

具身智能基础技术路线

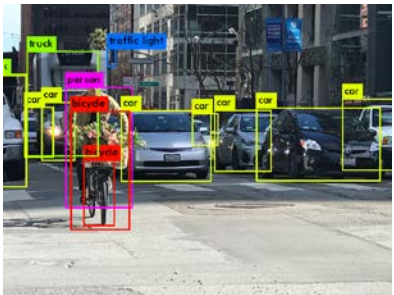
董云龙

2024.05

<https://github.com/yunlongdong/Awesome-Embodied-AI>

具身智能 (Embodied AI)

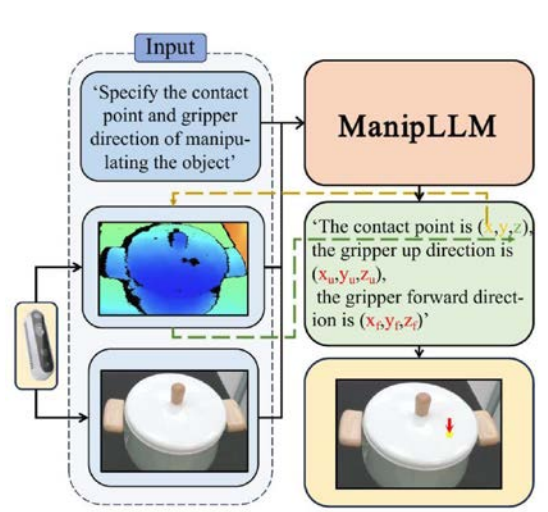
Embodied AI: 集成环境理解、智能交互、认知推理、规划执行于一体的系统化方案



环境理解



智能交互



Please pass me the blue empty plate.

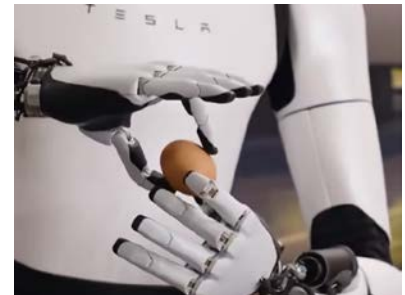


- 1. Pick up apple
- 2. Place apple on table



- 3. Pick up banana
- 4. Place banana on table

认知推理

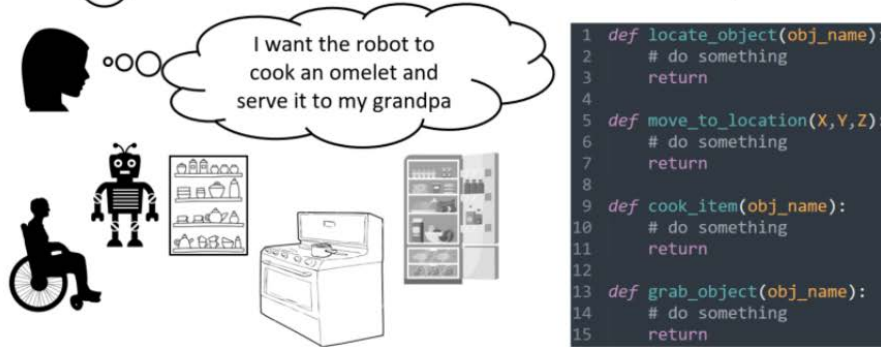


规划执行

LLM As General Planner

通过**包罗万象**的Tokens训练的LLM，其内蕴的**Common Sense**以及**思维逻辑**，能够成为面向机器人的General Planner

① Define a task-relevant robot API library*



A silhouette of a person's head is shown with a thought bubble containing the text: "I want the robot to cook an omelet and serve it to my grandpa". Below this, a robot is depicted in a kitchen environment with various appliances like a stove, refrigerator, and shelves. To the right, a code block lists the following Python functions:

```
1 def locate_object(obj_name):
2     # do something
3     return
4
5 def move_to_location(X,Y,Z):
6     # do something
7     return
8
9 def cook_item(obj_name):
10    # do something
11    return
12
13 def grab_object(obj_name):
14    # do something
15    return
```

*APIs should be easily implementable on the robot and have descriptive text names for the LLM. They can be chained together to form more complex functions.

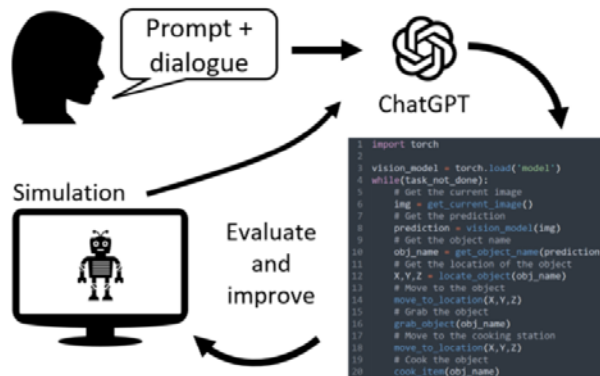
② Build prompt following engineering principles

Consider you are a home assistant robot. Your goal is to prepare an omelette for an elderly person. You are equipped with functions:

- `locate_object(obj_name)`: returns a X,Y,Z tuple representing the location of the desired object defined by string "obj_name";
- `move_to_location(X,Y,Z)`: moves the robot's hands to a specific X,Y,Z location in space. Returns nothing;
- `cook_item(obj_name)`: cooks a particular item defined by "obj_name". Returns nothing;
- `grab_object(obj_name)`: picks a particular object defined by "obj_name". Returns nothing;

Output python code with the sequence of steps that achieves your objective.

③ User on the loop: iterate on solution quality and safety



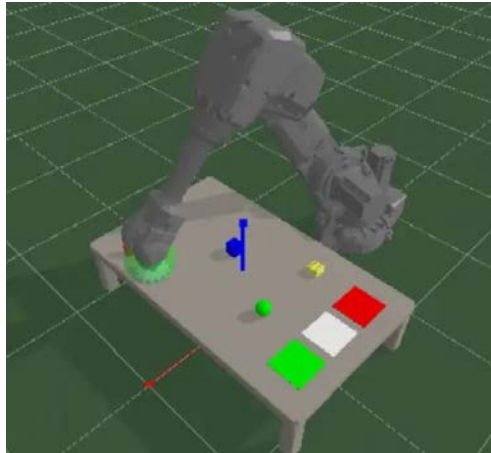
A diagram illustrating a feedback loop. A person's head is shown with a speech bubble labeled "Prompt + dialogue" pointing to the ChatGPT logo. An arrow from ChatGPT points to a code block containing Python code for a simulation. An arrow from the code block points to a computer monitor displaying a robot in a simulation environment. A speech bubble labeled "Evaluate and improve" points from the simulation back to ChatGPT, completing the loop.

```
1 import torch
2
3 vision_model = torch.load('model')
4 while(task_not_done):
5     # Get the current image
6     img = get_current_image()
7     # Get the prediction
8     prediction = vision_model(img)
9     # Get the object name
10    obj_name = get_object_name(prediction)
11    # Get the location of the object
12    X,Y,Z = locate_object(obj_name)
13    # Move to the object
14    move_to_location(X,Y,Z)
15    # Grab the object
16    grab_object(obj_name)
17    # Move to the cooking station
18    move_to_location(X,Y,Z)
19    # Cook the object
20    cook_item(obj_name)
```

④ Execute!



LLM As General Planner: TinyRobotBench



PyBullet演示环境

Prompt

你现在是一个负责给机器人生成规划代码的模块，以下是你能够直接调用的API:

1. `get_location_by_name(name)`: 根据输入的name获取其对应物体所在的xyz位置，用来找到对应物体的位置
2. `move_tool(xyz)`: 将末端夹爪移动到xyz位置。
3. `grasp()`: 末端夹爪执行抓取，能够抓住当前机器人夹爪附近的物体。
4. `ungrasp()`: 松开末端夹爪抓取，所夹住的物体会落在当前位置。
5. `get_names_on_table()`: 返回类型List，返回所有在桌面上物体的名字。
6. `get_box_postion()`: 返回可以暂时存放物体的box的xyz位置。

请你根据上述API，基于python只实现编码plan()函数，将桌面上所有的物体移动到目标位置，只输出代码，不需要输出其他描述。

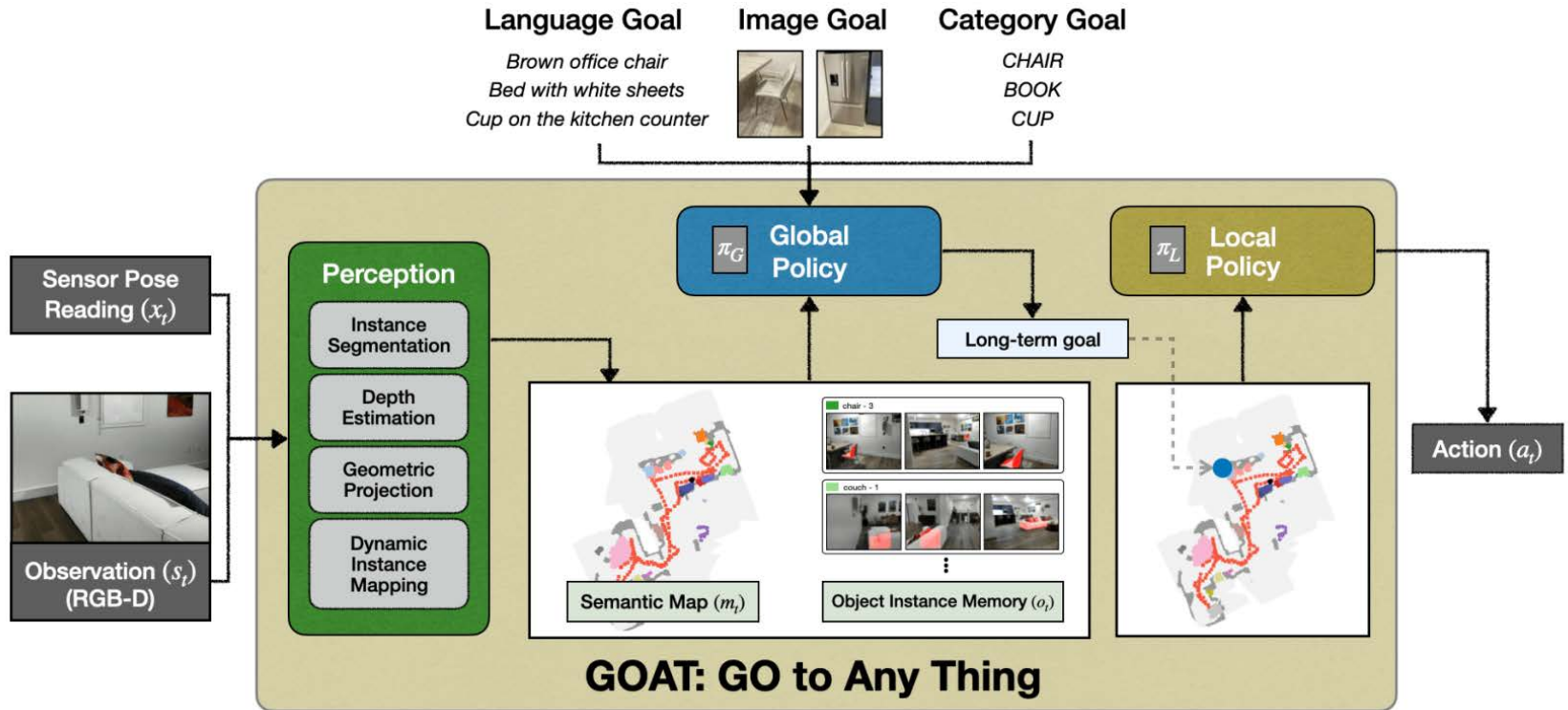


Code

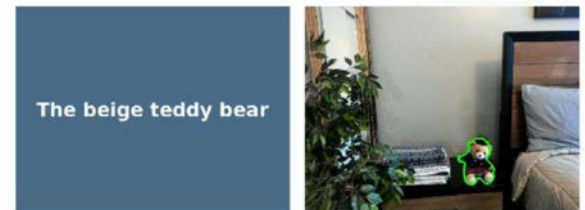
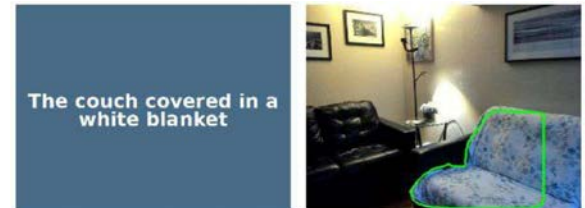
```
def plan():  
    table_objects = get_names_on_table()  
  
    for obj_name in table_objects:  
        obj_location = get_location_by_name(obj_name)  
        target_box_position = get_box_postion()  
  
        move_tool(obj_location)  
        grasp()  
        move_tool(target_box_position)  
        ungrasp()
```

LLM可以作为一个桥接人类指令意图到具体规划代码生成的媒介

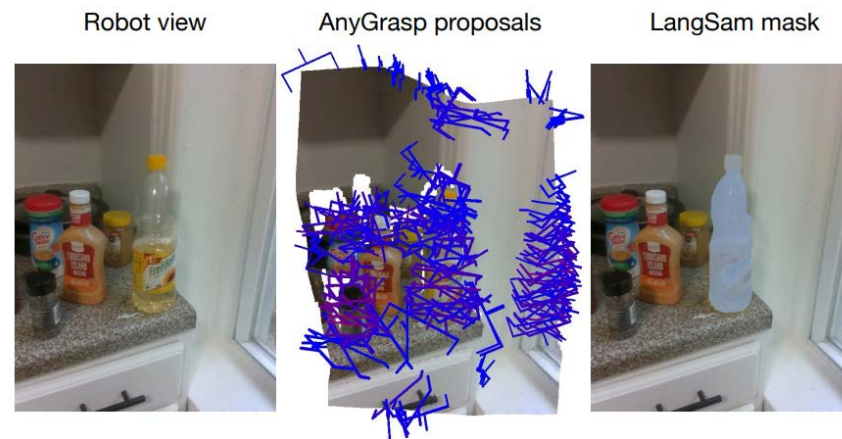
Embodied AI: A System View (Navigation)



- 使用**MaskRCNN**实例分割进行目标检测和像素分割
- 使用**MiDaS**单目深度估计进行RGBD传感器数据修复
- 分割后的RGBD投影**Semantic Map**进行环境建图
- 使用**SuperGLUE**进行图像与图像匹配
- 使用**CLIP**进行文本与图像匹配
- 使用**Mistral 7B**从复杂指令抽提Object Category

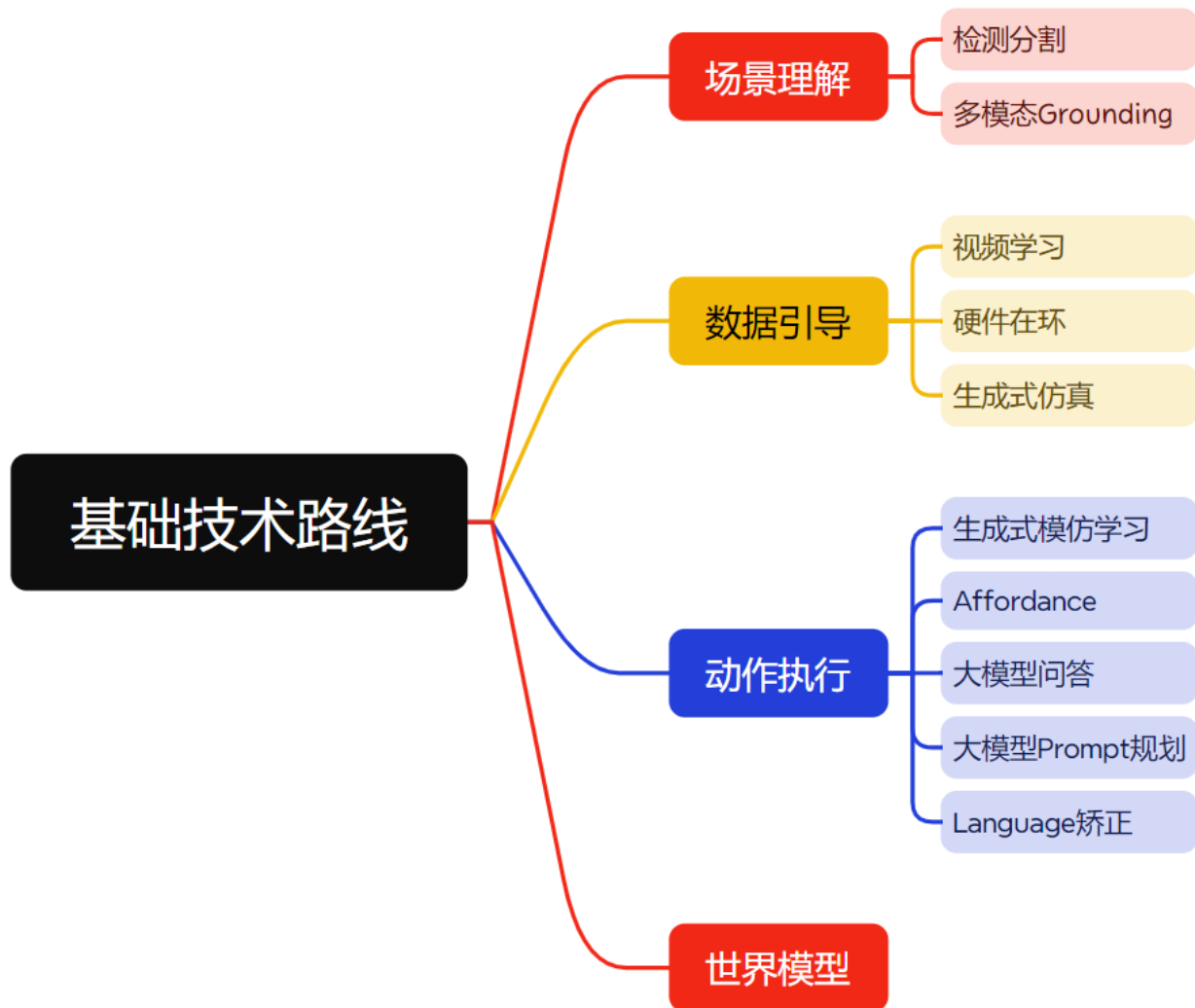


Embodied AI: A System View (Navigation & Grasping)



- 使用**AnyGrasp**生成Grasping Candidates
- 使用**Lang-SAM**, 分割特定文本物体Mask
- 基于规则在Mask内选择最终Grasping Pose

基础技术路线



从**场景理解**、**数据引导**、**动作执行**、**世界模型**这四个方面进行基础技术路线的拆解

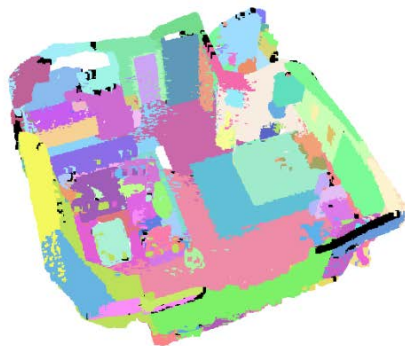
场景理解——分割、检测



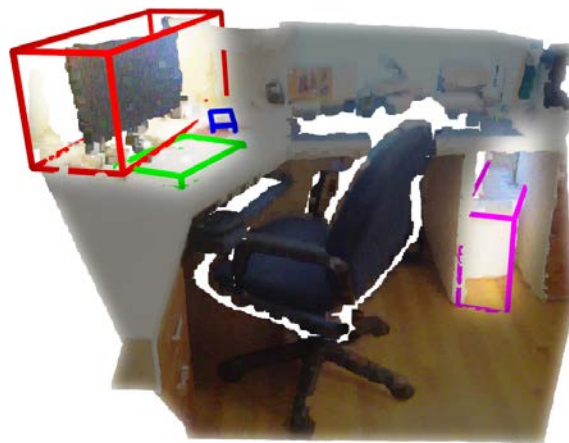
SAM



Open-Voc Detection



SAM3D



Open-Voc Detection in Point Cloud

场景理解——多模态Grounding (Image)



Describe the image in detail. Please output interleaved segmentation mask



The image showcases a large, white **building** with a **red roof**, surrounded by a well-manicured lawn and palm **trees**. The **sky** is visible over the building, the **pavement**, and the **grass**. The grass is also seen extending to the pavement.

多模态（文本、图像）大模型能够实现**像素级别**的细粒度Grounding，表征着大模型的理解能力大幅度提高

Can you please describe the region ?



A man sitting on a motorcycle

Why does he need this ?



It is a helmet and he needs it for protection while riding his motorcycle.

Prompt的**多样化**使得Grounding具有更好的**针对性**

场景理解——多模态Grounding (3D)

Scene Captioning



LEO: *In this room, there is a wooden floor that is clean and flat. A tall wardrobe ...*

3D Question Answering

Question: Where is the Printer located?

LEO: *Standing on the desk next to the wardrobe.*

3D Dialogue

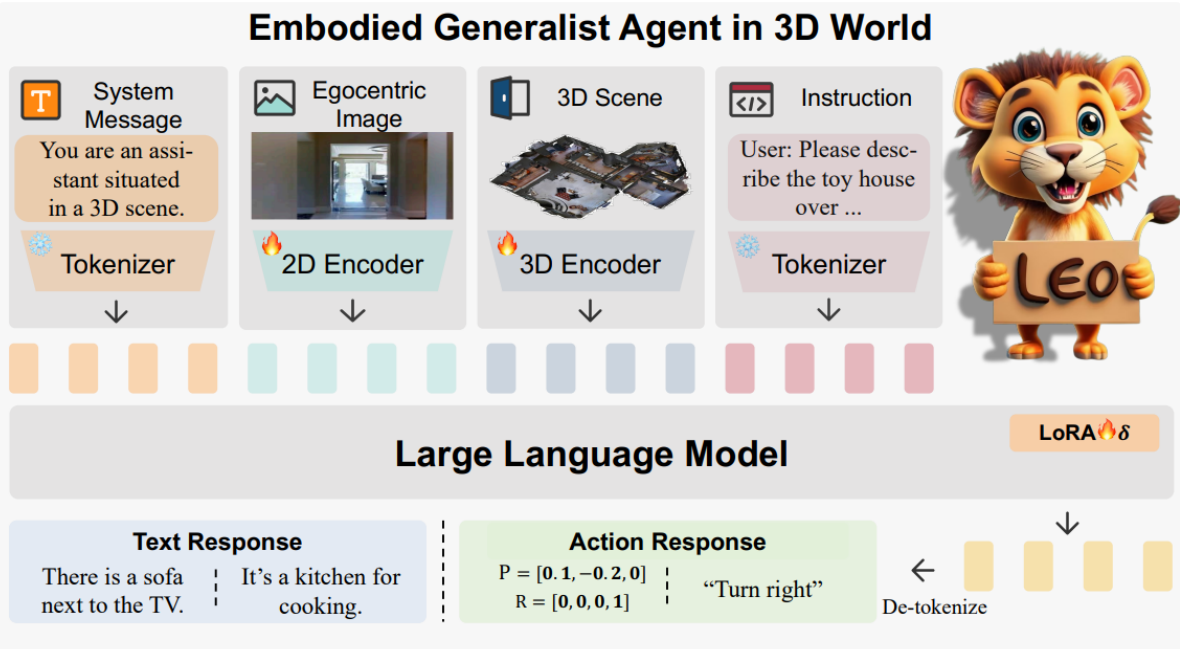
Human: I need help organizing my bedroom. Can you assist me?

LEO: *Of course! What specifically do you need help with?*

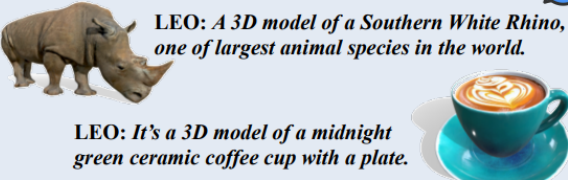
Human: Can you check if there are any clean clothes hanging on the wall?

LEO: *I'm sorry, but the clothes hanging on the wall are messy.*

Human: Thanks! Also, can you check if there are any clean clothes hanging on the wall?




3D Object Captioning



LEO: *A 3D model of a Southern White Rhino, one of largest animal species in the world.*

LEO: *It's a 3D model of a midnight green ceramic coffee cup with a plate.*

Embodied Reasoning

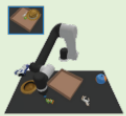


Situation: Sitting at the edge of the bed and facing the couch.

Question: Can I go straight to the coffee table in front of me?

LEO: *No.*

Robotic Manipulation



Human: pack the yoshi figure in the brown box

LEO:

P = [0.1, -0.2, 0]

R = [0, 0, 0, 1]

Embodied Navigation




Human: Find the pillow

LEO:

1. Move forward
2. Turn right
3. Move forward
4. ...

Task Planning



Human: Organize and tidy up the bedroom.

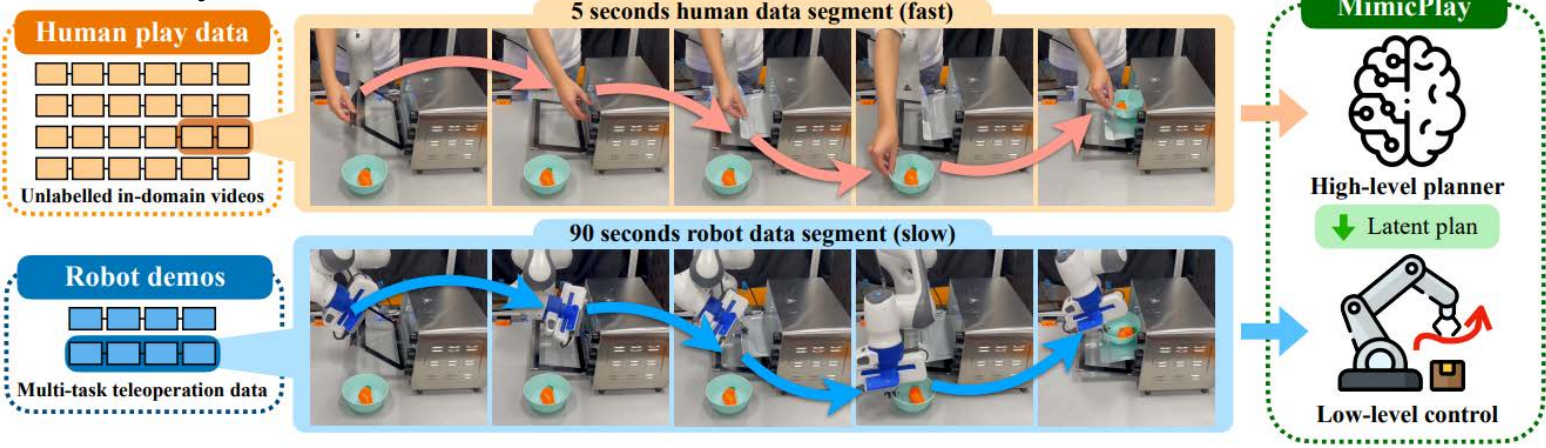
LEO:

1. Clean the floor by sweeping to remove ...
2. Make the bed ...
3. Fold and organize ...

多模态、多任务赋能的LLM具有更加通用的场景理解能力

数据引导——From Video

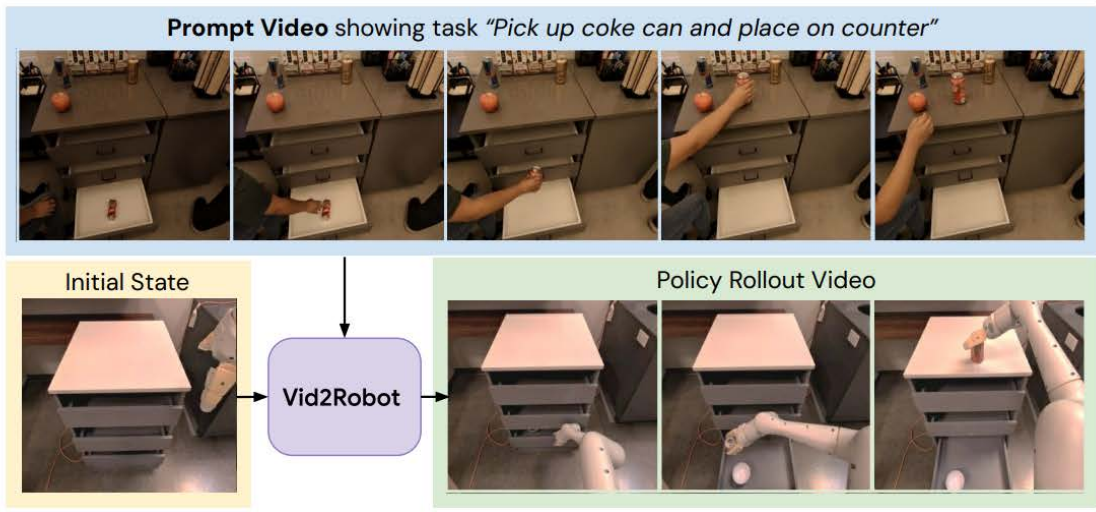
MimicPlay



Vid2Robot

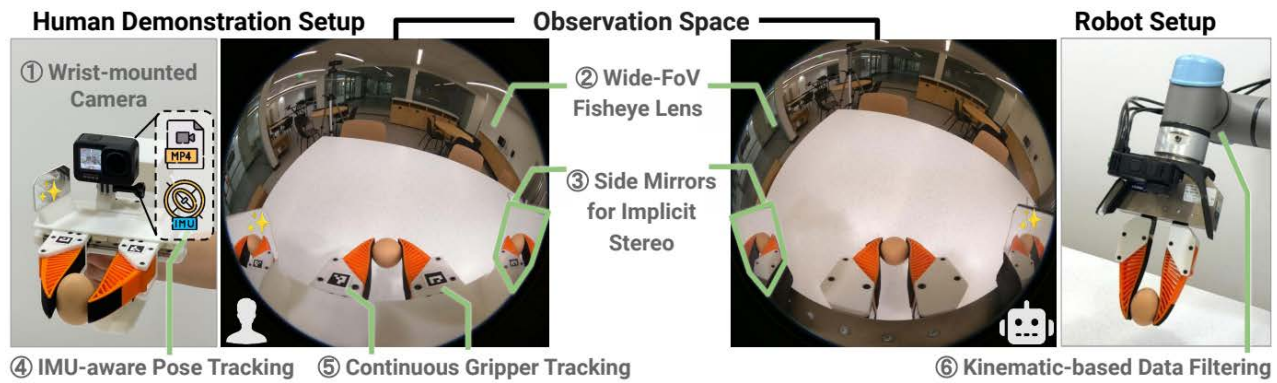
Robot observes human doing a task.

Vid2Robot outputs actions to complete shown task in its own environment.

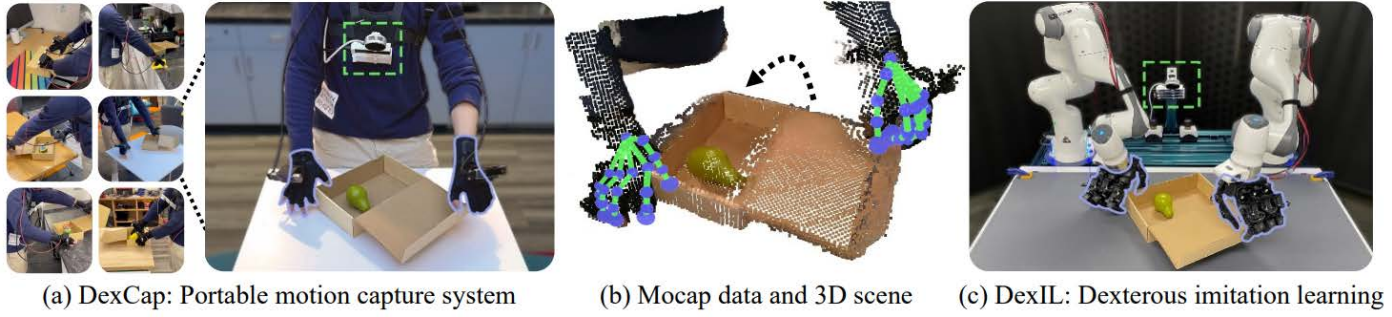


数据引导——Light-weight Hardware

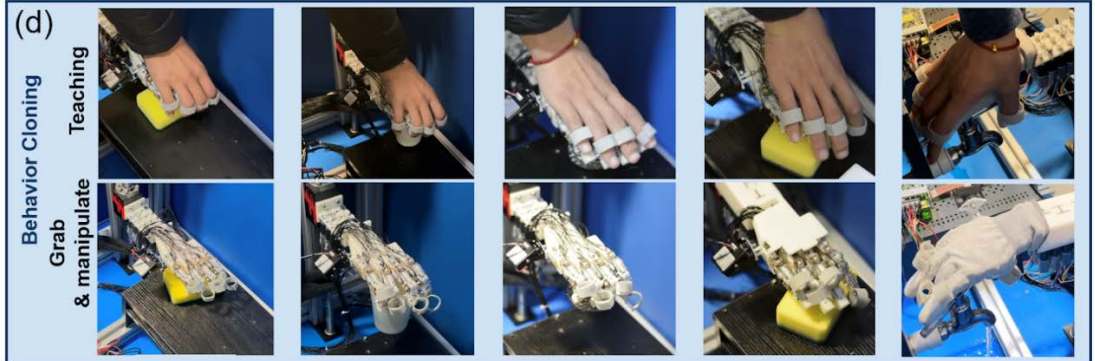
UMI



DexCap



HIRO Hand



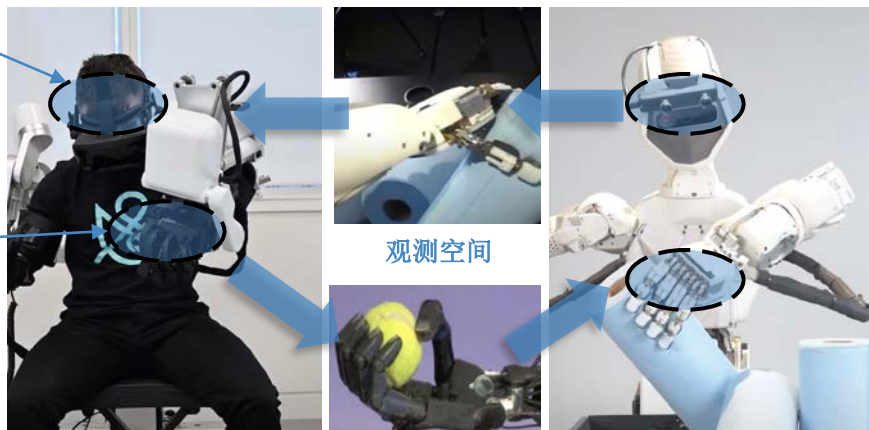
数据引导——Heavy Hardware

人形机器人灵巧操作数据

搭建灵巧操作数据采集平台，对齐人类和机器人的观测和操作空间

VR显示器

VR手套



人类专家

动作空间

人形机器人



Sanctuary AI



Tesla



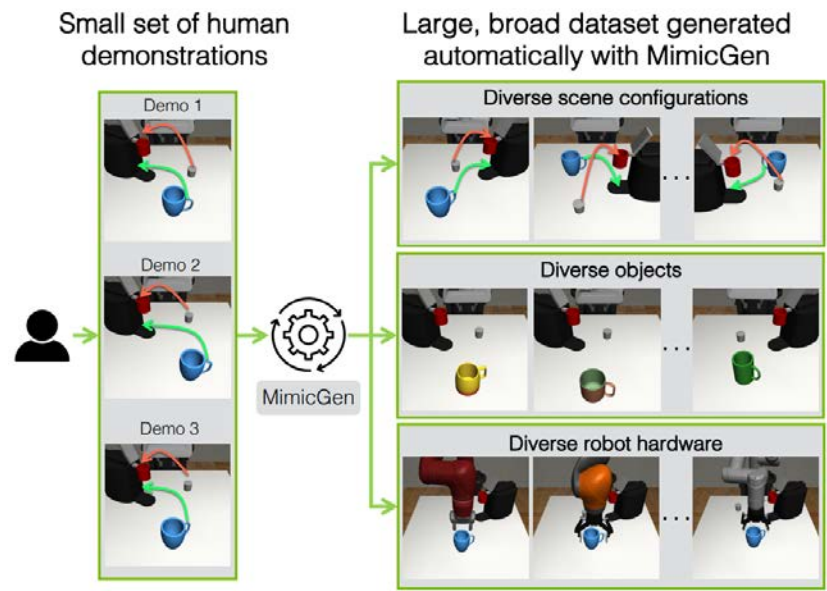
数据引导——Generative Simulation

RoboGen



Propose-generate-learn cycle

MimicGen

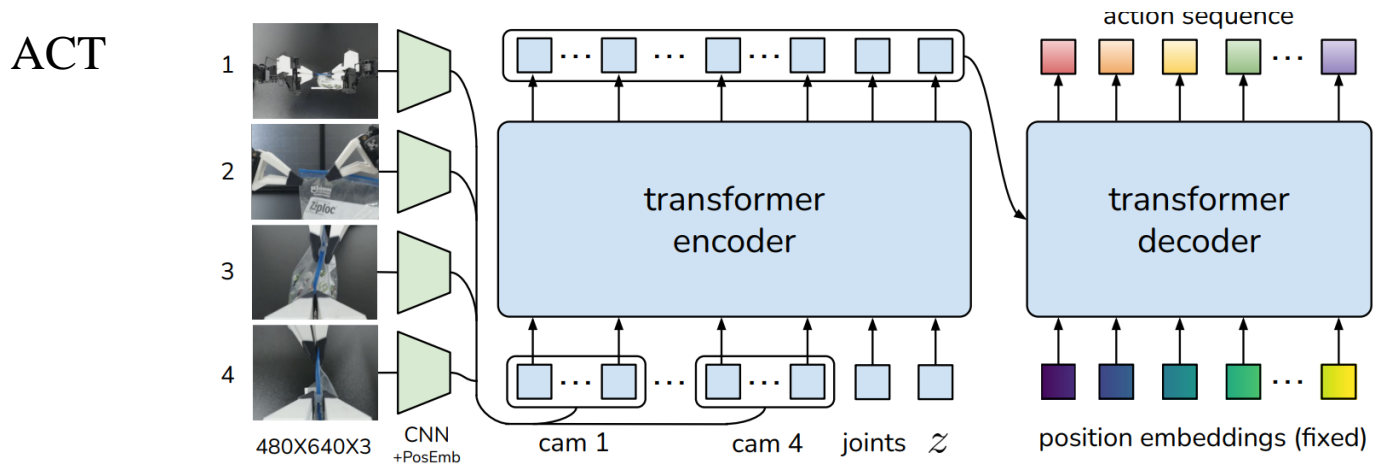


Demonstrations augmentation

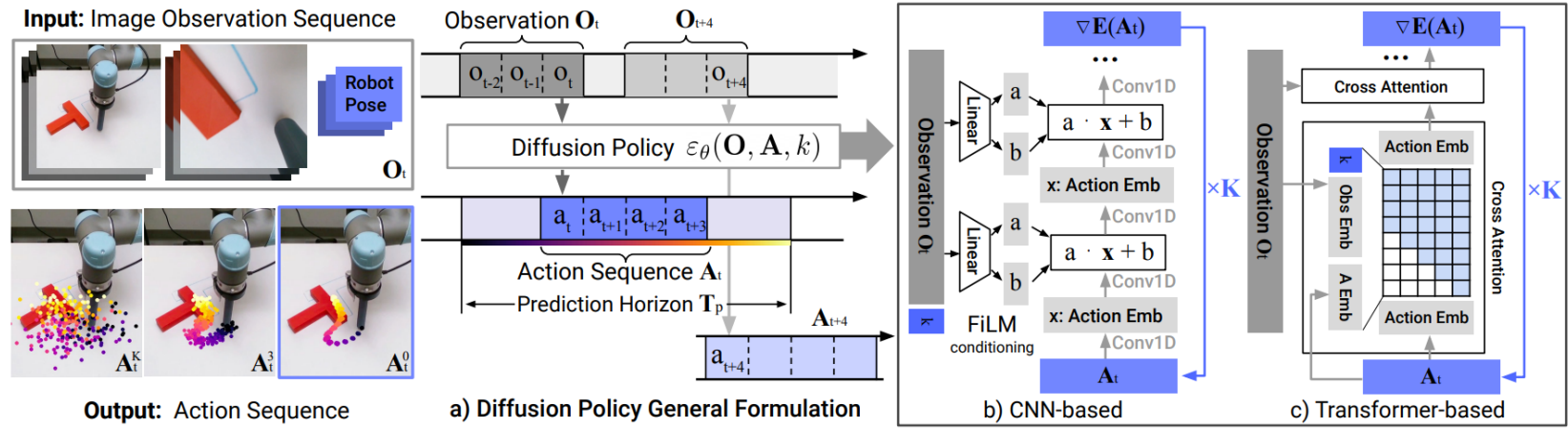


Ajay Mandlekar et. al., MimicGen: A Data Generation System for Scalable Robot Learning using Human Demonstrations, 2023
 Yufei Wang et. al., ROBOGEN: TOWARDS UNLEASHING INFINITE DATA FOR AUTOMATED ROBOT LEARNING VIA GENERATIVE SIMULATION, 2023

动作执行——Generative Imitation Learning

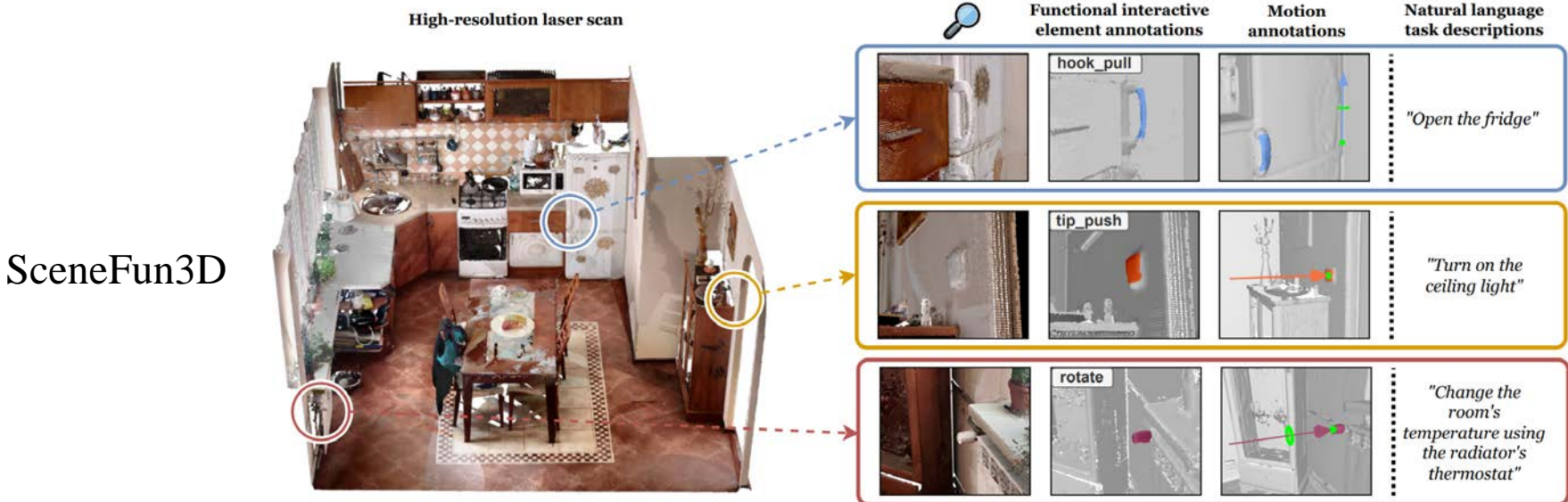
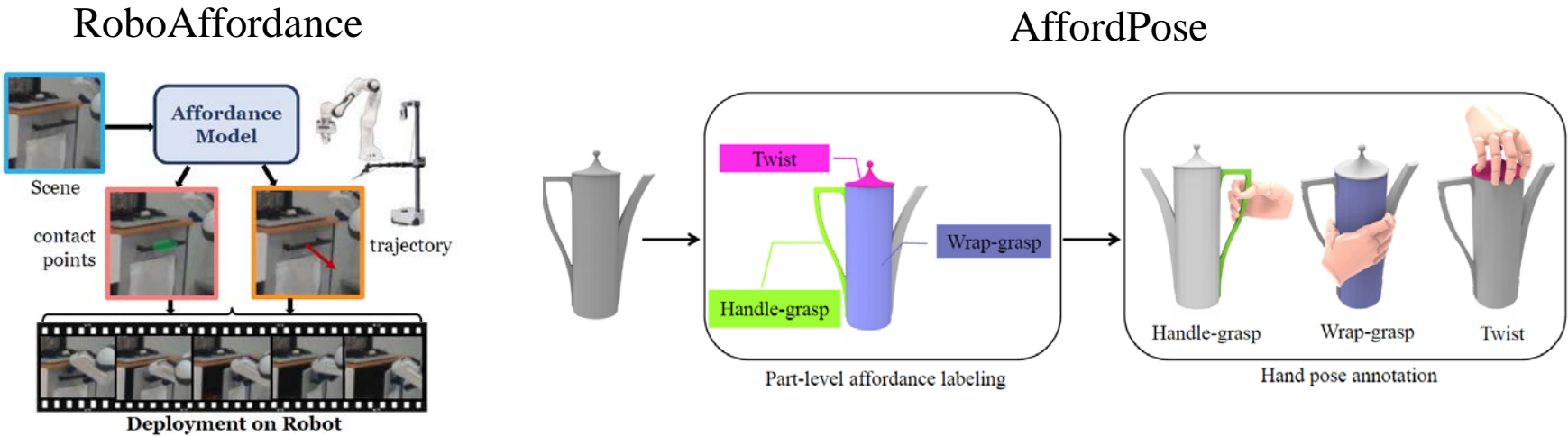


Diffusion Policy



Tony Z. Zhao et. al., Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware, 2023
 Cheng Chi et. al., Diffusion Policy: Visuomotor Policy Learning via Action Diffusion, 2023

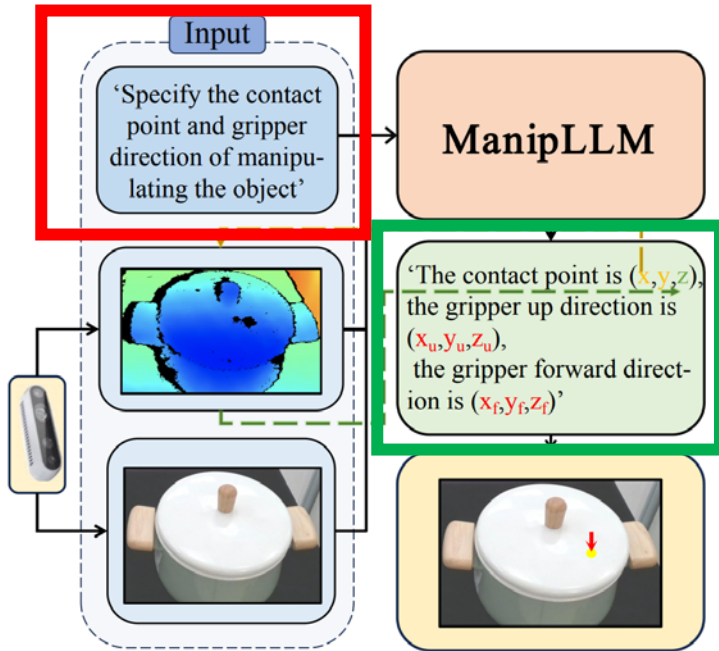
动作执行——Affordance



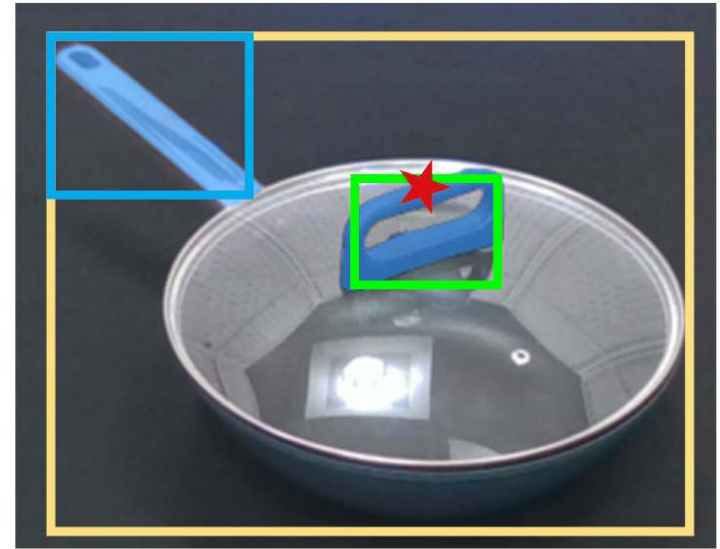
Shikhar Bahl et. al., Affordances from Human Videos as a Versatile Representation for Robotics, 2023
 Juntao Jian et. al., AffordPose: A Large-scale Dataset of Hand-Object Interactions with Affordance-driven Hand Pose, 2023
 SceneFun3D et. al., SceneFun3D: Fine-Grained Functionality and Affordance Understanding in 3D Scenes, 2024

动作执行——Q&A from LLM

ManipLLM



ManipVQA



User: Handle of the pan.
📏: [0.09, 0.11, 0.41, 0.42].

User: Can contain liquids?
📏: True.

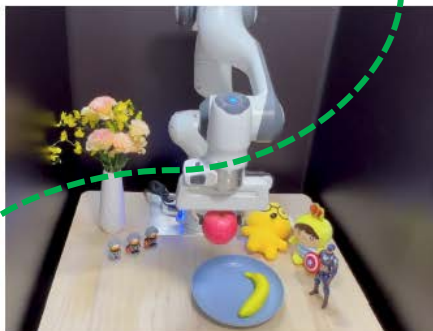
User: Grasp to open the lid.
📏: [0.40, 0.31, 0.71, 0.52], contact point ★.

动作执行——Prompt Planning from LLM



Please pass me the blue empty plate.

充分利用GPT-4V的**多模态理解**、**复杂规划**能力



1. Pick up apple
2. Place apple on table



3. Pick up banana
4. Place banana on table



5. Pick up blue plate
6. Place blue plate in human hand



We are having an art class, please prepare an area for the children.



1. Pick up screwdriver
2. Place screwdriver in box



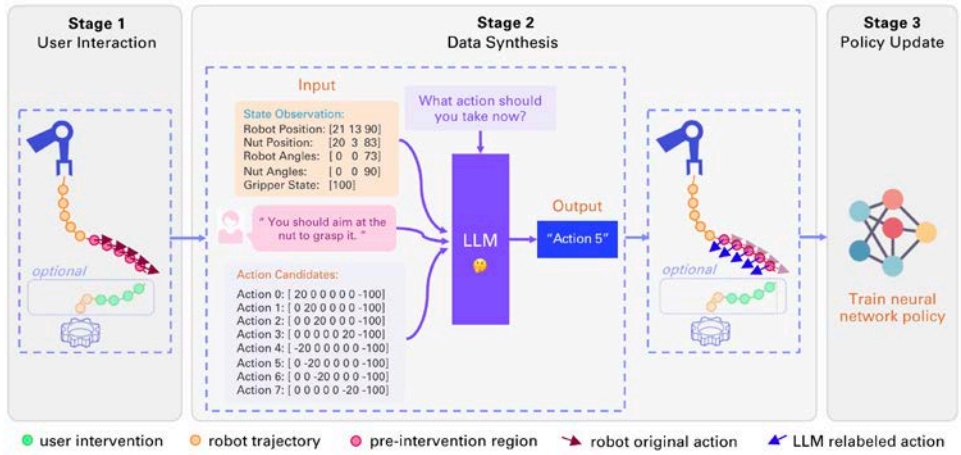
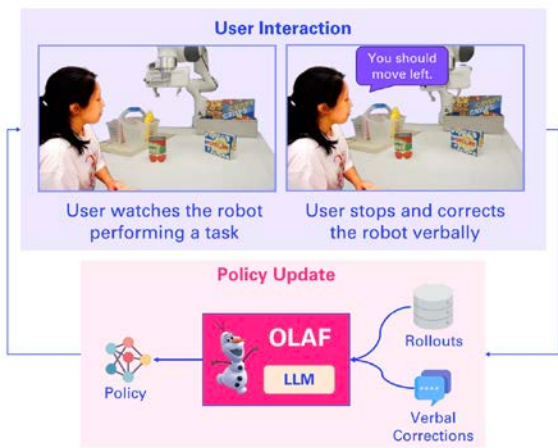
3. Pick up knife



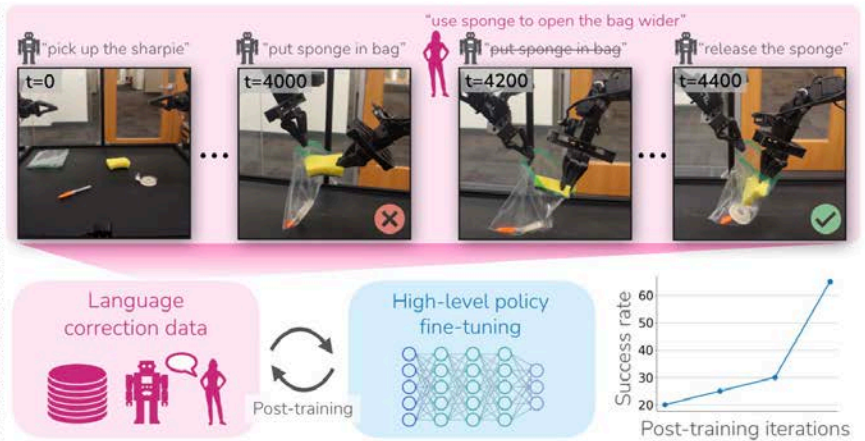
4. Place knife in box

Hard2Simple

动作执行——Language Corrections



OLAF



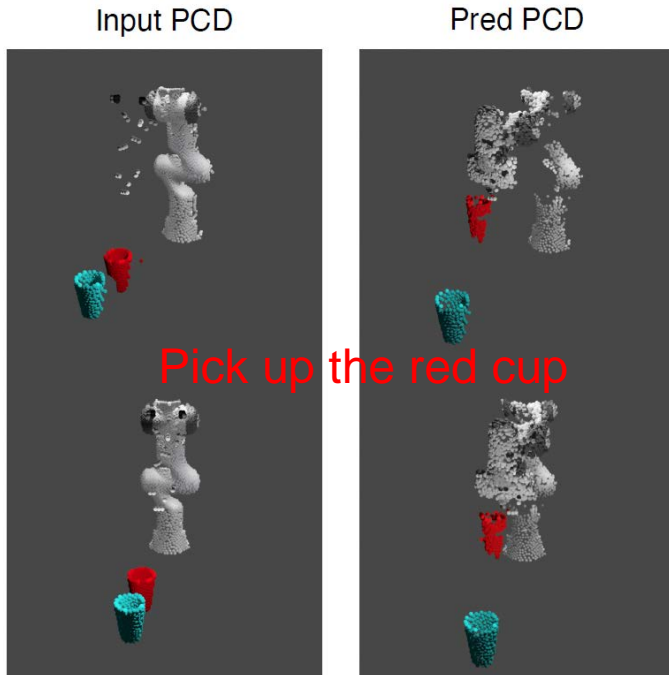
Yell At Your Robot

Shi L X et al. Yell At Your Robot: Improving On-the-Fly from Language Corrections, 2024.
Liu H et al. Interactive robot learning from verbal correction, 2023.

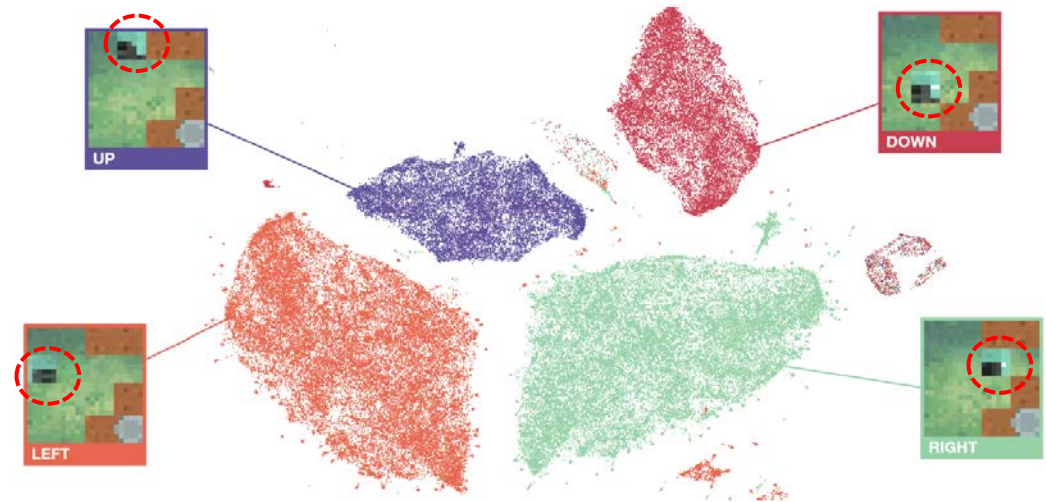
世界模型

FOCUS: Understanding the world in terms of objects and the possible interplays with them is an important **cognition ability**, especially in robotics manipulation, where many tasks require robot-object interactions.

$$\dot{x} = f(x, u) \quad s_t, a_t \rightarrow s_{t+1}$$



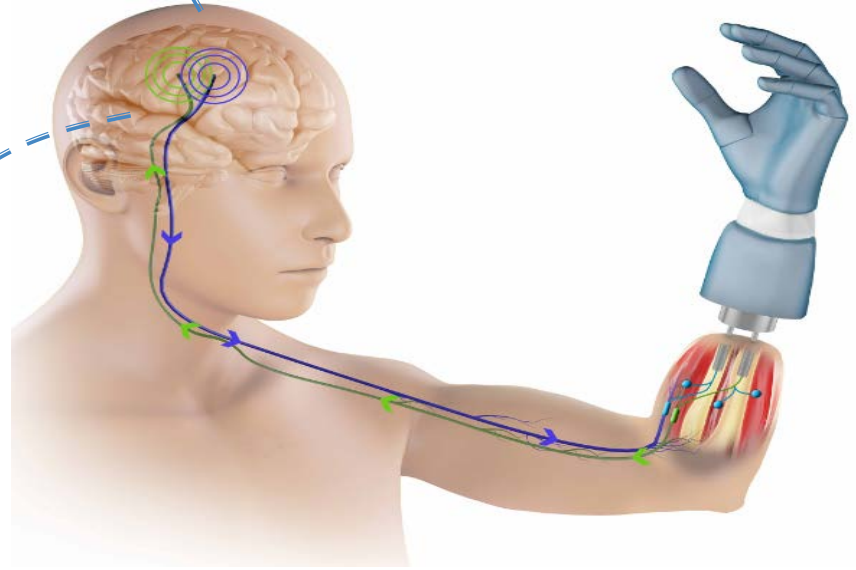
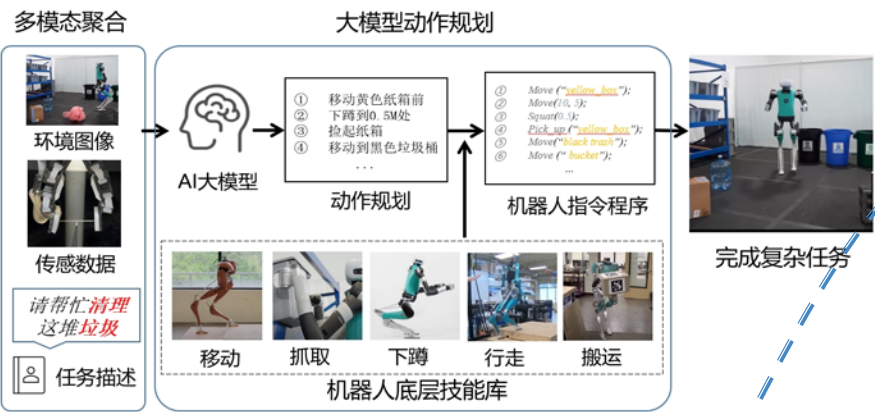
3D VLA



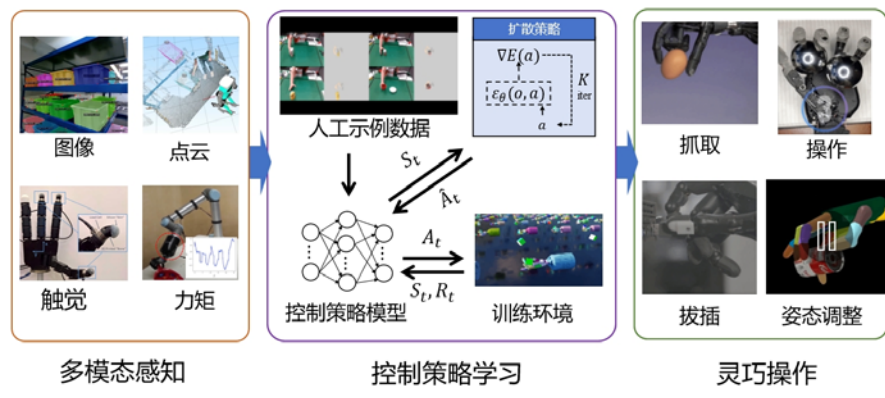
LAPO

机器人“大脑”、小脑发展不平衡

模拟人思维决策过程——“大脑”



模仿生物复杂运动控制——“小脑”



相较于**智能**“大脑”的智力快速提升，**灵巧**“小脑”能够实现的**灵巧操作**能力亟待加强

谢谢大家！