北京航空航天大學
BEIHANG UNIVERSITY

Embedded OS in China
国产嵌入式操作系统产业论坛

面向具身智能的

大小脑模型协同算法研究及实践

盛律 | 软件学院

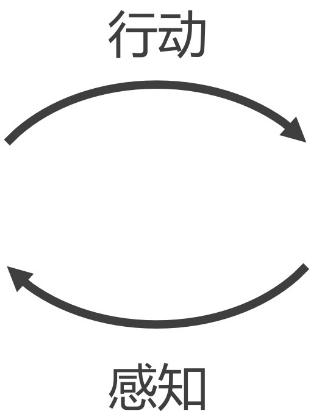
2025-08-23

具身智能的基本概念

具身智能

基于**物理载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**

智能体



外界环境

具身智能丨CCF专家谈术语

具身智能的基本概念

具身智能

基于**物理载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**

传统智能

只可远观，被动接受
别人告诉我这就是盒子

被动抽象接受

具身智能

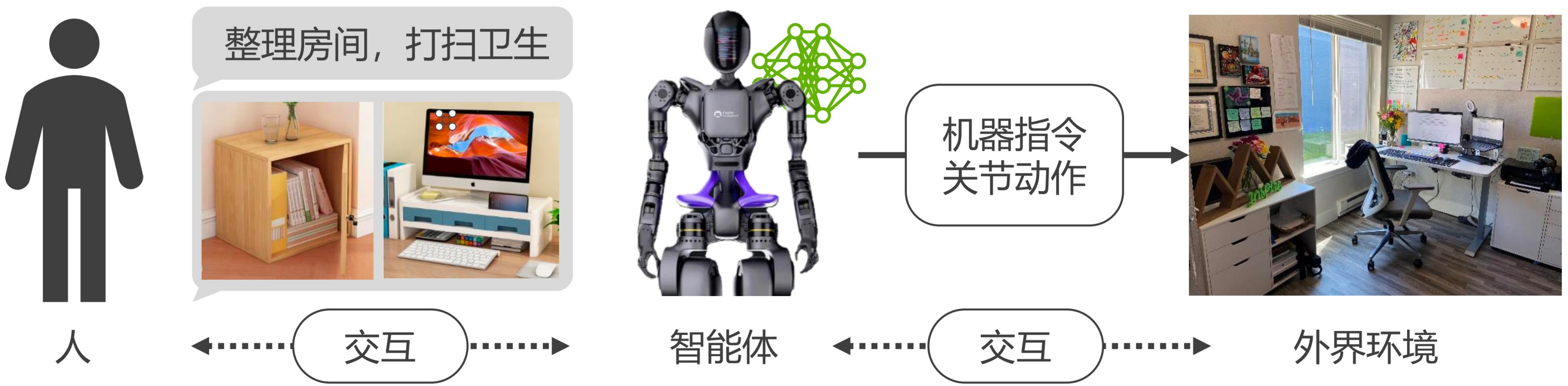
主动具体体验

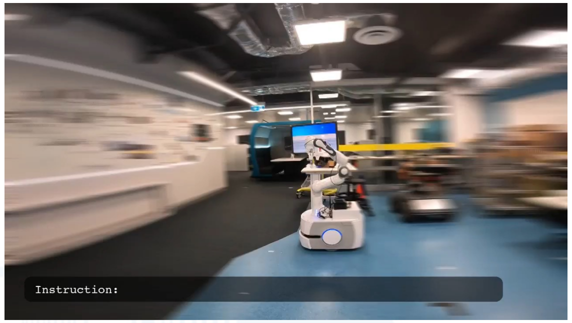
可以打开，可以装东西
我主动体验什么是盒子

重要意义

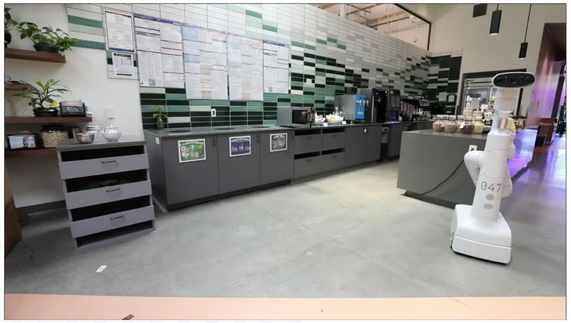
具身智能因其能自主产生智能行为和适应性，是通用人工智能的可能起点

具身智能的关键任务





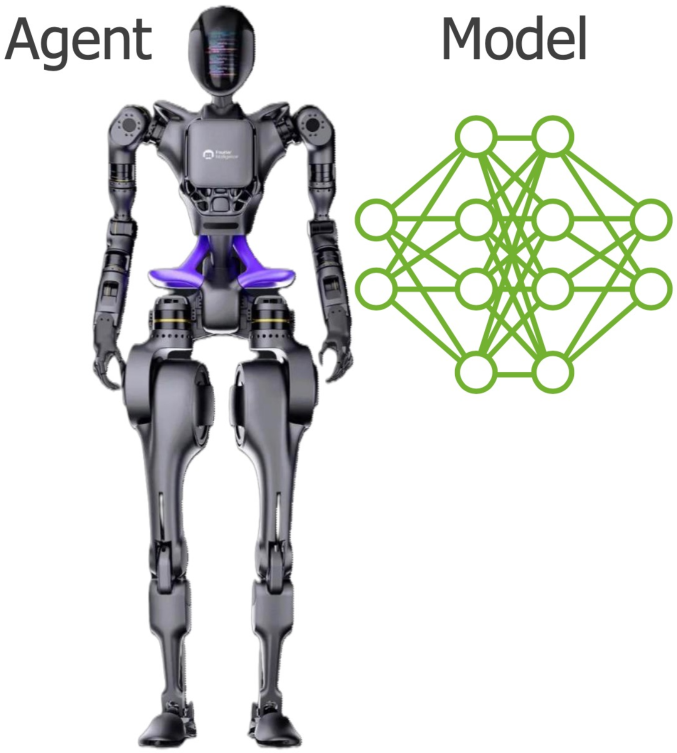
导航



问答

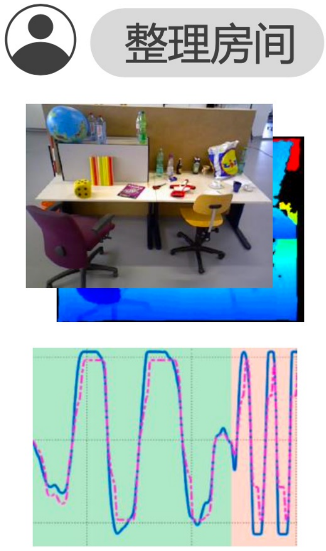
操作

具身智能的核心目标



通用泛化

模态泛化

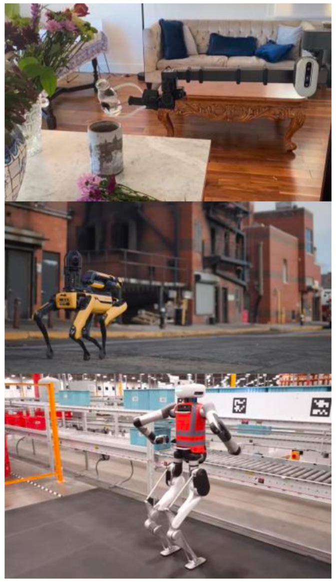


…

任务泛化

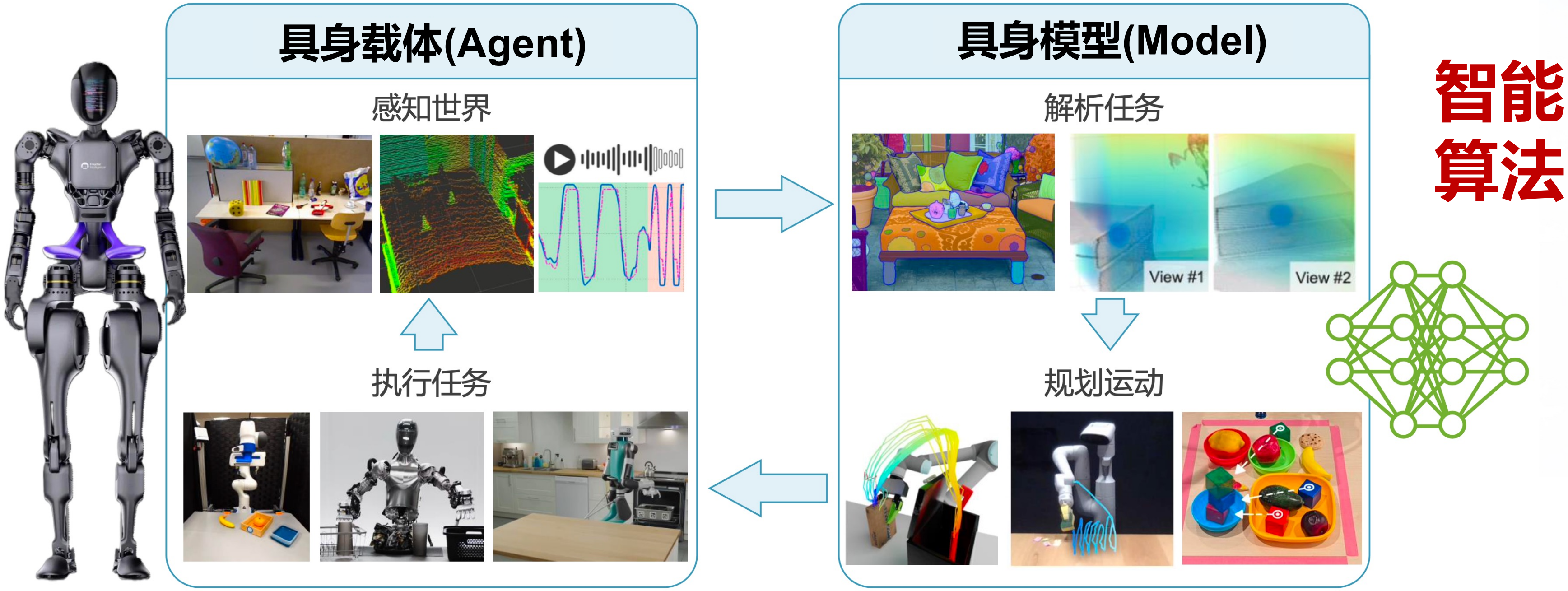
…

场景泛化



…

具身智能的核心要素



物理载体

具身载体(Agent)

感知世界

执行任务

具身模型(Model)

解析任务

规划运动

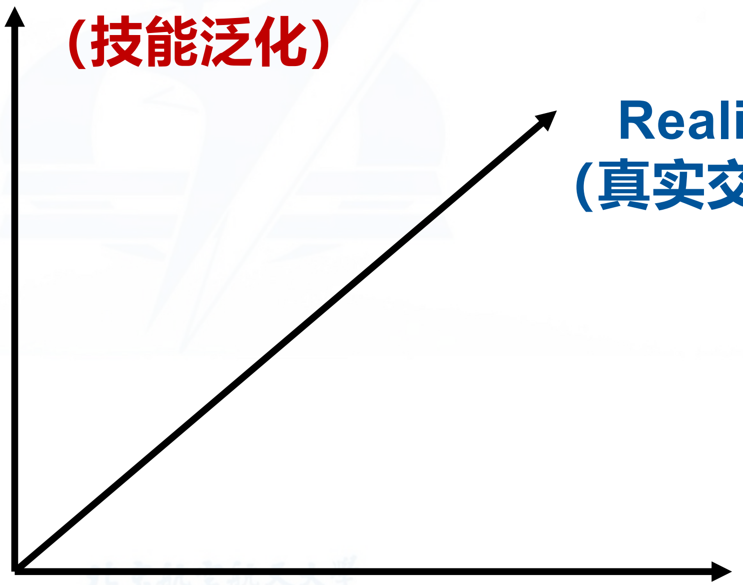
智能算法

现状 相比**具身载体**的**日趋成熟**，**具身模型**的算法研究**方兴未艾、挑战众多**

具身模型应该考虑哪些能力?

- **技能泛化**、**真实交互**、**本体扩展**

Adapted from Jim Fan’s talk

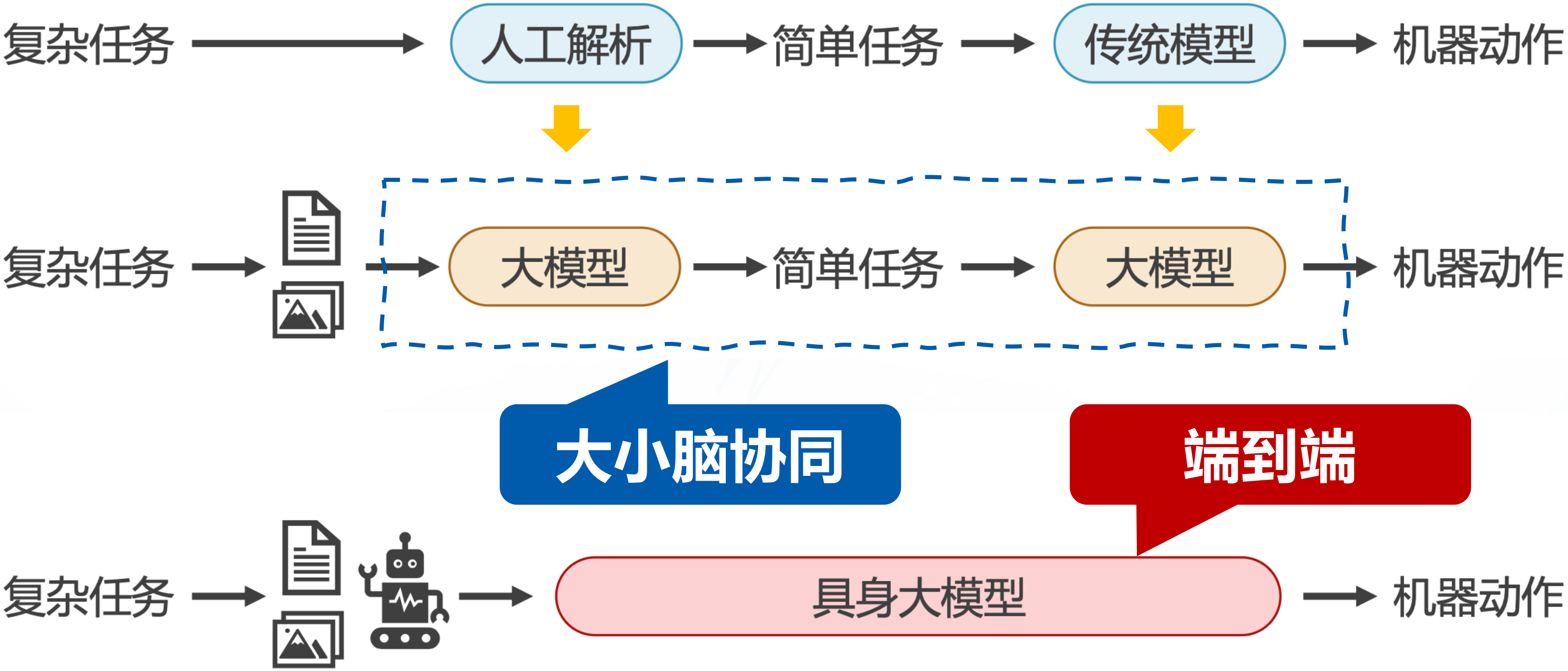


Skill
(技能泛化)

Reality
(真实交互)

Embodiment
(本体扩展)

具身模型的几种类型



复杂任务 → 人工解析 → 简单任务 → 传统模型 → 机器动作

复杂任务 → 大模型 → 简单任务 → 大模型 → 机器动作

大小脑协同

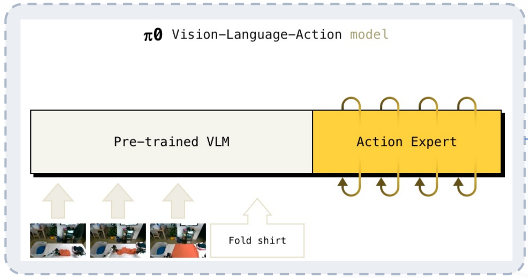
端到端

复杂任务 → 具身大模型 → 机器动作

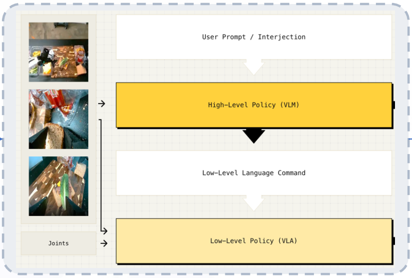
具身模型的最新进展：代表性新工作

Physical Intelligence (π)

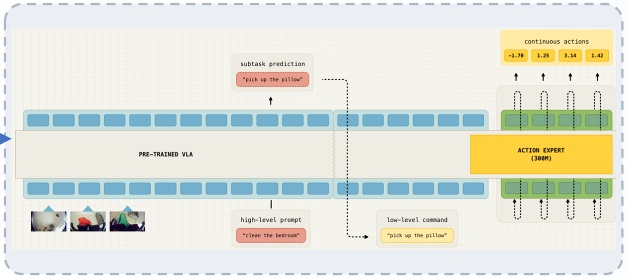
端到端VLA π_0 (2024.10)



大小脑 hi robot (2025.02)

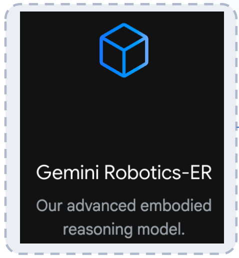


混合$\pi_{0.5}$ (2025.04)



Google DeepMind

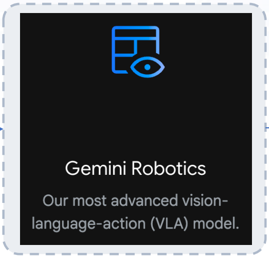
大脑-小脑



Gemini Robotics-ER

Our advanced embodied reasoning model.

端到端VLA



Gemini Robotics

Our most advanced vision-language-action (VLA) model.

端测SDK

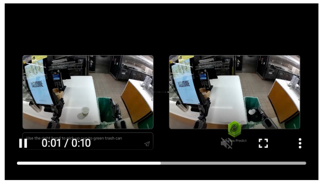
Gemini Robotics On-Device

Our vision-language-action model optimized to run locally on robotic devices.

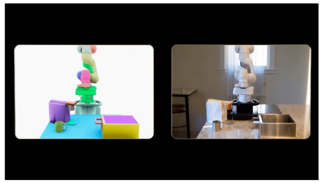
NVIDIA

具身大脑

(2025.03)



Cosmos Predict



Cosmos Transfer

Cosmos Reason

端到端VLA

GR00T N1

GR00T N1.5

具身大模型离实用还有差距

2023及之前

2024

2025及之后

基本能力

- 感知
- 操作
- 导航

单任务
单本体
单场景

Hand-Eye Coordination Robotic Arm

ROKAE

MECH MIND

多任务
单本体
单场景

感知和理解
决策和规划
执行和协作
评估和反馈

covariant

Stanford University

Berkeley

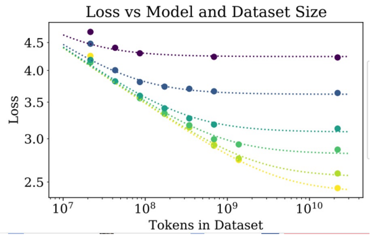
Physical Intelligence

大模型

大数据

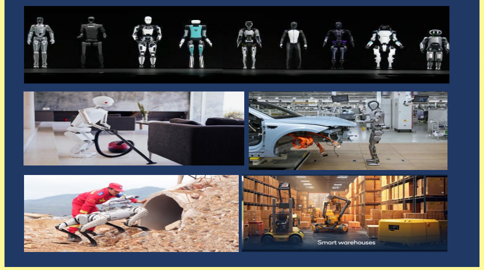
Scaling Law

在大语言模型和多模态大模型上都得到了验证



通用智能系统
多本体
多场景

端到端
多模态大模型机器人



不好用

模型能力弱，未达到具身智能的“ChatGPT时刻”

不易用

大脑、小脑、本体适配难度高

不通用

一个模型只适用于一种本体

需要“聪明”的大脑大模型和跨本体的大小脑协作框架，实现跨本体、跨场景、可泛化的具身智能

北京航空航天大学

大小脑模型协同的技术路线仍有机会

- ❑ 端到端模型虽决策高效，但**泛化性**和**扩展性**受限，受制于**环境交互与硬件适配**，难以适应多样场景。而模块化的大小脑协同框架凭借**强泛化、可解释**优势，正成为学界与业界的研究热点

模块化：大小脑协同框架赋予具身智能体模块化优势，具备可扩展架构、高效开发与强适应性三大特性

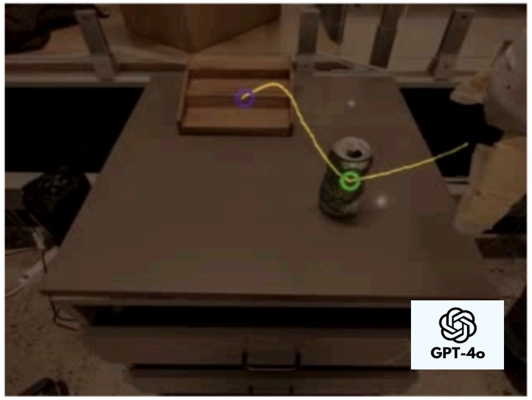
可泛化：基于VLM开发的大脑具备丰富的多模态认知能力，且不受小脑模型的影响

可解释：决策过程更加透明，提升人机协同效率

大小脑模型协同框架 **是当前实现具身智能体更易落地的技术路线**

传统多模态大模型能够作为“大脑”？

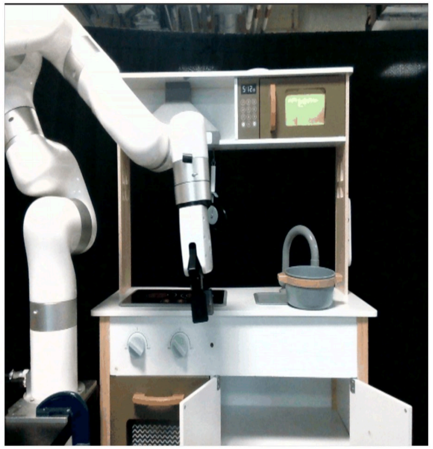
■ 传统VLMs在具身智能场景（长程闭环操作、时空智能等）中面临严峻挑战

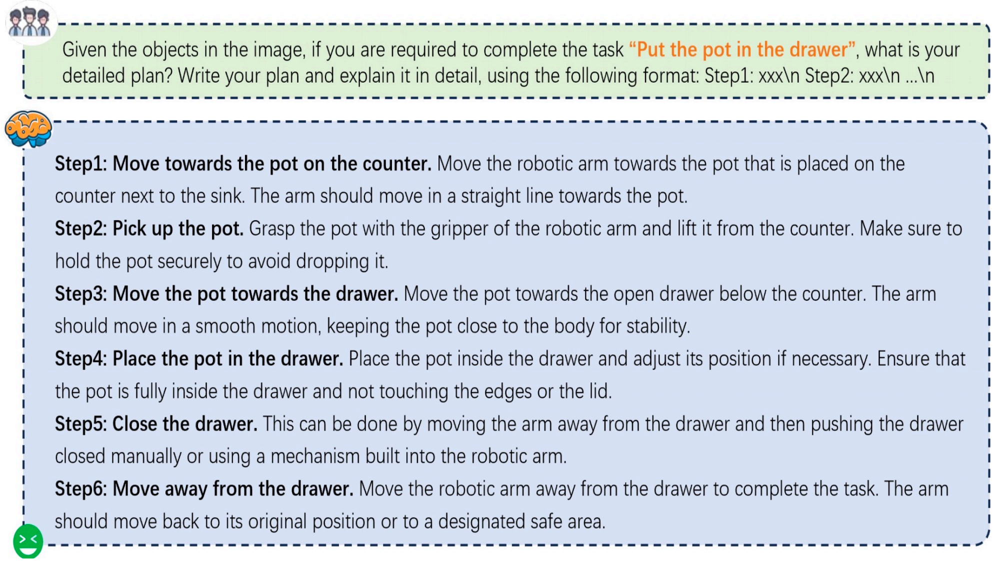


If the yellow robot gripper follows the yellow trajectory, what will happen?

A. Robot puts the soda on the wooden steps.
B. Robot moves the soda in front of the wooden steps.
C. Robot moves the soda to the very top of the wooden steps.
D. Robot picks up the soda can and moves it up.

GPT-4o在具身任务中表现欠佳





Given the objects in the image, if you are required to complete the task “Put the pot in the drawer”, what is your detailed plan? Write your plan and explain it in detail, using the following format: Step1: xxx\n Step2: xxx\n ...\n

Step1: Move towards the pot on the counter. Move the robotic arm towards the pot that is placed on the counter next to the sink. The arm should move in a straight line towards the pot.

Step2: Pick up the pot. Grasp the pot with the gripper of the robotic arm and lift it from the counter. Make sure to hold the pot securely to avoid dropping it.

Step3: Move the pot towards the drawer. Move the pot towards the open drawer below the counter. The arm should move in a smooth motion, keeping the pot close to the body for stability.

Step4: Place the pot in the drawer. Place the pot inside the drawer and adjust its position if necessary. Ensure that the pot is fully inside the drawer and not touching the edges or the lid.

Step5: Close the drawer. This can be done by moving the arm away from the drawer and then pushing the drawer closed manually or using a mechanism built into the robotic arm.

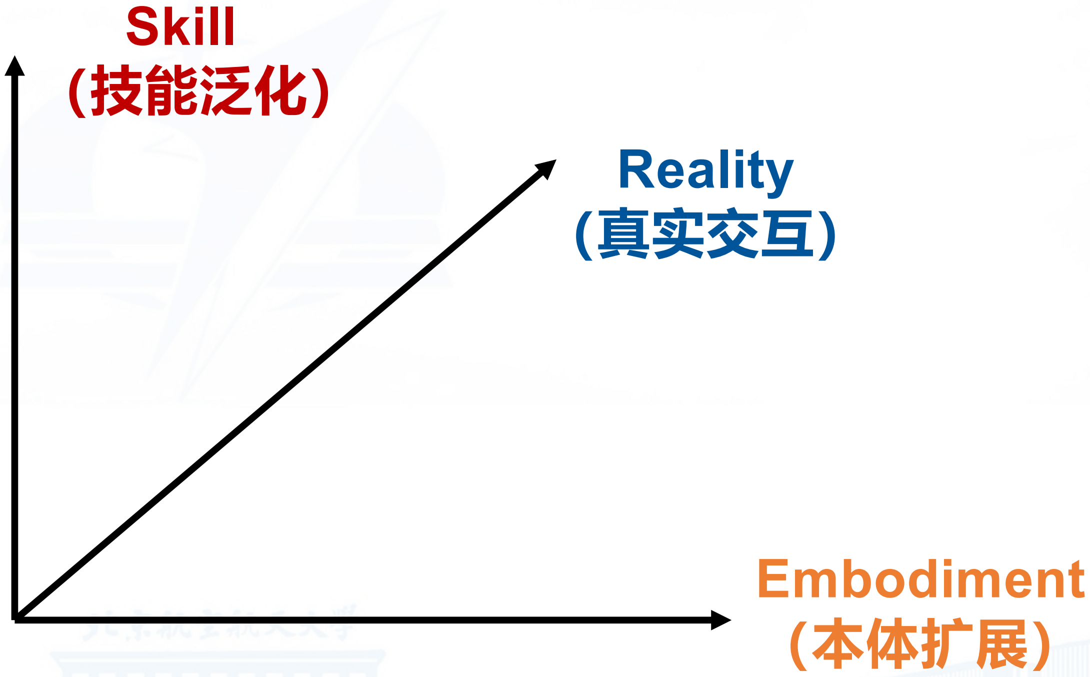
Step6: Move away from the drawer. Move the robotic arm away from the drawer to complete the task. The arm should move back to its original position or to a designated safe area.

以‘把锅放到抽屉里’为例，该任务涉及多步骤的长时间交互，包括移动、抓取、放置等操作，并需要与锅、抽屉等物体进行持续交互

回顾：具身模型应该考虑哪些能力?

- **技能泛化**、**真实交互**、**本体扩展**

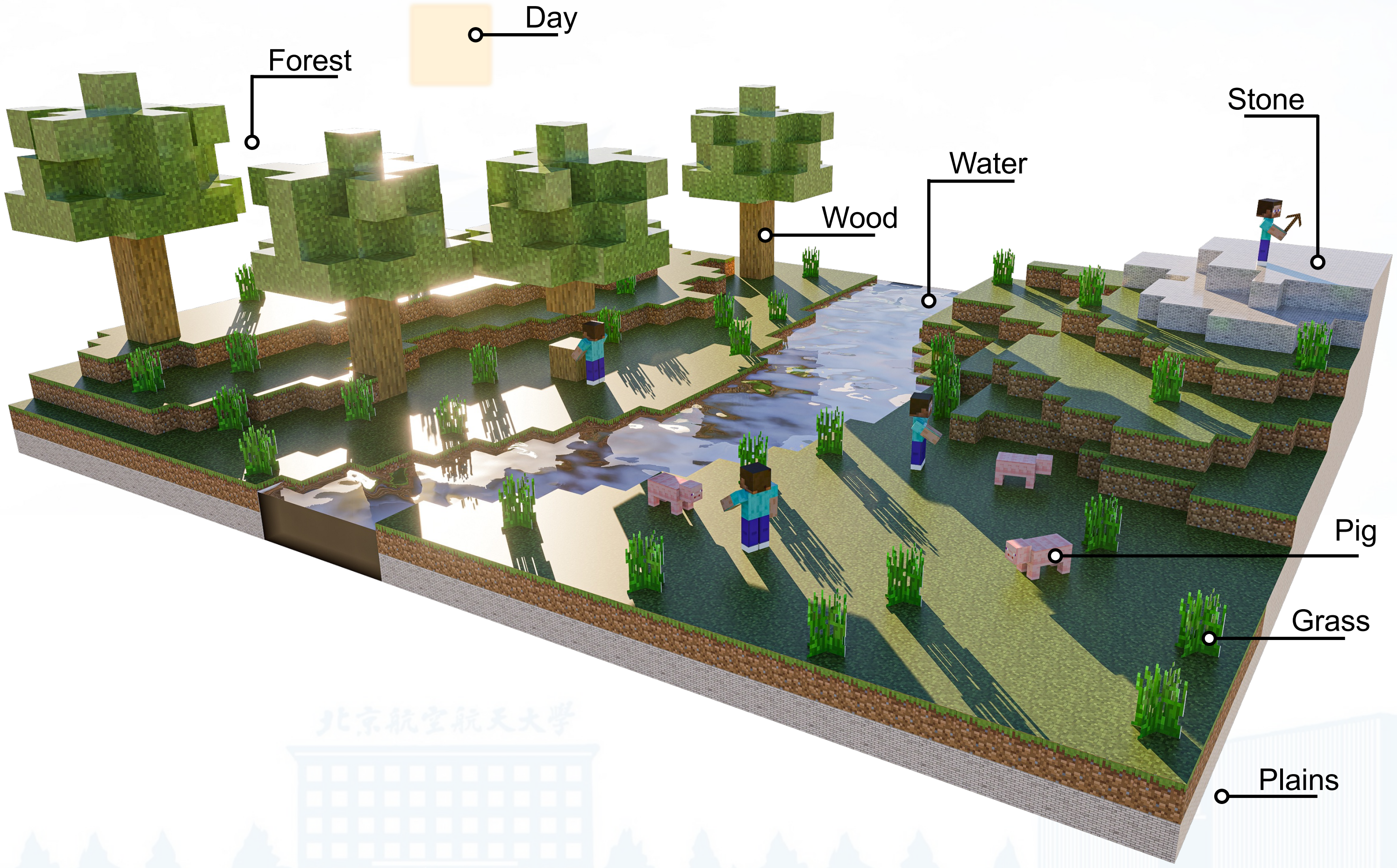
Adapted from Jim Fan’s talk



技能泛化：多智能体实现长时序开放具身任务解决

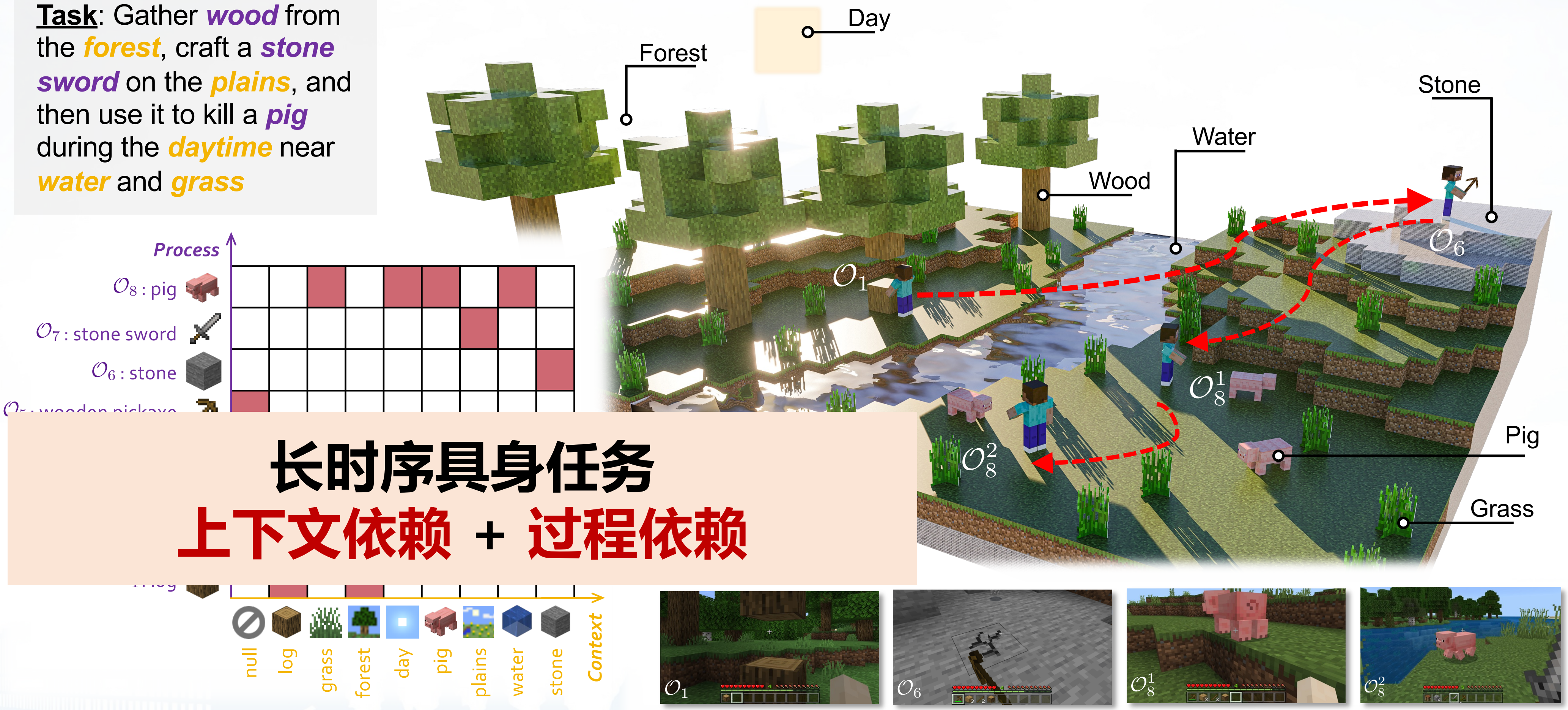
Long-horizon open-world embodied tasks

Task: Gather ***wood*** from the ***forest***, craft a ***stone sword*** on the ***plains***, and then use it to kill a ***pig*** during the ***daytime*** near ***water*** and ***grass***



技能泛化：多智能体实现长时序开放具身任务解决

<u>Task</u>: Gather ***wood*** from the ***forest***, craft a ***stone sword*** on the ***plains***, and then use it to kill a ***pig*** during the ***daytime*** near ***water*** and ***grass***

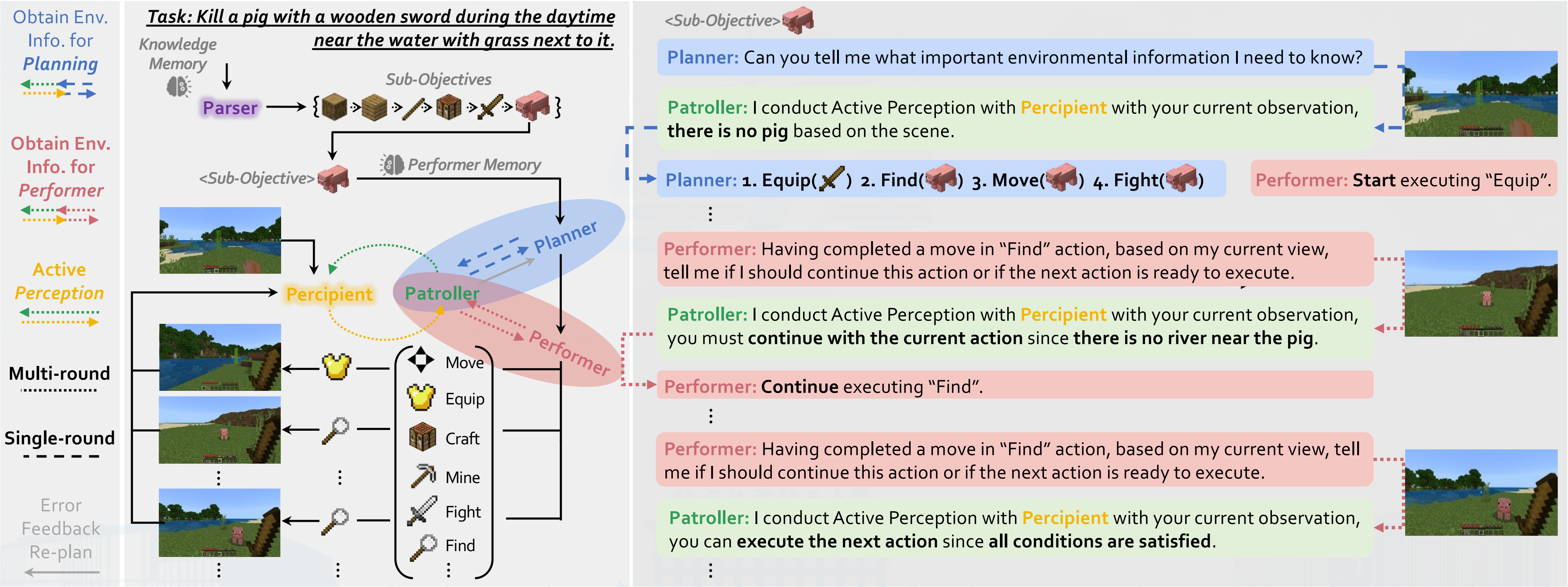


长时序具身任务

上下文依赖 + 过程依赖

技能泛化：多智能体实现长时序开放具身任务解决

- **MP5 (CVPR 2024):** 5 (M)LLMs with different roles, communicating for different purposes



Planner: Can you tell me what important environmental information I need to know?

Patroller: I conduct Active Perception with **Percipient** with your current observation, **there is no pig** based on the scene.

Planner: 1. Equip( ) 2. Find( ) 3. Move( ) 4. Fight( )

Performer: **Start** executing "Equip".

⋮

Performer: Having completed a move in "Find" action, based on my current view, tell me if I should continue this action or if the next action is ready to execute.

Patroller: I conduct Active Perception with **Percipient** with your current observation, you must **continue with the current action** since **there is no river near the pig**.

Performer: **Continue** executing "Find".

⋮

Performer: Having completed a move in "Find" action, based on my current view, tell me if I should continue this action or if the next action is ready to execute.

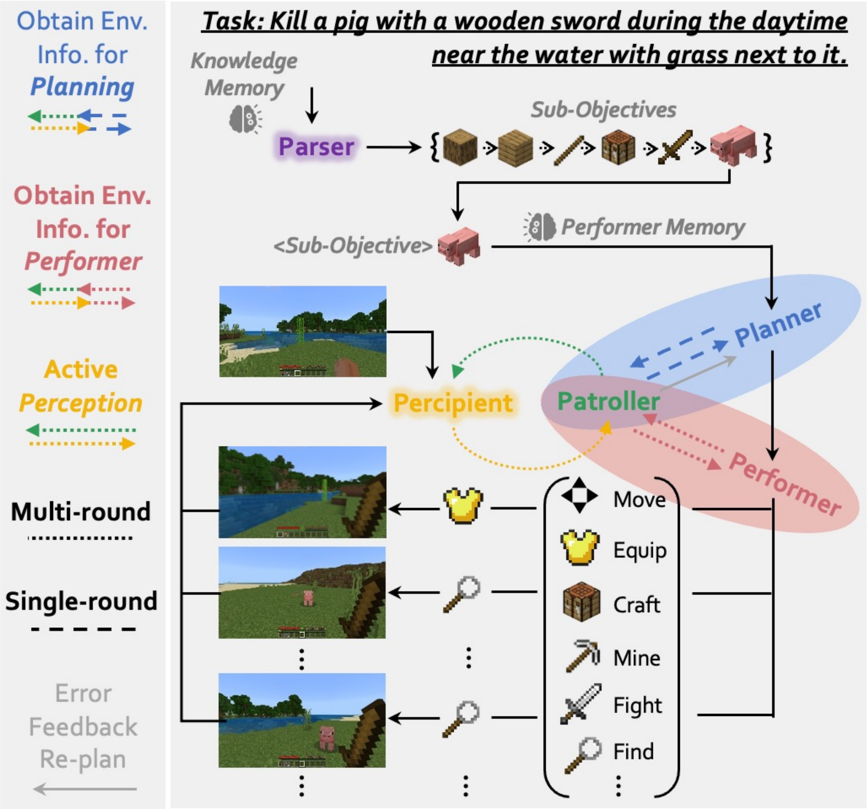
Patroller: I conduct Active Perception with **Percipient** with your current observation, you can **execute the next action** since **all conditions are satisfied**.

⋮

MP5: A multi-modal open-ended embodied system in minecraft via active perception, **CVPR 2024**

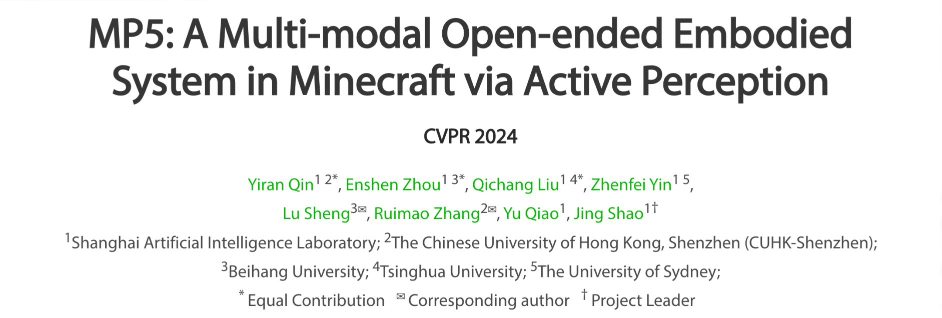
技能泛化：多智能体实现长时序开放具身任务解决

- **MP5 (CVPR 2024):** 5 (M)LLMs with different roles, communicating for different purposes



MP5: A multi-modal open-ended embodied system in minecraft via active perception, **CVPR 2024**

技能泛化：多智能体实现长时序开放具身任务解决



MP5: A Multi-modal Open-ended Embodied System in Minecraft via Active Perception

CVPR 2024

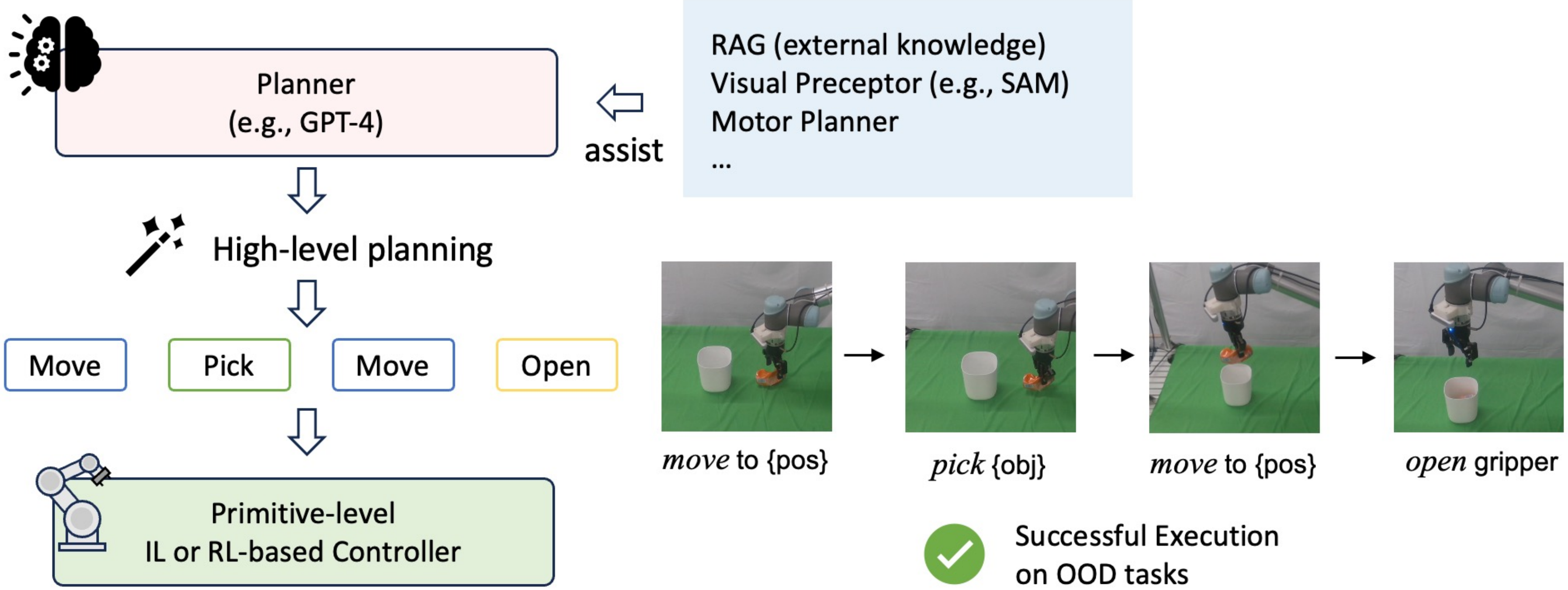
Yiran Qin[1 2*], Enshen Zhou[1 3*], Qichang Liu[1 4*], Zhenfei Yin[1 5],
Lu Sheng[3✉], Ruimao Zhang[2✉], Yu Qiao[1], Jing Shao[1†]
[1]Shanghai Artificial Intelligence Laboratory; [2]The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen);
[3]Beihang University; [4]Tsinghua University; [5]The University of Sydney;
[*] Equal Contribution [✉] Corresponding author [†] Project Leader

- 能精准理解环境上下文内容
- 能够解决钻石级难度任务
- 能持续执行开放式生存任务

Task: Dig the sand under the water with a wooden shovel during the day time in a sunny day.

技能泛化：组合泛化实现未知技能的学习

- **RA-P (IROS 2025, NeurIPS 2024 OWA):** composable generalizable agents in real world
  - Decompose complicated tasks into fine-grained primitive skills, generalizable to new physical skills



Planner
(e.g., GPT-4)

assist

RAG (external knowledge)
Visual Preceptor (e.g., SAM)
Motor Planner
...

High-level planning

Move
Pick
Move
Open

Primitive-level
IL or RL-based Controller

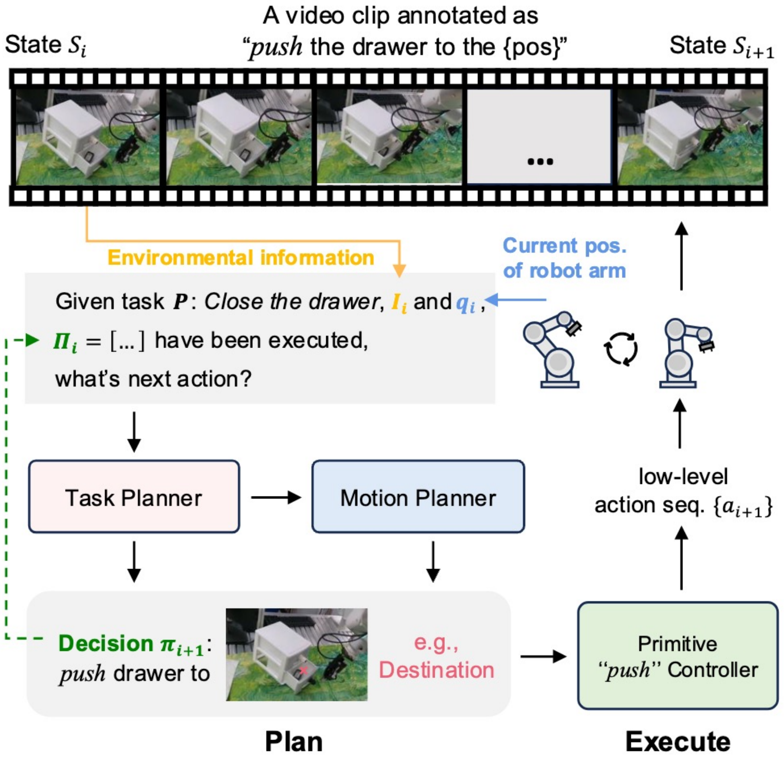
move to {pos} → *pick* {obj} → *move* to {pos} → *open* gripper

Successful Execution
on OOD tasks

RH20T-P: A Primitive-Level Robotic Dataset Towards Composable Generalization Agents, **IROS 2025**

技能泛化：组合泛化实现未知技能的学习

- **RA-P (IROS 2025, NeurIPS 2024 OWA):** composable generalizable agents in real world
  - Decompose complicated tasks into fine-grained primitive skills, generalizable to new physical skills



State S_i

A video clip annotated as "*push* the drawer to the {pos}"

State S_{i+1}

Environmental information

Current pos. of robot arm

Given task $\boldsymbol{P}$: *Close the drawer*, I_i and q_i, Π_i = [...] have been executed, what's next action?

Task Planner

Motion Planner

low-level action seq. $\{a_{i+1}\}$

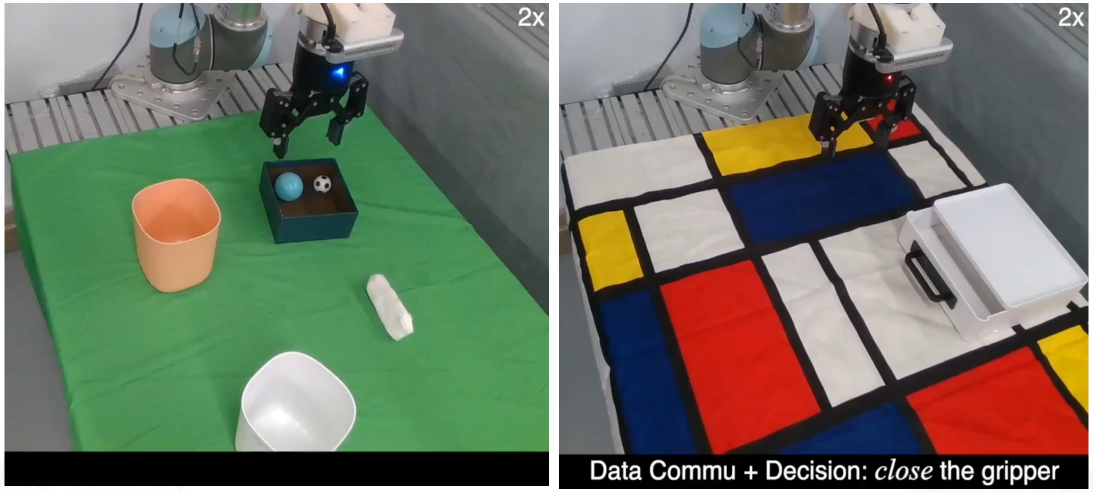
Decision π_{i+1}: *push* drawer to

e.g., Destination

Primitive "*push*" Controller

Plan

Execute

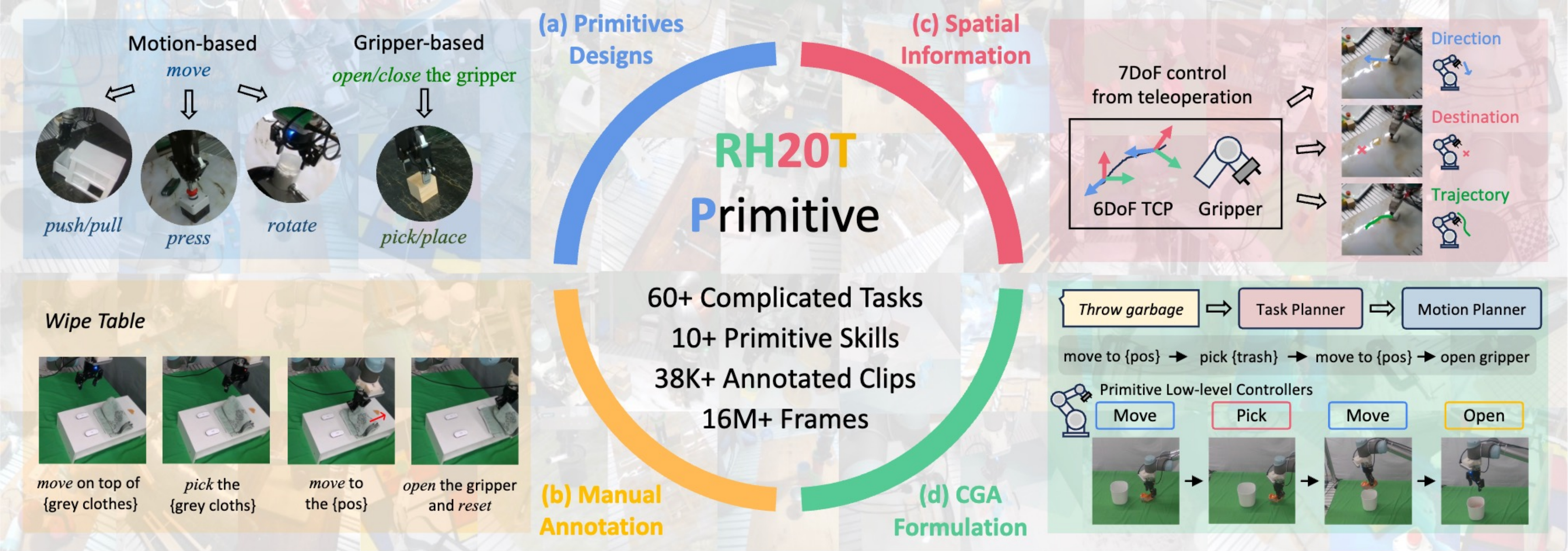


A baseline of RA-P

RH20T-P: A Primitive-Level Robotic Dataset Towards Composable Generalization Agents, **IROS 2025**

技能泛化：组合泛化实现未知技能的学习

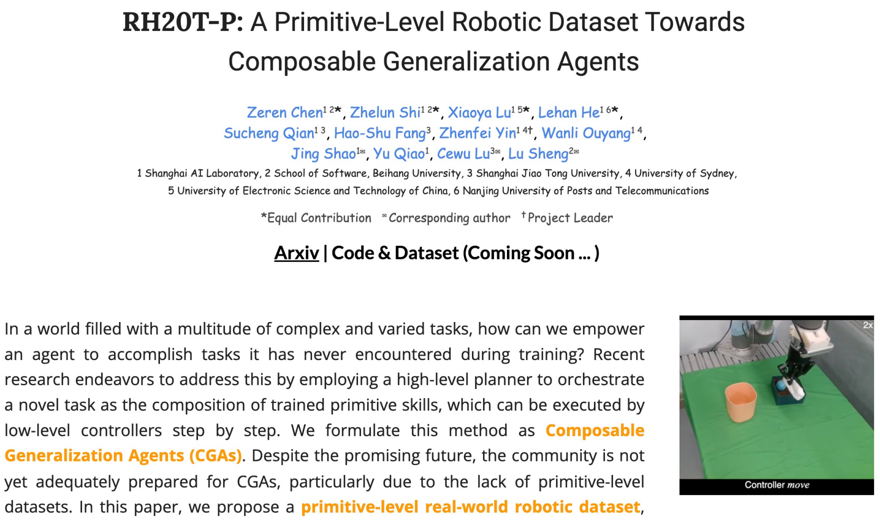
- **RA-P (IROS 2025, NeurIPS 2024 OWA):** composable generalizable agents in real world
  - Decompose complicated tasks into fine-grained primitive skills, generalizable to new physical skills
  - **A comprehensive dataset: RH20T-P**



RH20T-P: A Primitive-Level Robotic Dataset Towards Composable Generalization Agents, **IROS 2025**

技能泛化：组合泛化实现未知技能的学习

- More demos about the dataset and our RA-P? Please check the project page



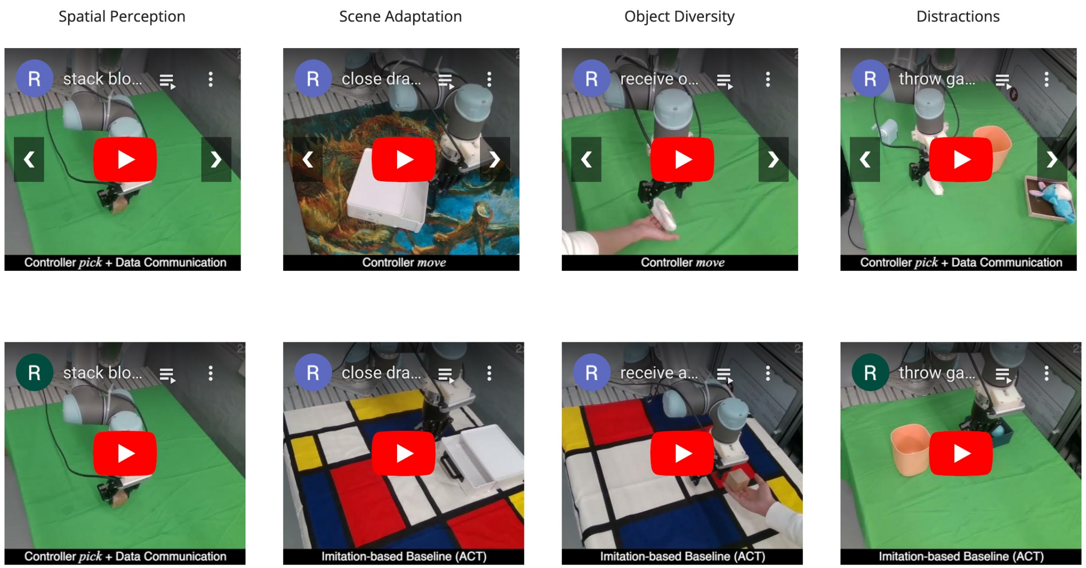
RH20T-P: A Primitive-Level Robotic Dataset Towards Composable Generalization Agents

Zeren Chen[1 2]*, Zhelun Shi[1 2]*, Xiaoya Lu[1 5]*, Lehan He[1 6]*,
Sucheng Qian[1 3], Hao-Shu Fang[3], Zhenfei Yin[1 4]†, Wanli Ouyang[1 4],
Jing Shao[1]✉, Yu Qiao[1], Cewu Lu[3]✉, Lu Sheng[2]✉
1 Shanghai AI Laboratory, 2 School of Software, Beihang University, 3 Shanghai Jiao Tong University, 4 University of Sydney,
5 University of Electronic Science and Technology of China, 6 Nanjing University of Posts and Telecommunications

*Equal Contribution ✉Corresponding author †Project Leader

Arxiv | Code & Dataset (Coming Soon ...)

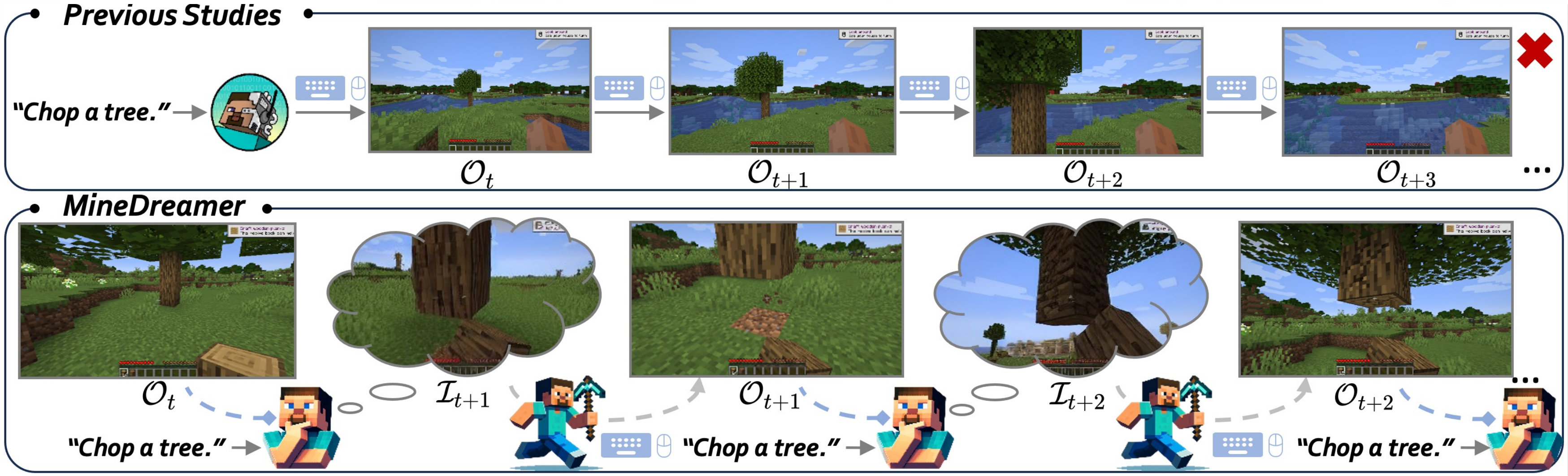
In a world filled with a multitude of complex and varied tasks, how can we empower an agent to accomplish tasks it has never encountered during training? Recent research endeavors to address this by employing a high-level planner to orchestrate a novel task as the composition of trained primitive skills, which can be executed by low-level controllers step by step. We formulate this method as **Composable Generalization Agents (CGAs)**. Despite the promising future, the community is not yet adequately prepared for CGAs, particularly due to the lack of primitive-level datasets. In this paper, we propose a **primitive-level real-world robotic dataset**,


Spatial Perception
Scene Adaptation
Object Diversity
Distractions
stack blo...
close dra...
receive o...
throw ga...
Controller *pick* + Data Communication
Controller *move*
Controller *move*
Controller *pick* + Data Communication
stack blo...
close dra...
receive a...
throw ga...
Controller *pick* + Data Communication
Imitation-based Baseline (ACT)
Imitation-based Baseline (ACT)
Imitation-based Baseline (ACT)

RH20T-P: A Primitive-Level Robotic Dataset Towards Composable Generalization Agents, **IROS 2025**

真实交互：想象链强化行动执行的环境动态适应性

- **MineDreamer (IROS 2025, NeurIPS 2024 OWA研讨会)**

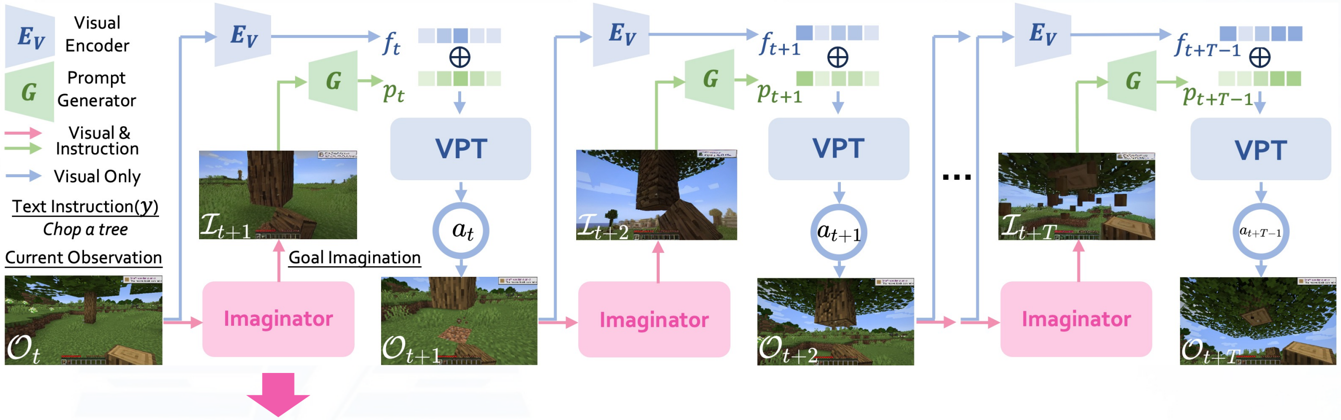


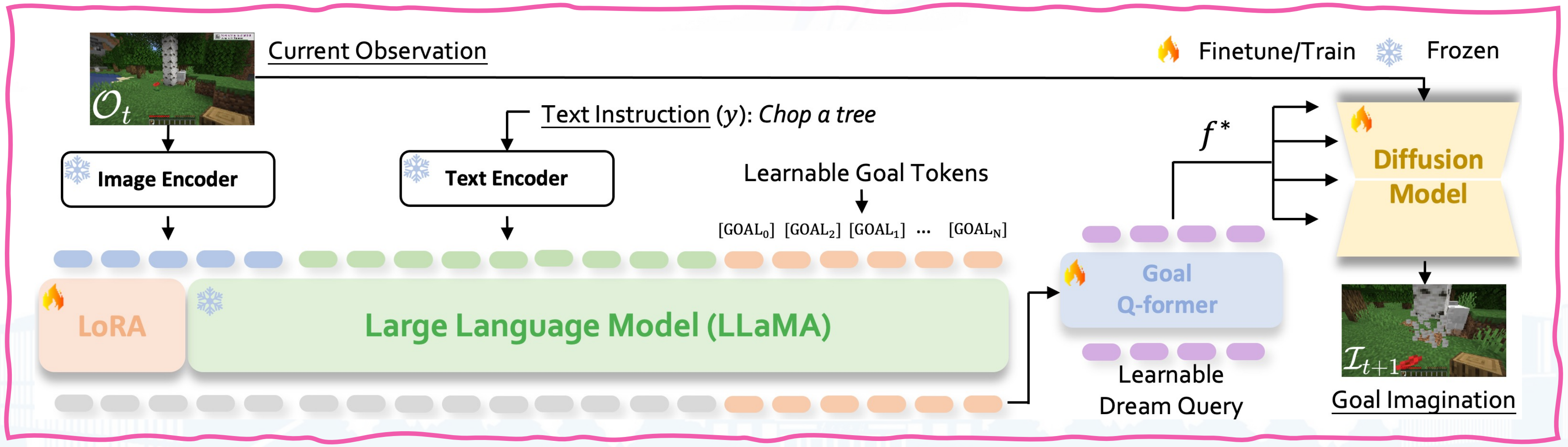
- 当处理困难问题时，一种可靠的思路是预测未来可能的执行效果，评估当前行动的可行性，以此来指导更可靠的行动执行
- **Chain-of-Imagination（想象链）可以强化具身行动执行的指令跟随能力**

MineDreamer: Learning to Follow Instructions via Chain-of-Imagination for Simulated-World Control, **IROS 2025**

真实交互：想象链强化行动执行的环境动态适应性

- **Chain-of-imagination**
  - **Imagination-conditional VPT in a sequential way**
  - 提供和**动态环境、语言指令、当前状态**更为相关、效果更为精准的视觉提示





真实交互：想象链强化行动执行的环境动态适应性

MineDreamer: Learning to Follow Instructions via Chain-of-Imagination for Simulated-World Control

Enshen Zhou[1 2*], Yiran Qin[1 3*]
Zhenfei Yin[1 4], Yuzhou Huang[3], Ruimao Zhang[3✉], Lu Sheng[2✉], Yu Qiao[1], Jing Shao[1 †]
[1]Shanghai Artificial Intelligence Laboratory; [2]Beihang University; [3]The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen); [4]The University of Sydney
* Equal Contribution ✉ Corresponding author † Project Leader

Arxiv | **PDF** | **Code** | **Dataset**

All Code, Datasets, and Checkpoints are released! Come on and enjoy it!

Demo: Programmatic Evaluation Following Text Instructions

😚 In the videos below, we demonstrate the performance of *MineDreamer* in Programmatic Evaluation, controlled through single-step text instruction.

Go explore

Maximum game duration: 3000 steps (1 minute and 40 seconds, FPS=30)

Maximum Travel Distance(Blocks): 640.27

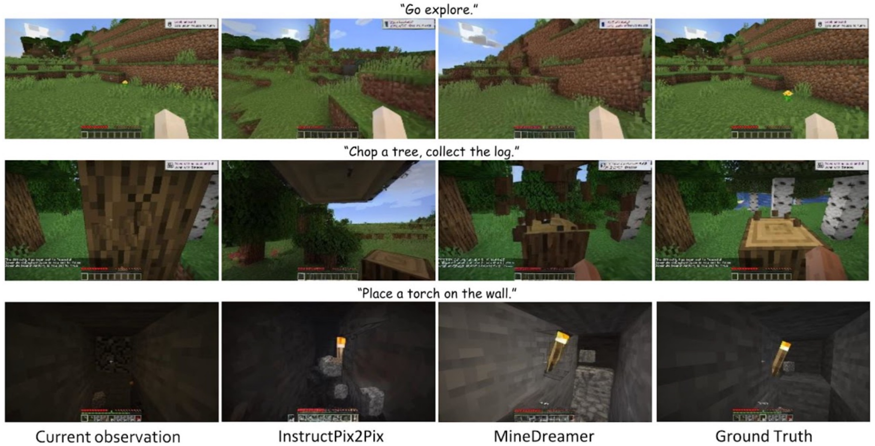
Collect seeds

Maximum game duration: 3000 steps (1 minute and 40 seconds, FPS=30)

Maximum Inventory Count about Seeds: 36

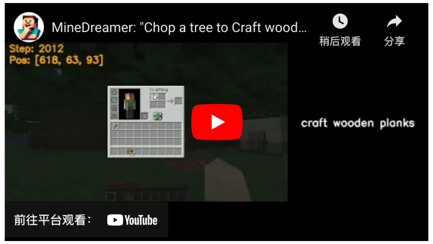
Imagination Visual Results on Evaluation Set Compared to the Baseline

🤩The images below demonstrate the generative quality of goal imagination compared to the baseline.



Demo: Command-Switching Evaluation for Long-Horizon Tasks Following Text Instructions

🥰 In the videos below, we demonstrate the performance of *MineDreamer* in Command-Switching Evaluation, controlled through multi-step text instructions.



Chop a tree -> Craft wooden planks

Maximum game duration: 3000 steps (2.5 minutes, FPS=20)

Switching time: at the 1500th step (1 minute and 15 seconds)

Gather dirt -> Build a tower

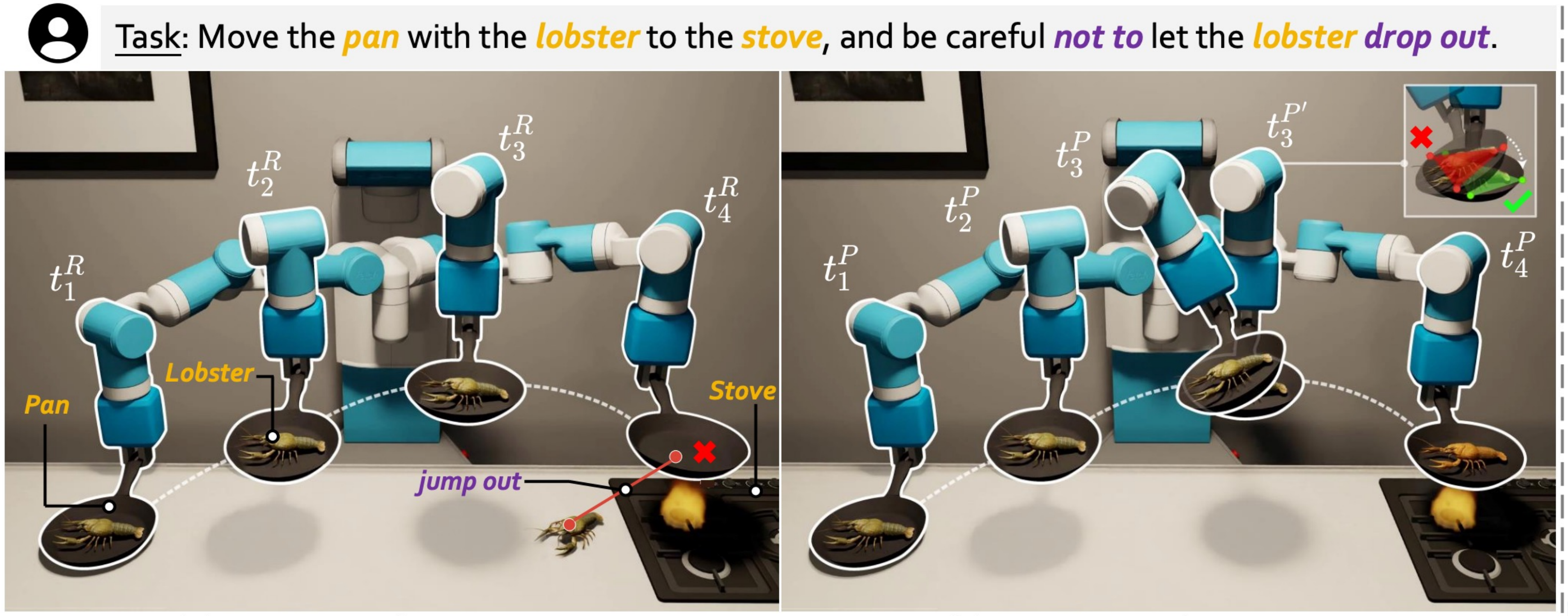
Maximum game duration: 3000 steps (2.5 minutes, FPS=20)

Switching time: at the 2000th step (1 minute and 40 seconds)

MineDreamer: Learning to Follow Instructions via Chain-of-Imagination for Simulated-World Control. **IROS 2025**

真实交互：实时监控提升具身任务执行的成功率

- **How to increase the success rate? → Reduce the rate of failure…**
- **Reactive（反应式） + Proactive（主动式）failure detections**



Task: Move the *pan* with the *lobster* to the *stove*, and be careful *not to* let the *lobster drop out*.

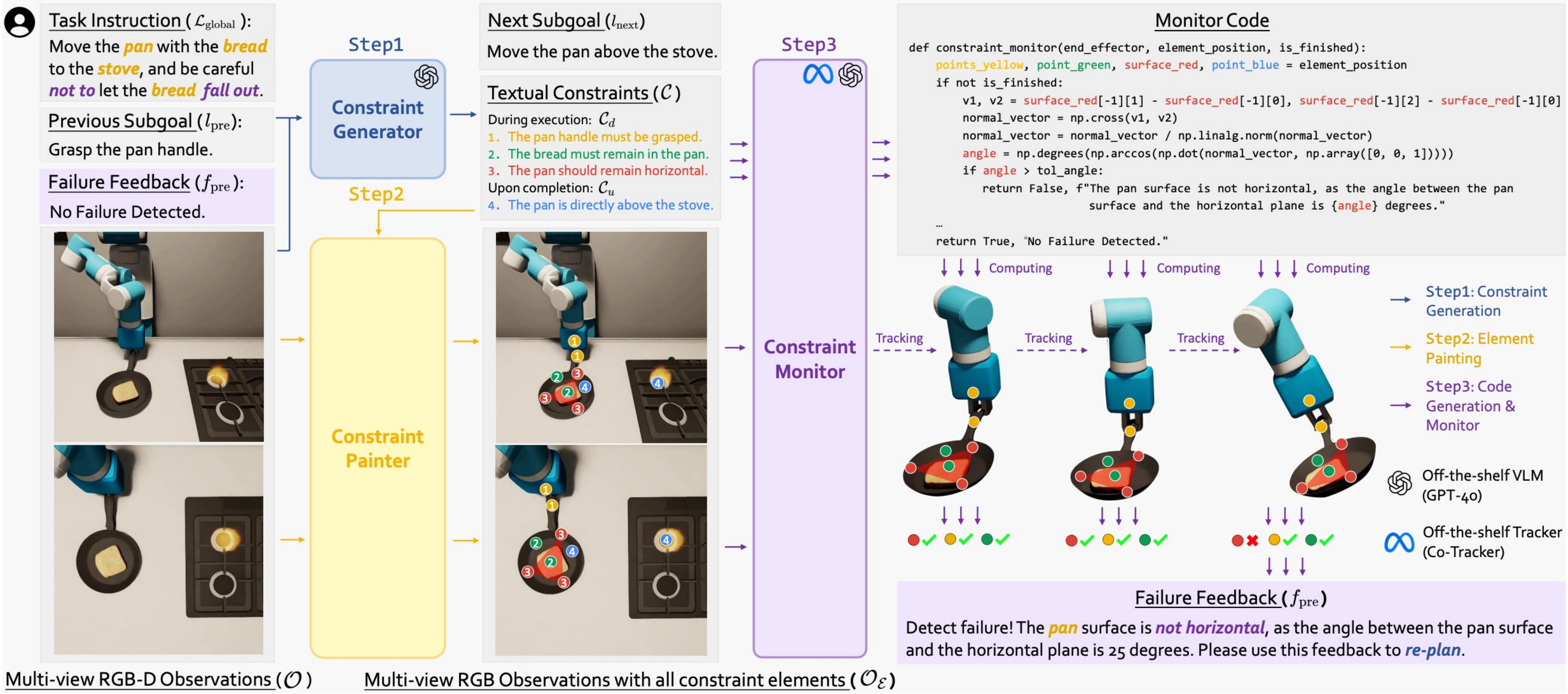
(a) Reactive Failure Detection

(b) Proactive Failure Detection

3D perception capability + Real-time efficiency VLM ✗

真实交互：实时监控提升具身任务执行的成功率

- **Code-as-Monitor (CVPR 2025):** Constraint-aware Visual Programming



Task Instruction ($\mathcal{L}_{\text{global}}$):
Move the *pan* with the *bread* to the *stove*, and be careful *not to* let the *bread* *fall out*.

Previous Subgoal (l_{pre}):
Grasp the pan handle.

Failure Feedback (f_{pre}):
No Failure Detected.

Next Subgoal (l_{next})
Move the pan above the stove.

Textual Constraints ($\mathcal{C}$)
During execution: $\mathcal{C}_d$
1. The pan handle must be grasped.
2. The bread must remain in the pan.
3. The pan should remain horizontal.

Upon completion: $\mathcal{C}_u$
4. The pan is directly above the stove.

Monitor Code

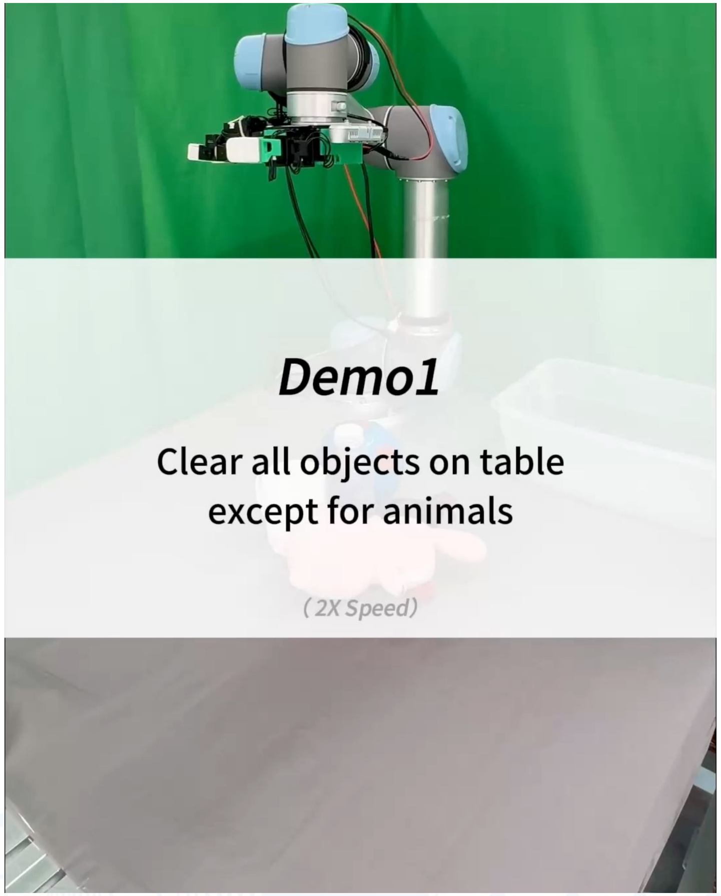
```
def constraint_monitor(end_effector, element_position, is_finished):
    points_yellow, point_green, surface_red, point_blue = element_position
    if not is_finished:
        v1, v2 = surface_red[-1][1] - surface_red[-1][0], surface_red[-1][2] - surface_red[-1][0]
        normal_vector = np.cross(v1, v2)
        normal_vector = normal_vector / np.linalg.norm(normal_vector)
        angle = np.degrees(np.arccos(np.dot(normal_vector, np.array([0, 0, 1]))))
        if angle > tol_angle:
           return False, f"The pan surface is not horizontal, as the angle between the pan
                                surface and the horizontal plane is {angle} degrees."
    …
    return True, "No Failure Detected."
```

Failure Feedback (f_{pre})
Detect failure! The *pan* surface is *not horizontal*, as the angle between the pan surface and the horizontal plane is 25 degrees. Please use this feedback to *re-plan*.

Multi-view RGB-D Observations ($\mathcal{O}$)

Multi-view RGB Observations with all constraint elements ($\mathcal{O}_{\mathcal{E}}$)

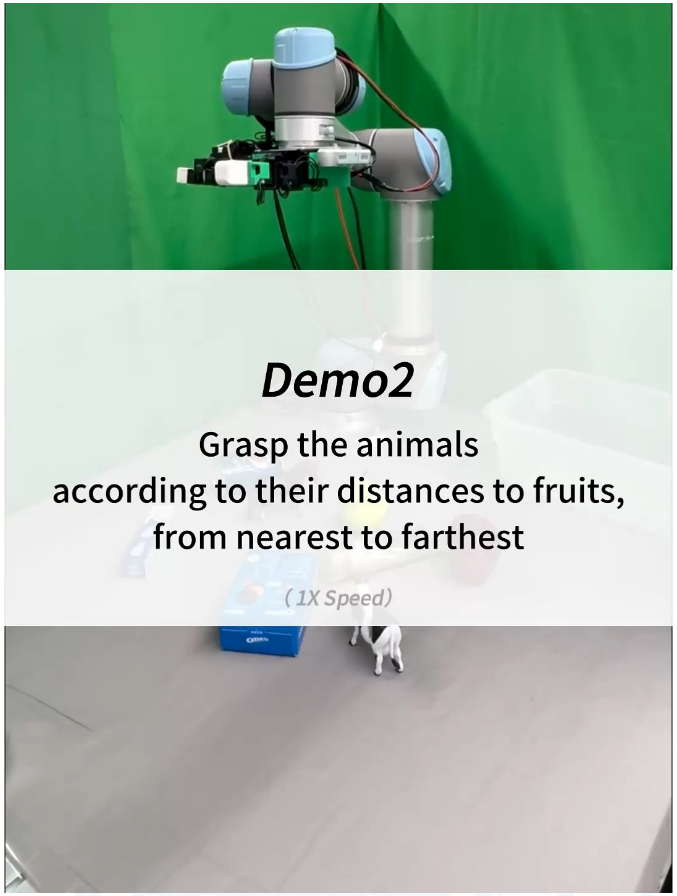
真实交互：实时监控提升具身任务执行的成功率



Demo1

Clear all objects on table
except for animals

(2X Speed)



Demo2

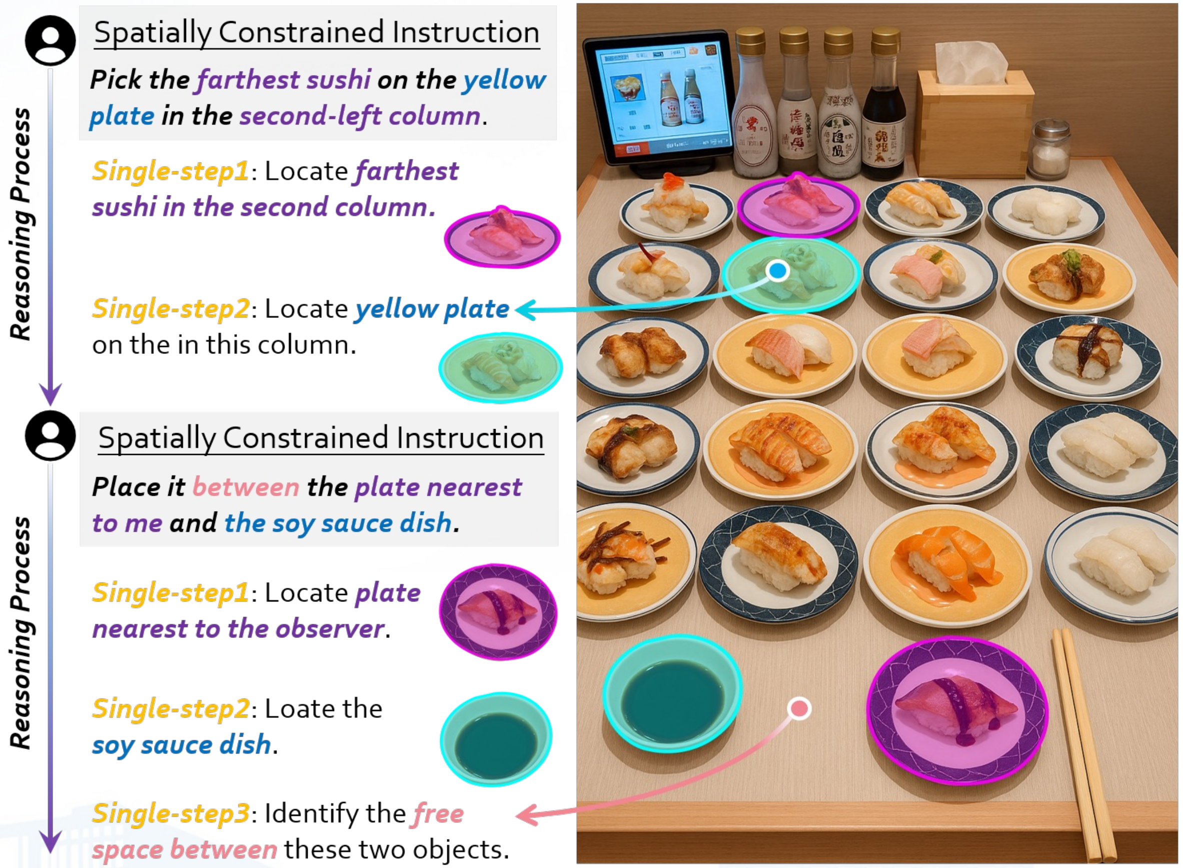
Grasp the animals
according to their distances to fruits,
from nearest to farthest

(1X Speed)

- The first framework to integrate both reactive and proactive failure detection
- Simplify real-time failure detection with high precision
- Achieves SOTA performance in both simulated and real-world environments
- Exhibits strong generalizability on unseen scenarios, tasks, and objects

Code-as-Monitor: Constraint-aware Visual Programming for Reactive and Proactive Robotic Failure Detection, **CVPR 2025**

具身大脑的基本能力提升：空间感知 + 深度思考



Spatially Constrained Instruction

Pick the farthest sushi on the yellow plate in the second-left column.

Single-step1: Locate ***farthest sushi in the second column.***

Single-step2: Locate ***yellow plate*** on the in this column.

Spatially Constrained Instruction

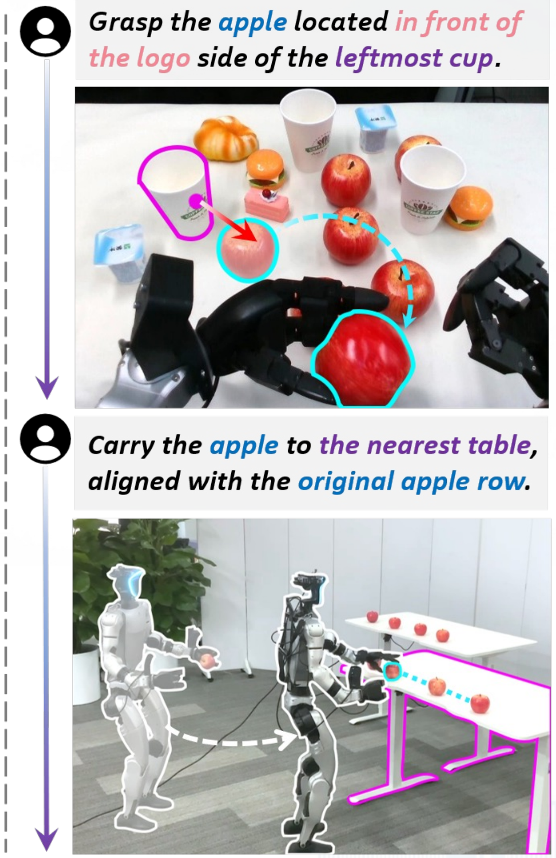
Place it between the plate nearest to me and the soy sauce dish.

Single-step1: Locate ***plate nearest to the observer.***

Single-step2: Loate the ***soy sauce dish.***

Single-step3: Identify the ***free space between*** these two objects.

Multi-Step Spatial Referring with Reasoning



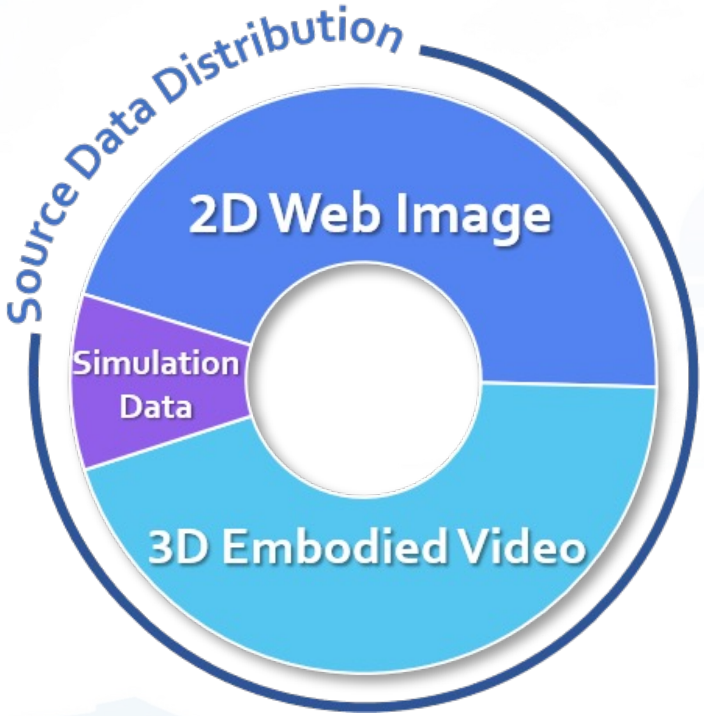
Grasp the apple located in front of the logo side of the leftmost cup.

Carry the apple to the nearest table, aligned with the original apple row.

Real-world Manipulation and Navigation

Zhou E, et al. RoboRefer: Towards Spatial Referring with Reasoning in Vision-Language Models for Robotics. (in Submission)

具身大脑的基本能力提升：空间感知 + 深度思考

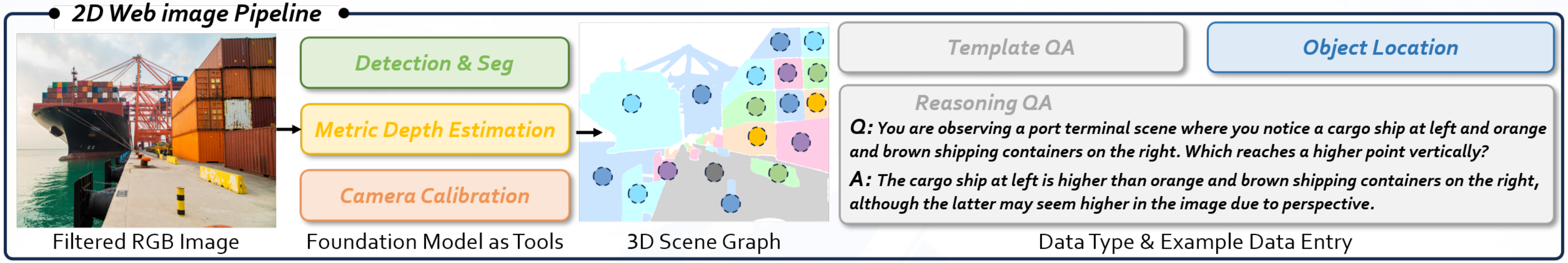


大规模数据提升能力提升

- 2D Web Images (OpenImages)
- 3D Embodied Videos (CA-1M)
- Simulation Data by Infinigen with generative assets

Zhou E, et al. RoboRefer: Towards Spatial Referring with Reasoning in Vision-Language Models for Robotics. (in Submission)

具身大脑的基本能力提升：空间感知＋深度思考



2D Web image Pipeline

Filtered RGB Image → Foundation Model as Tools (*Detection & Seg*, *Metric Depth Estimation*, *Camera Calibration*) → 3D Scene Graph

Data Type & Example Data Entry

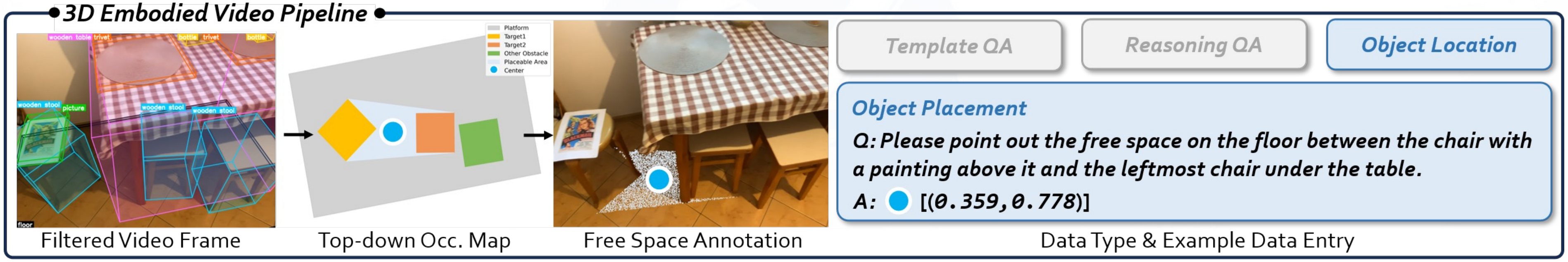
Template QA

Object Location

Reasoning QA

Q: You are observing a port terminal scene where you notice a cargo ship at left and orange and brown shipping containers on the right. Which reaches a higher point vertically?

A: The cargo ship at left is higher than orange and brown shipping containers on the right, although the latter may seem higher in the image due to perspective.



3D Embodied Video Pipeline

Filtered Video Frame → Top-down Occ. Map → Free Space Annotation

Data Type & Example Data Entry

Template QA

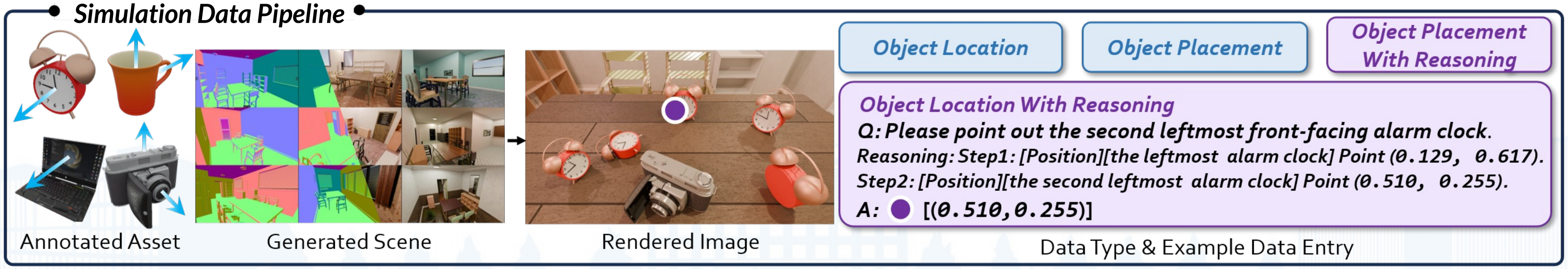
Reasoning QA

Object Location

Object Placement

Q: Please point out the free space on the floor between the chair with a painting above it and the leftmost chair under the table.

A: ● [(0.359,0.778)]



Simulation Data Pipeline

Annotated Asset → Generated Scene → Rendered Image

Data Type & Example Data Entry

Object Location

Object Placement

Object Placement With Reasoning

Object Location With Reasoning

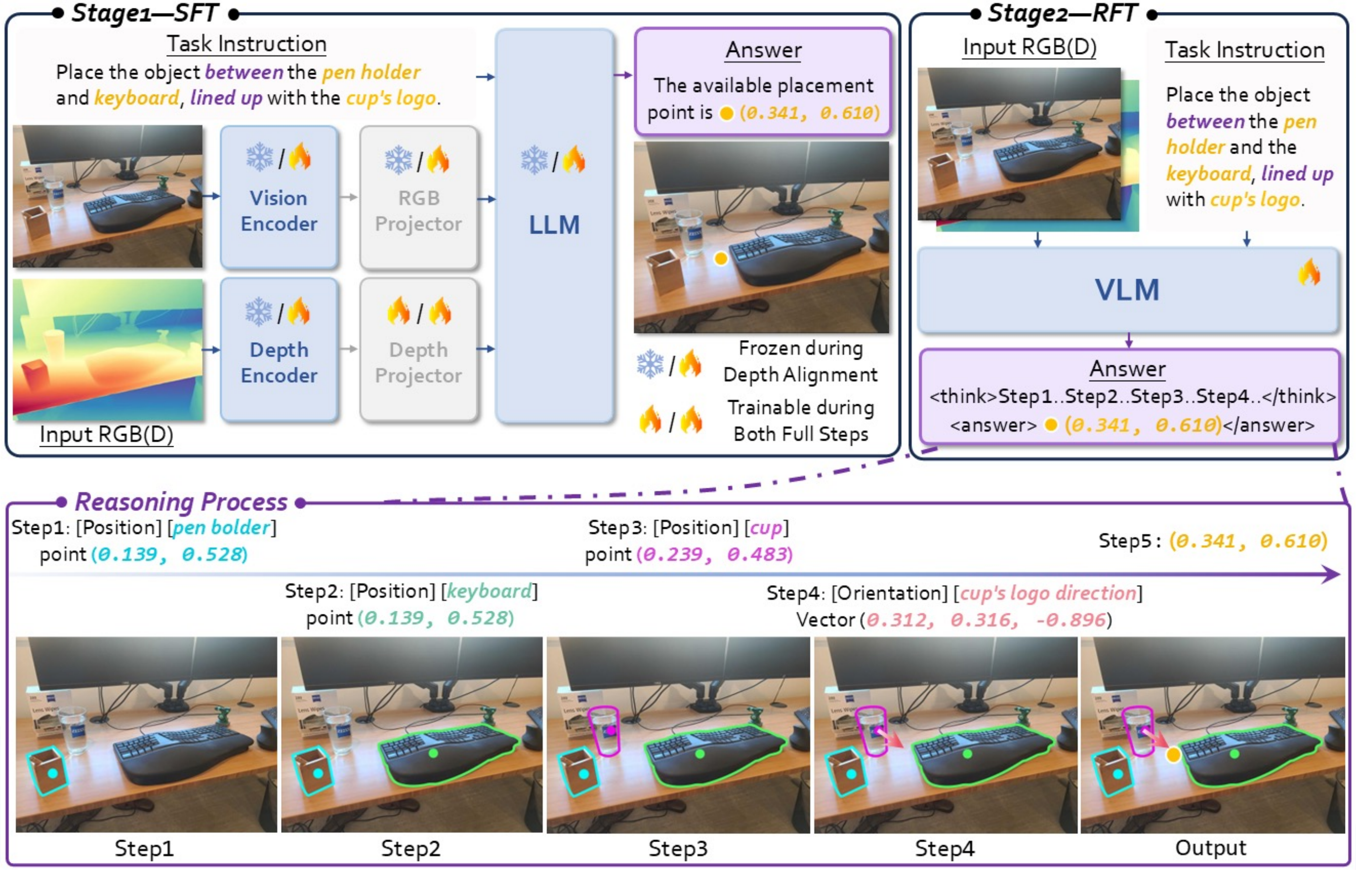
Q: Please point out the second leftmost front-facing alarm clock.

Reasoning: Step1: [Position][the leftmost alarm clock] Point (0.129, 0.617).
Step2: [Position][the second leftmost alarm clock] Point (0.510, 0.255).

A: ● [(0.510,0.255)]

具身大脑的基本能力提升：空间感知 + 深度思考

- **RoboRefer:** Accurate Spatial referring by VLMs that enables multi-step dynamic reasoning


Stage1—SFT
Task Instruction
Place the object between the pen holder and keyboard, lined up with the cup's logo.
Answer
The available placement point is ● (0.341, 0.610)
❄/🔥 Vision Encoder
❄/🔥 RGB Projector
❄/🔥 LLM
❄/🔥 Depth Encoder
🔥/🔥 Depth Projector
Input RGB(D)
❄/🔥 Frozen during Depth Alignment
🔥/🔥 Trainable during Both Full Steps
Stage2—RFT
Input RGB(D)
Task Instruction
Place the object between the pen holder and the keyboard, lined up with cup's logo.
VLM 🔥
Answer
<think>Step1..Step2..Step3..Step4..</think>
<answer> ● (0.341, 0.610)</answer>
Reasoning Process
Step1: [Position] [pen bolder] point (0.139, 0.528)
Step2: [Position] [keyboard] point (0.139, 0.528)
Step3: [Position] [cup] point (0.239, 0.483)
Step4: [Orientation] [cup's logo direction] Vector (0.312, 0.316, -0.896)
Step5: (0.341, 0.610)
Step1
Step2
Step3
Step4
Output

Zhou E, et al. RoboRefer: Towards Spatial Referring with Reasoning in Vision-Language Models for Robotics. (in Submission)

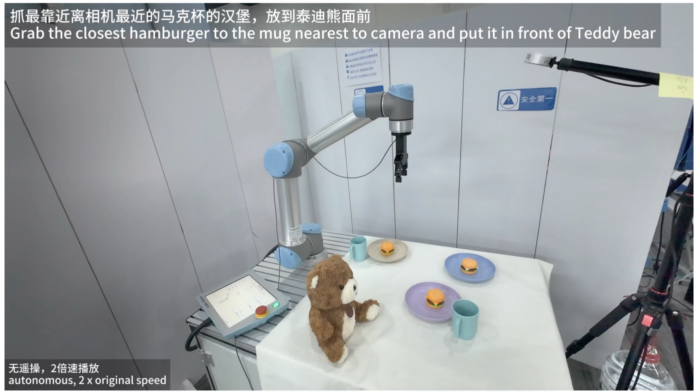
具身大脑的基本能力提升：空间感知 + 深度思考

Demo3: G1 Humanoid Mobile Manipulation

Pick the apple in front of the logo side of the leftmost cup and place it on the nearest table, aligned with the apple row.

(1x Speed)

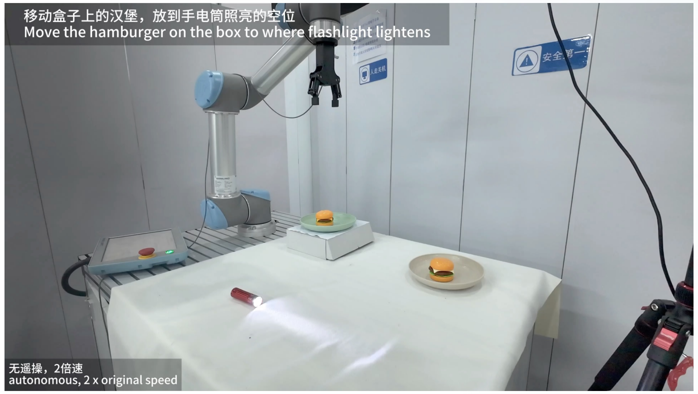
展示人形机器人（宇树G1）在**移动操作任务**中的效果，展示了模型**判断物体远近、识别朝向、距离的能力。**



抓最靠近离相机最近的马克杯的汉堡，放到泰迪熊面前
Grab the closest hamburger to the mug nearest to camera and put it in front of Teddy bear

无遥操，2倍速播放
autonomous, 2 x original speed

展示机械臂（UR5）在场景关键要素变化下完成抓取放置，展示了模型**快速的场景适应能力**，以及模型**判断物体远近、识别朝向、距离的能力。**



移动盒子上的汉堡，放到手电筒照亮的空位
Move the hamburger on the box to where flashlight lightens

无遥操，2倍速
autonomous, 2 x original speed

展示机械臂（UR5）抓取指定高度物体并放置在光线照射区域，展示模型**物体空间高度识别**与**光照区域识别能力。**

抓取橙子左边的苹果，放在柜子第一层的空闲区域
Pick up the apple to the left of the orange and place it in the vacant space on the first shelf of the cabinet.

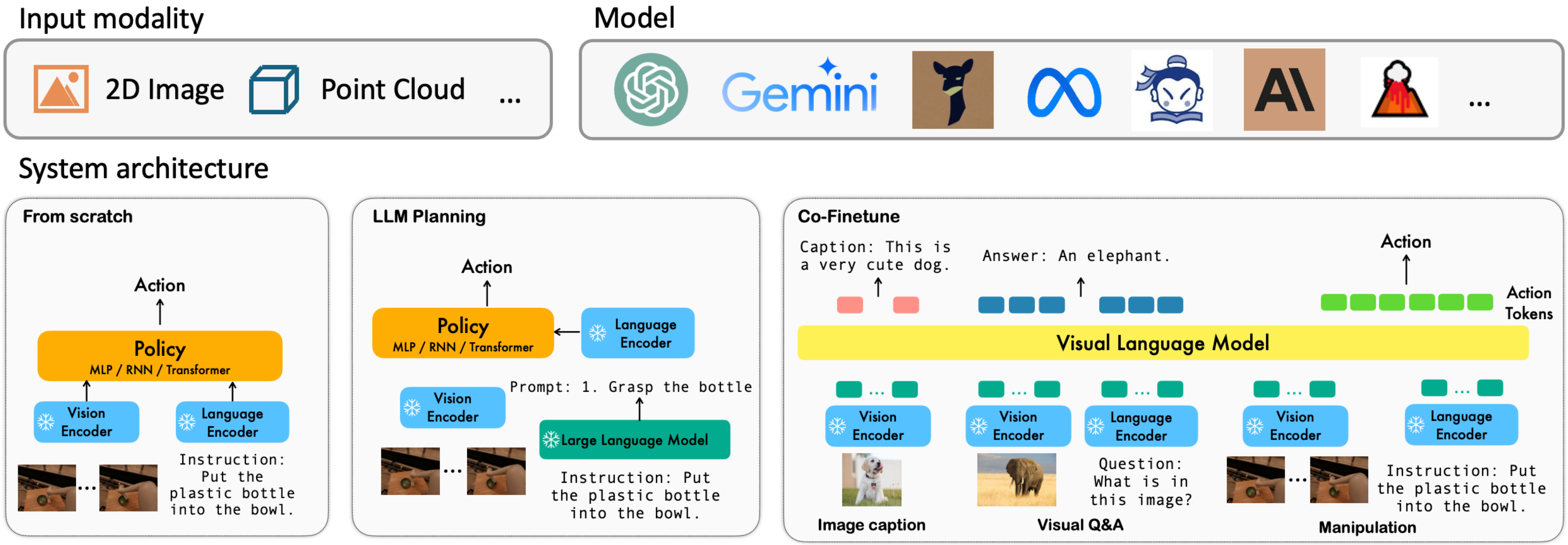
展示机械臂（Franka）对物体的抓取放置，展示了模型基于**空间关系进行物体指代**的能力，以及在**三维空间中定位空闲区域**的能力

任务指令：“我要喝右边的饮料”，展示人形机器人（宇树G1）在**灵巧手**操作任务中的效果，体现了顶层模型判断相对方向的能力，以及灵巧手模型精准控制能力

任务指令：“我要吃肉汉堡”，展示双臂机器人（松灵）在**夹爪**操作任务中的效果，体现了顶层模型对任务拆解以及执行的能力

Limitations still met for embodied models?

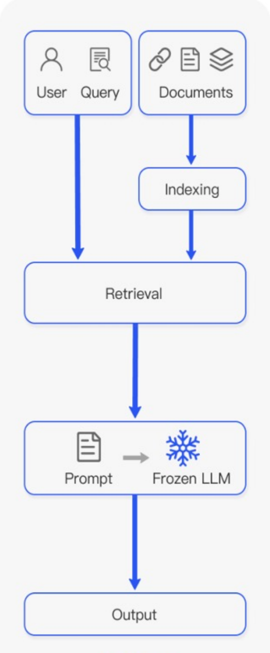
- Semantic and spatial perception?



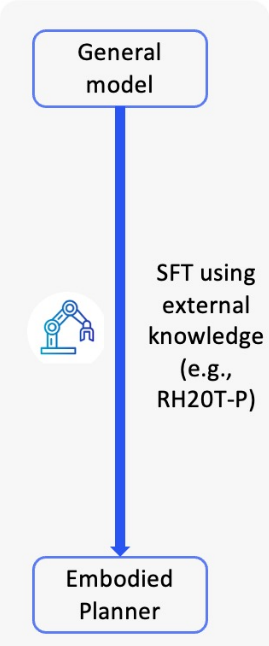
[RoboFlamingo, Li et al., 2024]

Limitations still met for embodied models?

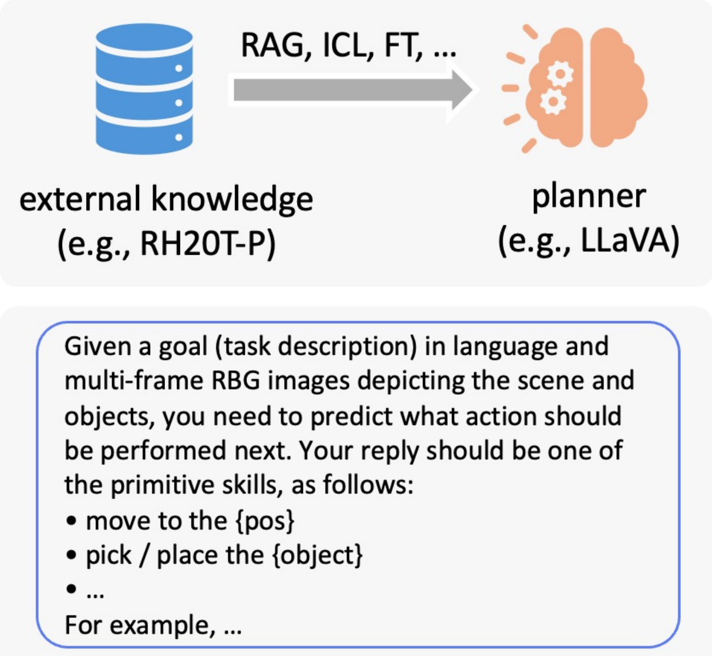
- Semantic and spatial perception?
- Reliable long-horizon planning?



RAG



Fine-tuning

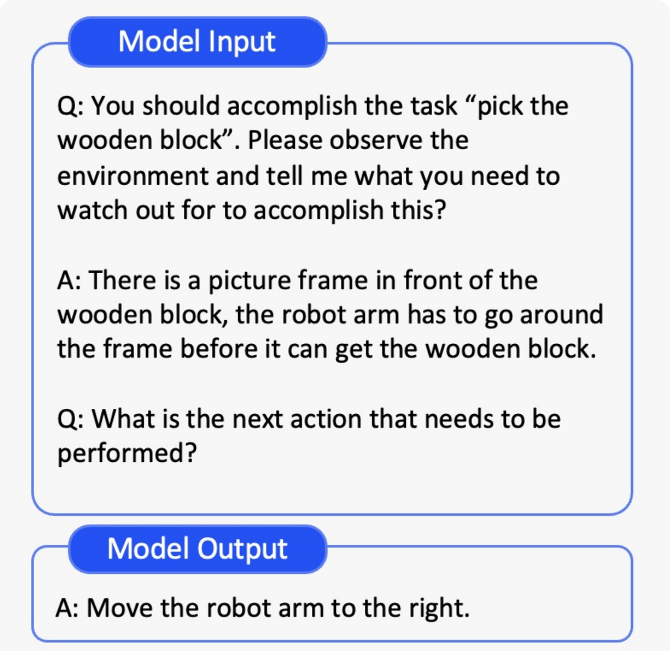


Given a goal (task description) in language and multi-frame RBG images depicting the scene and objects, you need to predict what action should be performed next. Your reply should be one of the primitive skills, as follows:

- move to the {pos}
- pick / place the {object}
- ...

For example, ...

In-context Learning



Q: You should accomplish the task “pick the wooden block”. Please observe the environment and tell me what you need to watch out for to accomplish this?

A: There is a picture frame in front of the wooden block, the robot arm has to go around the frame before it can get the wooden block.

Q: What is the next action that needs to be performed?

A: Move the robot arm to the right.

Chain-of-Thought

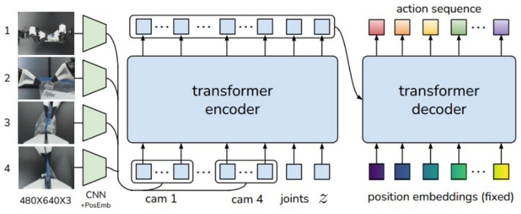
Limitations still met for embodied models?

- Semantic and spatial perception?
- Reliable long-horizon planning?
- Universally drive multiple specialized controllers for diverse skills?

Specialized Controller

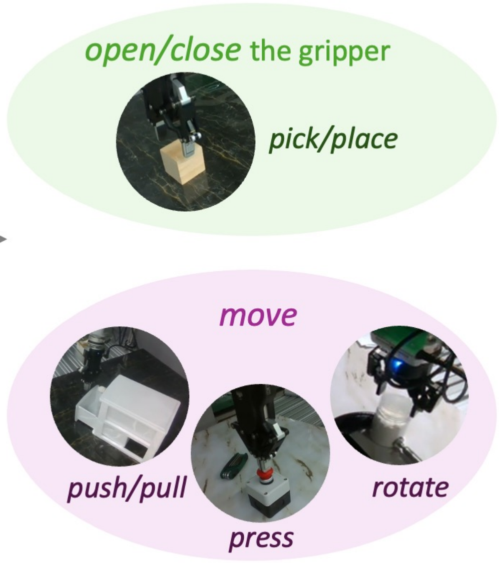
Diverse Primitive Skills

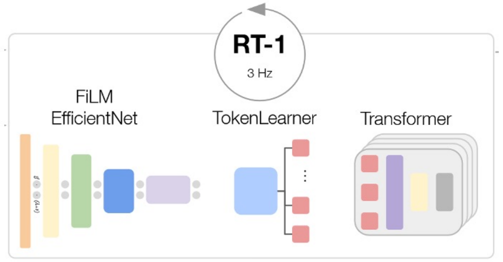
How to make the LLM drive the controller?



[ACT, Zhao et al., 2023]

How to make the controller execute primitive skills?





[RT-1, Brohan et al., 2023]

Thank You!

盛律，北京航空航天大学

Homepage : https://lucassheng.github.io/

Email : lsheng@buaa.edu.cn