



東南大學
SOUTHEAST UNIVERSITY

Panasonic



RAP: Retrieval-Augmented Planning with Contextual Memory for Multimodal LLM Agents

NeurIPS 2024

**Tomoyuki Kagaya * 1 Thong Jing Yuan * 2 Yuxuan Lou * 3
Jayashree Karlekar 2 Sugiri Pranata 2 Akira Kinose 1 Koki Oguri 1
Felix Wick 4 Yang You 3**

*Equal contribution

1Panasonic Connect Co., Ltd., Japan

2Panasonic R&D Center, Singapore

3National University of Singapore, Singapore

4Panasonic R&D Center Germany, Germany.

汇报人：张兴才

2025年 6月 20日

1

背景

2

方法

3

实验

4

总结

1

背景

2

方法

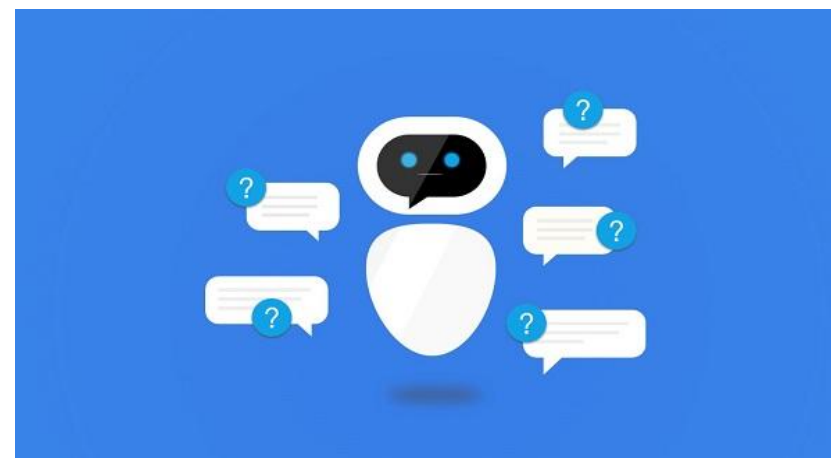
3

实验

4

总结

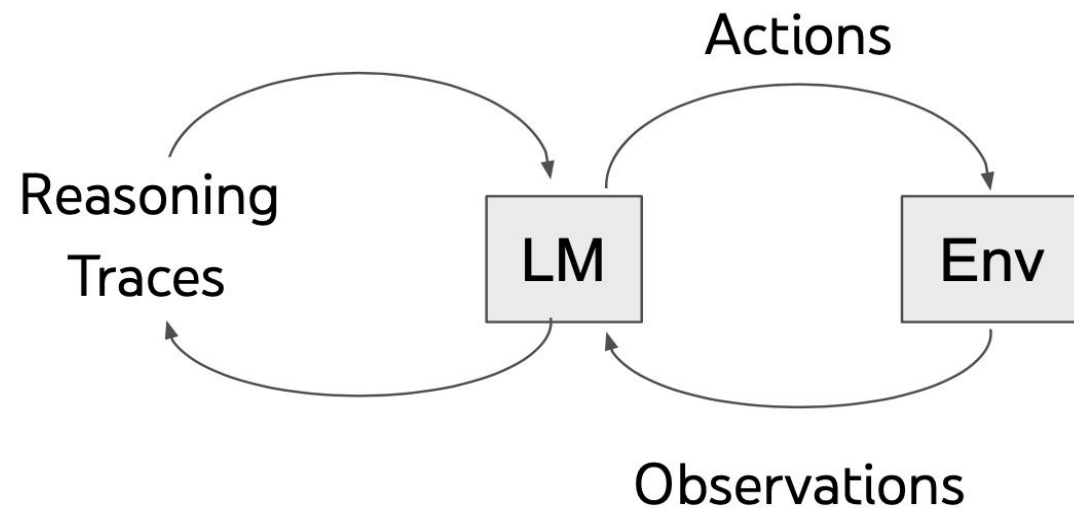
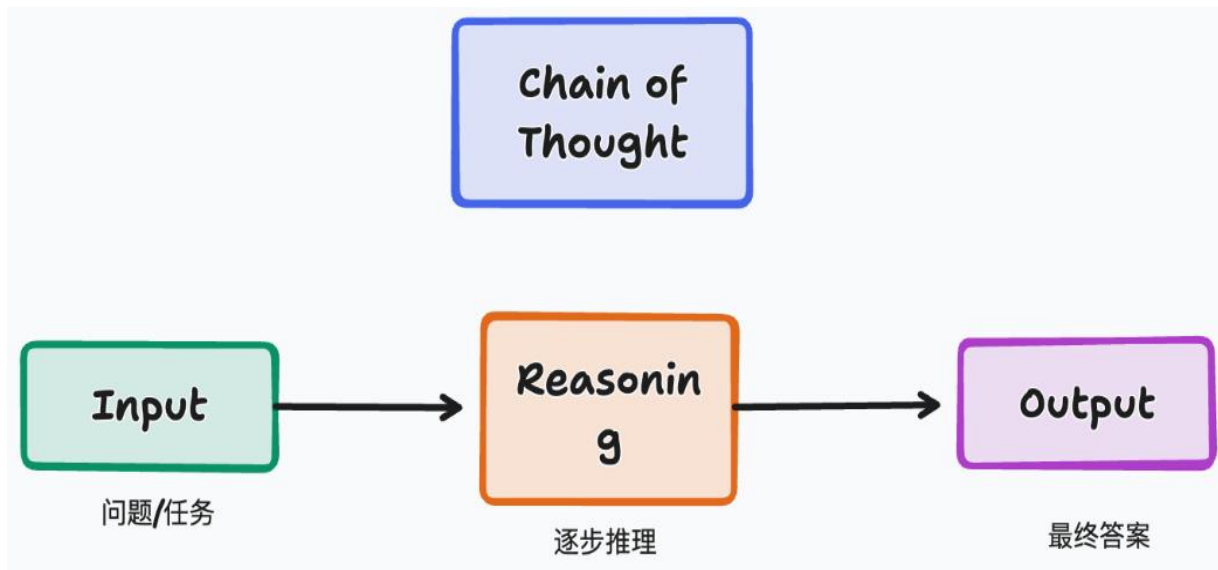
背景：大模型决策场景



LLM能够作为智能体部署在复杂多样的决策场景中

背景：如何提高大模型决策能力

思考方式：

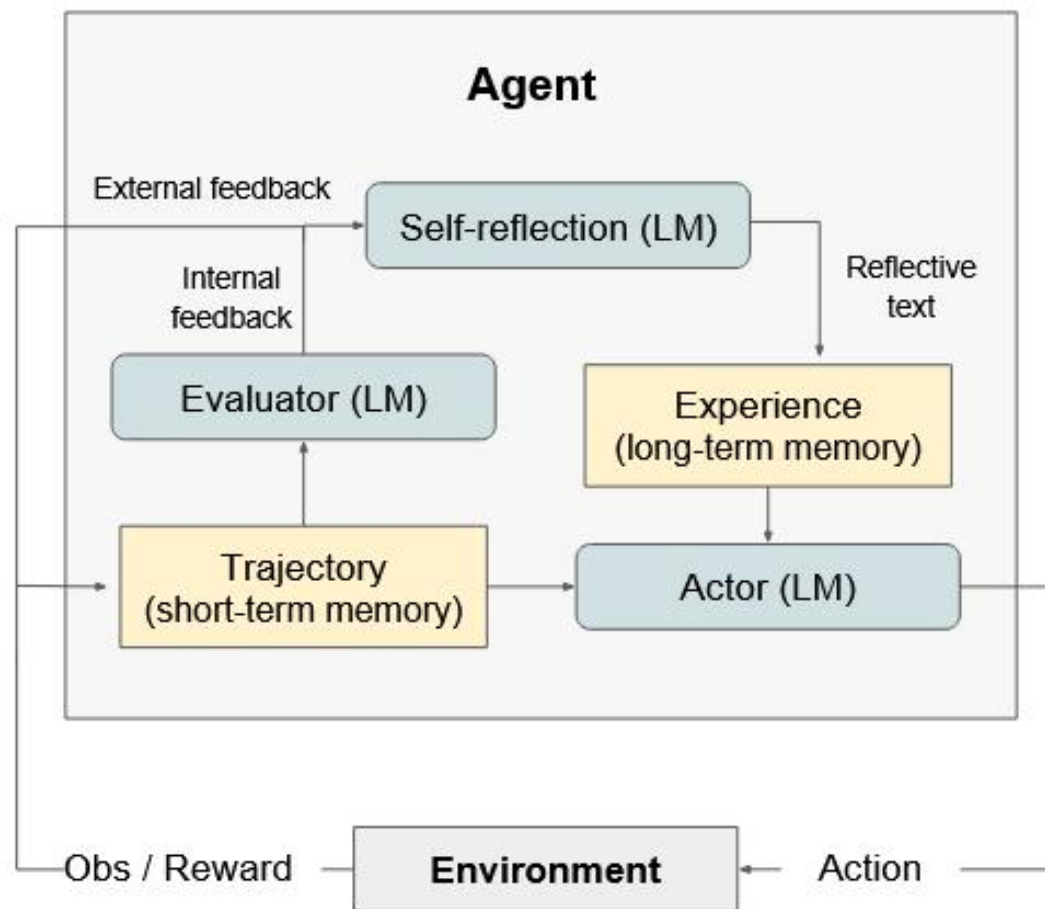


ReAct (Reason + Act)

引导LLM在复杂任务场景中使用不同的思考方式

背景：如何提高大模型决策能力

自我反思：



LLM对未成功的任务进行自我反思，从而逐步提升解决问题的能力

局限性：单个任务轨迹

思考方式		自我反思
CoT [1]	React [2]	Reflexion [3]

[1] Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models[J]. Advances in neural information processing systems, 2022, 35: 24824-24837.

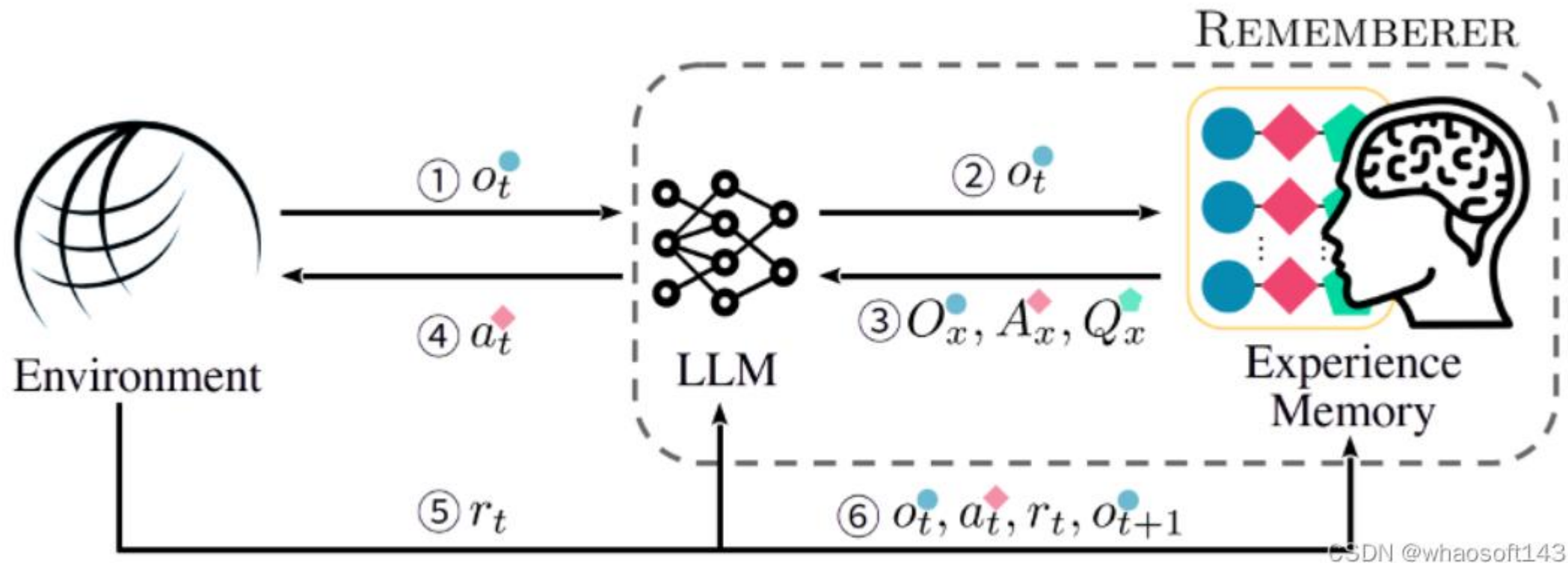
[2] Yao S, Zhao J, Yu D, et al. React: Synergizing reasoning and acting in language models[C]//International Conference on Learning Representations (ICLR). 2023.

[3] Shinn N, Cassano F, Labash B, et al. Reflexion: Language agents with verbal reinforcement learning, 2023[J]. URL <https://arxiv.org/abs/2303.11366>, 2023, 1.

局限于对单个任务轨迹的反思，所获得的见解无法有效传递给后续任务

机遇：利用记忆存储经验

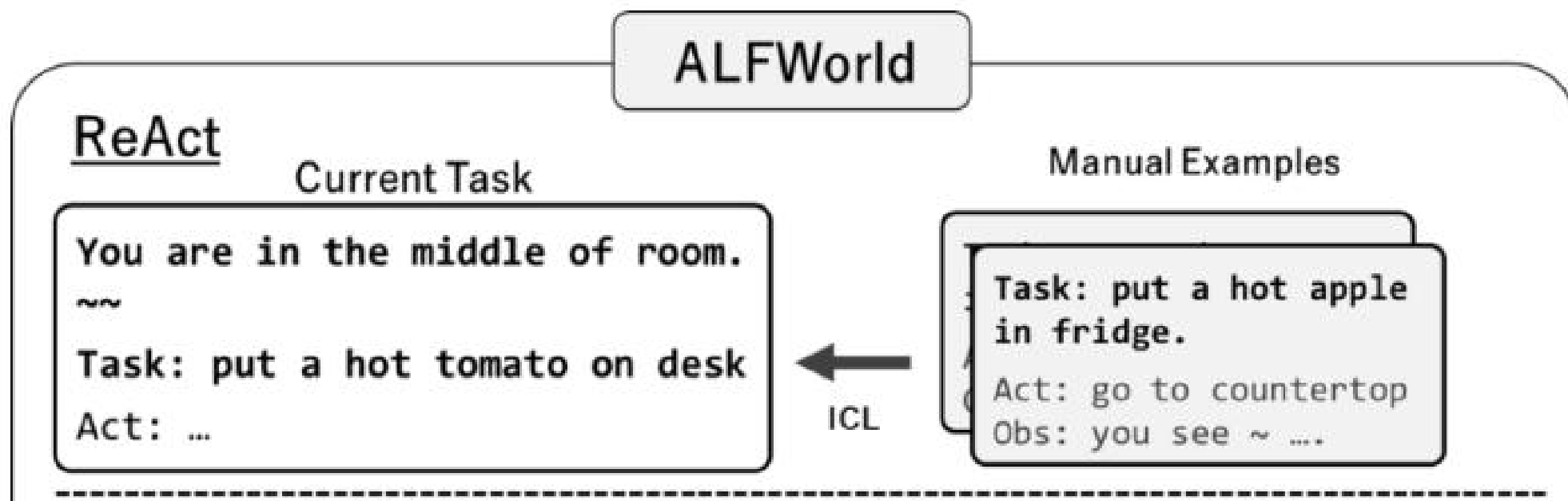
记忆：



通过记忆的形式将过往经验进行存储，并将其他任务中的成功经验迁移到当前任务

挑战：如何将过去的经验融入当前决策过程

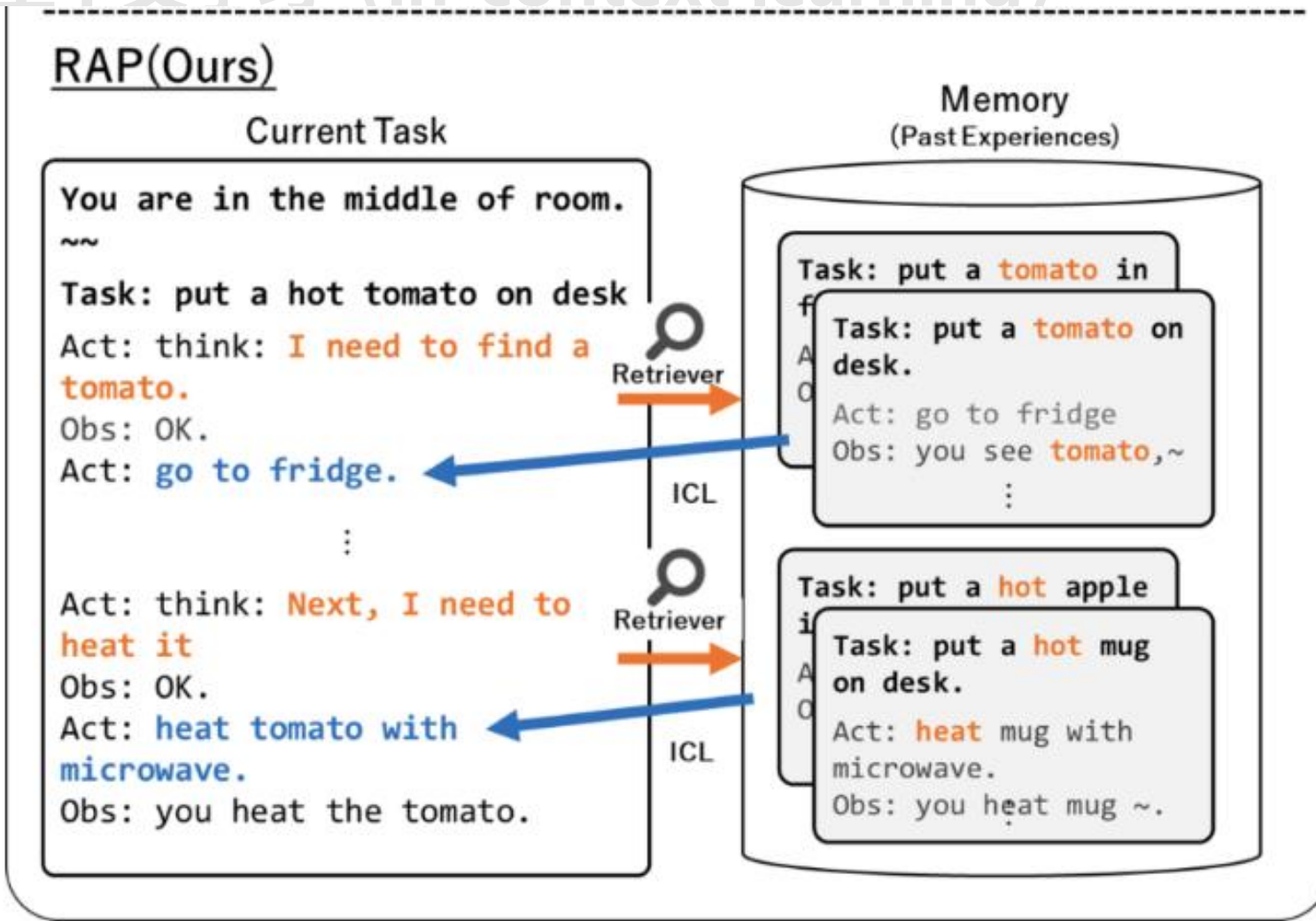
上下文学习 (in-context learning)



手动的示例学习无法覆盖所有任务情况

挑战：如何将过去的经验融入当前决策过程

上下文学习 (in-context learning)



1. 更加广泛多样的示例

2. 在不同时间点动态获取最符合当前情境的经验

务情况

提纲

1

背景

2

方法

3

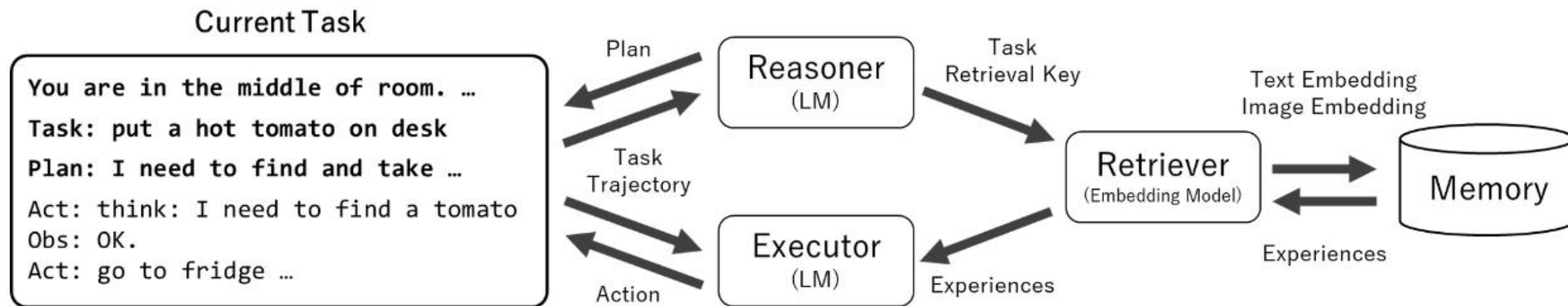
实验

4

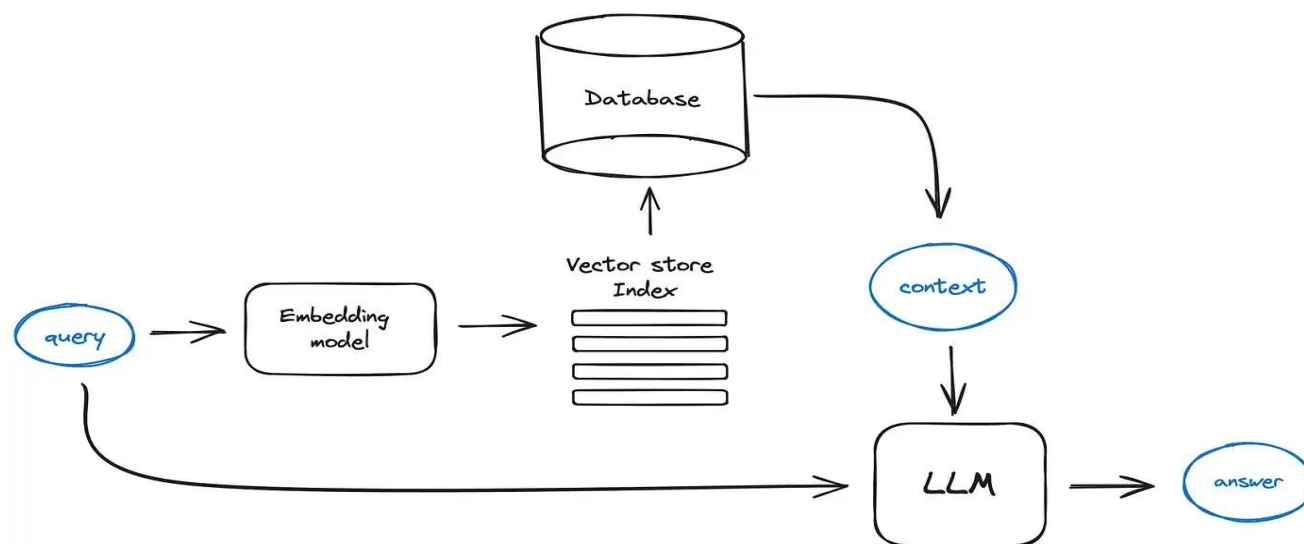
总结

方法: RAP (Retrieval-Augmented Planning)

整体思路:

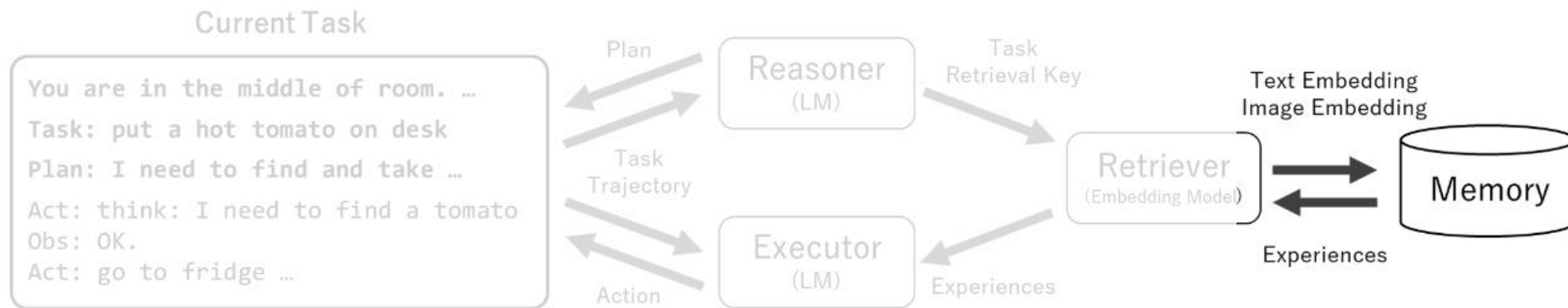


Naive RAG



方法：RAP (Retrieval-Augmented Planning)

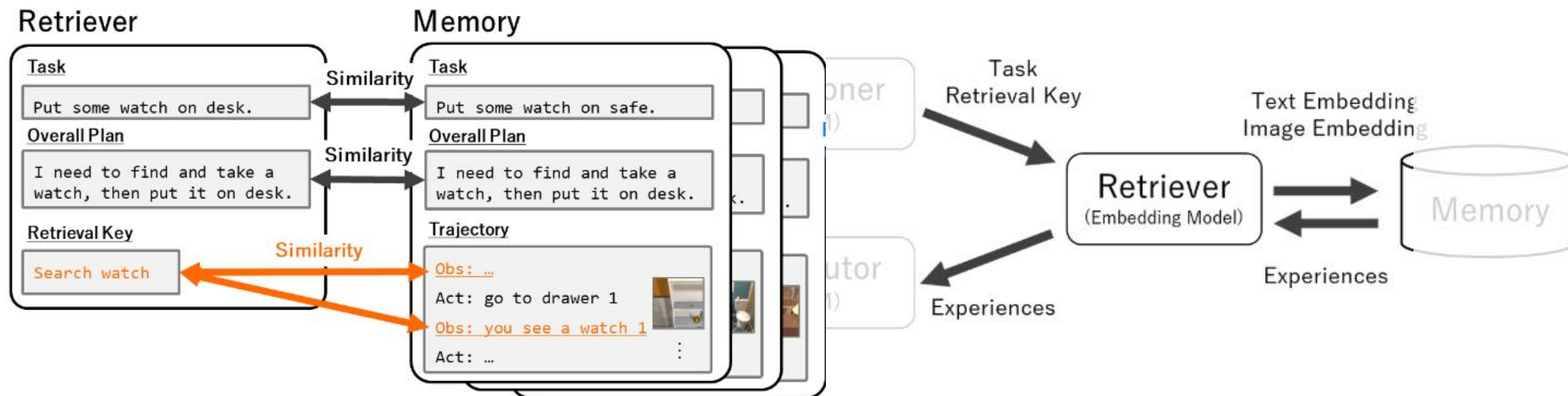
Memory:



- **记忆数据库**: 存储了先前**成功完成**的任务日志
- **详细内容**: 任务信息、整体计划、每个动作以及动作后的场景观测信息
- **文本环境**: 观测信息是环境的文本描述
- **多模态环境**: 每次动作后摄像头拍摄的环境照片
- **仅保存成功任务的理由**: 通过提供成功的记忆样本，智能体能够做出更明智的决策，并避免过去的失败

方法：RAP (Retrieval-Augmented Planning)

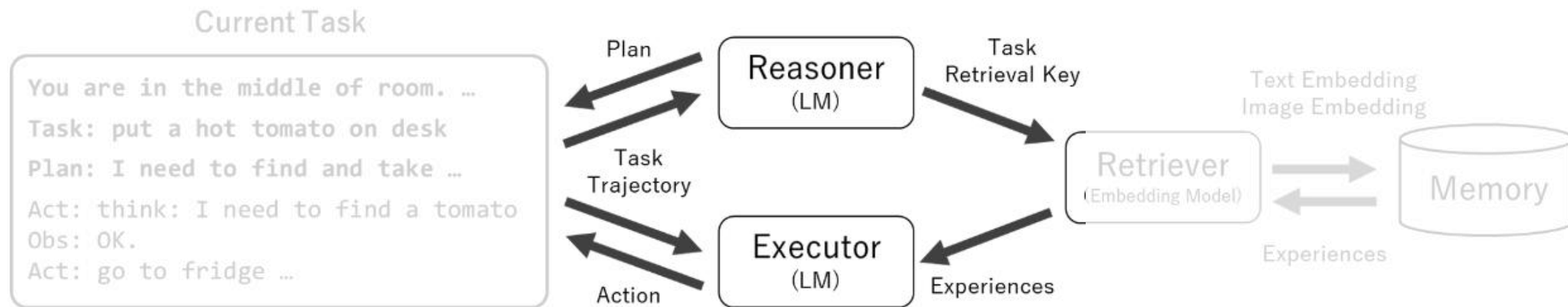
Retriever:



- **检索方式：** 向量余弦相似度匹配
- **文本检索：** 任务信息、整体计划、关键词三者相似度分数的加权排序（从高到低）
- **图片检索：** 当前环境图片与历史图片相似度分数排序（从高到低）

方法：RAP (Retrieval-Augmented Planning)

Reasoner&Executor:



Reasoner:

- **输入：** 当前环境信息、任务指令
- **输出：** 任务计划、动作计划、检索键（用于检索）

Executor:

- **输入：** 检索后的经验、任务指令、任务环境
- **输出：** 动作

Executor(Prompt)

Here is the task information.
[Task information]

Retrieve Here are examples.
[Top-k experiences from Memory]

Here is the task.
Task: put some watch on desk.
Plan: I need to find and take ~.
Act 1: think: I need to find a watch.
Obs 1: OK.
Act 2: go to drawer 1
Obs 2: drawer 1 is closed
Act 3: ...



实验：实验设置

Benchmark:

文本多步任务：ALFWorld、WebShop

■ ALFWorld:

是一个合成的文本游戏，它要求智能体在多种交互式环境中完成多步任务。

任务类型：拾取（Pick）、清洁（Clean）、加热（Heat）、冷却（Cool）、查看（Look）以及双重拾取（Pick2）。

使用134个游戏对智能体进行评估。

■ WebShop:

一个模拟网络购物的应用，涵盖了大量结构化与非结构化文本，智能体需要根据用户指令选择合适的商品进行购买。

使用100 条用户指令对智能体进行评估。

实验：实验设置

Benchmark:

多模态具身智能环境：Franka Kitchen、Meta-World

- 具身智能体：

基于两种 VLM 构建了具身智能体：LLaVA 和 CogView，比较了基础模型与加入 RAP 检索系统后的性能。

- Franka Kitchen：

包含复合任务，例如物体摆放与烹饪准备。

- Meta-World:

提供了 50 种精细机器人技能任务，侧重于操作控制。

- 策略网络训练：

使用25条数据训练了一个策略网络，使其可以将vlm的高层规划映射成动作序列

每个benchmark中跑5个任务，每个任务跑50次，计算成功率。

评估: ALFWorld

- RAP: 在任务执行过程中进行成功任务经验的记忆
- RAPtrain: 先在1000个任务中进行训练, 有了一定的知识储备

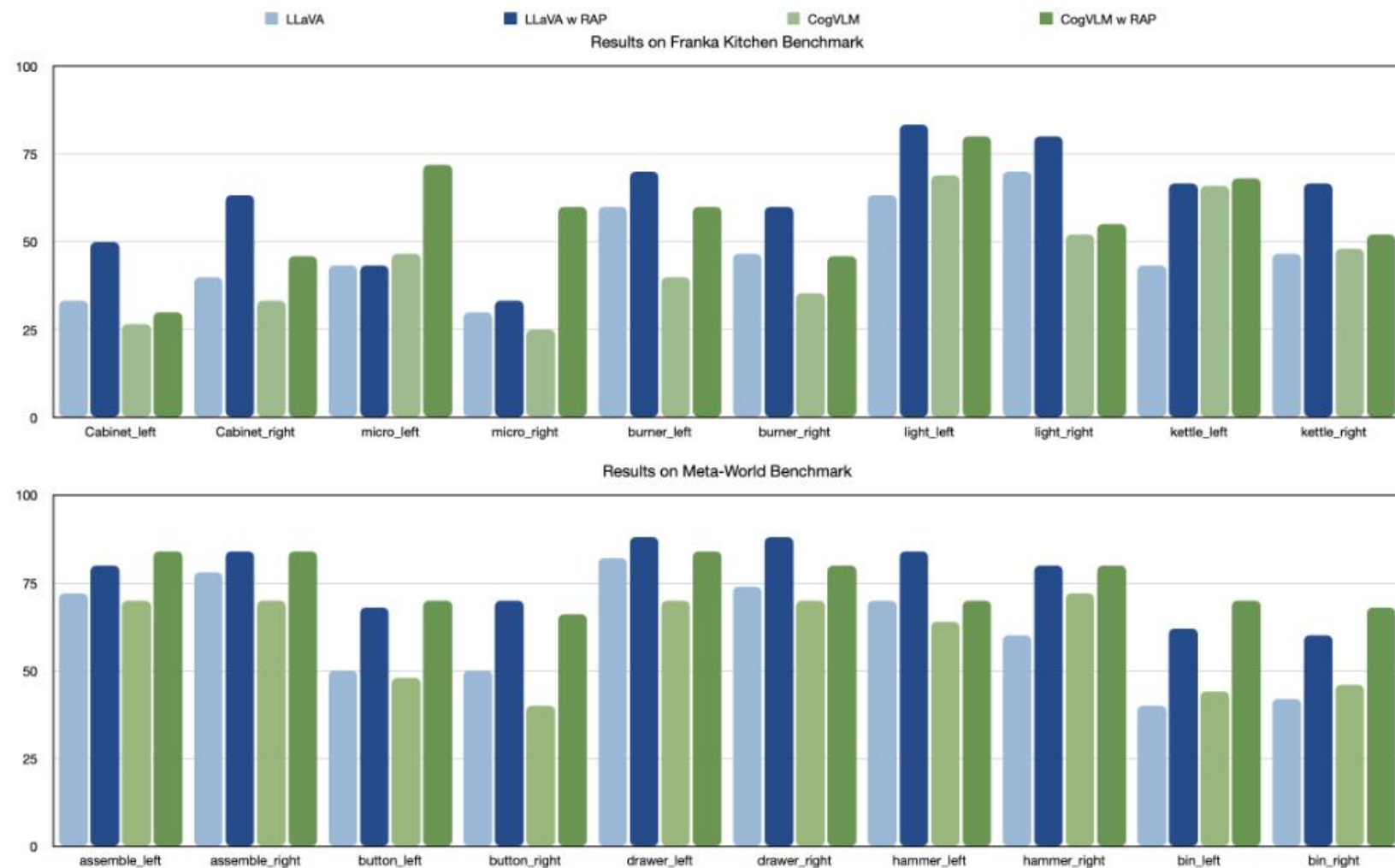
Table 1. ALFWorld task-specific success rate(%).

Method($d_{max}=3$)	Model	Pick	Clean	Heat	Cool	Look	Pick2	All
Act	GPT-3.5	66.7	51.6	73.9	61.9	38.9	17.6	53.7
ReAct	GPT-3.5	50.0	41.9	73.9	66.7	55.6	23.5	52.2
Reflexion	GPT-3.5	75.0	77.4	65.2	76.2	83.3	<u>70.6</u>	74.6
ADaPT*	GPT-3.5	87.5	80.6	60.8	76.2	61.1	52.9	71.6
RAP(Ours)	GPT-3.5	95.8	<u>87.1</u>	<u>78.3</u>	90.5	<u>88.9</u>	<u>70.6</u>	<u>85.8</u>
RAP _{train} (Ours)	GPT-3.5	95.8	100.0	82.6	<u>85.7</u>	100.0	76.5	91.0
ReAct	GPT-4	83.3	71.0	95.7	81.0	100.0	94.1	85.8
RAP(Ours)	GPT-4	95.8	90.3	100.0	95.2	100.0	88.2	94.8
ReAct	Llama2-13b	29.2	41.9	34.8	52.4	38.9	17.6	36.6
RAP(Ours)	Llama2-13b	62.5	61.3	56.5	61.9	44.4	17.6	53.0

* We use the performance reported by (Prasad et al., 2023)



实验：具身智能场景



Method	Franka Kitchen	Meta World
LLaVA	43.4	65.4
LLaVA with RAP	61.6	79.2
CogVLM	44.2	61.5
CogVLM with RAP	56.9	68.8

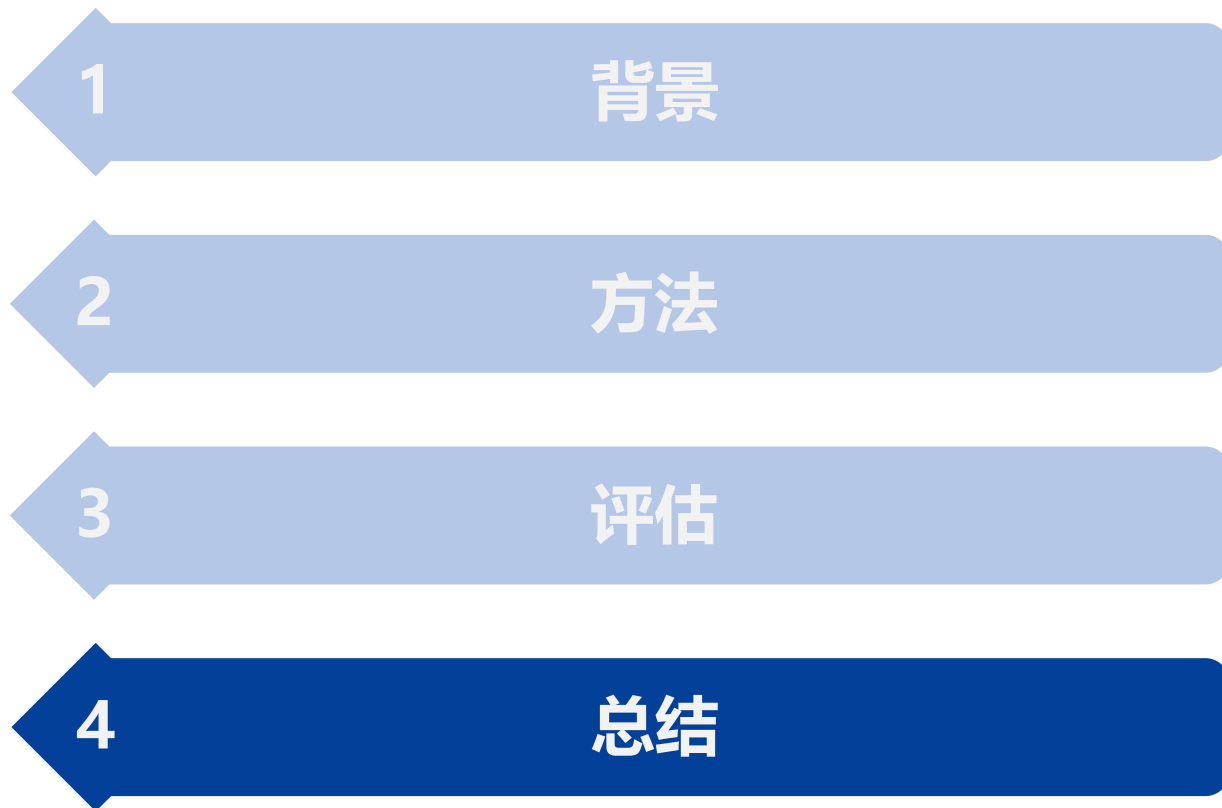
- 横轴：5个任务类型和2个不同视角摄像头的组合（共10组）
- 效果：加上了RAP的效果每组表现都优于baseline

实验：消融实验

Model	Model _{Memory}	Success Rate
GPT-3.5	-	44.0
GPT-3.5	GPT-3.5	63.4
Llama2-13b	-	20.9
Llama2-13b	GPT-3.5	27.6

- 验证解决任务的经验独立于具体模型
- 作者先对基础模型为GPT-3.5的agent在ALFWorld中跑了1000次训练，总结了一定的经验
- 将基础模型替换成Llama2-13b
- 替换后的agent与没有记忆的智能体表现仍有提升

提纲



Conclusion

- 作者提出了一个 RAP 框架，用于增强 LLM 智能体的规划能力。
- 通过存储过往经验并基于与当前情境的相似性进行智能检索，丰富了决策过程。
- RAP 不仅适用于文本环境，还能应用于多模态具身任务。

Thinking

➤ 能否提高

- 文本记忆和多图片记忆目前还是在不同的任务中使用，并没有实现两个模态的记忆融合使用，相互弥补。其实在具身智能场景中，是可以同时保存文本和图片记忆的（例如文本记忆存储vlm面对不同任务时的规划经验，图片可以存储任务完成时的状态图片，这样方便vlm从图片模态理解任务完成时的样子）

➤ 泛化：

- 文章的这个记忆思想泛化性目前已经很多人考虑到了，我自己的工作是在视频问答场景，也有在gui上和代码生成等。

➤ 用到我们的idea中：

- 方法上：在具体的实现上，我们的工作和这篇工作的实现方式有些类似，但是在具体的检索经验模块，作者使用了不同的query来加权的方式排序可以考虑借鉴。
- 多模态：目前我们的视频问答场景，由于做的是agent场景，最终的回答还是会回到文本模态的输出，所以进行多模态的记忆存储确实不太好讲通故事；文中的具身智能场景其实是一个比较好的需要多模态记忆存储的场景，在我们的智能体工具中添加一个vlm，专门用来做具身智能任务，面对这部分任务时就可以使用多模态记忆



東南大學
SOUTHEAST UNIVERSITY

恳请各位老师与同学批评指正！