多媒体智能课程汇报

2024.7.1

中国科学院计算技术研究所
Institute of Computing Technology,Chinese Academy of Sciences

题目选择

- 论文：EmbodiedGPT: Vision-Language Pre-Training via Embodied Chain of Thought
- 作者： Yao Mu、Qinglong Zhang、Mengkang Hu、Wenhai Wang、Mingyu Ding、Jun Jin、Bin Wang、Jifeng Dai、Yu Qiao、Ping Luo
- 单位：香港大学、上海人工智能实验室、诺亚方舟实验室
- 录用：NeurIPS 2023（CCF-A）
- 领域：多媒体智能、具身智能

汇报内容

汇报
内容

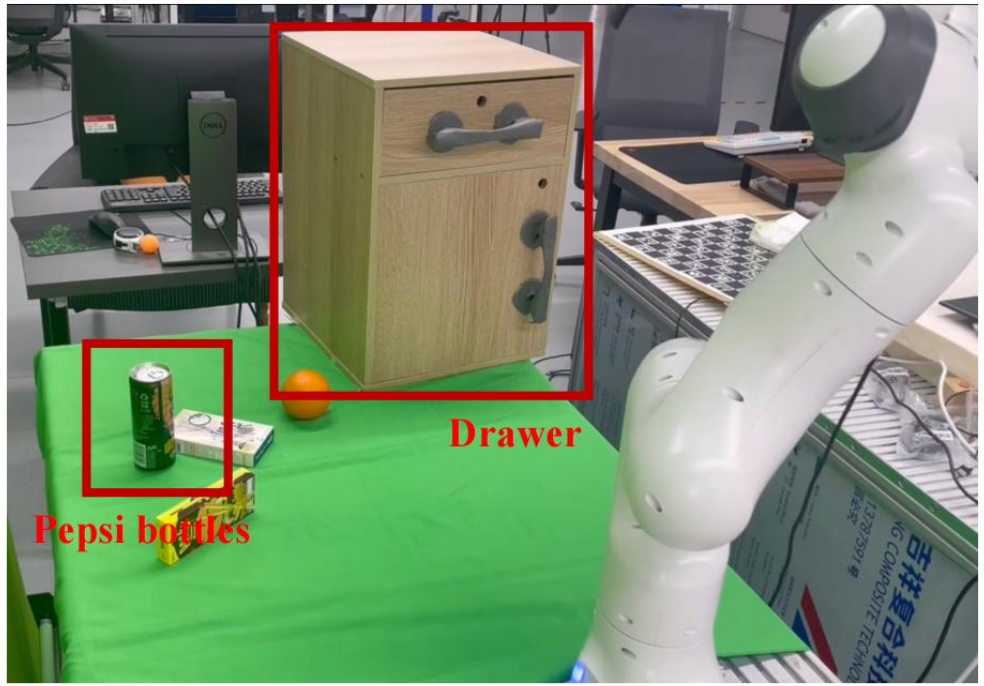
一 研究背景

二 研究方法

三 实验评估

研究背景

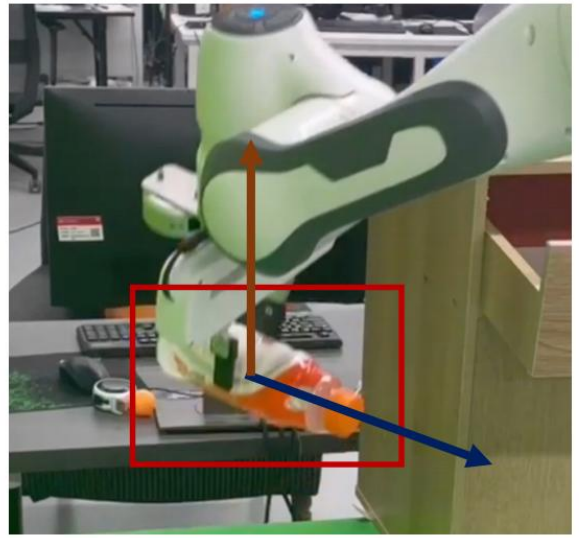
Perception: Classification & Grounding



Embodied Cognition

Where to interact?
How to interact?

Interactive preferences
Physical Constrains

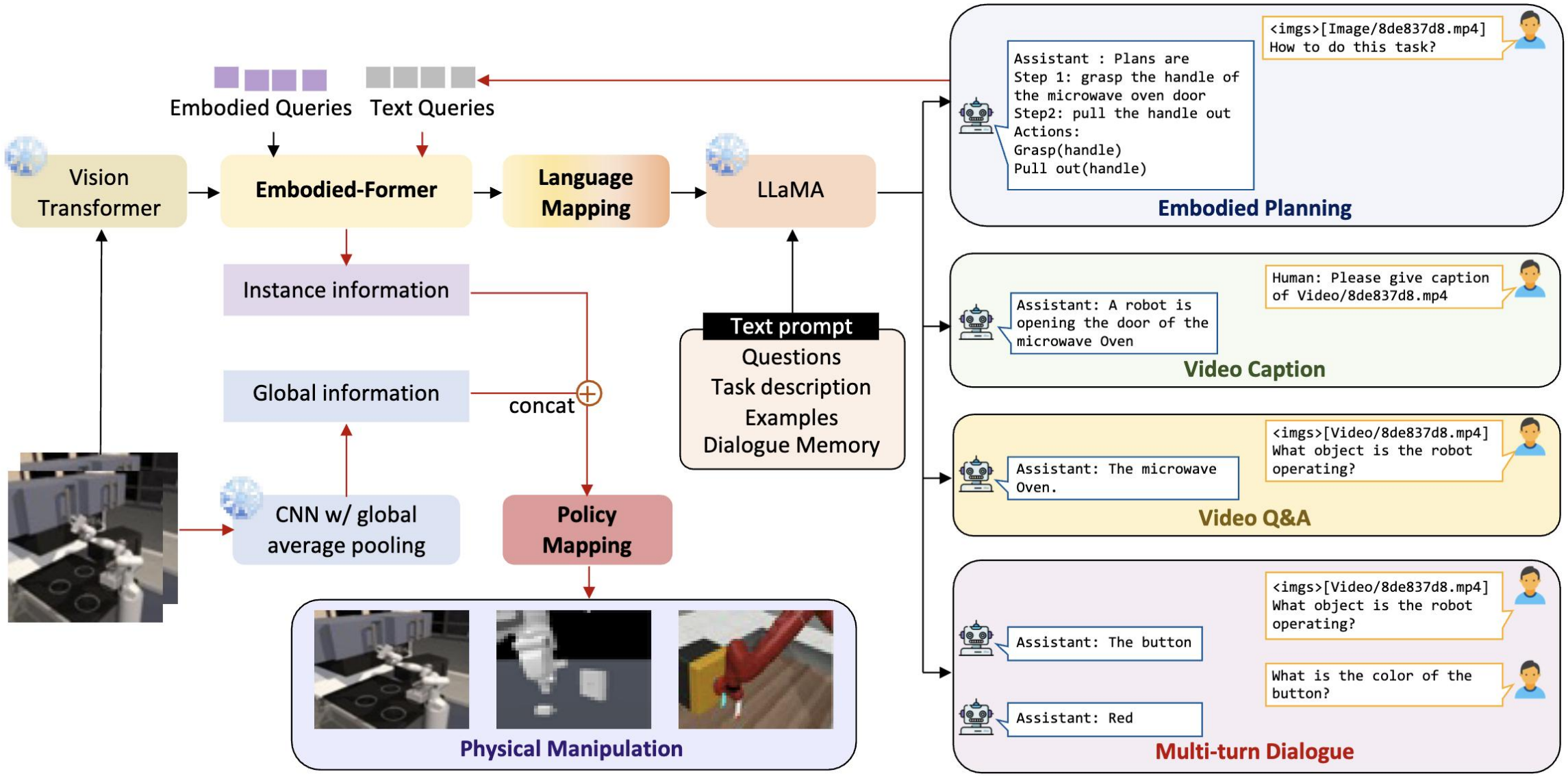


I know it → **I know how to deal with it**

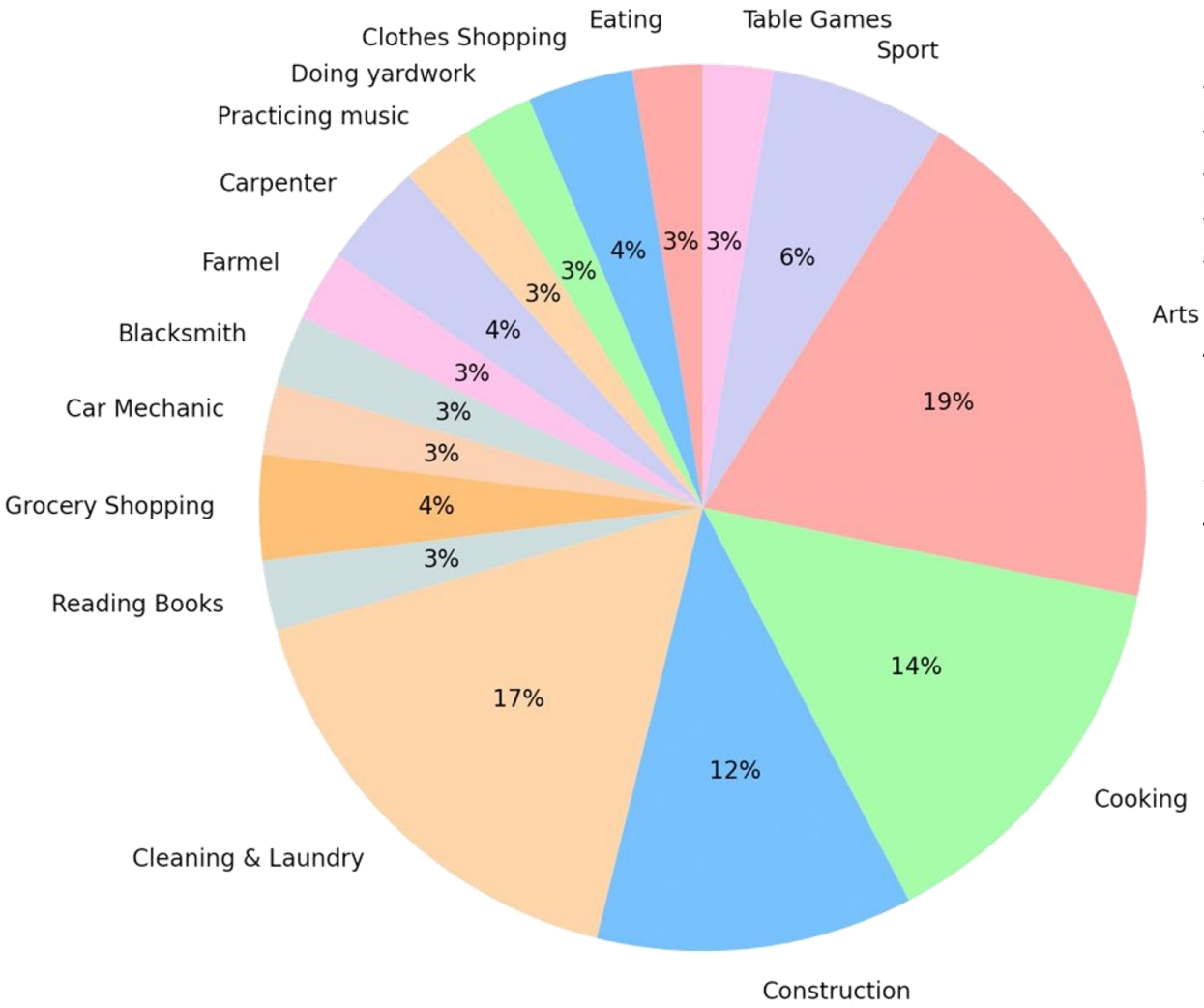
研究背景

- **LLM的进步：** LLM具备在语言理解、推理和建立逻辑链的能力
- **大规模数据集：** EAI任务需要以自我为中心的机器人领域的数据、精确规划的结构化语言指令
- **现有数据集的局限性：** 成本高、规模较小且特定于特定领域， Sim2Real

研究方法：Embodied GPT



研究方法：EgoCot Dataset



Task: Move sliced meat onto a plate

Plans:

Step 1: First, grasp the handle of the knife.

Step 2: Pick up the knife.

Step 3: Use your left hand to position the meat onto the knife.

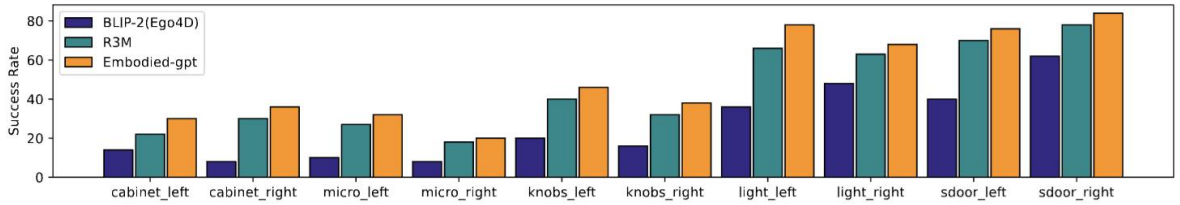
Step 4: Then, move the knife to the position over the plate.

Step 5: Transfer the meat slices from the knife to the plate.

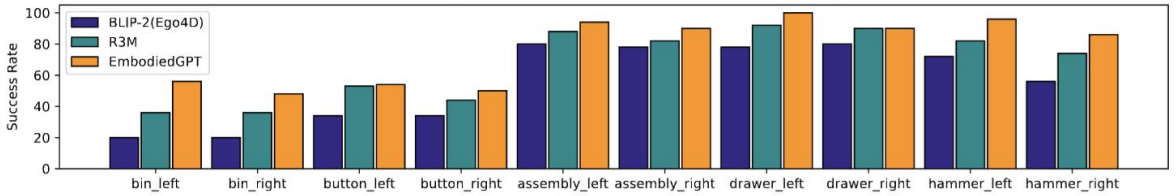
Actions:

1. grasp(handle of the knife)
2. pick up(knife)
3. position(meat, onto the knife)
4. move(knife, position over the plate)
5. transfer(meat slices, plate)

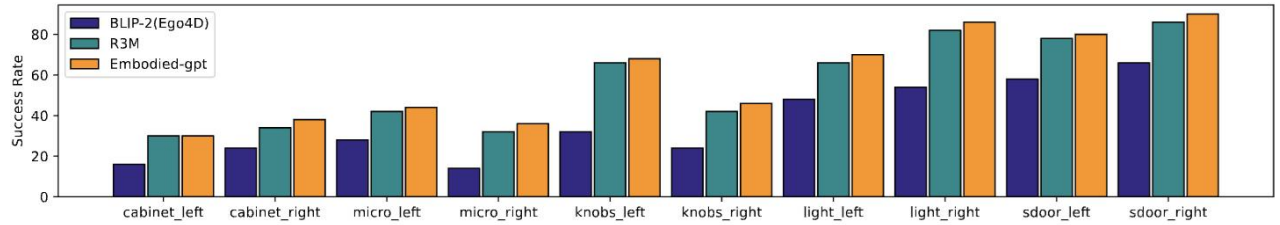
实验评估



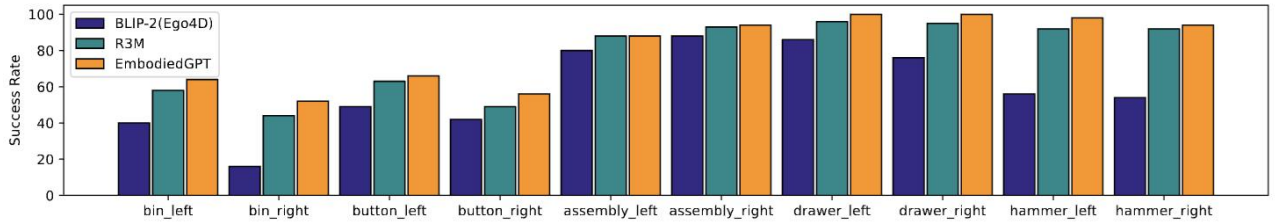
(a) Performance comparison in *Franka Kitchen* with only 10 demos.



(b) Performance comparison in *Meta-World* with only 10 demos.



(a) Performance comparison in *Franka Kitchen* with 25 demos.



(b) Performance comparison in *Meta-World* with 25 demos.

Model	Object(↑)	Spatial(↑)	Redundancy(↓)	Plan Reasonable(↑)	Plan Executable(↑)
Minigpt4	5.6	4.8	4.4	4.5	4.8
LLaVA-7B	7.3	7.4	3.9	7.5	6.6
LLaVA-13B	**8.5**	8.6	3.4	8.4	7.6
EmbodiedGPT	8.4	**8.8**	**2.6**	**8.8**	**8.4**

Table 1: Generate Quality Evaluation on image input tasks.

Figure 6: Performance of EmbodiedGPT in low-level control tasks with 25 demonstration demos.

Model	Franka(10 demos)	Franka(25 demos)	Meta-World(10 demos)	Meta-World(25 demos)
EmbodiedGPT	**50.8**% ±2.8	**58.5**% ±2.7	**76.4**% ±2.2	**81.2**%±2.0
- Close-loop	38.6% ±2.9	47.3% ±2.5	62.7% ±2.2	64.9% ±2.0
- COT	26.2% ±3.2	36.4% ±2.7	55.2% ±2.4	58.7% ±2.0

谢谢！
请各位批评指正！