

# 目录

---

- ChatGPT之前的视觉语言预训练
- 大视觉语言模型的架构和训练
- 大视觉语言模型的评测
- 大视觉语言模型的能力扩充
- **大语言模型支撑的具身智能（视觉导航）**

# 具身智能

目前，通过具身智能来解决任务的研究要求 AI 具备以下的能力

- 观测 (usually in an egocentric view)
- 交流 (via texts or audios)
- 推理 (understand surroundings and plan)
- 行动 (through motor controls or high-level actions).



## 视觉语言导航 (Vision-and-Language Navigation)

构建真实机器人的一个好的原型

### Instruction

Turn right and head towards the *kitchen*. Then turn left, pass a *table* and enter the *hallway*. Walk down the hallway and turn into the *entry way* to your right *without doors*. Stop in front of the *toilet*.



Initial Position



Target Position



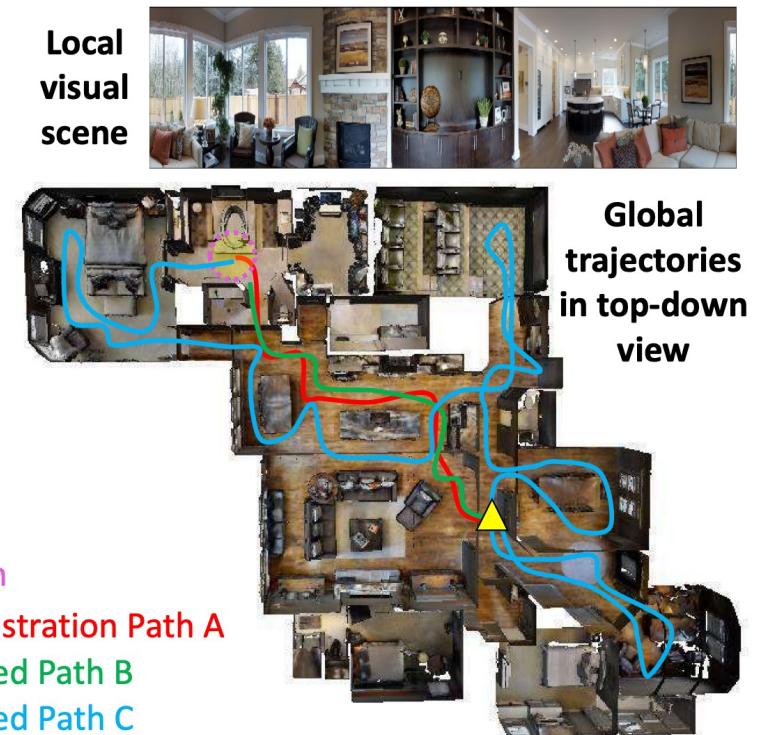
Demonstration Path A



Executed Path B



Executed Path C



一个典型的VLN样例

Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments, 2018

Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation, 2019.

# 基于大语言模型的机器视觉导航

- LLM展示了强大的文本理解和生成能力。
- LLM具有良好的泛化能力，展示出卓越的零样本性能。

Title: Star's Tux Promise Draws Megyn Kelly's Sarcasm  
Subtitle: Joaquin Phoenix pledged to not change for each awards event  
Article: A year ago, Joaquin Phoenix made headlines when he appeared on the red carpet at the Golden Globes wearing a tuxedo with a paper bag over his head that read, "I am a shape-shifter. I can't change the world. I can only change myself." It was a promise to not change to fit into the Hollywood mold: "I think that's a really special thing, to not change yourself. I think it's a really special thing to say, 'This is what's inside of me, I'm proud of it, and I'm not going to be ashamed because of the way that someone else thinks I should be.'" Now, it's the Oscars, and Phoenix is at it again. But this time, his publicist is saying he'll be wearing a tux no matter what.  
Megyn Kelly was not impressed, and she let him have it on The Tonight Show. "You know, I feel like, I feel like you could have worn the tux," she says. "But you're saying you're a shape-shifter. I don't know if you can change your tux, but you can change your mind. You can change your mind. You can change your mind." Phoenix says he did, but it didn't stick. "I was like, 'Okay, I'm going to wear a tuxedo to this thing.' And then I thought, 'I don't want to wear a tuxedo to this thing.'" Kelly goes on to encourage him to change his mind again, but Phoenix says it's too late: "I'm committed to wearing this."

Figure 3.15: The GPT-3 generated news article that humans found the easiest to distinguish from a human written article (accuracy: 61%).

## Sample question using prompt template

The following are multiple choice questions (with answers) about medical knowledge.  
\*\*Question\*\*: A 40-year-old woman has had hypercalcemia for 1 year and recently passed a renal calculus. Serum parathyroid hormone and calcium concentrations are increased, and serum phosphate concentration is decreased. Parathyroid hormone most likely causes an increase in the serum calcium concentration by which of the following mechanisms?

- (A) Decreased degradation of 25-hydroxycholecalciferol
  - (B) Direct action on intestine to increase calcium absorption
  - (C) Direct action on intestine to increase magnesium absorption
  - (D) Increased synthesis of 25-hydroxycholecalciferol
  - (E) Inhibition of calcitonin production
  - (F) Stimulation of 1,25-dihydroxycholecalciferol production
- \*\*Answer:\*\*(F)

## Zero-shot medical analysis (GPT-4)

→ 使用大模型来构建更好的VLN智能体！

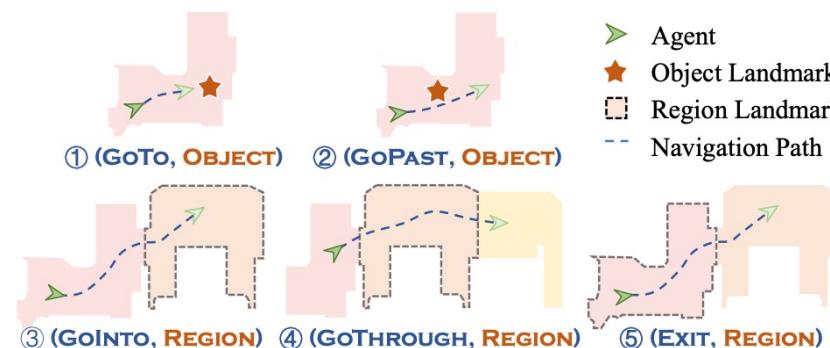
Language models are few-shot learners, 2020

A comprehensive capability analysis of gpt-3 and gpt-3.5 series models, 2023

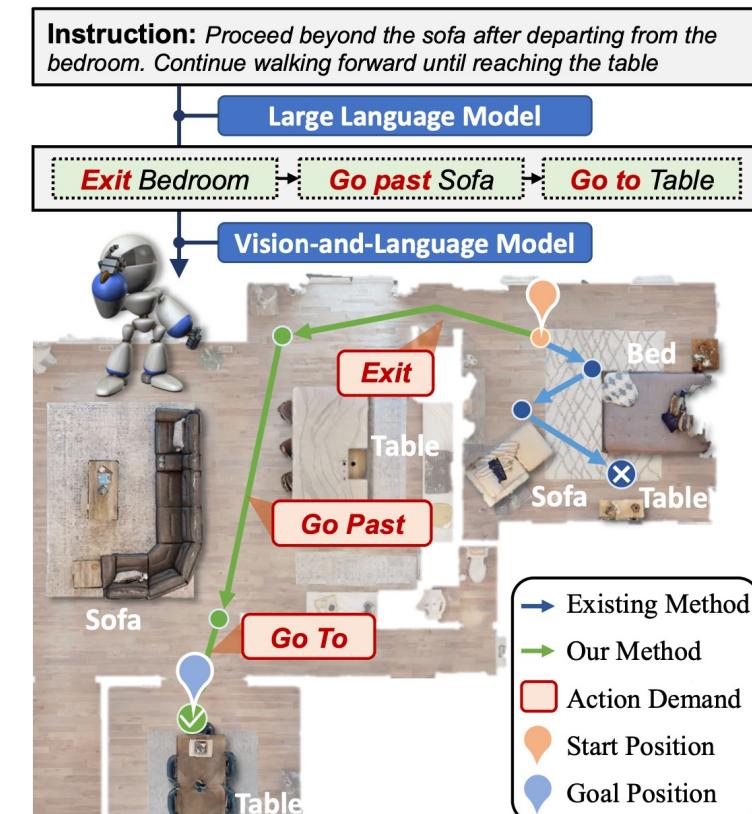
# A<sup>2</sup> Nav: 大语言模型作为指令解析器

利用LLM优越的文本理解和推察能力将指令分解成若干个子任务。

- 每个子任务包含
  - 一个地标
  - 一个与该地标相关的特定动作



- 动作感知的导航策略：
  - 零样本物体导航器 (Zero-Shot Object Navigator, ZSON)
  - 每个ZSON负责一个特定类型的子任务

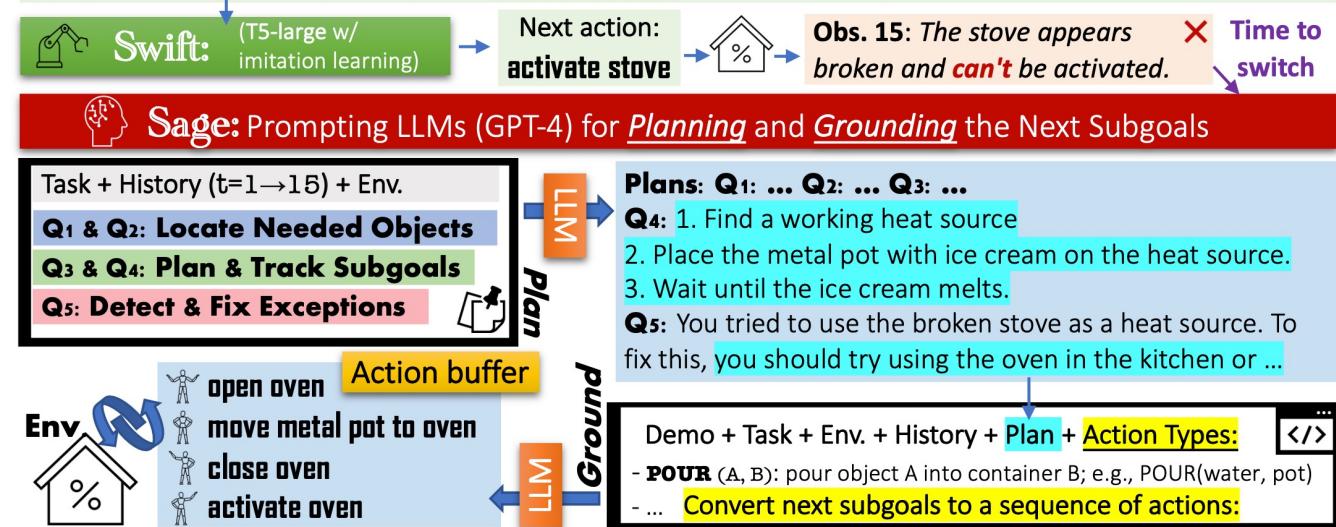


# SWIFTSAGE：引入大语言模型进行决策反思

让LLM对失败进行反思，然后进行纠错（给出下一步的目标）

- SWIFT 模块 (T-5 Large):
  - 通过模仿学习进行简单决策。
- SAGE 模块 (GPT-4):
  - 纠正SWIFT模块的错误决策。
- SAGE的两阶段策略:
  - 规划: 通过回答问题对历史和任务执行情况进行总结
  - 匹配: 促使LLM专注于下一步目标，并将其转换成一系列行动。

Task: Your task is to melt ice cream.; Time: 14; Score: 35; Action history: <extra\_id\_0> Action 5 (+5): open fridge --> You opened fridge. In it, you see an ice cream... [...] <extra\_id\_9> Action 14 (+0): move metal pot to stove --> You move the metal pot to the stove. </s> Current environment: This room is kitchen. You see: a fridge (closed) | a sink | an oven (closed, turned off) | a stove (turned off; on it: a metal pot containing ice cream ...). | [...] | Inventory: an orange, ... </s> Visited: workshop, hallway, kitchen </s> What should be the next action?



- 1) There are five consecutive time steps with zero reward ( $\sum_{i=t-5}^{t-1} R_i = 0$ ).
- 2) The SWIFT's prediction for the next action ( $A'_t$ ) is invalid in the current environment.
- 3)  $A'_t$  can result in a critical decision, such as giving the final answer for the experiment.
- 4) The observation of  $A'_t$  suggests that an exception is encountered.

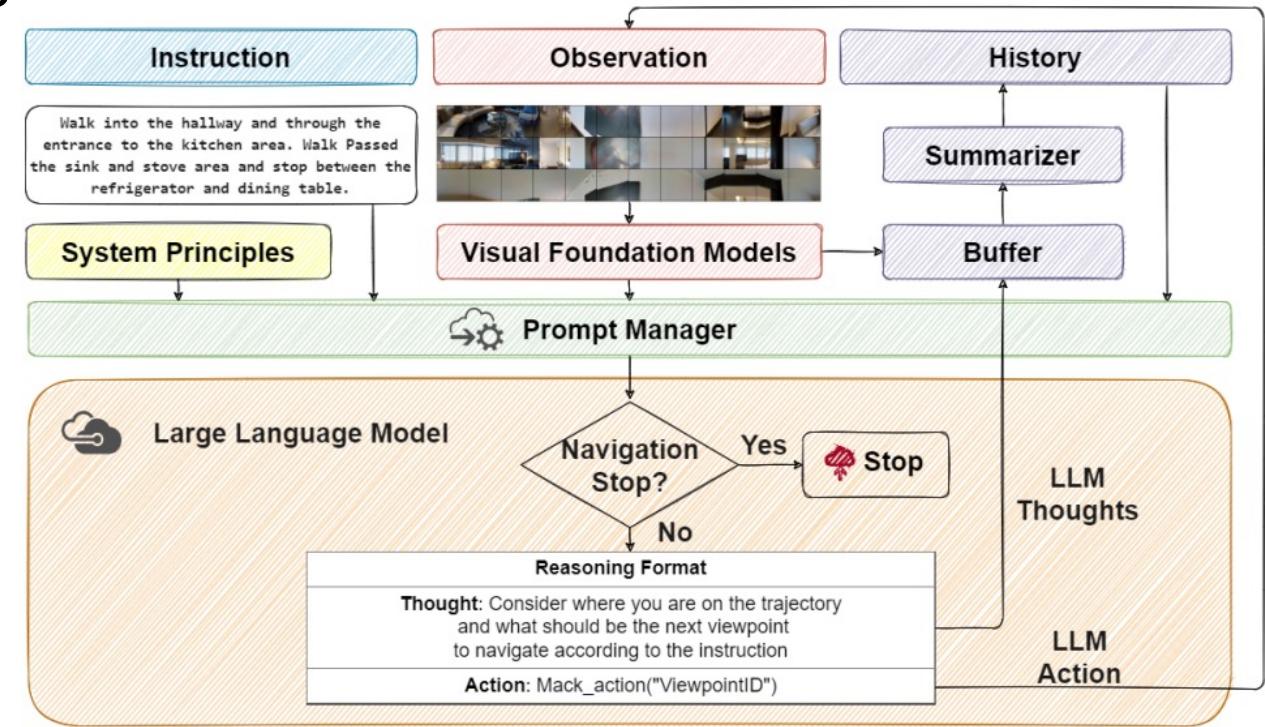
何时从SWIFT模块切换到SAGE模块？

# NavGPT：大语言模型作为动作决策模块

- 一个零样本视觉语言导航框架，使用大语言模型作为动作决策模块。
- 以文本的形式表示当前的视觉观测和过去的历史轨迹
- 在决策时，使用 chain-of-thought，融合思维（推理）和行动（决策）

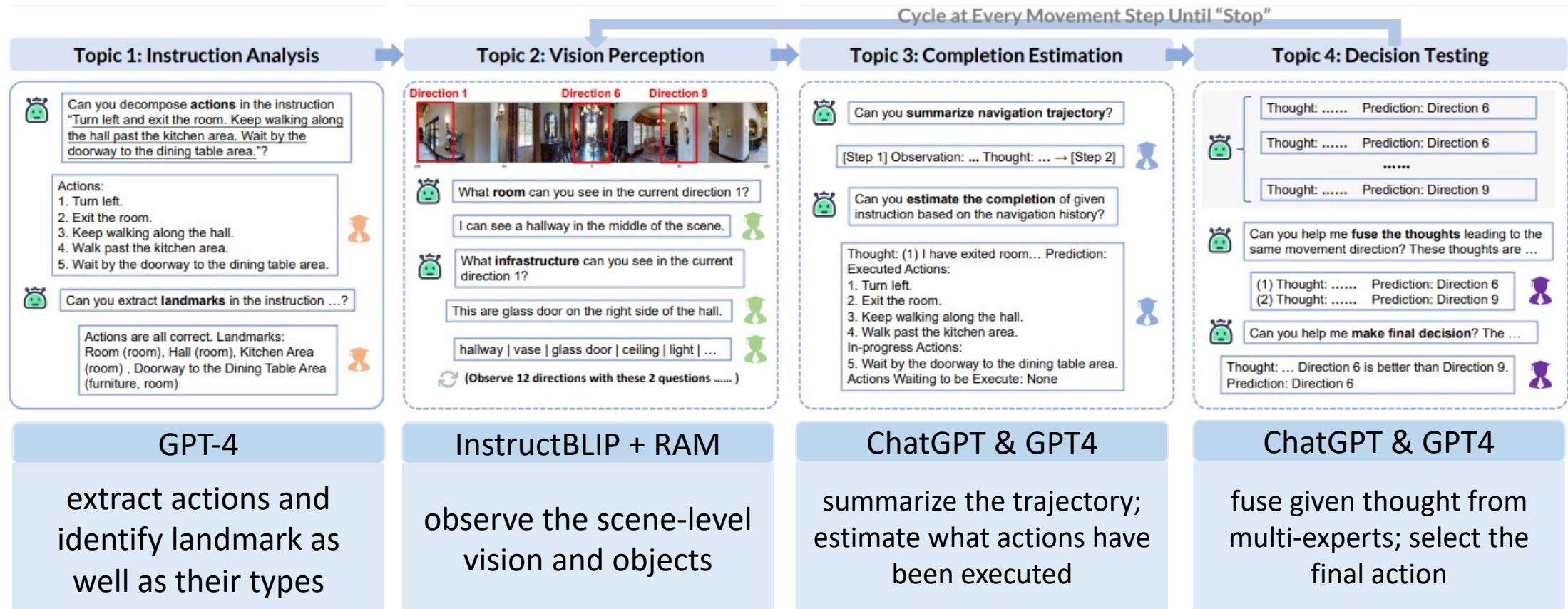
模型架构

- **Visual Foundation Models:**
  - 视觉描述: BLIP-2
  - 物体检测: Faster-RCNN
  - 方位归纳: GPT-3.5
- **History Summarizer:** GPT-3.5
- **LLM:** GPT-4



# DiscussNav : 大语言模型作为领域专家

- 将导航过程转换为基于多位专家讨论的决策过程（文本形式）。



# DiscussNav : 大语言模型作为领域专家

RESULTS ON R2R VALIDATION UNSEEN SPLIT.

Training Schema	Method	TL	NE↓	OSR↑	SR↑	SPL↑
Train Only	Seq2Seq [1]	8.39	7.81	28	21	-
	Speaker Follower [5]	-	6.62	45	35	-
	EnvDrop [6]	10.70	5.22	-	52	48
Pretrain + Finetune	PREVALENT [7]	10.19	4.71	-	58	53
	VLN $\odot$ BERT [8]	12.01	3.93	69	63	57
	HAMT [9]	11.46	2.29	73	66	61
	DuET [10]	13.94	3.31	81	72	60
Zero-shot	DuET (Init. LXMERT [51])	22.03	9.74	7	1	0
	NavGPT [11]	11.45	6.46	42	34	29
	DiscussNav (Ours)	9.69	5.32	61	43	40

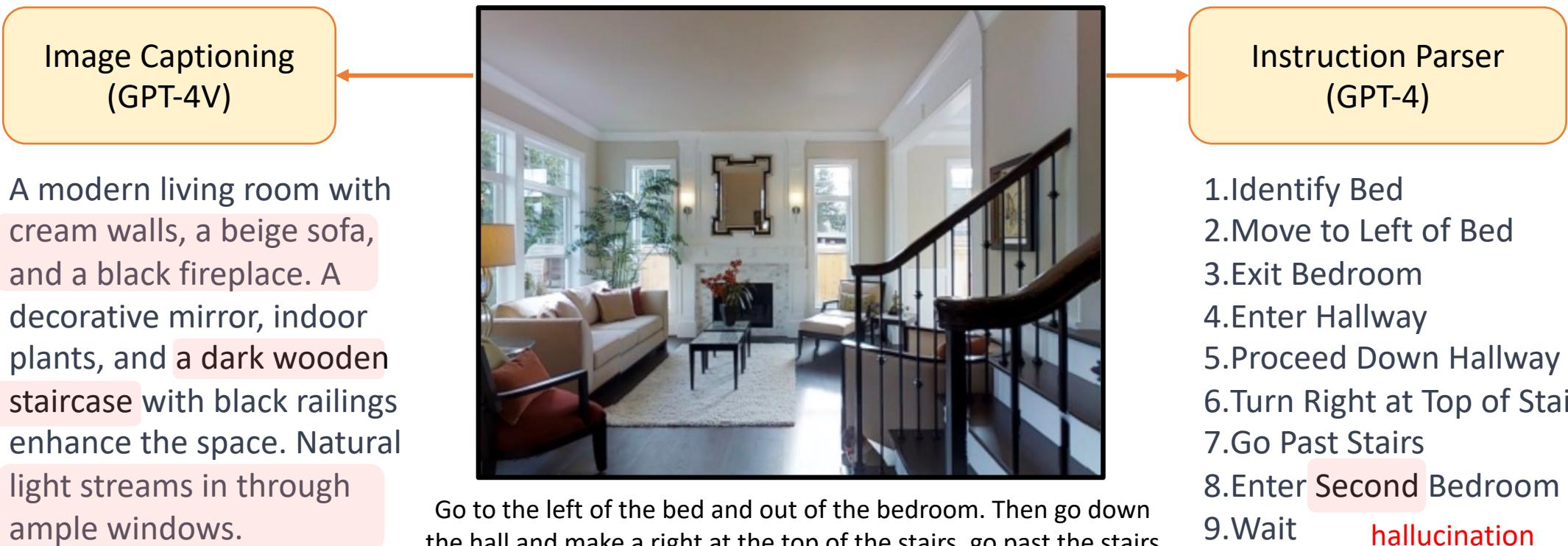
REAL ROBOT EXPERIMENTS IN INDOOR SCENE.

Training Schema	Method	SR↑
Pretrain + Finetune	DuET [10]	0
Zero-shot	NavGPT [11]	10
	DiscussNav (Ours)	25

- 在zero-shot 设定下， DiscussNav相比其他方法的性能更好
- 但未能超过专有的小模型
  - Sim-2-Real Transfer实验：结果表明 NavGPT 单轮 CoT 的能力受限，
  - 而 DiscussNav 在理解指令、观察细粒度地标和做出决策方面表现更好

# 使用大语言模型导致信息丢失&误差累积

- 将图像表示为文本将会丢失许多细节。



- 如果任何模块出错，错误将会累积并传播到策略决策模块中。

# 如何在具身智能任务中使用多模态大模型？

- 使用多模态模型 (LVLMs) 来理解视觉和语言已经是常用的做法。
- 然而，LVLMs并非为具身人工智能任务而设计。

➤ Image Text Retrieval

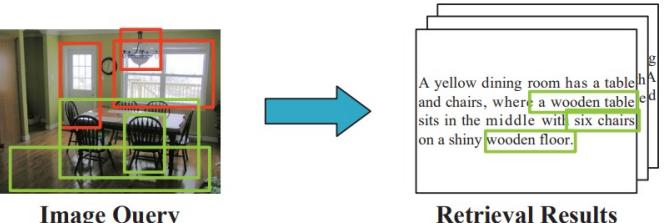


Image Query

Retrieval Results

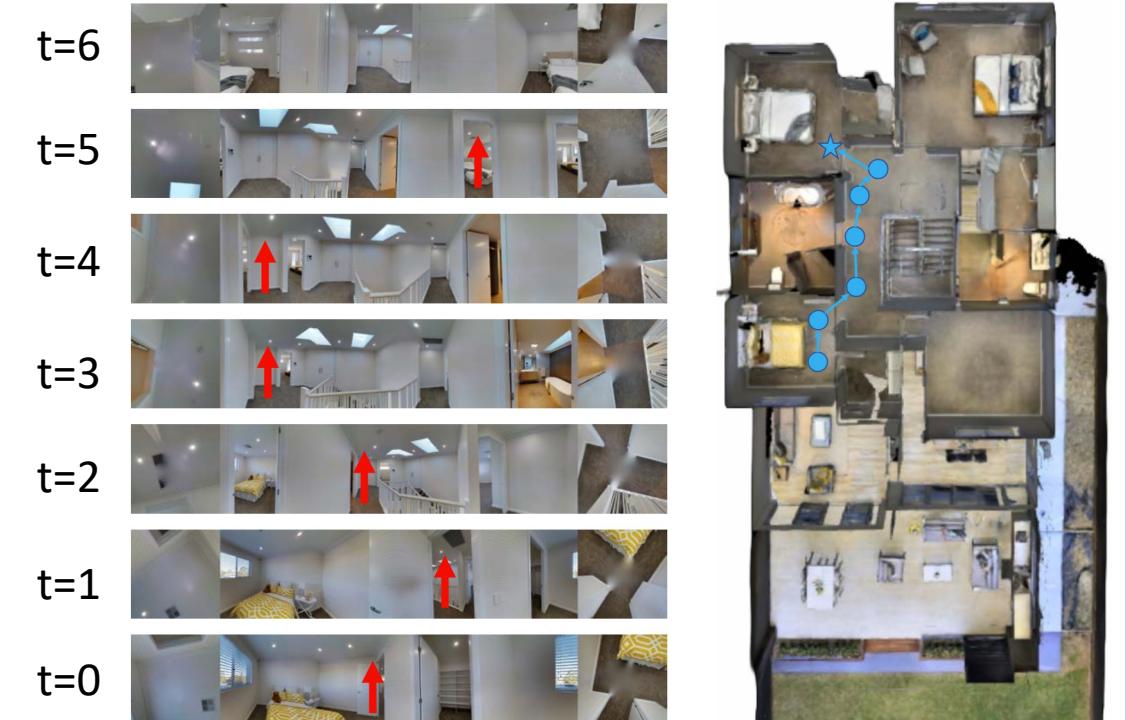
➤ Vision-based Text Generation



A person riding a motorcycle on a dirt road.

Two dogs play in the grass.

单图的、单步的推理、理解、生成任务



t=6

t=5

t=4

t=3

t=2

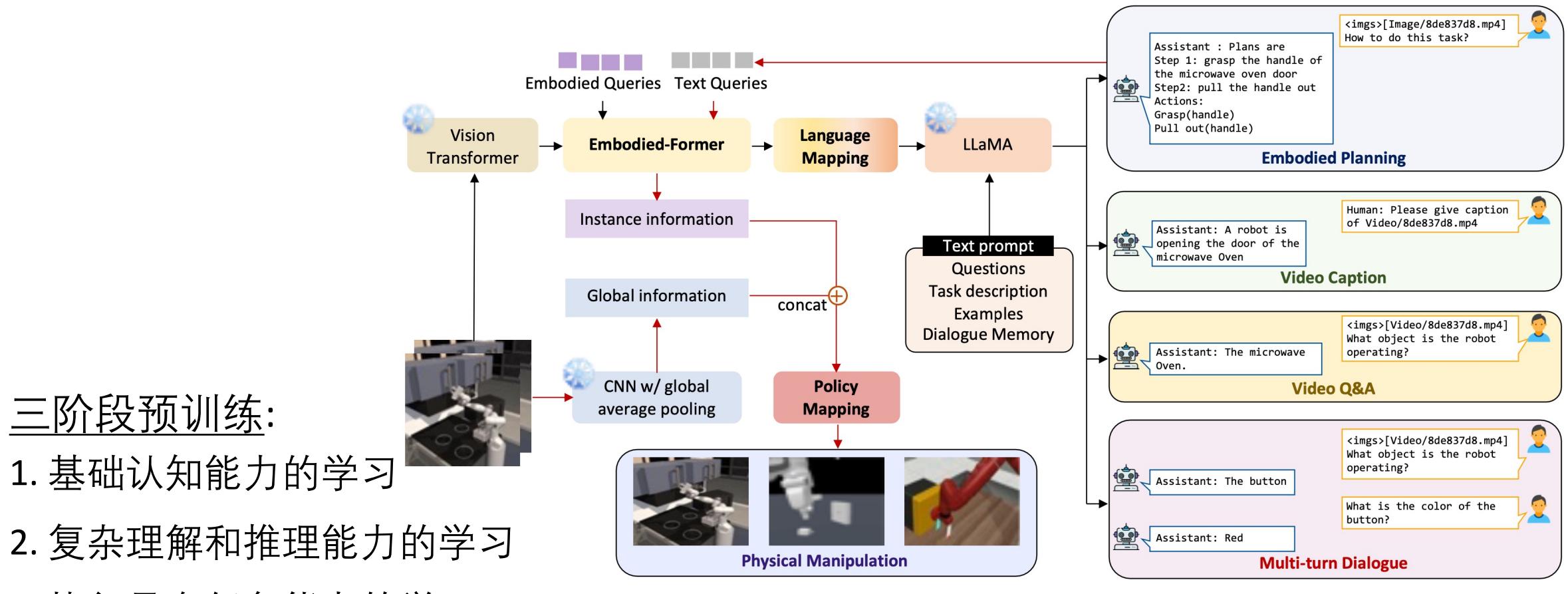
t=1

t=0

多图的、多步的推理、时间顺序、空间方位关系

# EmbodiedGPT: 采用大规模预训练

对LVLMs进行预训练，以构建一个**端到端的具身智能基础模型**。



三阶段预训练：

1. 基础认知能力的学习
2. 复杂理解和推理能力的学习
3. 执行具身任务能力的学习

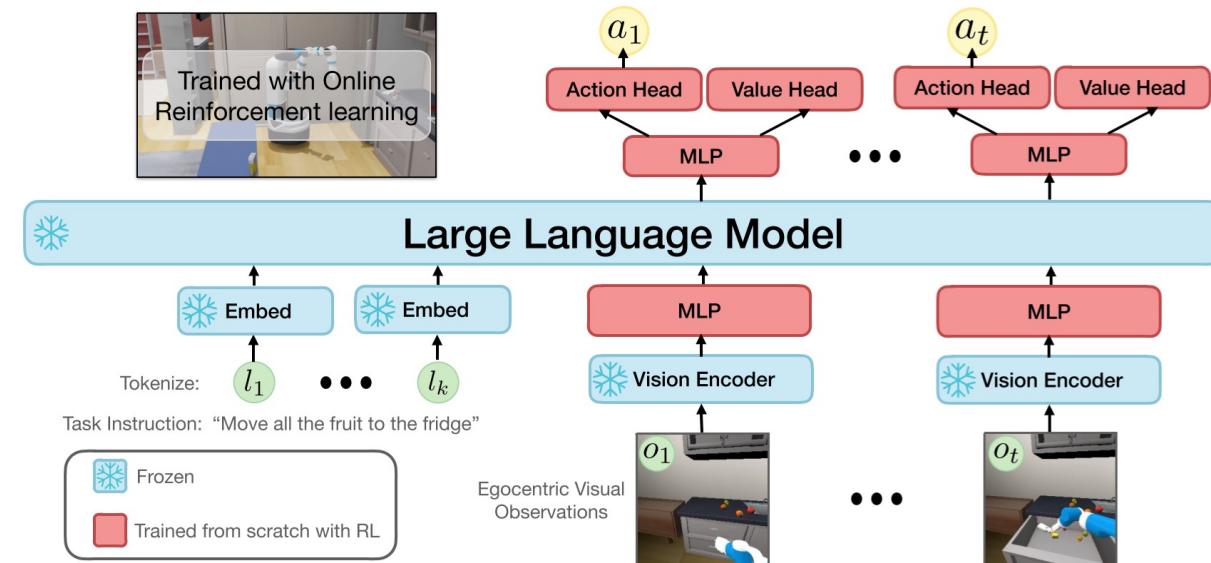
# LLaRP：通过强化学习训练 LVLMs

- 通过在线强化学习将LVLMs训练为适配多种具身任务的视觉-语言策略

## 模型架构

- Visual Encoder:** VC-1
- LLM:** LLaMA-7B V1
- Connection:** Linear
- Action Decoder:** MLP with ReLU

强化学习采用DD-PPO算法。



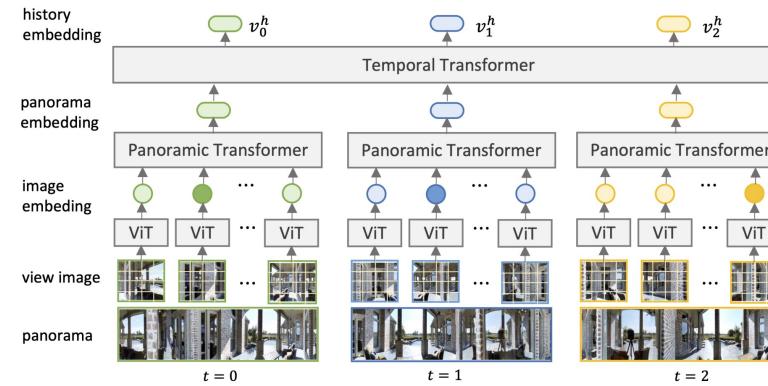
模型在一个新的基准数据集 Language Rearrangement 上进行训练，该数据集包含 150k 个训练任务和 1k 个测试任务，例如复杂的操控、导航和探索。

由于数据集部署在Habitat 2.0模拟器中，智能体动态的与环境进行交互，从而进行在线强化学习。

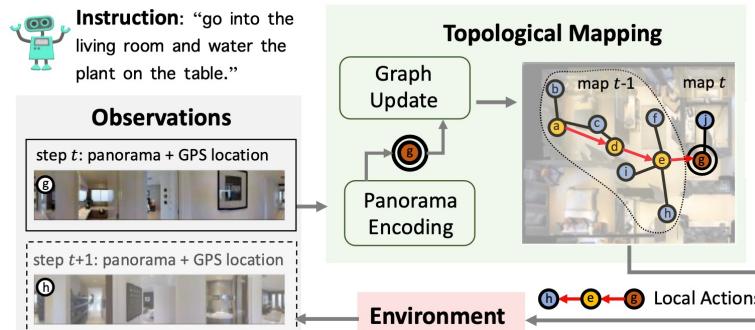
# 挑战：对长期依赖性的建模

导航是一个顺序决策问题，因此历史信息至关重要。

Transformer-based Agent



History as sequential observation features



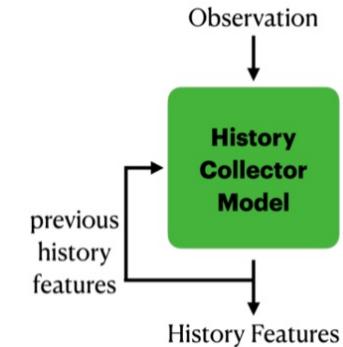
History as Topological graph

LLM-based Agent

Instruction: Walk into the house, and go through the kitchen. Leave the kitchen through the door on the right of the large black bookcase. Go down the stairs. Stop on the third stair from the top.  
Init observation: ...With hardwood floors and a dining table...  
Thought: First, I need to go through the kitchen. The Kitchen is in the front left direction, so I should move to the viewpoint 'e8aaaf779b9c41eca16156cfddcedd2b'.  
Action: action\_maker("e8aaaf779b9c41eca16156cfddcedd2b")  
Observation: ...A dining room and kitchen with hardwood floors.  
Thought: I am now in the kitchen. Next, I need to find the door on the right of the large black bookcase. I should move to the viewpoint 'e685b29863794f00b09c23e1536412d8' to get a better view of the kitchen and find the door.  
Action: action\_maker("e685b29863794f00b09c23e1536412d8")  
Observation: ...A kitchen with white cabinets, varying flooring and counter top materials, and white ceiling lights.  
Thought: I can see the kitchen clearly now, but I still can't find the large black bookcase and the door on its right. I should move to the viewpoint '5b01ffa96874e60a9dec1fbac7e4ee' to explore the kitchen further.  
Action: action\_maker("5b01ffa96874e60a9dec1fbac7e4ee")  
Observation: ...A kitchen with a marble top island, counter top, and fireplace.  
Thought: I found the large black bookcase in the rear left direction, and the door on its right. I should move to the viewpoint 'e62661ec94d04ada7fee9a81d429d2b' to get closer to the door.  
Action: action\_maker("e62661ec94d04ada7fee9a81d429d2b")  
Observation: ...A hallway with hardwood floors and a door leads to a room with a bookcase and a ceiling fan, while another hallway features a bookcase and a bookshelf.  
Thought: I am now closer to the door on the right of the large black bookcase. I should move to the viewpoint 'ea374bef649549d8a12bb0d6d67ed206e' to go through the door and reach the stairs.  
Action: action\_maker("ea374bef649549d8a12bb0d6d67ed206e")  
Observation: ...A hallway with hardwood floors, a staircase, a painting.  
Thought: I am now in the hallway with the staircase. I need to go down the stairs and stop on the third stair from the top. I should move to the viewpoint 'a8311575bf3d485a81af8350aa25bdb' to start going down the stairs.  
Action: action\_maker("a8311575bf3d485a81af8350aa25bdb")  
Observation: ...A view of a staircase with a black railing, a hallway with a white railing and art, and a bathroom with a white ceiling and a light fixture.  
Thought: I am now on the stairs. I should move to the viewpoint '998dfb266ae84564b136c862fae09dd5' to go down the stairs and stop on the third ...

History as Long Context

LVLM-based Agent



History as summarized features

Any better methods to modelling the long-term dependency of this task ?

# 挑战：目前缺乏对模型空间理解能力的探索

Model	Space-based Perception		Spatial Relation Judgment				Avg.	
	CLEVR		VSR		MP3D-Spatial			
	Acc	Instability	Acc	Instability	Acc	Instability	Acc	Instability
<b>Generation Evaluation</b>								
BLIP-2 <sub>F</sub>	42.67	0.28	46.95	0.21	39.87	0.32	43.16	0.27
InstructBLIP <sub>F</sub>	44.84	0.39	52.37	0.25	41.01	0.37	46.07	0.34
InstructBLIP <sub>V</sub>	<b>46.32</b>	0.51	52.37	0.49	34.59	0.50	44.43	0.50
LLaVA <sub>V</sub>	19.01	1.24	40.00	0.88	27.19	1.13	28.73	1.08
LLaVA <sub>L<sub>2</sub></sub>	36.52	0.61	<b>52.54</b>	0.21	34.67	0.64	41.24	0.49
MiniGPT4	33.74	0.84	36.44	0.81	33.62	0.84	34.60	0.83
mPLUG-Owl	27.48	1.01	28.81	0.97	24.23	1.04	26.84	1.01
PandaGPT	29.65	0.90	35.76	0.86	34.50	0.80	33.30	0.85
IB-LLM	31.45	0.96	40.00	0.94	35.22	0.83	35.56	0.91
LA-V2	21.39	1.05	23.05	1.04	27.06	1.01	23.83	1.03
mmGPT	22.26	1.13	28.98	1.01	29.30	0.98	26.85	1.04
Shikra	23.82	0.77	46.27	0.60	29.77	0.84	33.29	0.74
Lynx	40.58	0.68	45.76	0.66	34.38	0.78	40.24	0.71
Cheetor <sub>V</sub>	24.72	1.03	35.76	0.77	31.21	0.88	30.56	0.89
Cheetor <sub>L<sub>2</sub></sub>	29.10	0.77	40.85	0.69	33.53	0.73	34.49	0.73
BLIVA	30.64	0.85	35.25	0.61	34.12	0.59	33.34	0.68

- LViLM对空间信息的理解较差 (<50%)。
- 很少有模型报告它们在与空间建模相关的数据集上的表现。
- 缺乏使模型掌握空间理解能力的相关探索。



ChatGPT-4 :

在图片的左侧，有一个高大的不锈钢冰箱。冰箱上方是橱柜。还有一个内置烤箱，嵌入橱柜中。台面似乎是浅色的，与浅色调的地砖相得益彰。此外，还有一个红色圆柱形物体，可能是一个香料研磨器或厨房用具架。在它旁边，台面上放着一份可能是报纸或杂志的印刷品。