
  - …
  - 环境输入
    - 奖励
    - 电机反馈信号
    - 多视角相机
  - 输出
    - 电机动作



[1] Joshi et al Robotic Grasping using Deep Reinforcement Learning.2020 CoRR

3-3 未来方向

通用执行模型


指令文本、环境图片
技术路线一
技术路线二
人类演示
强化学习
模仿学习
规划大模型
文本大模型（只拆解任务）
结构化的规划结果（调用哪个小模型，参数是什么）
语言短指令（非结构化的）
VS
组成
技能库
洗碗
切菜
炒菜
通用执行器
具身执行大模型
技能小模型
动作（一般就是轨迹）
动作（一般就是轨迹）

未来方向

- 鉴于大模型的成功案例，开发一个通用执行模型变得尤为重要
  - 一个高度泛化的通用执行模型，能够理解各种人类文本命令，适应不同的场景配置、物体位置和形状，以及不同机器人的动作空间和操作模式
  - 这样的模型可以持续适应新技能，提供更好的泛化能力，并且学习速度更快
- 此外，模型的性能极大地受到数据质量、模型架构和训练方法的影响
  - 实现卓越的泛化能力需要依赖多样化和高质量的训练数据
  - 通过跨多个任务联合训练，模型可以在不同的任务技能中展示广泛的适应性
  - 当前，像Octo、RT-X和OpenVLA等模型已经在高质量数据集Open X Embodiedemnt的基础上表现出卓越的性能，实现了包括**物体类型、位置、任务场景、机器人类型和人类命令**等多个维度的泛化

什么最重要？泛化！泛化！泛化！以RT-1实验为例

- 第一行从左到右干扰物（Distractors）难度逐渐加大
- 第二行从左到右分别是初始环境，变换桌布图案，新的厨房环境

Model	Seen Tasks	Unseen Tasks	Distractors	Backgrounds
Gato (Reed et al., 2022)	65	52	43	35
BC-Z (Jang et al., 2021)	72	19	47	41
BC-Z XL	56	43	23	35
RT-1 (ours)	**97**	**76**	**83**	**59**

具体效果：RT-1

- 转到真实厨房中的真实场景中，分为三个等级：
  - L1 用于泛化到新的桌面布局和照明条件
  - L2 用于泛化到未见过的干扰对象
  - L3 用于泛化到新的任务设置、新任务对象或未见过的位置（例如水槽附近）

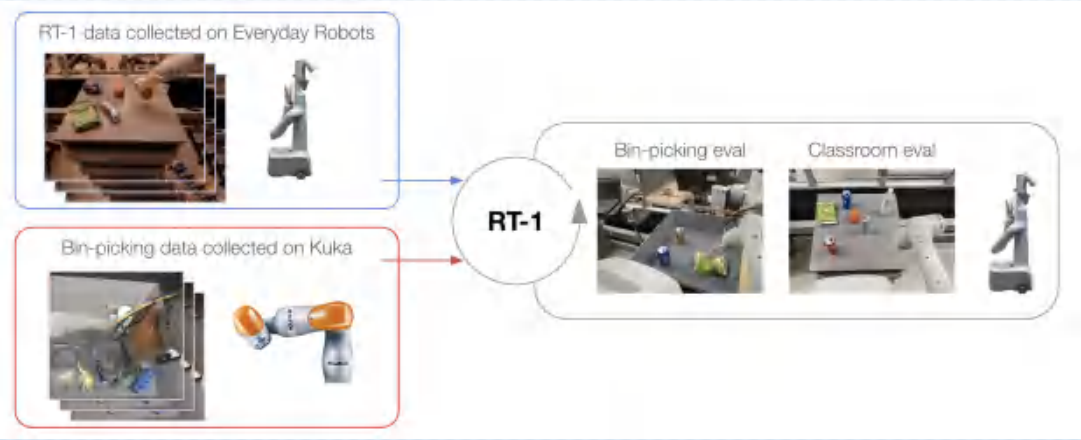
Models	Generalization Scenario Levels			
	All	L1	L2	L3
Gato Reed et al. (2022)	30	63	25	0
BC-Z Jang et al. (2021)	45	38	50	**50**
BC-Z XL	55	63	**75**	38
RT-1 (ours)	**70**	**88**	**75**	**50**

具体效果：RT-1

- 加入仿真数据会降低真实环境下见过的物体和技能上的表现，大幅提高仿真环境下的表现

Models	Training Data	Real Objects	Sim Objects (not seen in real)	
		Seen Skill w/ Objects	Seen Skill w/ Objects	Unseen Skill w/ Objects
RT-1	Real Only	92	23	7
RT-1	Real + Sim	90(-2)	**87(+64)**	**33(+26)**

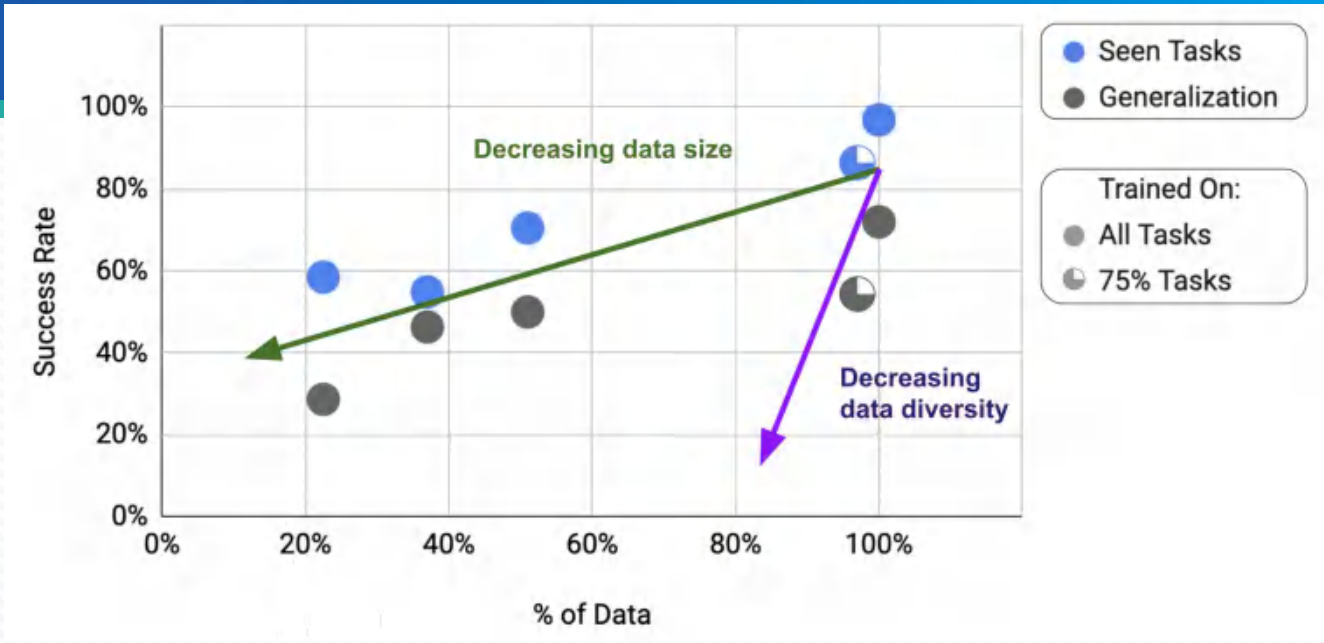
- **加入新类型机械臂在新任务的数据不会造成严重的性能下降，且会提高在新加入任务的表现**



Models	Training Data	Classroom eval	Bin-picking eval
RT-1	Kuka bin-picking data + EDR data	90(-2)	**39(+17)**
RT-1	EDR only data	92	22
RT-1	Kuka bin-picking only data	0	0

具体效果：RT-1

- **数据多样性更重要**
- 减少3%的数据量和25%见过的任务量，
- 几乎等同于减少50%数据量和0%任务量



Models	% Tasks	% Data	Seen Tasks	Generalization			
				All	Unseen Tasks	Distractors	Backgrounds
Smaller Data							
RT-1 (ours)	100	100	97	73	76	83	59
RT-1	100	51	71	50	52	39	59
RT-1	100	37	55	46	57	35	47
RT-1	100	22	59	29	14	31	41
Narrower Data							
RT-1 (ours)	100	100	97	73	76	83	59
RT-1	75	97	86	54	67	42	53

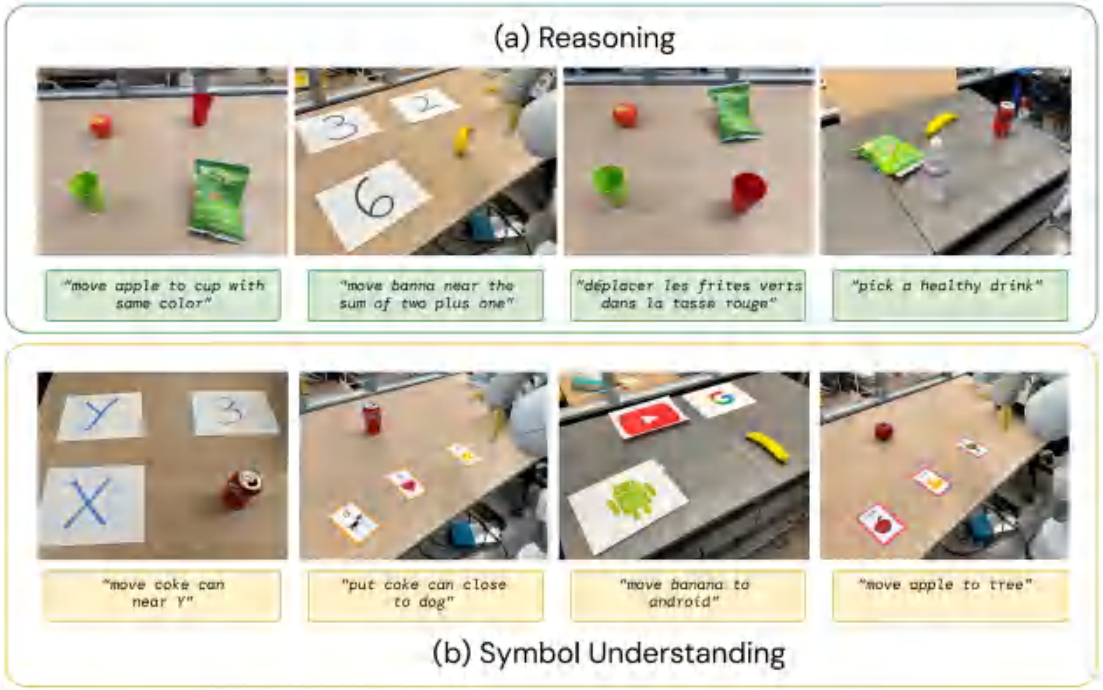
数据规模减少50%

数据多样性降低25%

具体效果：RT-2

- 保持对图片、文字的语义理解能力、推理能力，并与动作策略相结合。**基座大的模型表现更好**

Task Group	Tasks
Symbol Understanding: Symbol 1	move coke can near X, move coke can near 3, move coke can near Y
Symbol Understanding: Symbol 2	move apple to tree, move apple to duck, move apple to apple, move apple to matching card
Symbol Understanding: Symbol 3	put coke can close to dog, push coke can on top of heart, place coke can above star
Reasoning: Math	move banana to 2, move banna near the sum of two plus one, move banana near the answer of three times two, move banana near the smallest number
Reasoning: Logos	move cup to google, move cup to android, move cup to youtube, move cup to a search engine, move cup to a phone
Reasoning: Nutrition	get me a healthy snack, pick a healthy drink, pick up a sweet drink, move the healthy snack to the healthy drink, pick up a salty snack
Reasoning: Color and Multilingual	move apple to cup with same color, move apple to cup with different color, move green chips to matching color cup, move apple to vaso verde, Bewegen Sie den Apfel in die rote Tasse, move green chips to vaso rojo, mueve la manzana al vaso verde, déplacer les frites verts dans la tasse rouge
Person Recognition: Celebrities	move coke can to taylor swift, move coke can to tom cruise, move coke can to snoop dog
Person Recognition: CelebA	move coke can to person with glasses, move coke can to the man with white hair, move coke can to the brunette lady



Model	Symbol Understanding				Reasoning					Person Recognition			Average
	Symbol 1	Symbol 2	Symbol 3	Average	Math	Logos	Nutrition	Color/Multilingual	Average	Celebrities	CelebA	Average	
VC-1 (Majumdar et al., 2023a)	7	25	0	11	0	8	20	13	10	20	7	13	11
RT-1 (Brohan et al., 2022)	27	20	0	16	5	0	32	28	16	20	20	20	17
RT-2-PaLI-X-55B (ours)	**93**	**60**	**93**	**82**	25	52	**48**	**58**	**46**	**53**	**53**	**53**	**60**
RT-2-PaLM-E-12B (ours)	67	20	20	36	**35**	**56**	44	35	43	33	**53**	43	40

具体效果：RT-2

- **参数量大效果更好，混合预训练数据联合微调可以提高泛化性**
- From scratch：仅使用机器人数据，不基于VLM
- fine-tuning：使用机器人数据微调VLM
- Co-fine-tuning：机器人数据和VQA数据一起训练VLM

混合预训练数据分别带来2%和11%的成功率增幅

Model	Size	Training	Unseen Objects		Unseen Backgrounds		Unseen Environments		Average
			Easy	Hard	Easy	Hard	Easy	Hard	
RT-2-PaLI-X	5B	from scratch	0	10	46	0	0	0	9
RT-2-PaLI-X	5B	fine-tuning	24	38	79	50	36	23	42
RT-2-PaLI-X	5B	co-fine-tuning	60	38	67	29	44	24	44
RT-2-PaLI-X	55B	fine-tuning	60	62	75	38	57	19	52
RT-2-PaLI-X	55B	co-fine-tuning	70	62	96	48	63	35	63

增加参数量分别带来10%和20%的增幅

具体效果：OpenVLA

- 定义了**视觉泛化、运动泛化、物理泛化和语义泛化**四种评估任务，OpenVLA仅在语义上落后
  - **visual**：未见过的的背景、干扰物体、物体的颜色/外观
  - **motion**：未见过的物体位置/方向
  - **physical**：未见过的大小和形状
  - **semantic**：未见过的目标物体、指令、概念
  - 额外增加一个多对象的任务，测试能否做到语言和对应物体的grounding

Category	Task	# Trials	RT-1-X # Successes	Octo # Successes	RT-2-X # Successes	OpenVLA (ours) # Successes
Visual gen	Put Eggplant into Pot (Easy Version)	10	1	5	7	**10**
Visual gen	Put Eggplant into Pot	10	0	1	5	**10**
Visual gen	Put Cup from Counter into Sink	10	1	1	0	**7**
Visual gen	Put Eggplant into Pot (w/ Clutter)	10	1	3.5	6	**7.5**
Visual gen	Put Yellow Corn on Pink Plate	10	1	4	8	**9**
Motion gen	Lift Eggplant	10	3	0.5	6.5	**7.5**
Motion gen	Put Carrot on Plate (w/ Height Change)	10	2	1	**4.5**	**4.5**
Physical gen	Put Carrot on Plate	10	1	0	1	**8**
Physical gen	Flip Pot Upright	10	2	6	5	**8**
Physical gen	Lift AAA Battery	10	0	0	2	**7**
Semantic gen	Move Skull into Drying Rack	10	1	0	**5**	**5**
Semantic gen	Lift White Tape	10	**3**	0	0	1
Semantic gen	Take Purple Grapes out of Pot	10	**6**	0	5	4
Semantic gen	Stack Blue Cup on Pink Cup	10	0.5	0	**5.5**	4.5
Language grounding	Put {Eggplant, Red Bottle} into Pot	10	2.5	4	**8.5**	7.5
Language grounding	Lift {Cheese, Red Chili Pepper}	10	1.5	2.5	8.5	**10**
Language grounding	Put {Blue Cup, Pink Cup} on Plate	10	5	5.5	8.5	**9.5**
		Mean Success Rate	18.5±2.7%	20.0±2.6%	50.6±3.5%	**70.6±3.2%**

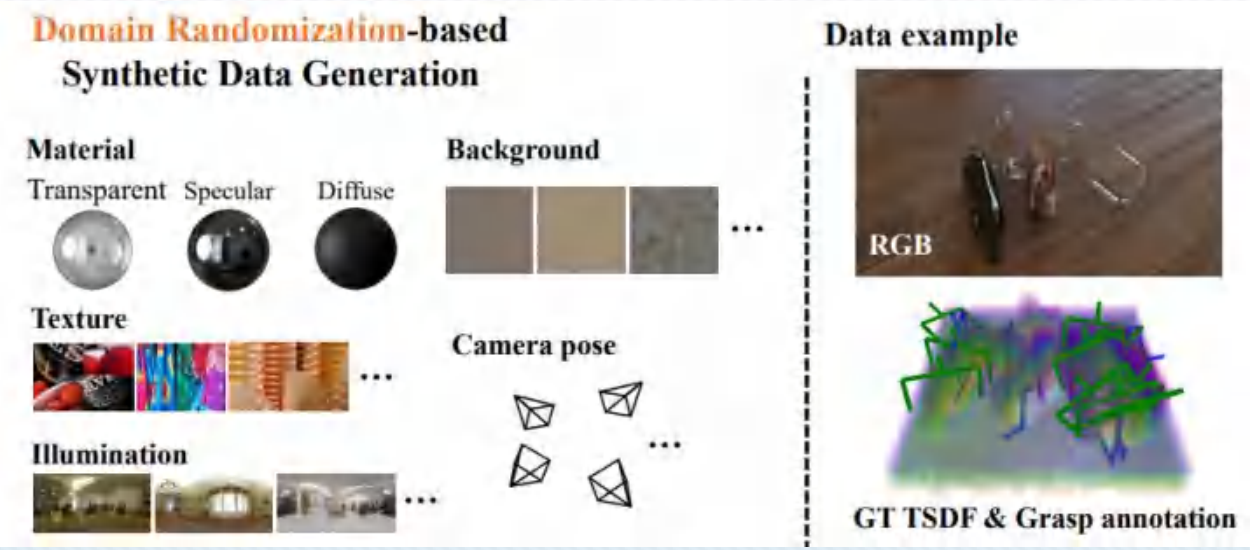
数据 – 大规模收集人工数据

- **Open-X-Embodiment数据集**
  - 汇编了来自22个不同机器人系统的超过一百万个操作轨迹
  - 涵盖了34个机器人研究实验室的60个独立数据集
  - 记录了从拾取到组装等简单到复杂的行动，涵盖了包括家电、食物在内的多种家庭物品
- 研究表明，使用Open X Embodiment数据集训练的模型如RT-X、Octo和OpenVLA，与先前模型相比，在泛化能力上有显著提升
  - RT-X的研究显示，使用足够大的模型，更广泛的数据集显著提高了模型在多个领域的性能
  - Octo在对象定位、照明、背景和干扰等泛化条件下表现出色，与从零开始训练或使用传统微调基线方法相比，展示了更好的学习效率
  - OpenVLA在视觉、动力、物理和语义泛化的各种任务中设定了新的评估标准，虽然在语义泛化方面略有不足，但在将微调转移到新机器人的大多数任务中，相比其他算法表现更佳

[1] Padalkar et al. Open x-embodiment: Robotic learning datasets and rt-x models. 2023 arXiv
[2] Team et al. Octo: An open-source generalist robot policy. 2024 arXiv
[3] Kim et al. OpenVLA: An Open-Source Vision-Language-Action Model. 2024 arXiv

数据 – 自动化构造数据

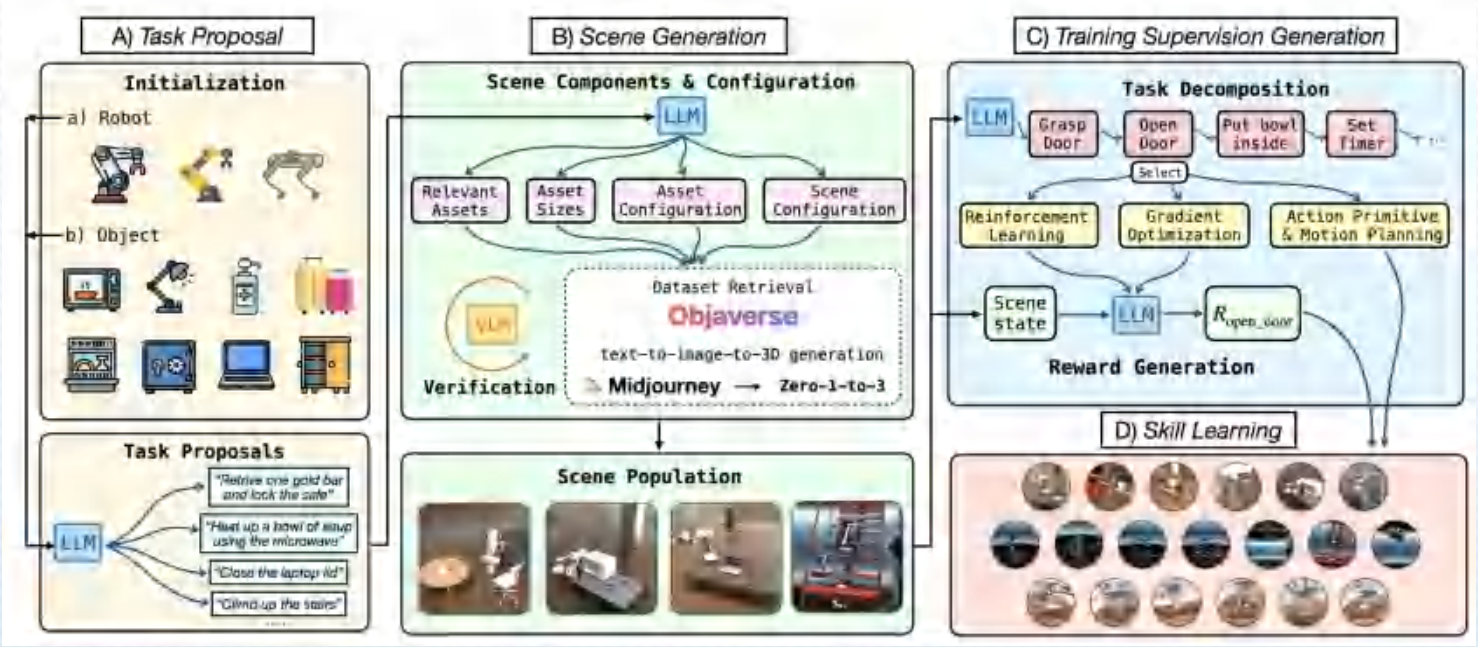
- **GraspNeRF**
  - 基于可泛化的神经辐射场（NeRF）的多视角6自由度（DoF）抓取检测方法
  - 透明物体的特殊材质给深度相机准确感知其几何结构带来挑战，而此方法有效应对了这一挑战
  - 生成大规模、光学真实的领域随机合成数据集
  - 系统能直接从仿真转移到现实世界并泛化出稳健的材料特定表现



[1] Dai et al. Graspnerf: Multiview-based 6-dof grasp detection for transparent and specular objects using generalizable nerf. 2023 IEEE International Conference on Robotics and Automation

数据 – 自动化构造数据

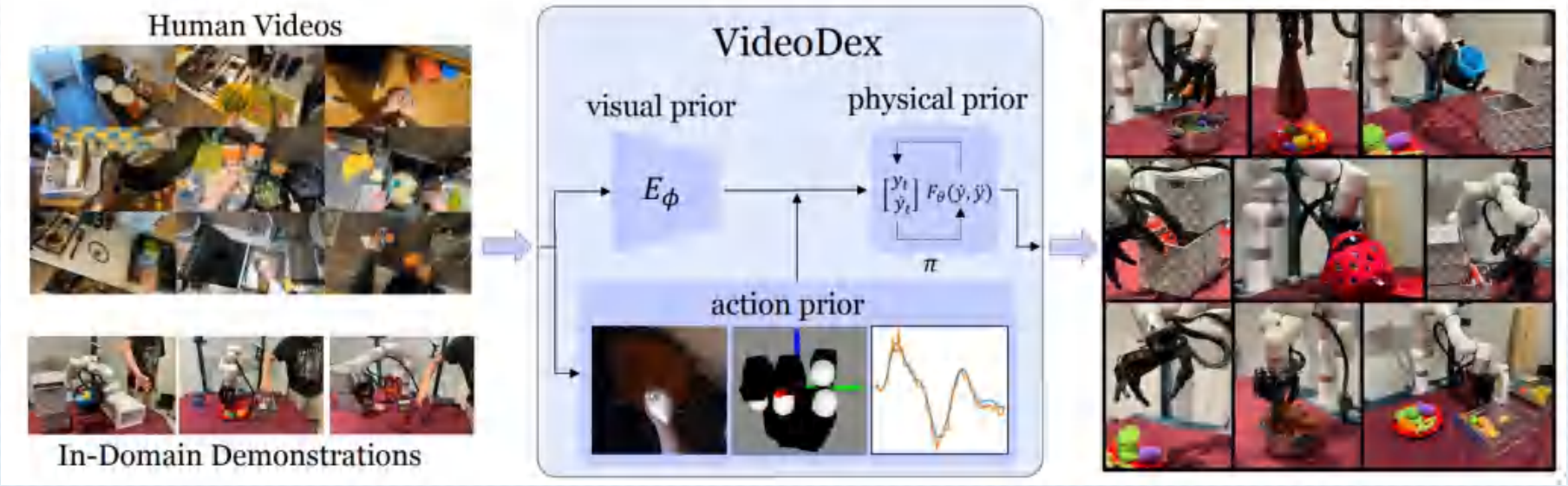
- **RoboGen**
  - 自主提出有趣的任务和技能，然后生成相应的模拟环境以在此基础上学习和获取技能
  - 任务提案：通过随机抽样机器人类型和物体，使用基于GPT-4的语言模型生成任务提案
  - 场景生成：系统为任务生成相应的场景，通过聚合必要的物体增加场景的复杂性和多样性
  - 训练监督生成：RoboGen计划并分解生成的任务为更短的子任务，并选择适当的算法（如强化学习、基于梯度的轨迹优化或带有运动规划的动作原语）进行学习



[1] Wang et al. Robogen: Towards unleashing infinite data for automated robot learning via generative simulation. 2023 arXi

数据 – 无标注数据

- **VideoDex**
  - 使用大量未标记人类视频数据训练机器人的方法，数据源为从互联网收集的大量手部动作视频
  - 通过重新定位和调整运动数据，该方法自动细化动作和视觉先验
  - 随后指导机器人在复杂环境中进行视觉感知和运动规划
  - 特别是，这种方法使机器人能够模仿人类对小型或复杂物体的精细操控，无需额外的训练负担



[1] Shaw et al. Videodex: Learning dexterity from internet videos. 2023 Conference on Robot Learning

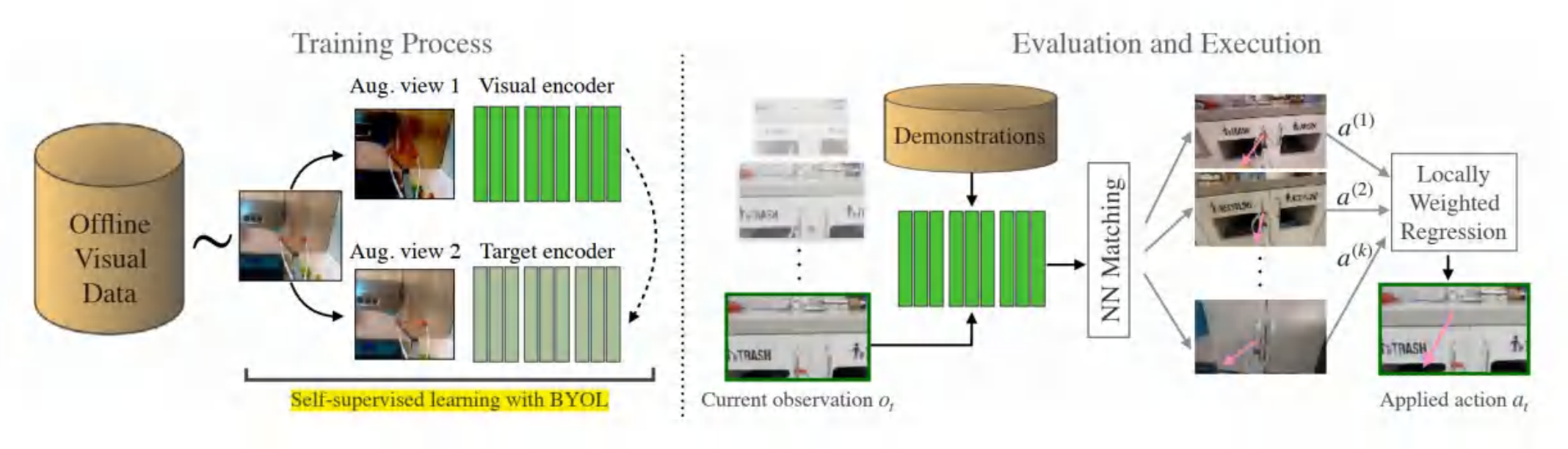
数据 – 无标注数据

- **H2R**
  - 通过观察日常环境中人类行为的视频来学习执行任务
  - 利用基于视觉表示技术的动作预测模型，从单一视点拍摄的初始图像预测人类行为的未来序列
  - 通过采用先进的视觉处理技术，该模型将人类运动轨迹转换为机器人世界坐标系中的动作序列
  - 整个学习过程仅依赖于观察自然人类互动，无需针对机器人任务的先前注释。

[1] Bharadhwaj et al. Zero-shot robot manipulation from passive human videos. 2023 arXiv

Recall：最简单的策略算法 和 有效的视觉表示

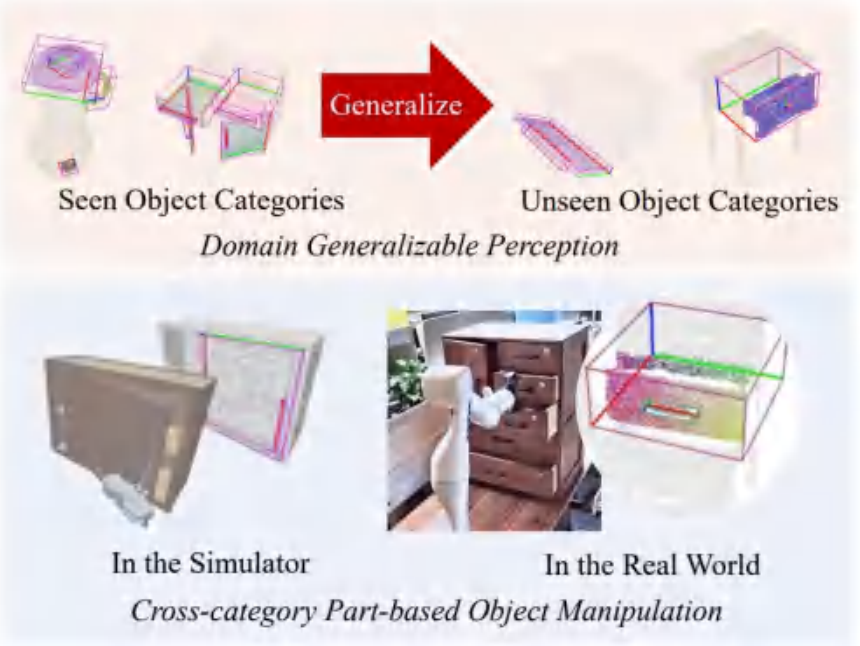
- **无参数**的策略学习算法：对于给定图片，寻找和它最相似的一组图片，取这组图片对应的动作进行加权平均，即为输出动作
- 证明了视觉编码的重要性。使用好的视觉编码，结合简单的策略学习算法，效果可以达到与行为克隆同等水平



[1] Pari et al. The Surprising Effectiveness of Representation Learning for Visual Imitation. RSS 2022

模型 - 视觉表示与动作策略解耦

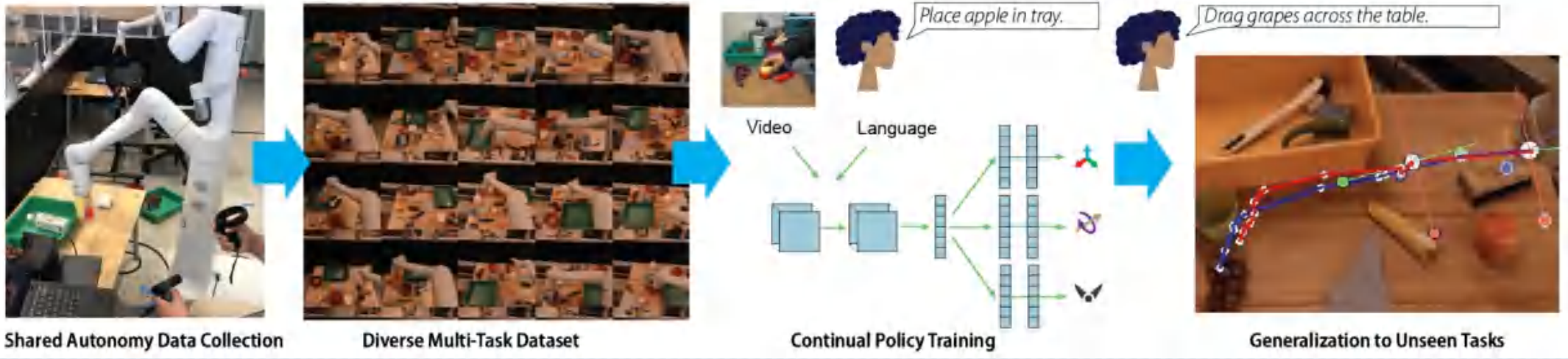
- **GAPartNet**
  - 首先对观察到的物体进行分割和分类，然后根据物体的类别选择特定的操作
  - 使用大规模合成数据训练视觉处理部分，显著增强了模型在真实世界环境中的适用性
  - 特别是对于透明物体或具有特殊姿态的物体，在抓取和操控方面实现了高成功率



[1] Geng et al. Gapartnet: Cross-category domain-generalizable object perception and manipulation via generalizable and actionable parts. 2023 CVPR

训练 - 多任务联合训练

- **BC-Z**
  - 一个基于视觉的机器人操控系统,通过模仿学习实现对新任务的零样本泛化
  - 能够根据不同的任务信息进行条件化，例如来自自然语言的预训练嵌入或人类执行任务的视频
  - 通过将数据收集扩展至涵盖100多项独特任务，研究发现该系统能够在没有任何机器人示范的情况下完成24项以前未见过的操控任务，平均成功率达到44%
  - 该研究展示了模仿学习在处理未知任务中的潜力，并为零样本泛化提供了有效的解决方案



[1] Jang et al. Bc-z: Zero-shot task generalization with robotic imitation learning. 2022 Conference on Robot Learning

3-4 具身执行小结

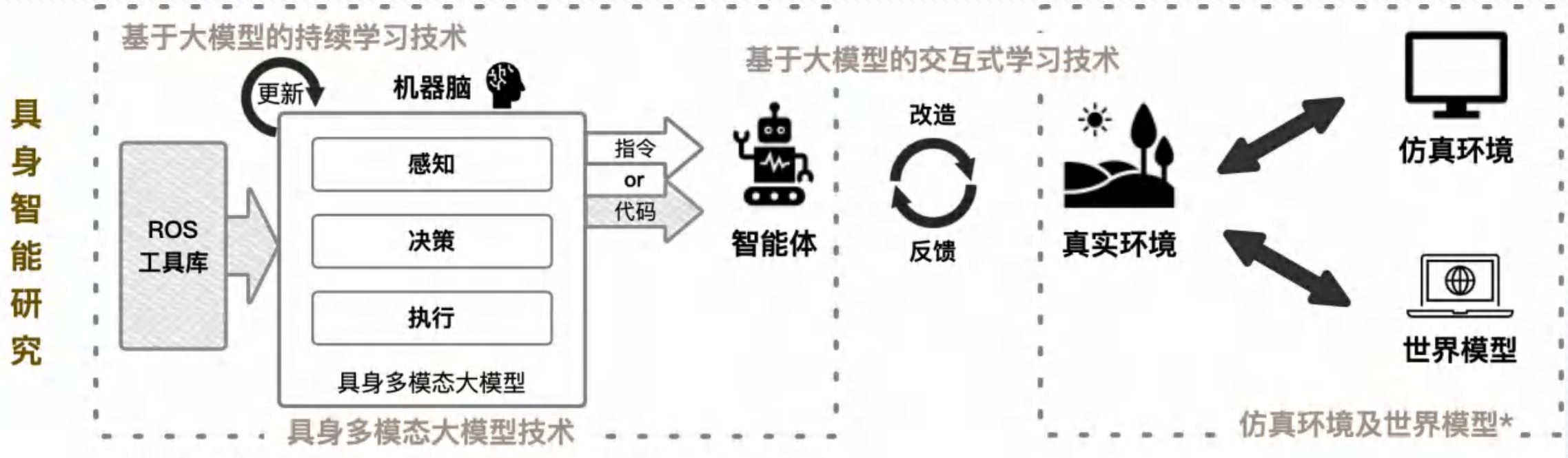
具身执行

- 大模型用于具身执行会存在很多问题：**推理速度慢、数量需求大、可解释性差**
- 但是具身执行强调**泛化性**，对物体位置、形状、场景、技能、机器人类别各种维度上的泛化性，泛化性也是目前最主要的挑战
- 因此大模型仍然是具身执行未来的趋势

- 目前使用大模型压缩大量数据，实现一个比较好的拟合效果，在真实场景数据上有很好的泛化性，仍然是最有可能实现通用执行模型的方式
- 虽然目前10B量级以下大模型的能力不断增强，但是1B以上的模型规模应该有必要的
- 如果不在具身执行上取得突破，实现一个通用执行模型，那么科幻电影中的智能人形机器人就永远不会到来，人工智能也只存在于命令行和对话界面上

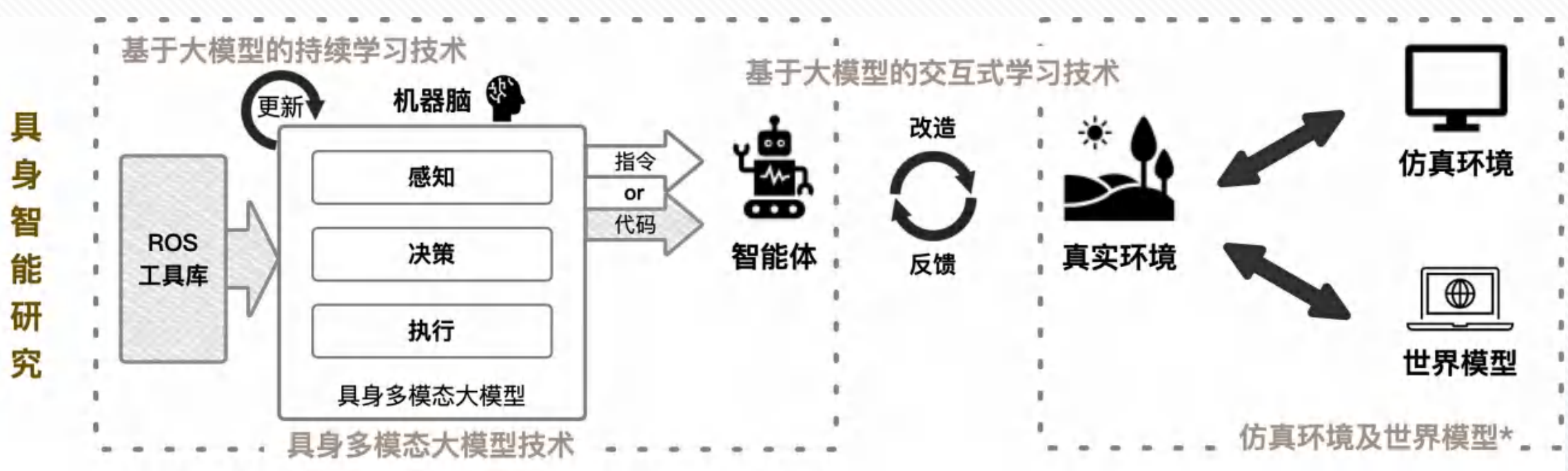
具身智能今后如何发展?

需要深入研究的具身智能技术栈



- 多模态具身智能大模型构建技术
- 基于大模型的交互式学习技术
- 基于大模型的持续学习技术
- 仿真环境及世界模型的构建技术

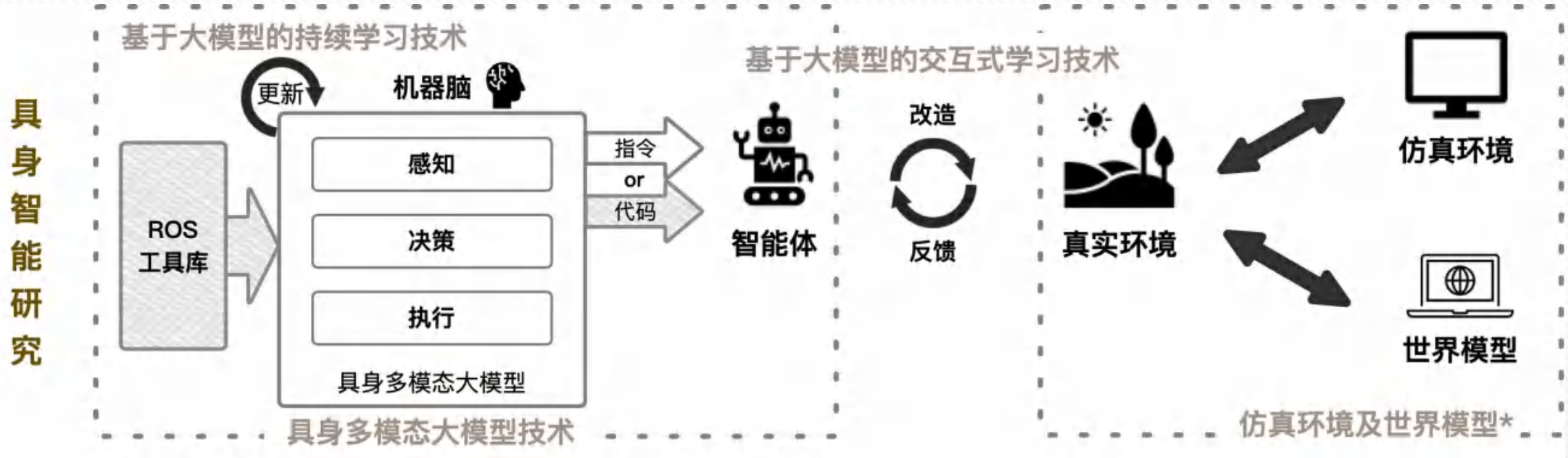
需要深入研究的具身智能技术栈



➢ 多模态具身智能大模型构建技术

- 如何解决数据问题?
- 如何处理复杂和多模态的输入数据?
- 如何输出稳定、像人类的执行动作?
- 如何让模型具备更丰富的世界知识、常识知识?
- 如何像人类一样有逻辑的分析和决策规划?
- 决策规划模型和执行模型能否统一?

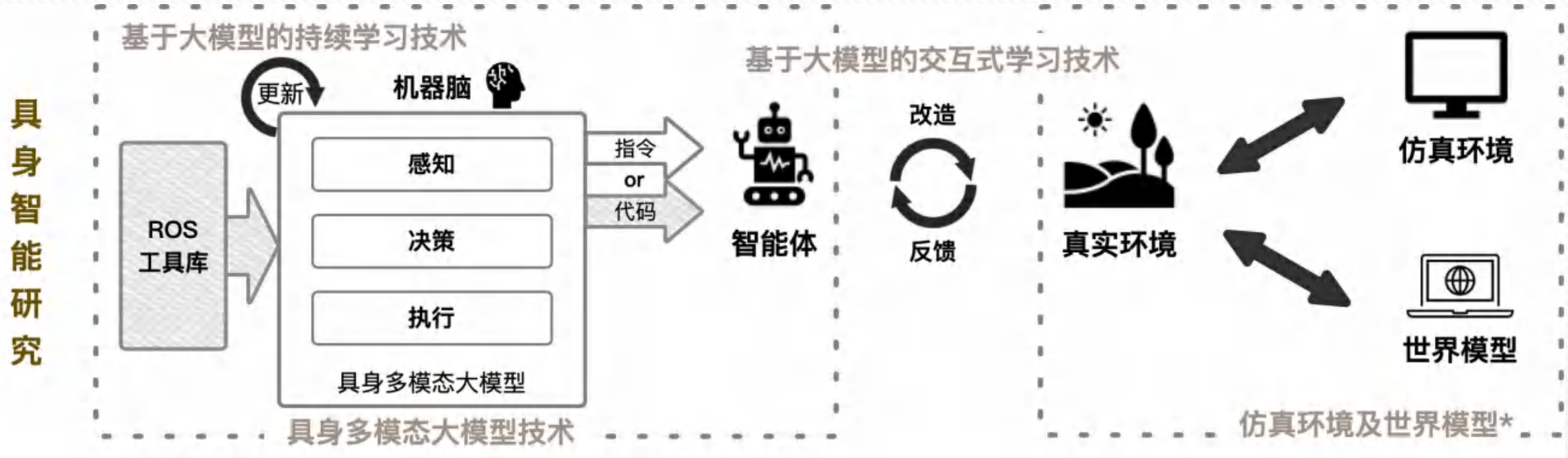
需要深入研究的具身智能技术栈



➢ 基于大模型的持续学习技术

- 人类在一生当中需要不断学习，智能机器人是否也需要这样?
- 如何让智能机器人在不断学习时，避免“狗熊掰棒子”？

需要深入研究的具身智能技术栈



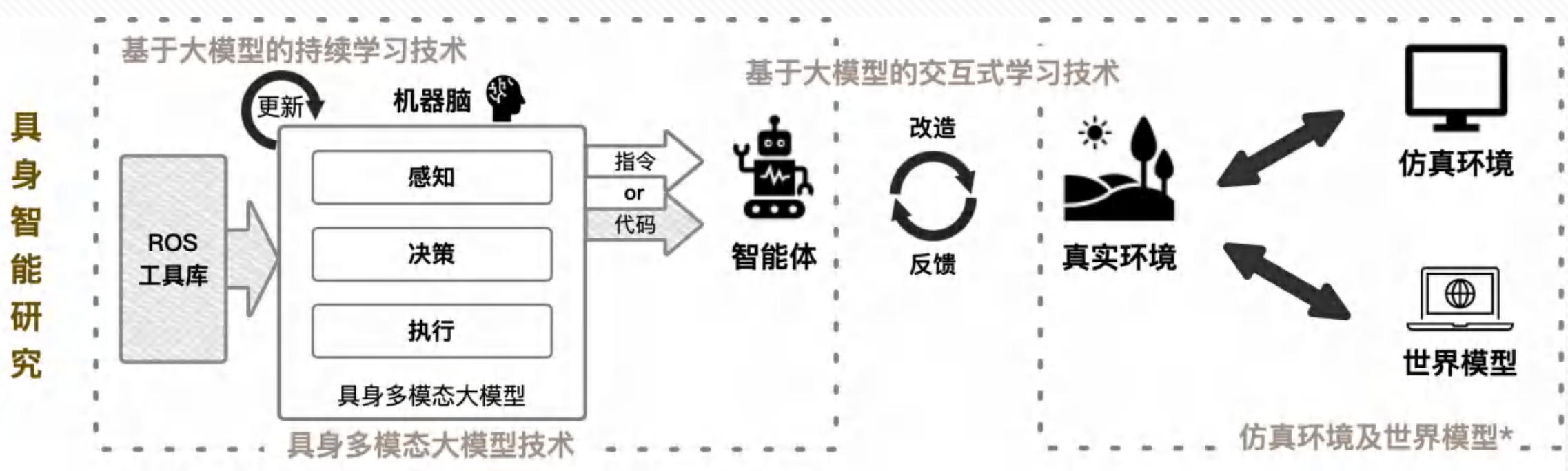
➢ 基于大模型的交互式学习技术

- **人类的学习过程绝大部分是通过与环境交互，智能机器人是否也需如此?**

  人类一直在通过接收环境反馈信息，纠正自己的行为和认知

- **如何让智能机器人在真实世界中自主的去学习?**

需要深入研究的具身智能技术栈



➢ 仿真环境及世界模型的构建技术

- **如何构建可以媲美真实环境的仿真环境或世界模型?**

Real2Sim：模型算法、机器人硬件效果都需要在仿真环境中测试验证，仿真越逼真越有效

世界模型：模型的训练也需要环境反馈怎么办？需要仿真快速给出准确反馈

为什么研究具身智能?

新质生产力

新质生产力，是创新起主导作用，摆脱传统经济增长方式、生产力发展路径，具有高科技、高效能、高质量特征，符合新发展理念的先进生产力质态 。新质生产力作为先进生产力的具体体现形式，是马克思主义生产力理论的中国创新和实践，是科技创新交叉融合突破所产生的根本性成果 。

中华人民共和国中央人民政府
www.gov.cn

发展新质生产力成为“十五五”规划基本思路研究重点

2024-03-27 21:34 来源： 新华社 字号：默认 大 超大 | 打印 |

新华社北京3月27日电（记者 严赋憬、陈炜伟）国家发展改革委主任郑栅洁日前表示，要发挥中长期规划和年度计划的导向作用，做好发展新质生产力的战略部署和任务分解，把发展新质生产力作为“十五五”规划基本思路的研究重点。

人工智能将成发展新质生产力重要引擎

学习强国 xuexi.cn

【沪上新论】人工智能将成发展新质生产力重要引擎

2024-05-24　作者：殷德生

新质生产力是创新尤其是以颠覆性技术和前沿技术创新起主导作用，具有高科技、高效能、高质量特征，符合新发展理念的先进生产力质态。其中，人工智能（AI）技术是颠覆性技术和前沿技术的主要载体之一。

人工智能催生和引领新一轮科技革命和产业变革，也将成为加快培育发展新质生产力的重要引擎。当前，人工智能技术已从“小模型+判别式”转向“大模型+生成式”，参数量由数亿级迅速增至万亿级，模态由单模态训练快速升级到多模态融合。在数字经济的基础上，通过数据、算法、算力的集成创新，生成式人工智能将为新质生产力提供不竭动力。一端是大数据和大型语料，中间是生成式人工智能，另一端就是产品或服务。生成式人工智能应用形成新质生产力的基本路径是：基于更大规模、更高量级的参数，多模态和更加智能的通用大模型，应用持续增长、拥有行业大数据并具备场景的行业大模型（垂类大模型），使用具身智能且成本日益下降的AI机器人，形成“机器替代人”的显著优势，最后实现人工智能技术催生新产业、新业态、新模式蓬勃发展。

通用人工智能的关键分支——具身智能与机器人

（一）指导思想

以习近平新时代中国特色社会主义思想为指导，深入贯彻落实党的二十大精神，完整、准确、全面贯彻新发展理念，加快构建新发展格局，统筹发展和安全，以大模型等人工智能技术突破为引领，在机器人已有成熟技术基础上，坚持应用牵引、整机带动、软硬协同、生态培育的路径，采取技术分级、产品分代、任务分期的方式，发挥制造业门类齐全、应用场景丰富、市场规模庞大以及新型举国体制优势，加快推动我国人形机器人产业创新发展，为建设制造强国、网络强国和数字中国提供支撑。

（二）发展目标

到 2025 年，人形机器人创新体系初步建立，“大脑、小脑、肢体”等一批关键技术取得突破，确保核心部组件安全有效供给。整机产品达到国际先进水平，并实现批量生产，在

二、突破关键技术

（一）打造人形机器人“大脑”和“小脑”

开发基于人工智能大模型的人形机器人“大脑”，增强环境感知、行为控制、人机交互能力，推动云端和边缘端智能协同部署。建设大模型训练数据库，创新数据自动化标注、清洗、使用等方法，扩充高质量的多模态数据。科学布局人形机器人算力，加速大模型训练迭代和产品应用。开发控制人形机器人运动的“小脑”，搭建运动控制算法库，建立网络控制系统架构。面向特定应用场景，构建仿真系统和训练环境，加快技术迭代速度，降低创新成本。

——工信部《人形机器人创新发展指导意见》

具身智能的研究意义

- **宏观意义：**具身智能的发展，可以推动我国工业生产智能化，为建设制造强国、 网络强国和数字中国提供支撑，促进我国实体经济的发展。
- **科研意义：**“具身”意味着主动性和交互性，而这也是目前最接近“智能”的大模型所欠缺的。因此，我们认为，具身智能是通用人工智能未来的发展方向。
- **应用意义：**

家务机器人

四足机器狗

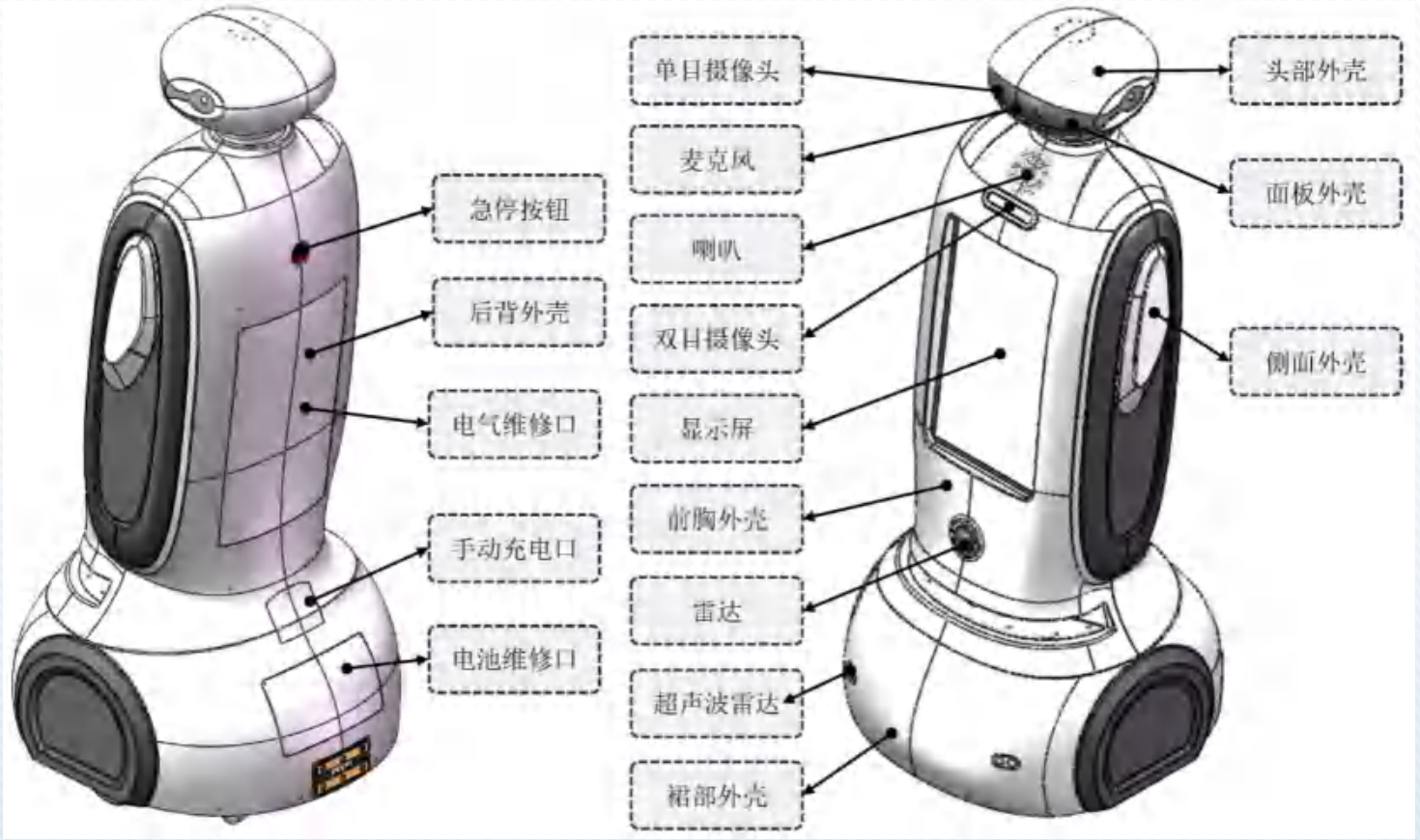
物流机器人

工业机器人

具身智能的行业应用

产业分析——我们的尝试

- 为解放展厅人类讲解员的重复劳动，团队开发可全自动导览或人机协同半自动导览的**展厅讲解机器人**



轮式机器人“小红”

产业分析——我们的尝试

- **语音识别准确率高**：在35dB，45dB，55dB的环境音测试中，用户整句的识别成功率达到了99%（即录音的起始点判断准确率极高）
- **人机交互覆盖范围全**：人工设计了20个常见的多轮问答和多个导航指令，该机器人的回答准确率高达95%、而导航指令均能命中对应的展区名称
- **检索增强稳定高效**：针对两万字的数据设计了30个问题，实现了100%的命中率

Tesla — Optimus Gen2

- Tesla推出Optimus的二代版本，其能力比一代有了长足的进步。但目前其技术细节并未公布

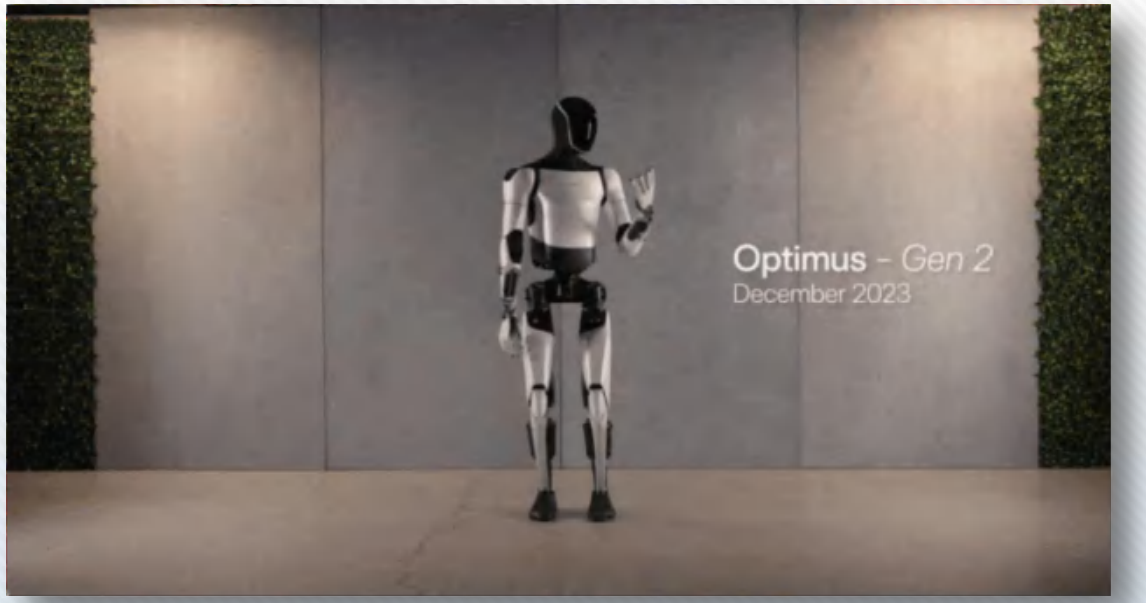
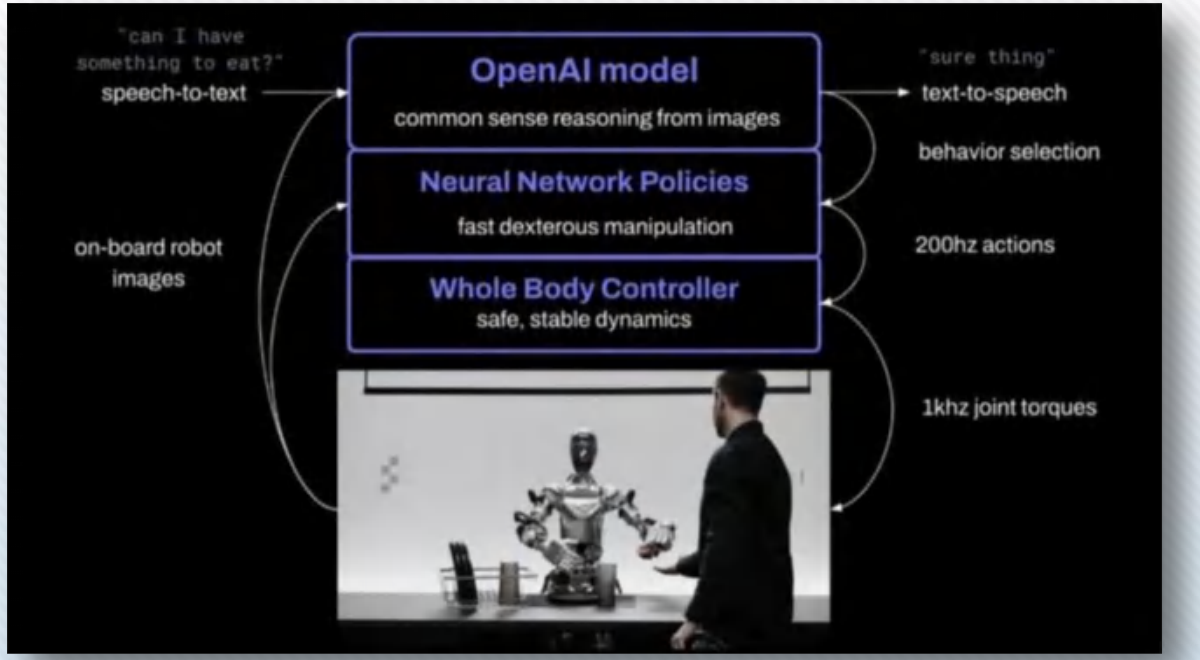


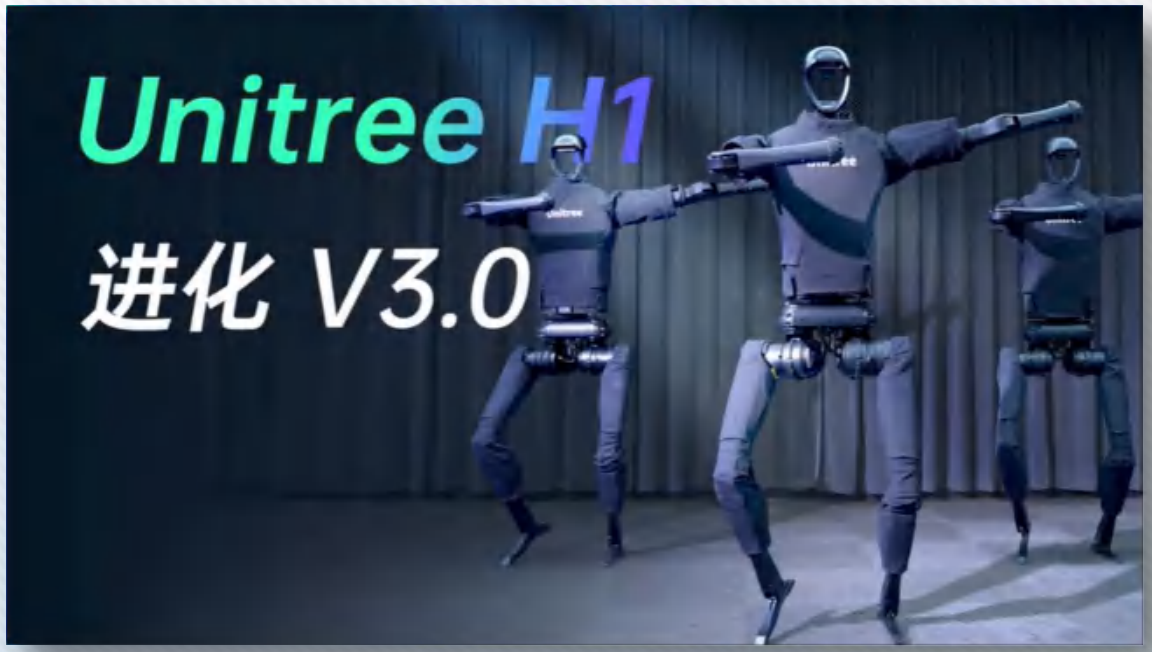
Figure — Figure 01

- OpenAI ChatGPT与机器人结合的产物，Figure 01具备强大的与人深入交流，同时独立做出决策和执行命令的能力
- Figure 01通过OpenAI多模态大模型提供视觉和文本的理解能力，并通过策略神经网络提供快速，低级，灵巧的机器人动作



宇树 — H1

- 宇树科技打造全国第一台能跑的全尺寸通用人形机器人，在机动性、灵活性等方面具备优势，移动速度达到世界领先水平
- H1机器人的机器脑智能化水平暂未公布



行业应用机器人

宇树B2机器狗

优必选清洁机器人

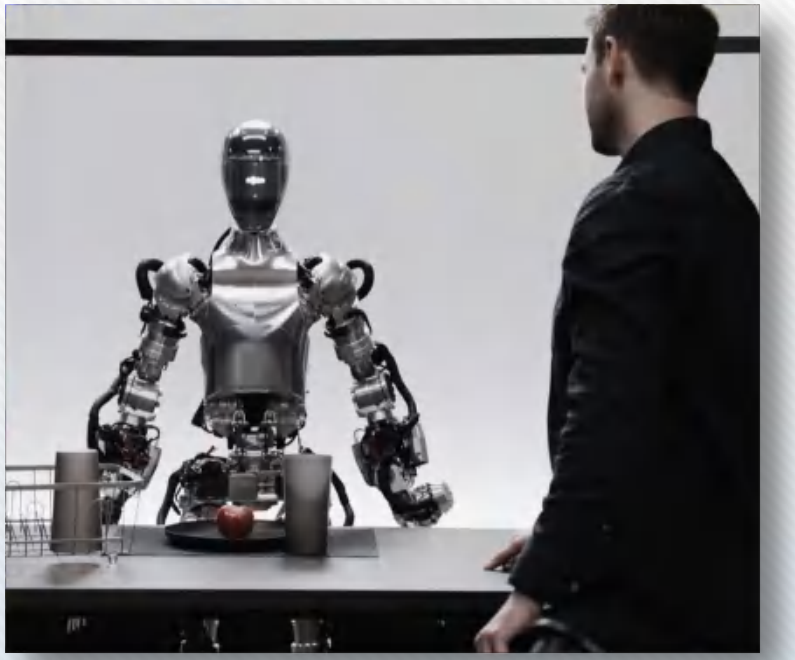
优必选递送机器人

优必选导览机器人

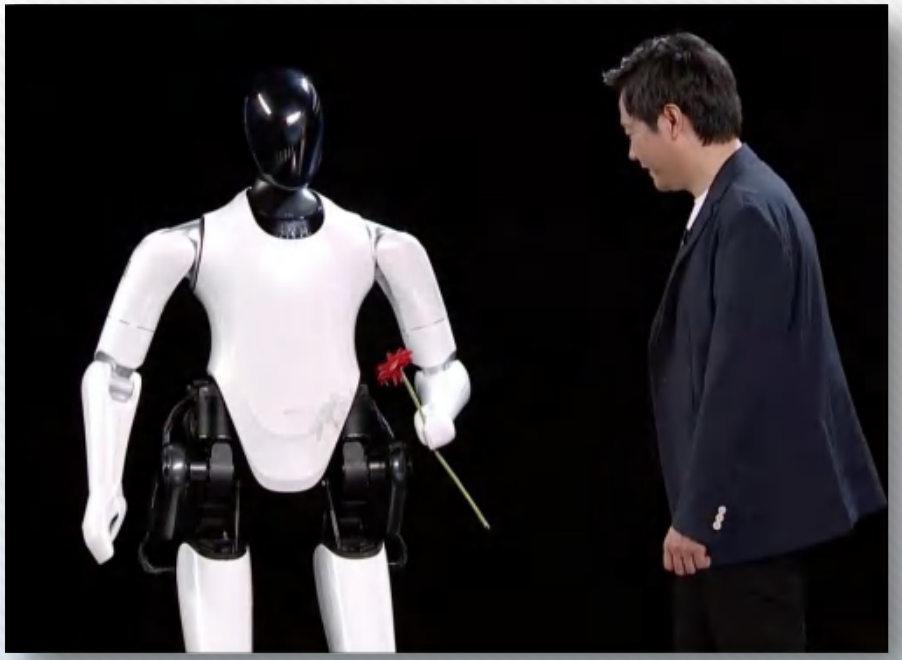
产业分析——人形机器人行业热潮

国际投资银行高盛援引的一份报告指出，如果人形机器人的产品设计、用例、技术、负担能力和广泛公众接受度方面的障碍得到彻底克服，**预计到 2035 年，市场规模将高达 1540 亿美元。**

优必选（中国深圳） Walker S

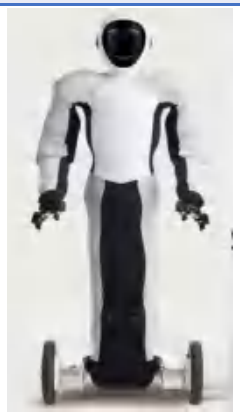
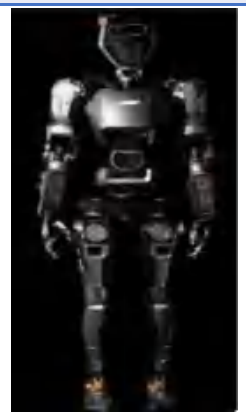


美国 Figure(OpenAI) Figure 01

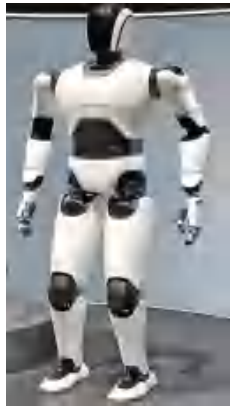
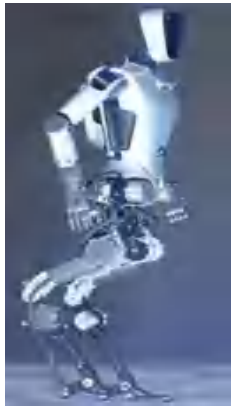


小米 CyberOne

产业分析——海外代表性公司和产品

公司	本田	波士顿动力	Agility Robot	特斯拉	1X	Figure	Sanctuary AI
产品名称	Asimo	Atlas	Digit	Optimus	Eve	Figure 01	phoenix
产品图片							
时间	2011	2013	2018	2021	2019	2024	2023
应用领域	交互服务	未明确	物流、工业制造	工业制造	医疗护理	通用机器人	通用机器人
最新进展	20年终止研发		融资1.5亿美元	小规模量产	融资1.25亿美元	融资6.75亿美元	结合LLM/人工智能

产业分析——国内代表性公司和产品

公司	优比选	小米	追觅	智元	傅立叶智能	宇树	达阔
产品名称	WalkerX	Cyber one	通用人形机器人	远征A1	GR1	H1	XR4
产品图片							
时间	2021	2022	2023	2023	2023	2023	2023
应用领域	生活服务	生活服务	家务	工业制造	生活、医疗	C端服务	C端交互
最新进展	市值超过1000亿			最新融资6亿元	最新融资4亿元	最新融资10亿元	最新融资超10亿元

产业分析——关注度较高的创业公司

公司名称	成立时间	高管团队	产品范围	估值
银河通用	2023-05-19	王鹤，博士斯坦福，北大助理教授，北京智源具身大模型中心主任 姚腾洲，北航机器人硕士，曾就职于ABB机器人研发中心	聚焦**药店、商超**等**零售场景**，专注研发**双臂轮式机器人**，预计2024年发布，2026年量产	美团为公司第一大外部股东 估值：数亿美金
星海图	2023-09-05	许华哲，本科清华，博士加州伯克利，博后斯坦福，清华叉院助理教授 高继扬，本科清华，博士南加州大学，多家无人驾驶企业就职，引用量4000+	坚持AI算法与本体协同研发的技术路线，从需求出发，**自主设计并制造本体**	天使轮：融资额千万级美元
穹彻智能	2023-11-02	卢策吾，上海交大教授，师从李飞飞王世全，本科浙江大学机械，博士斯坦福大学	**智能机器人；服务消费机器人；通用应用系统；硬件销售；**	融资金额千万级别，估值数亿

产业分析——关注度较高的创业公司

公司名称	成立时间	高管团队	产品范围	估值
星动纪元	2023-08-04	陈建宇，博士加州伯克利，清华叉院助理教授	人形机器人**硬件本体**；以大语言模型和力控算法构建的人形机器人**智能模块。**	获超亿元天使轮融资
千寻智能	2024-01-16	高阳，博士加州伯克利，清华叉院助理教授 **韩峰涛**，珞石机器人联合传世人，浙江大学	智能机器人研发；硬件销售、人工智能应用软件开发、互联网销售等	——
加速进化	2023-06-20	程昊，清华大学硕士，字节跳动游戏部门朝夕光年创始人	**人形机器人本体制造**和**运控开发平台**	千万元天使轮融资

产业分析总结——技术落地还需时间

- 通用机器人产业分析
  - 头部聚集效应严重，仅排名靠前的头部企业才容易存活
  - 需要大量的供应链优势，以加速机器人量产
  - 需要长期积累，短期难以获得收益
    - 难以商业化落地到特定场景
    - 难以通过技术突破获得融资

通用机器人具有重要的研究意义和应用价值，但短期内难以收获效益。且该行业竞争激烈，需要足够的资金和快速的技术迭代才能取得优势

具身智能的机遇与挑战

具身智能研究的机遇

- **模型和算法创新：**具身智能需要新的模型与算法，通过交互提升机器的感知、认知和决策能力

非交互的算法：学习有遮挡物体的识别 → **交互**的算法：学习如何移开遮挡识别物体

- **实验平台的发展：**机器人为各种感知、认知、决策算法提供了落地平台，研究人员可以在真实环境中对算法进行测试
- **多学科交叉：**具身智能是人工智能、机器人学、人机交互等多学科融合的研究方向，有些问题或许换个视角就能解决
- **技术转移潜力：**具身智能的研究可以促进技术的转移，推动本方向的发展，如CV为Robotics提供视觉功能

AI & Robotics：我们希望打造儿童情感陪伴机器狗，如何才能更智能？

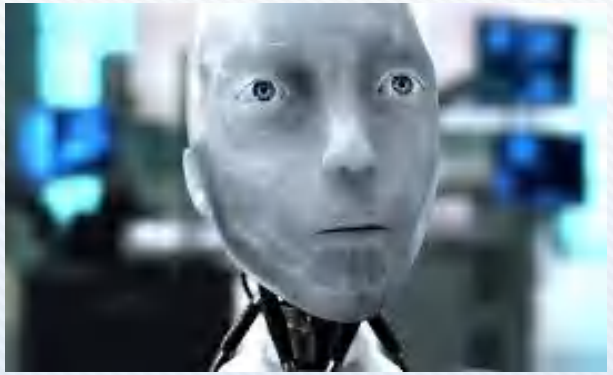
人机交互：低智能的陪伴机器狗也能满足需求，过于智能会导致恐怖谷效应

具身智能研究的挑战

机器人复杂系统实现的挑战：
智能化机器人包括感知、决策和行动，系统设计和实现的复杂性极高

机器人持续学习进化的挑战：
人类社会在发展，机器人也要不断的学习新工具、提高自身能力

机器人量产和商业化的挑战：
智能化算法需要达到低资源、低成本、高可控性、高稳定性的商业化、产品化需求



机器人伦理安全的挑战：
确保智能系统的行为符合人类价值观并且不构成威胁

机器人伦理安全的挑战

能够精准杀伤人类目标的智能无人机——AI 杀人蜂

机器人伦理安全的挑战

美国前任参谋长联席会议主席马克·米利

2024年7月15日，马克·米利称：

2039年，美国作战部队将有1/3是机器人

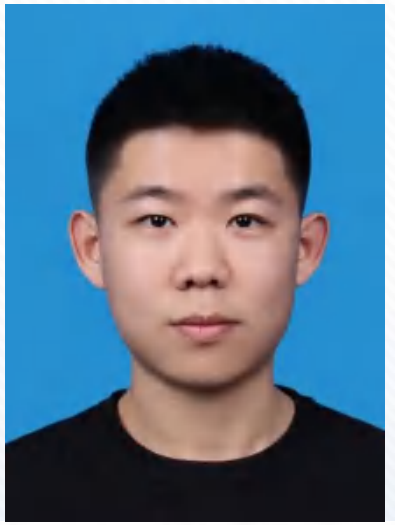
超级士兵机器人

机器人伦理安全的挑战

美国前任参谋长联席会议主席马克·米利

“从技术角度出发，你可以想象在未来，人工智能驱动的机器、人工智能驱动的机器人可以自己做出决定。这是世界想要的吗？”

谢谢！



张伟男

长聘教授/博士生导师

哈工大人工智能学院执行院长

兼计算学部副主任

国家级青年人才

邮箱：wnzhang@ir.hit.edu.cn

主页：https://homepage.hit.edu.cn/zhangweinan

王雪松 博士在读

陈一帆 硕士在读