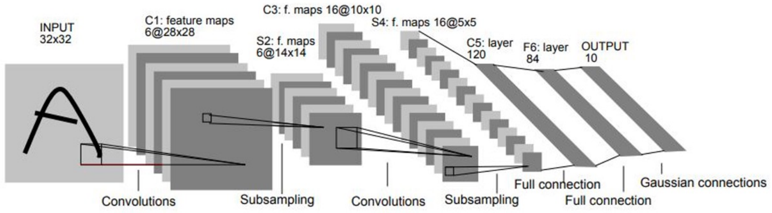
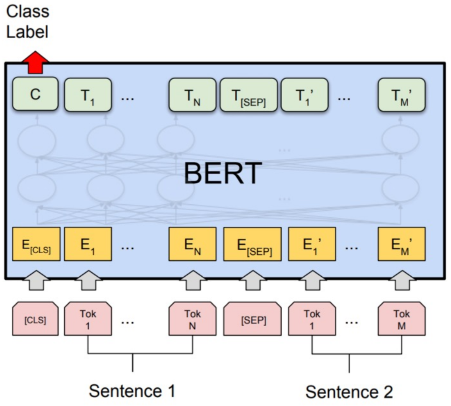
北京通用人工智能研究院
Beijing Institute for General Artificial Intelligence

Generalist Embodied AI in an Open World

Xiaojian Ma
Machine Learning @ BIGAI
11/24/2023

ML is stepping into a new era





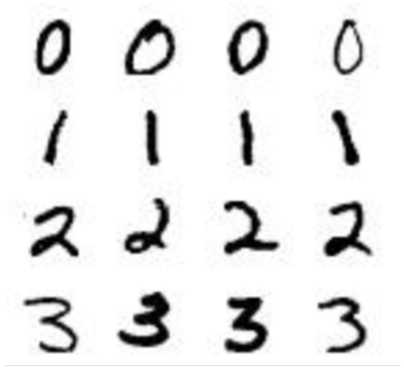
1990s

2010s

2020s

- More data, O(10k) -> O(10M) -> O(1T);
- More parameters, O(1M) -> O(1b) -> O(100b);
- More computation, GFLOPS -> TFLOPS

ML is stepping into a new era



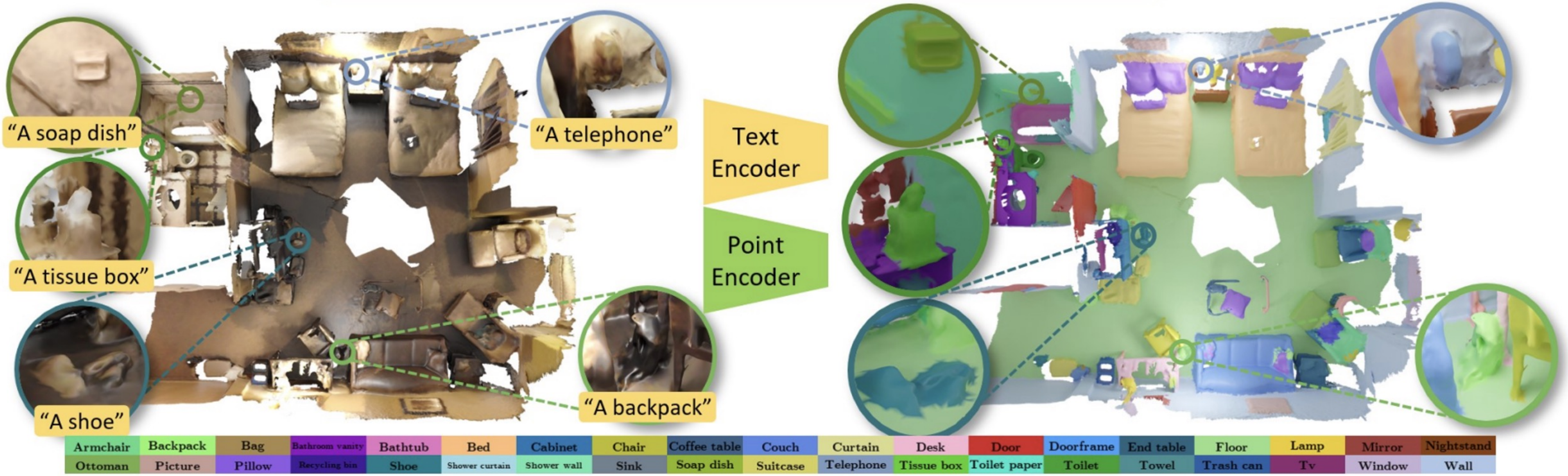
1990s

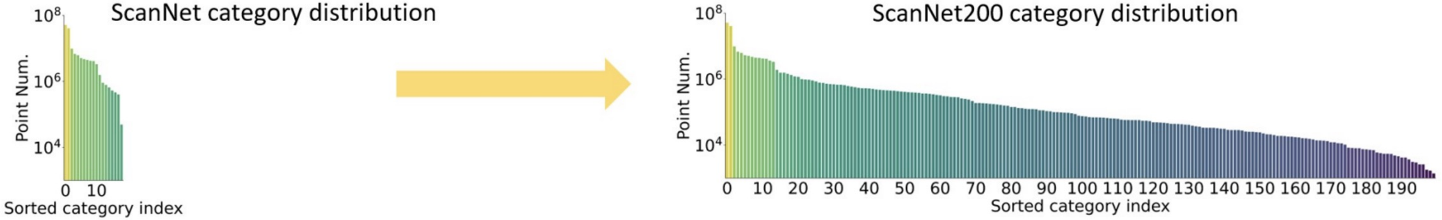
2010s

2020s

- Complex domains and semantics
- Close world (vocabulary) -> open world (vocabulary)

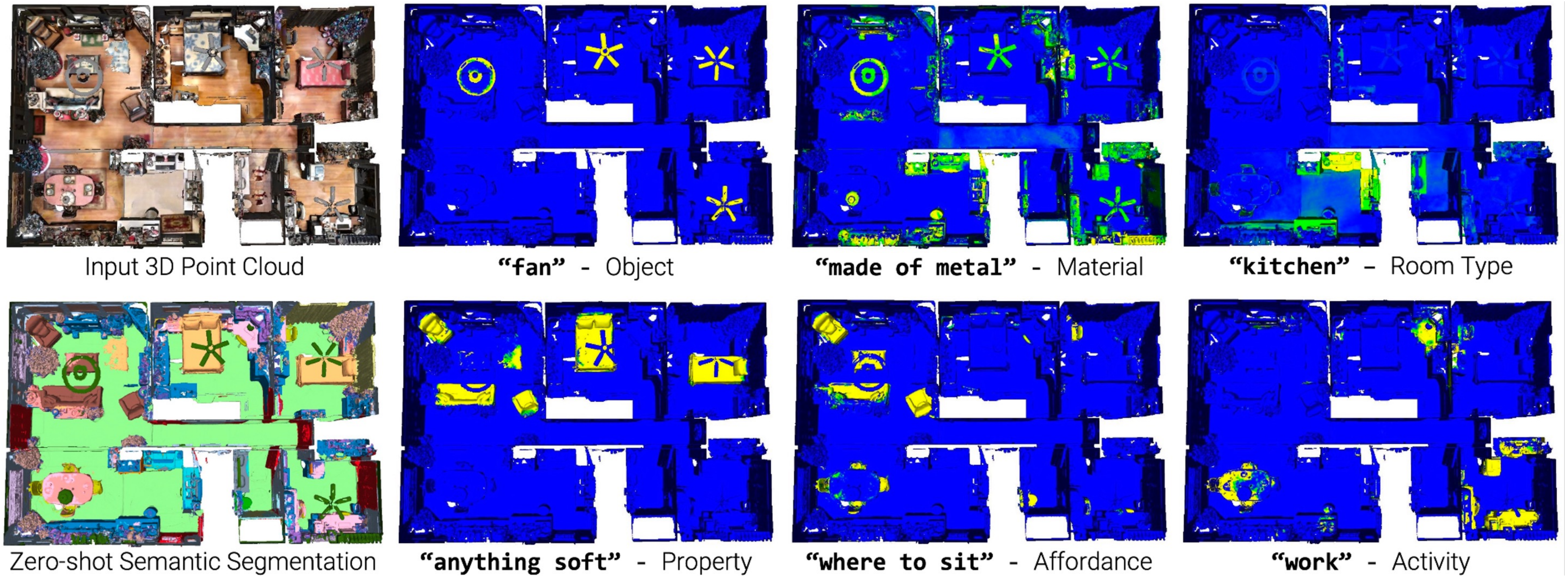
…and so is embodied AI





ScanNet-200: Language-Grounded Indoor 3D Semantic Segmentation in the Wild

...and so is embodied AI

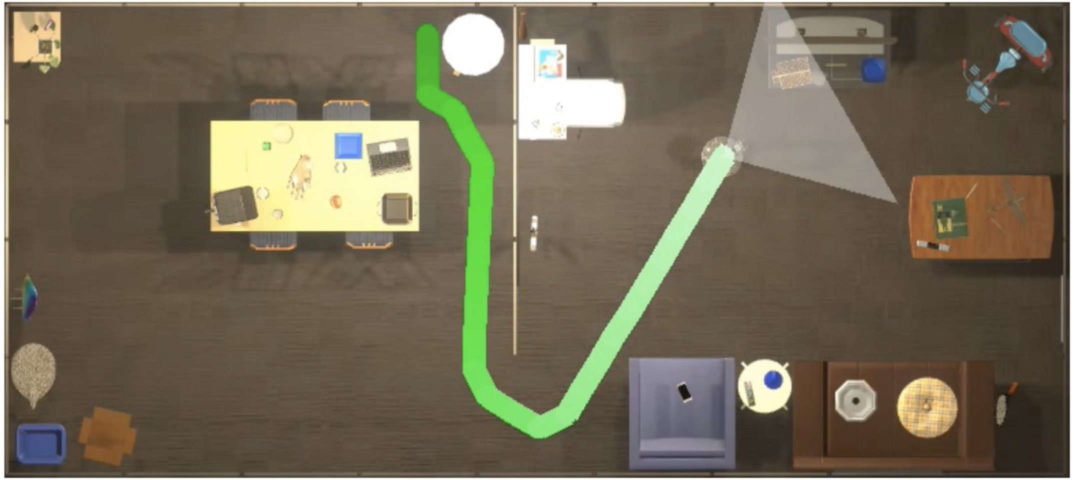


OpenScene: 3D Scene Understanding with Open Vocabularies

…and so is embodied AI

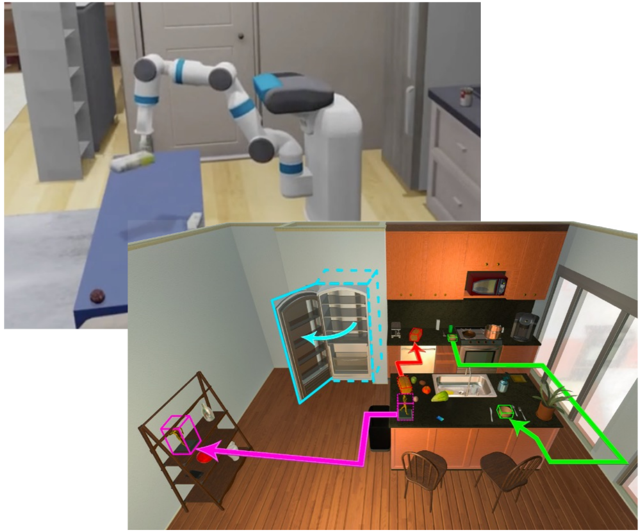
Top-down visualization

Task: Find the gingerbread house



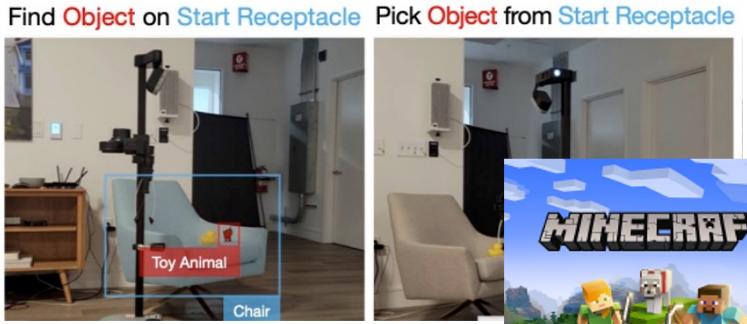
CoWs on Pasture:
Baselines and Benchmarks for Language-Driven Zero-Shot Object Navigation

A paradigm shift for embodied AI



contrived; limited tasks;
static; close world…

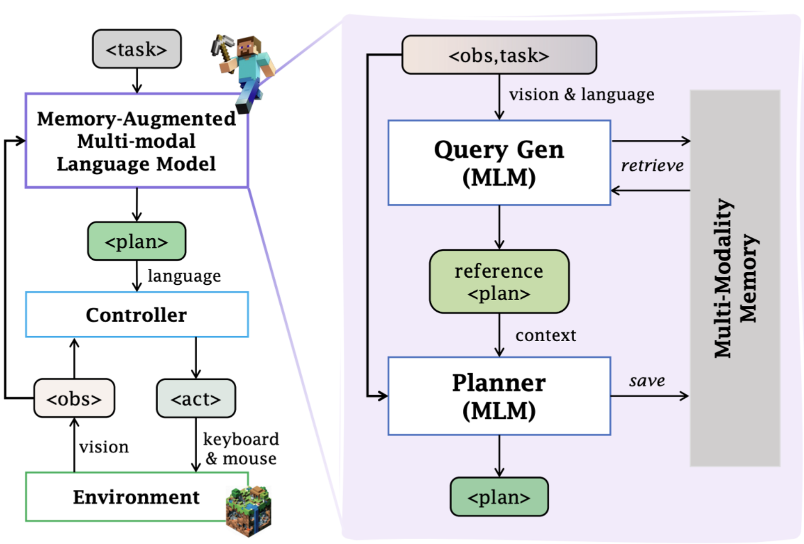
NeurIPS 2023 HomeRobot: Open Vocabulary Mobile Manipulation (OVMM) Challenge



realistic; massive tasks;
dynamic open world…

Embodied AI

Generalist Embodied AI in an Open World



?

LEO: An Embodied Generalist Agent in 3D World

CraftJarvis: Multi-task Embodied Agents in an Open World

CraftJarvis

stay tuned…

?

LEO: An Embodied Generalist Agent in 3D World

CraftJarvis: Multi-task Embodied Agents in an Open World

stay tuned…

LEO: An Embodied Generalist Agent

An Embodied Generalist Agent in 3D World,
arXiv preprint 2023

3D-VisTA: Pre-trained Transformer for 3D Vision and Text Alignment,
ICCV 2023

SQA3D: Situated Question Answering in 3D Scenes,
ICLR 2023

embodied-generalist.github.io

Embodied Generalist Agent

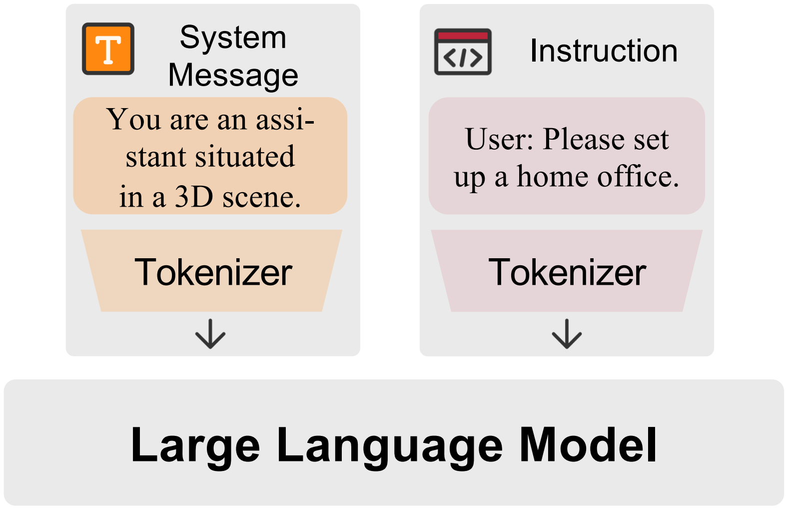
Capabilities: ***Perception***, ***Grounding***, ***Reasoning***, *Planning*, ***Acting***

Tasks

3D Object Captioning	**3D Question Answering**	**3D Dialogue**	**Embodied Navigation**
Scene Captioning	**Embodied Reasoning**	**Task Planning**	**Robotic Manipulation**

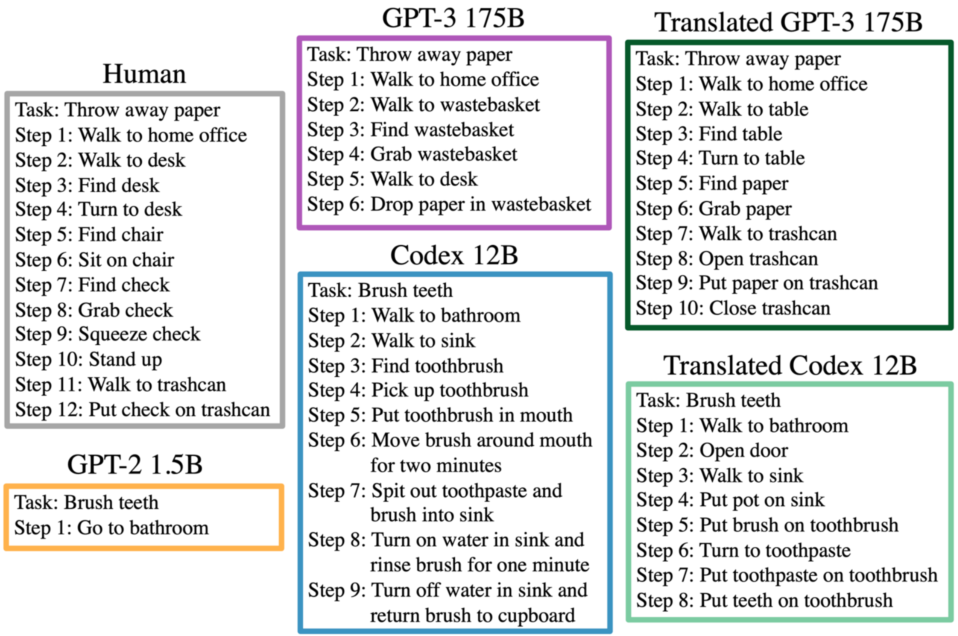
3D World

Single-modal agent

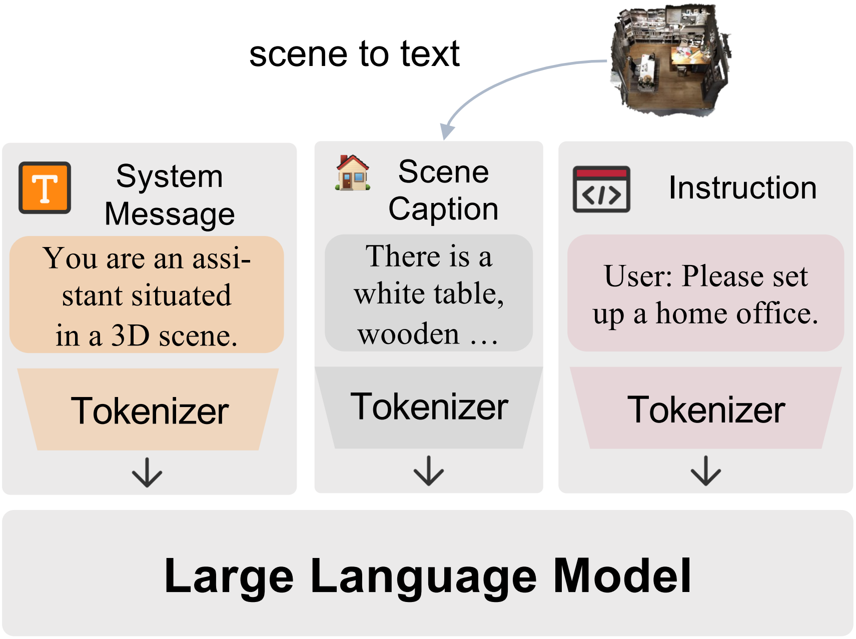


1. Choose a room…
2. Plan the layout…
3. Create a workspace…

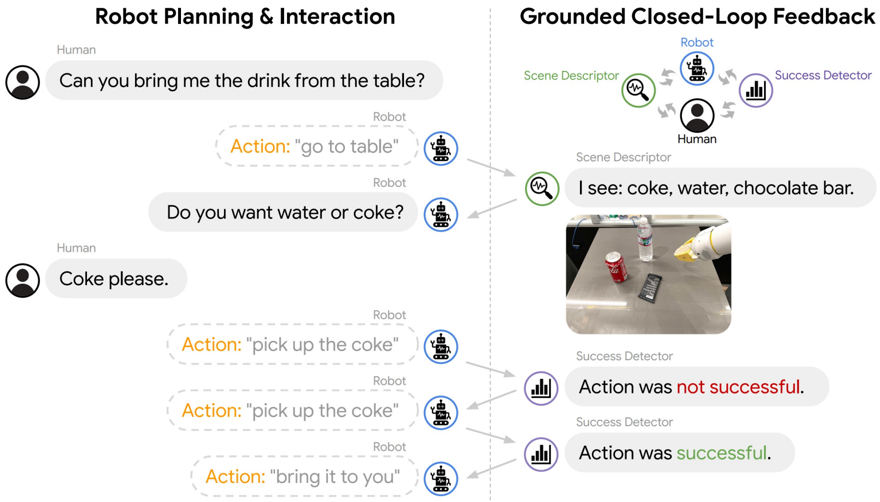
Without scene awareness:
ambiguous, hallucination



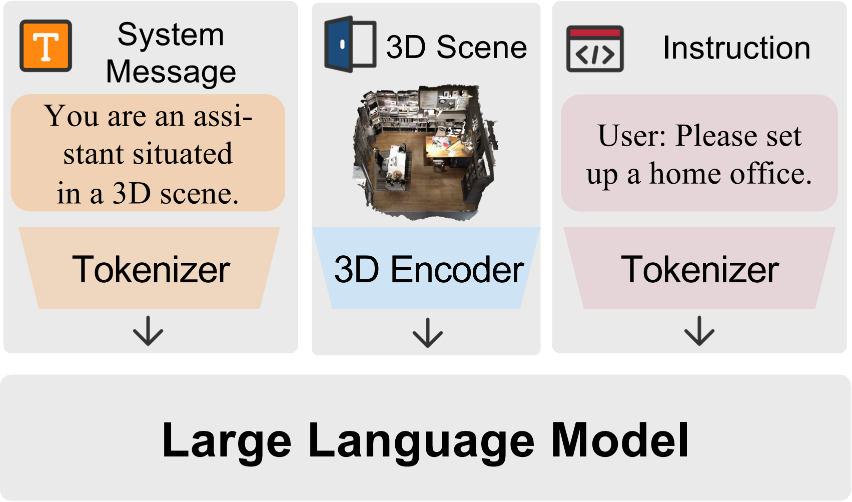
Language Models as Zero-Shot Planners: Extracting Actionable Knowledge for Embodied Agents



Tedious text, intractable to embed complex 3D information



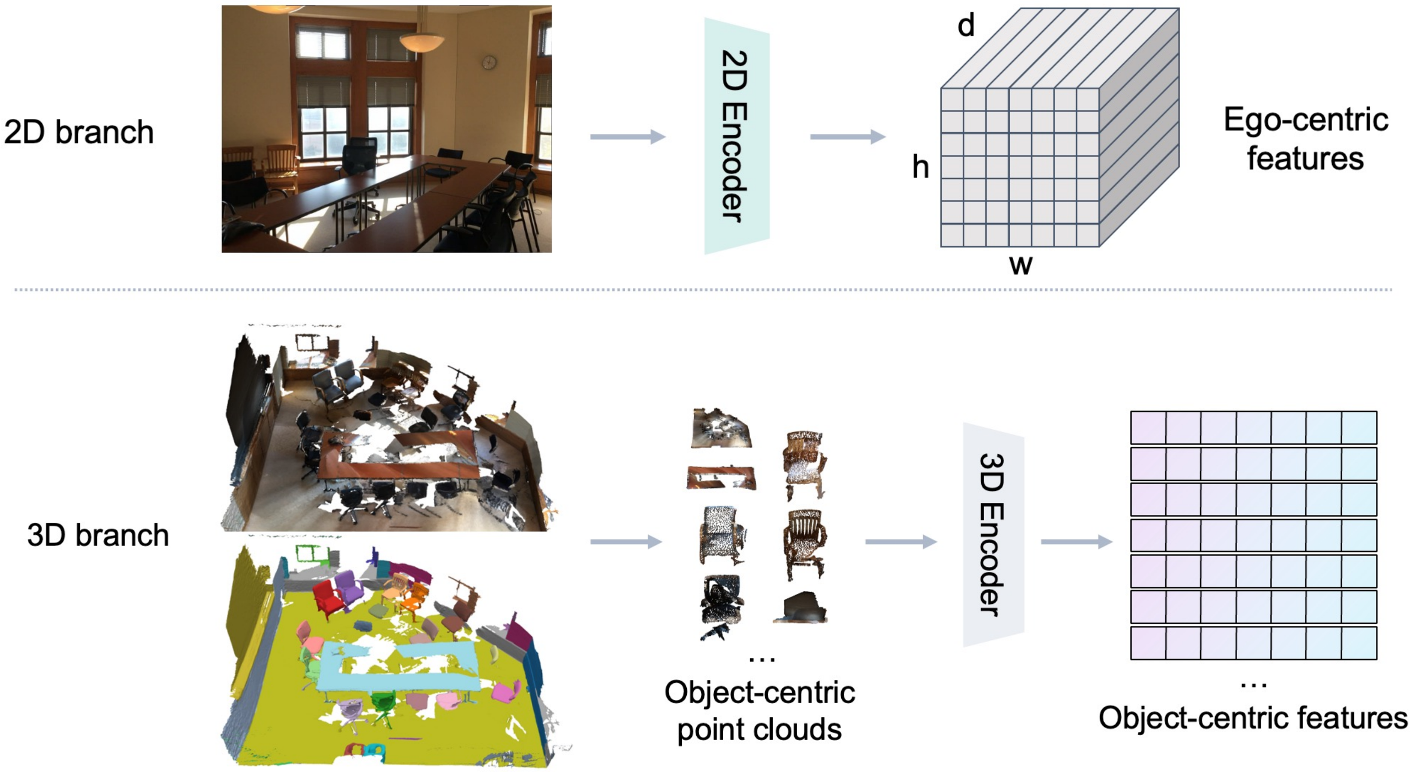
Inner Monologue: Embodied Reasoning through Planning with Language Models

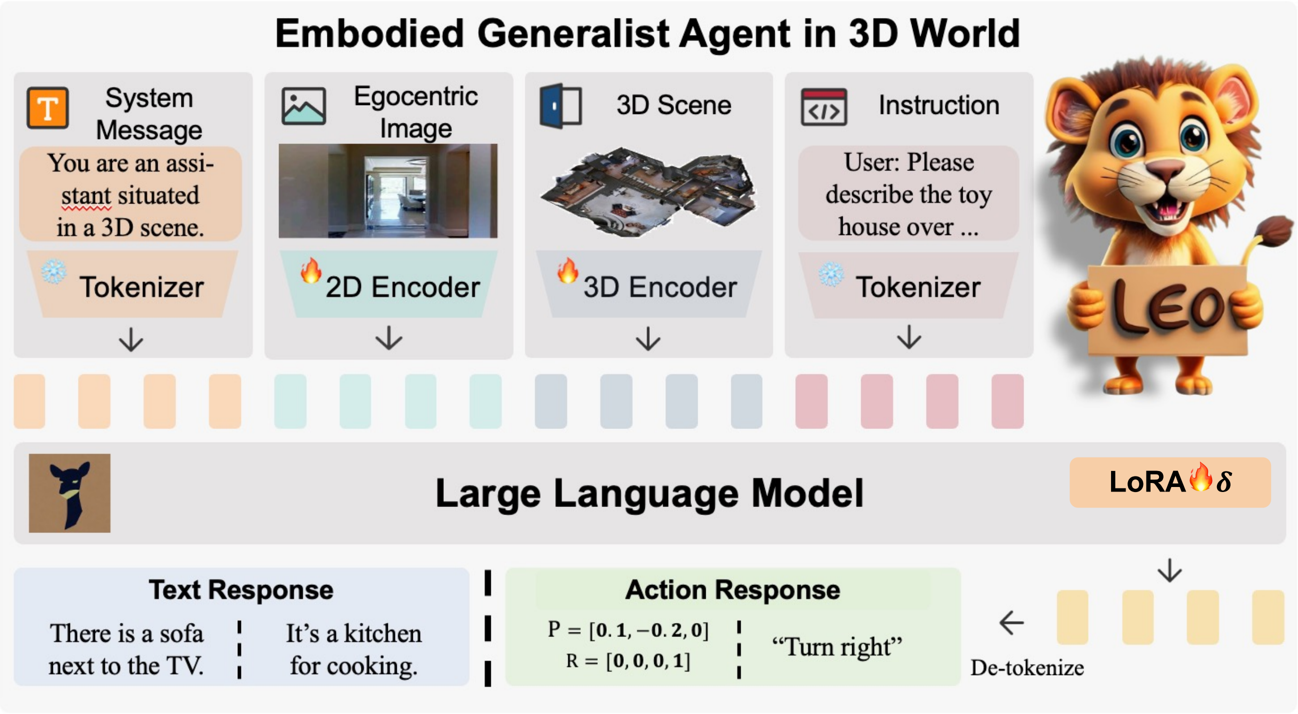


1. Place the desk in the desired position in the room…
2. Position the chair next to the desk, to the right of it.
3. Set up the shelf to the left of the desk…
4. Place the lamp on the desk…
5. Arrange the showcase to the right of the desk.
6. Place the plants on the shelf…
7. Hang the curtains on the wall behind the desk…

Scene-aware agent with capacity of perceiving (3D) scenes

Scene representation



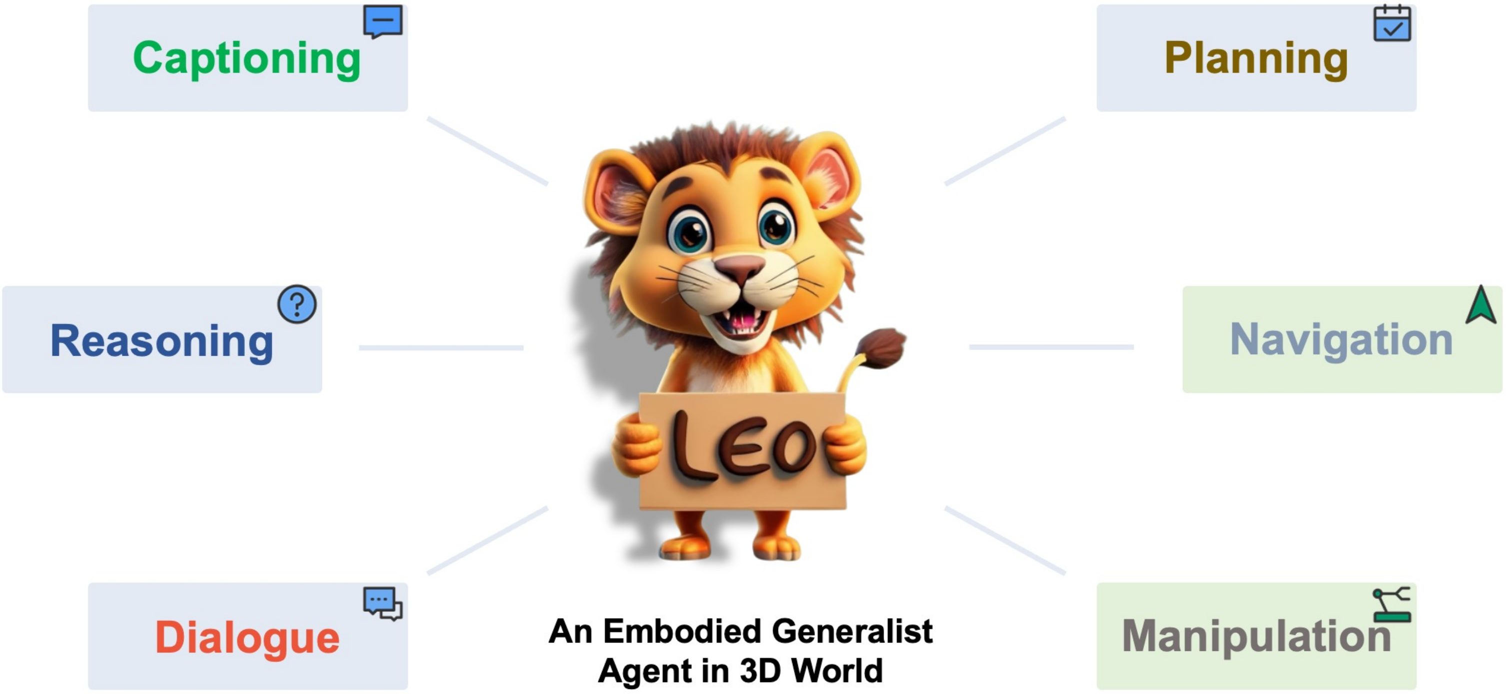


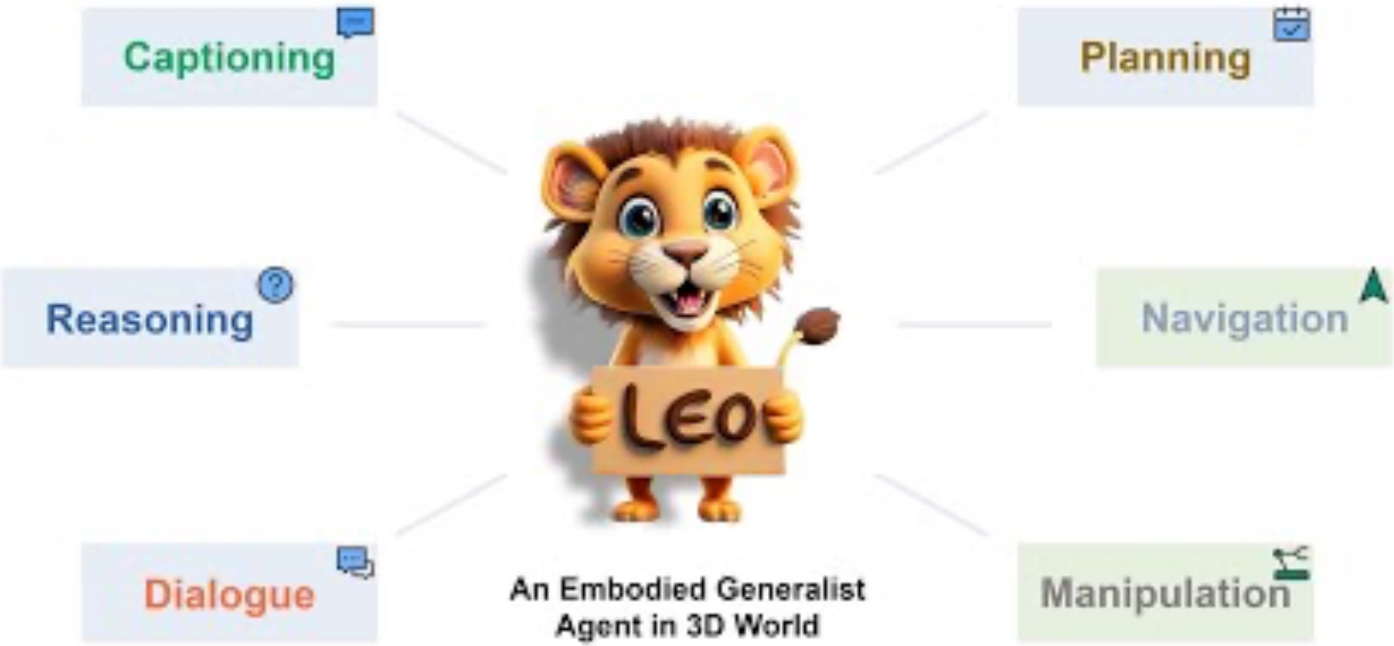
Unified task sequence

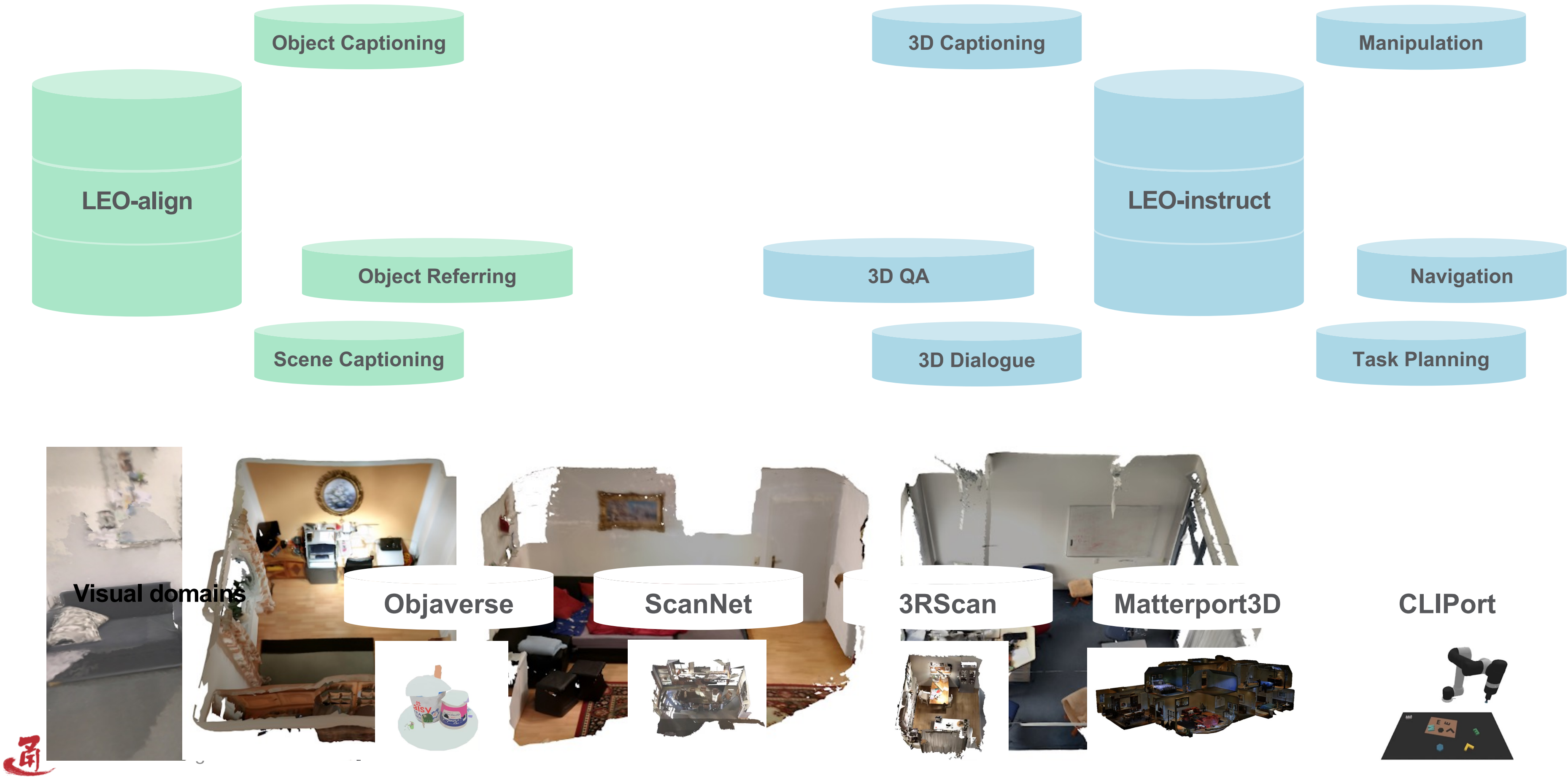
$$\underbrace{\texttt{You are...}}_{\text{system message}}\ \underbrace{s_{\text{2D}}^{(1)},\ldots,s_{\text{2D}}^{(M)}}_{\substack{\text{2D image tokens}\\ \text{(optional)}}}\ \underbrace{s_{\text{3D}}^{(1)},\ldots,s_{\text{3D}}^{(N)}}_{\substack{\text{object-centric}\\ \text{3D tokens}}}\ \underbrace{\texttt{USER:...}\quad\texttt{ASSISTANT:}}_{\text{instruction}}\ \underbrace{s_{\text{res}}^{(1)},\ldots s_{\text{res}}^{(T)}}_{\text{response}}.$$

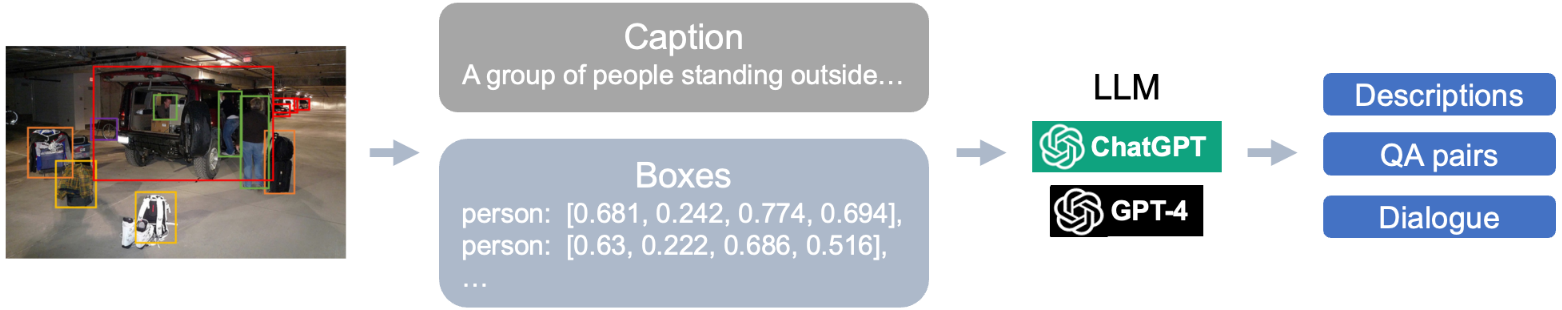
Auto-regressive objective

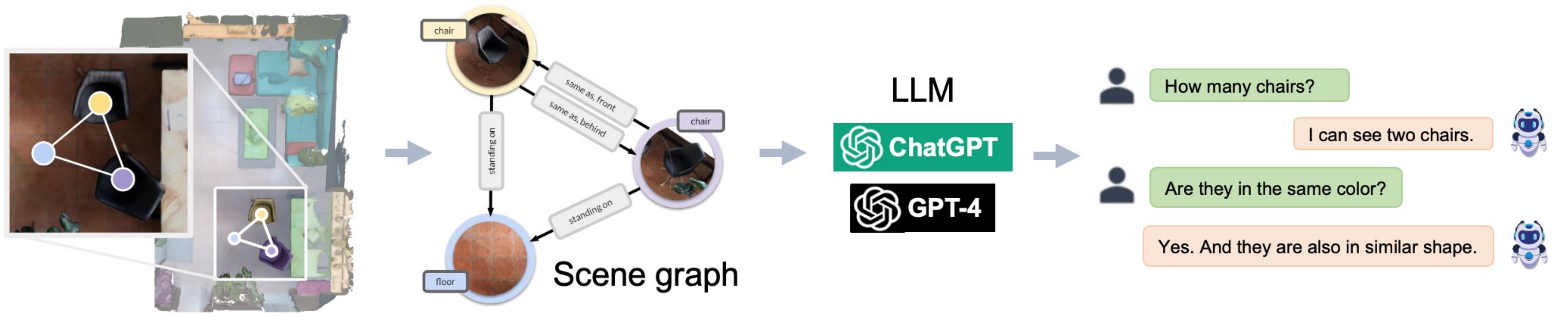
$$\mathcal{L}(\theta,\mathcal{B}) = -\sum_{b=1}^{|\mathcal{B}|}\sum_{t=1}^{T}\log p_\theta\left(s_{\text{res}}^{(b,t)} \middle| s_{\text{res}}^{(b,<t)}, s_{\text{prefix}}^{(b,1)},\ldots,s_{\text{prefix}}^{(b,L)}\right)$$


Captioning
Planning
Reasoning
Navigation
Dialogue
Manipulation
LEO
An Embodied Generalist
Agent in 3D World


Captioning
Planning
Reasoning
Navigation
LEO
Dialogue
Manipulation
An Embodied Generalist
Agent in 3D World


Object Captioning
LEO-align
Object Referring
Scene Captioning
3D Captioning
Manipulation
LEO-instruct
3D QA
Navigation
3D Dialogue
Task Planning
Visual domains
Objaverse
ScanNet
3RScan
Matterport3D
CLIPort


Caption
A group of people standing outside…
Boxes
person: [0.681, 0.242, 0.774, 0.694],
person: [0.63, 0.222, 0.686, 0.516],
…
LLM
ChatGPT
GPT-4
Descriptions
QA pairs
Dialogue


chair
same as, front
same as, behind
chair
standing on
standing on
floor
Scene graph
LLM
ChatGPT
GPT-4
How many chairs?
I can see two chairs.
Are they in the same color?
Yes. And they are also in similar shape.

Scene-Graph-based Prompting

Messages

1. System Message

You are an AI visual assistant in a 3D scene…

2. Demonstrations

Scene Graph Context:

{'sofa-1': {'attributes': {'color': 'red'}, 'relations': ['to the right of chair-2', 'in front of table-3']}, 'chair-2': {'attributes': {'color': 'brown'}, 'relations': []}}

Human-labeled Responses

3. Query (a new scene)

Scene Graph Context

Responses

1) Object scene caption
2) Scene caption
3) Dialogue (O-CoT)
4) QA (O-CoT)
5) Planning

Potential Issues:

1) Wrong Answers
2) Negative Responses
3) Responses with IDs
4) Others

Refinement Procedures

Raw Responses	Refined Responses
1. Wrong Answers There are two curtains in the room.	1. Answer Refinement There are 3 curtains in the room.
2. Negative Responses Question: Where is the football table? Thought: football table-17 Answer: unknown	2. Removing Negative Responses The negative responses will be removed.
3. Responses with IDs The kitchen cabinet-9 and kitchen counter-10 are parts of the kitchen.	3. GPT-based rewriting The kitchen features a cabinet and a counter.
4. Others Unnecessary contents (Thought, etc.)	4. Others Remove unnecessary contents.

Dialogue (O-CoT):
Dialogue Context: high level task: organize the bedroom. low level task: check some objects.
Human: I need help organizing my bedroom. Can you assist me?Thought: no specific objects
Assistant: Of course! What specifically do you need help with?
Human: Can you check if there are any clean clothes hanging on the wall? Thought: clothes-13
Assistant: I'm sorry, but the clothes hanging on the wall are messy.
Human: Thanks for your help! Thought: no specific object
Assistant: You're welcome! If you need any further assistance, feel free to ask.

Scene Caption:
In this room, there is a wooden floor that is clean and flat. A tall wardrobe stand on the right side of a desk, close to a basket. The wardrobe is in front of a chair, a bed, and behind a bedside table. Clothes are hanging on a white wall, to the right of a closed curtain. The bed is wide, clean, and covered with blue blanket. The room has a comfortable and organized setting with functional furniture.

Object Scene Caption:
The showcase is supported by the wall and positioned behind, close to, and to the left of the shelf. The wall is flat, architectural, and dark in color. The shelf is made of wood and has a low size.

QA (O-CoT):
Question: Where is the printer located? Thought: printer-8
Answer: standing on the desk
Question: How many blankets are on the bed? Thought: blanket-16, blanket-17 Answer: 2
Question: What is the type of the room? Thought:wardrobe-2, desk-7, chair-11, bed-15 Answer: bedroom

Planning:
High-Level Task: Organize and tidy up the bedroom.
Low-Level Actions:
1. Clean the floor by sweeping to remove any dirt.
2. Make the bed by arranging the blanket and pillows.
3. Place any loose items or belongings into the basket.
4. Arrange items on the shelves and showcase in a tidy way.

3DVL

	Scan2Cap (val)					ScanQA (val)					SQA3D
	C	B-4	M	R	Sim	C	B-4	M	R	EM@1	EM@1
Task-specific models											
Scan2Cap (GPT-3) (Chen et al., 2021)	35.2	22.4	21.4	43.5	-	-	-	-	-	-	$41.0^{\dagger}$
3DJCG (Cai et al., 2022)	47.7	31.5	24.3	51.8	-	-	-	-	-	-	-
Vote2Cap-DETR (Chen et al., 2023)	61.8	34.5	26.2	54.4	-	-	-	-	-	-	-
ScanRefer+MCAN (Chen et al., 2020)	-	-	-	-	-	55.4	7.9	11.5	30.0	18.6	-
ClipBERT (Lei et al., 2021)	-	-	-	-	-	-	-	-	-	-	43.3
ScanQA (Azuma et al., 2022)	-	-	-	-	-	64.9	10.1	13.1	33.3	21.1	47.2
Task-specific fine-tuned											
3D-VisTA (Zhu et al., 2023c)	66.9	34.0	27.1	54.3	53.8	69.6	10.4	13.9	35.7	22.4	48.5
3D-LLM (FlanT5) (Hong et al., 2023)	-	-	-	-	-	69.4	**12.0**	14.5	35.7	20.5	-
LEO	**68.4**	**36.9**	**27.7**	**57.8**	**54.7**	**80.0**	11.5	**16.2**	**39.3**	**36.6**	**53.7**

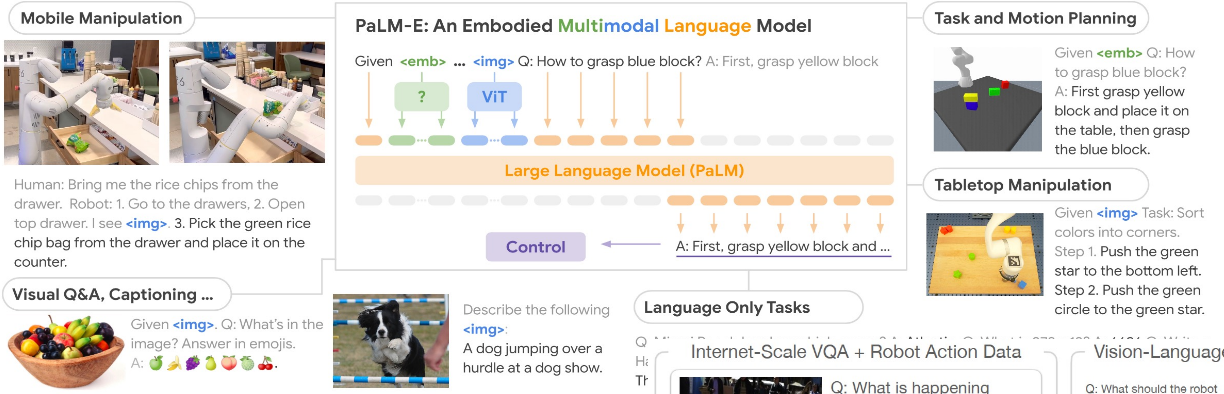
CLIPort Manipulation

	separating-piles		packing-google -objects-seq		put-blocks-in -bowls	
	seen	unseen	seen	unseen	seen	unseen
Transporter	48.4	52.3	46.3	37.3	64.7	18.7
CLIP-only	90.2	71.0	95.8	57.8	97.7	44.5
RN50-BERT	46.5	44.9	94.0	56.1	91.8	23.8
CLIPort (single)	98.0	**75.2**	**96.2**	71.9	**100**	25.0
CLIPort (multi)	89.0	62.8	84.4	70.3	**100**	**45.8**
LEO	**98.8**	**75.2**	76.6	**79.8**	86.2	35.2

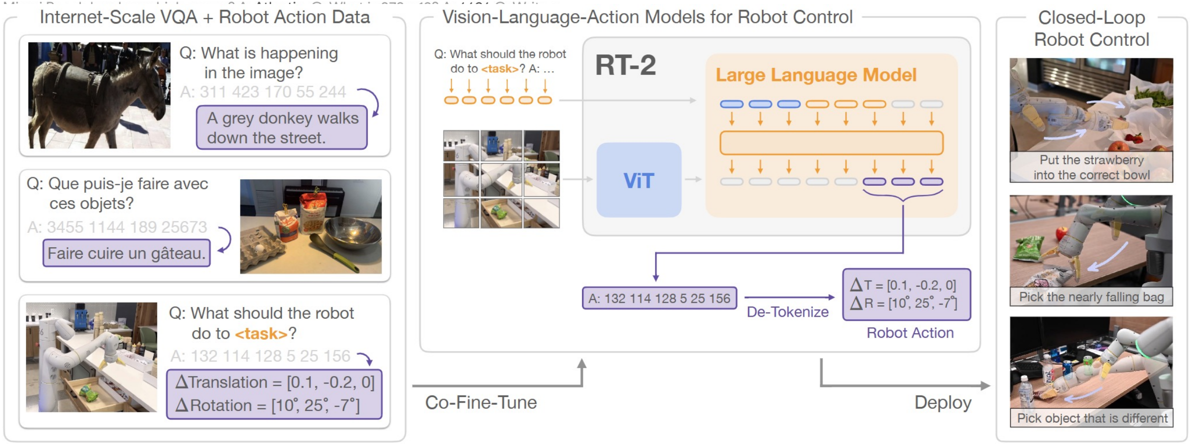
ObjNav Navigation

	MP3D-val		HM3D-val	
	S(↑)	L(↑)	S(↑)	L(↑)
H.w. (shortest)	4.4	2.2	-	-
H.w. (70k demo)	**35.4**	10.2	-	-
VC-1 (ViT-B)	-	-	**57.1**	**31.4**
LEO	23.1	**15.2**	$23.1^{\dagger}$	$19.1^{\dagger}$

Related research

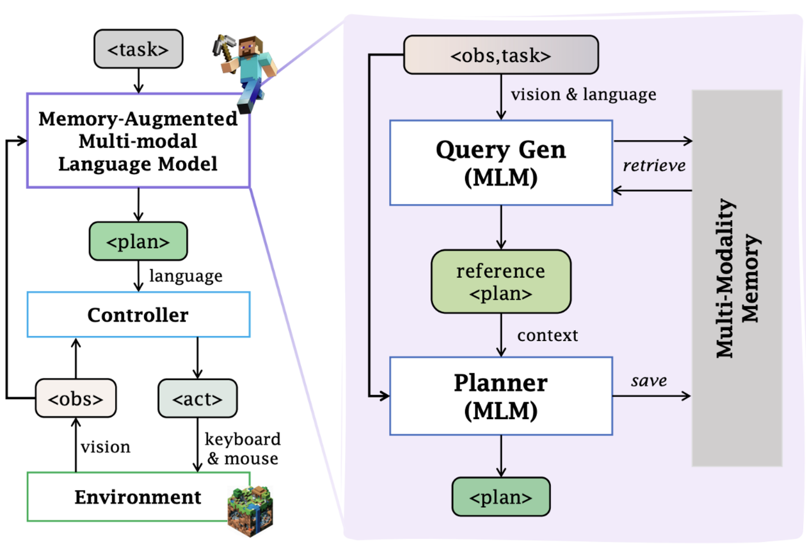


PaLM-E: a comprehensive generalist exceling in multi-modal reasoning and planning



RT-2: bridging the gap between vision, language and action

LEO: An Embodied Generalist Agent in 3D World



CraftJarvis: Multi-task Embodied Agents in an Open World

CraftJarvis

?

stay tuned…

CraftJarvis: Embodied Agents in an Open World

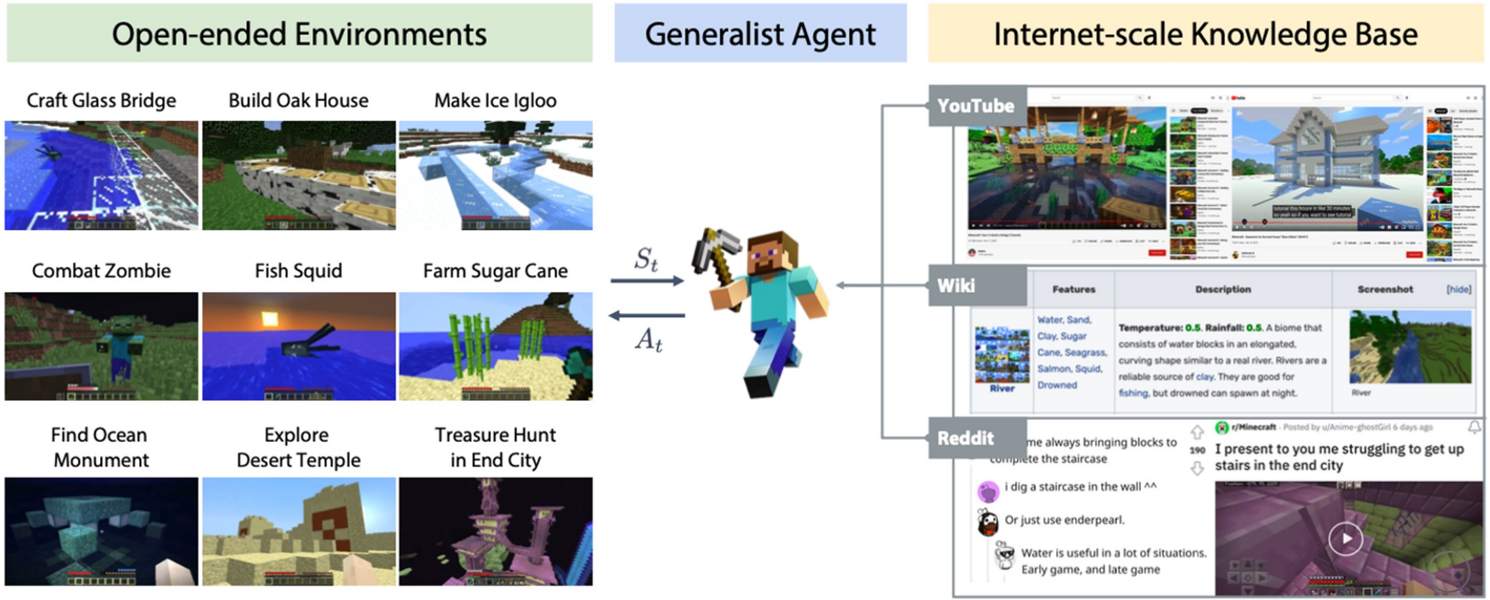
Open-World Multi-Task Control Through Goal-Aware Representation Learning and Adaptive Horizon Prediction,
CVPR 2023

Describe, Explain, Plan and Select: Interactive Planning with Large Language Models Enables Open-World Multi-Task Agents,
Best paper award, ICML ‘23 TEACH Workshop
NeurIPS 2023

JARVIS-1: Open-world Multi-task Agents with Memory-Augmented Multimodal Language Models,
arXiv 2023

craftjarvis-jarvis1.github.io

Minecraft: embodied AI in an open world



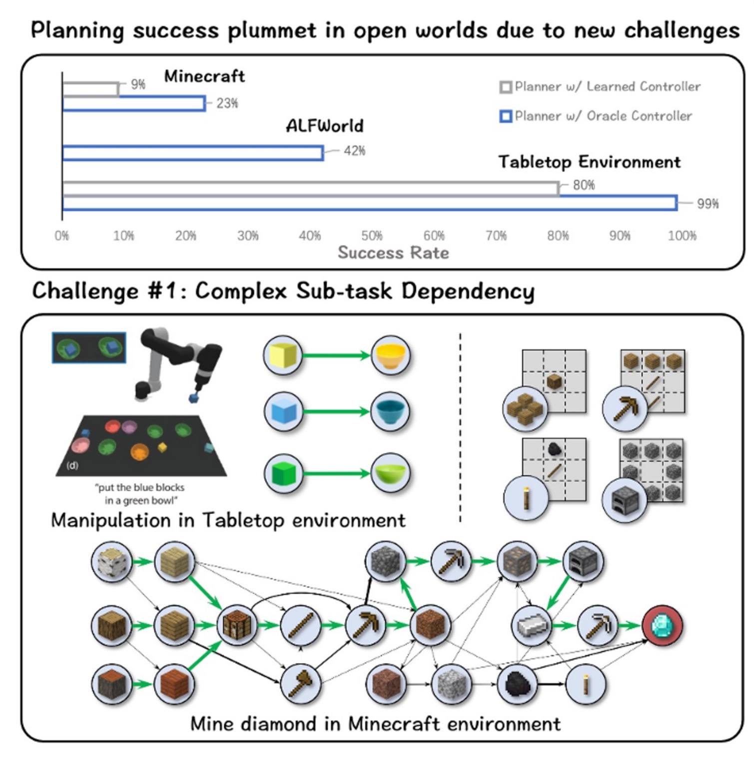
Today's embodied AI

- Restrictive objectives
- Very few tasks
- Limited knowledge

Embodied AI in an open world

- Open-ended objectives
- Massively multitask
- Web-scale knowledge

Challenges in open world environments

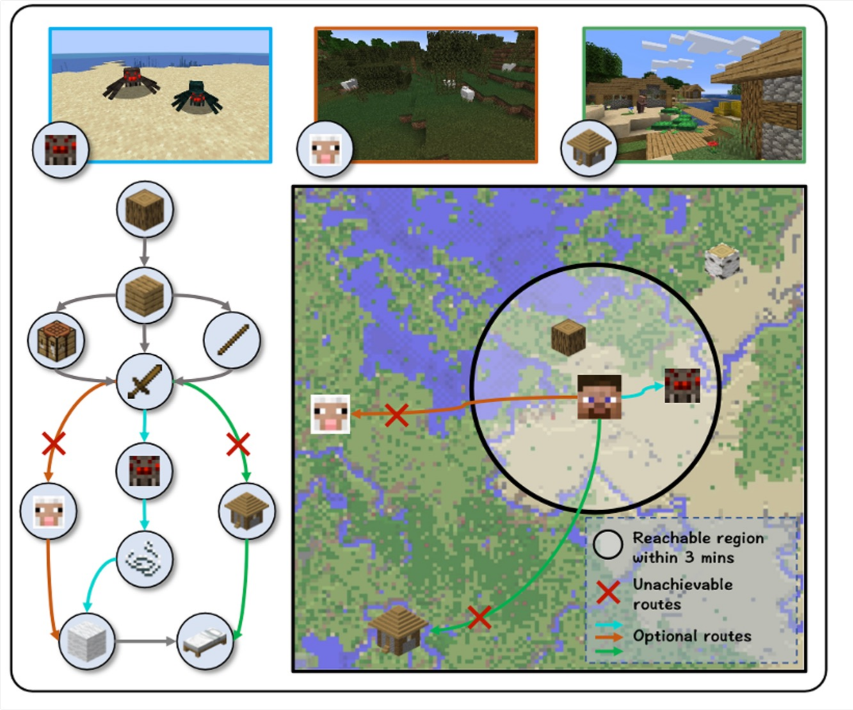


Challenge #1: long-horizon planning

Open worlds have highly abundant object types with complex dependency and relation. As a result, ground-truth plans typically involve a long sequence of sub-goals with strict dependencies.

=> Planning Success Rate will drops significantly on long-horizon tasks.

Challenges in open world environments

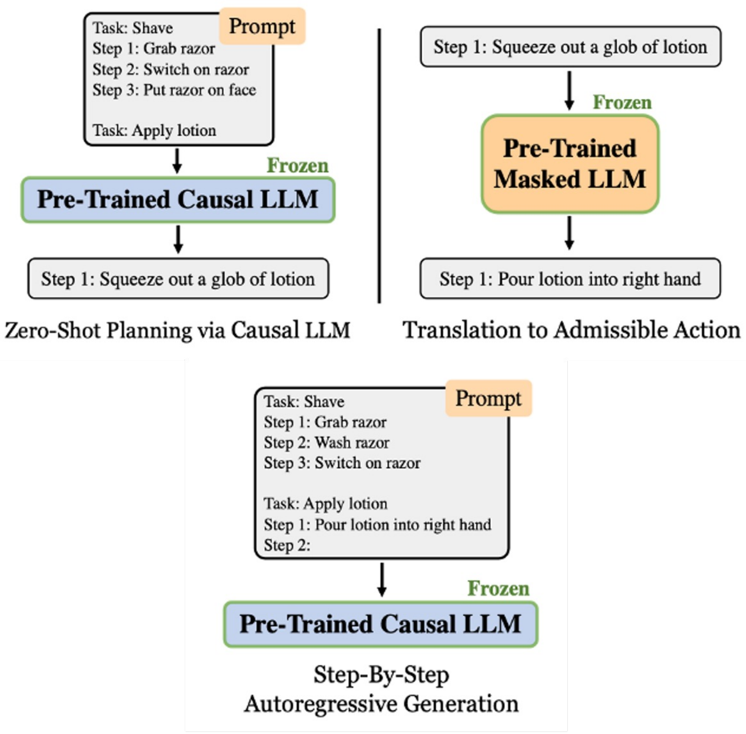


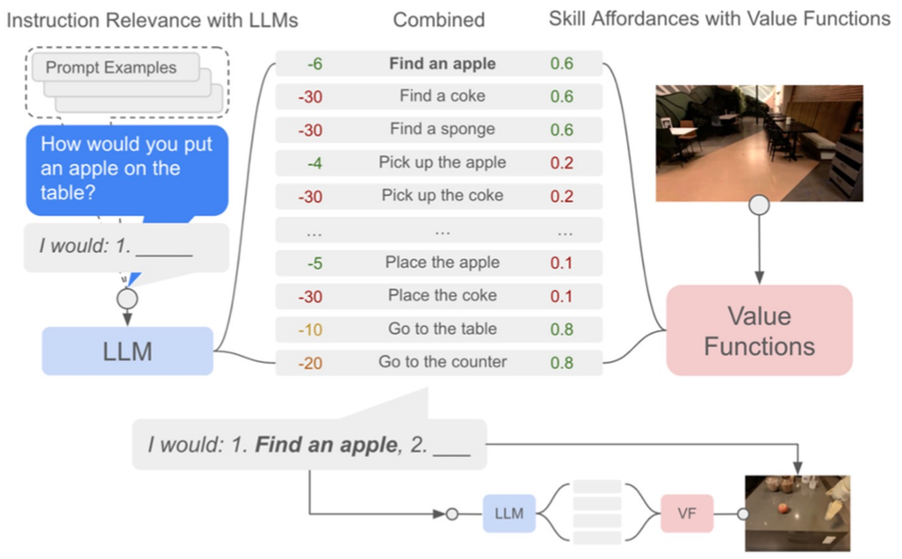
Challenge #2: state-aware planning

When dealing with a task that can be completed by executing multiple possible sequences of sub-goals, the planner should be able to select the best route base on the current state of the agent.

=> the complex and diverse state distribution of open-world environments makes state-awareness hard to achieve.

Challenges in open world environments

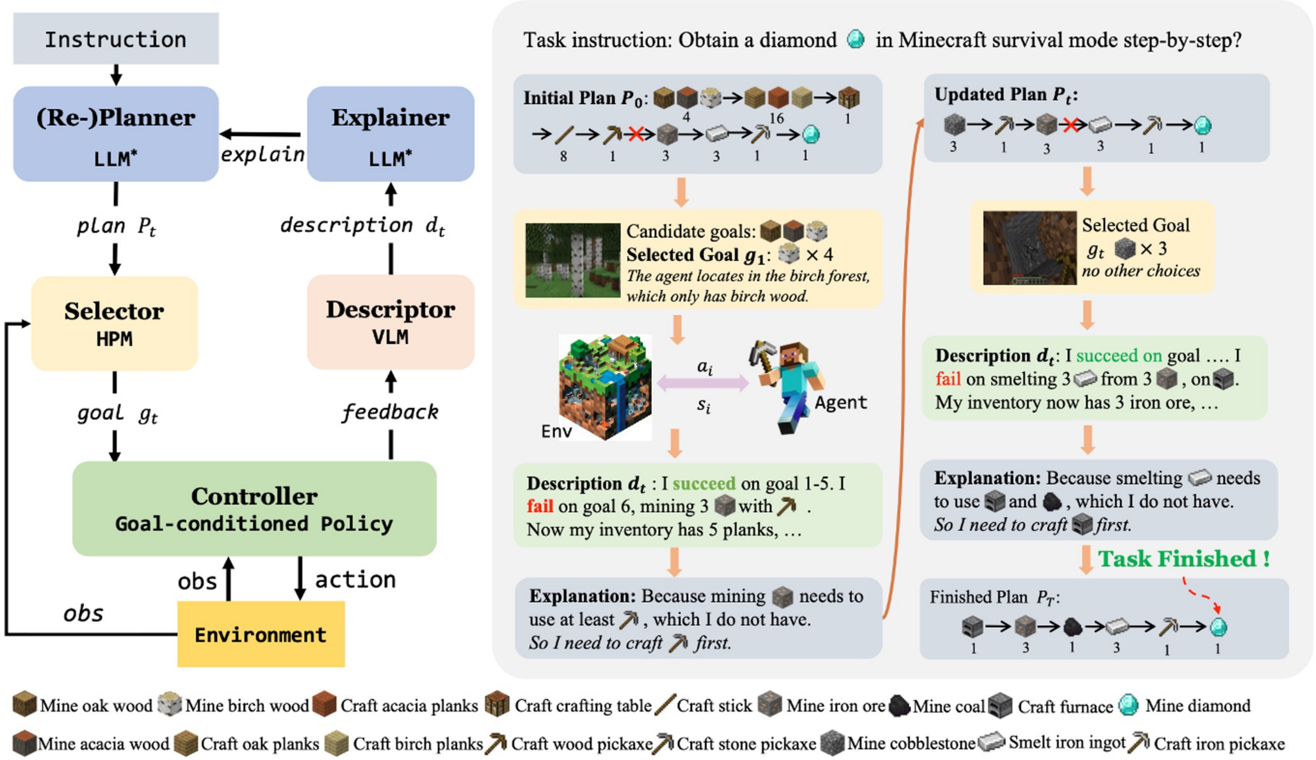


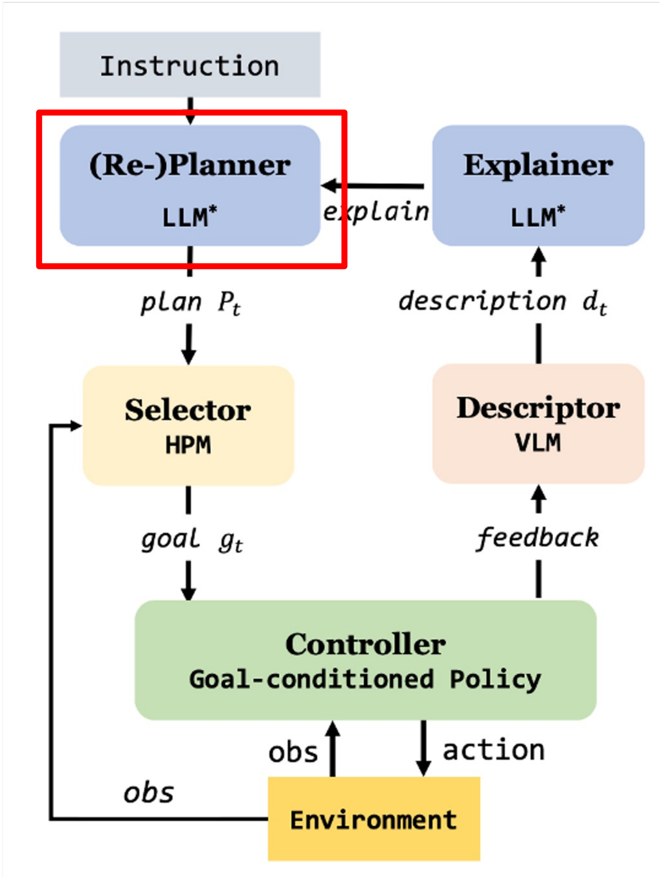


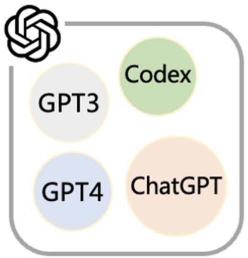
LLM for planning in close worlds

Language Models as Zero-Shot Planners: Extracting Actionable Knowledge for Embodied Agents
Do As I Can, Not As I Say: Grounding Language in Robotic Affordance

CraftJarvis: an embodied agents in Minecraft



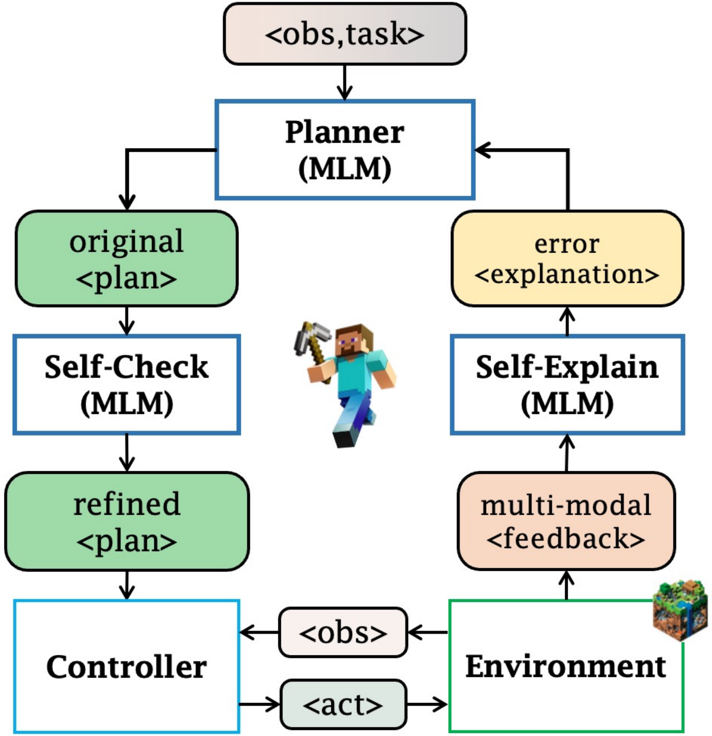


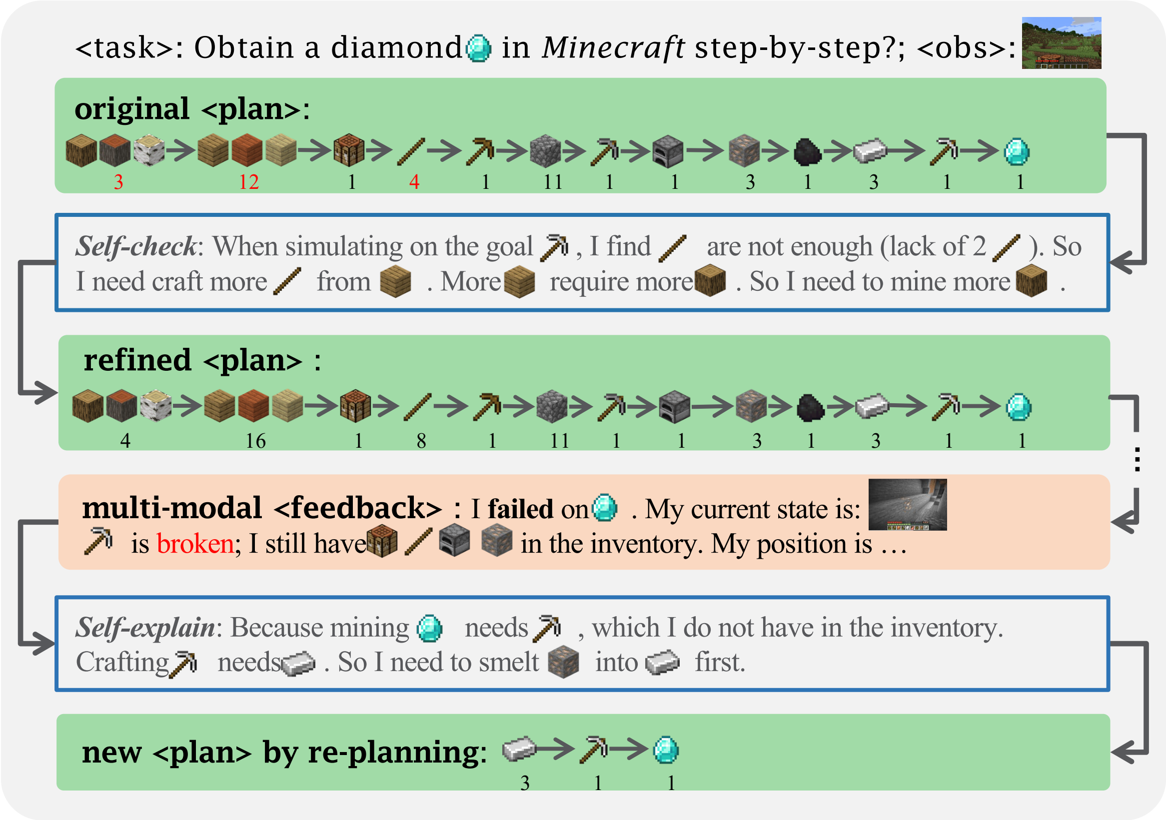



```python
def craft_wooden_axe(initial_inventory={}):
    # step 1: mine 3 logs
    mine(obj = {"log":3}, tool = None)
    # step 2: craft 12 planks from 3 logs
    craft(obj = {"planks":12}, materials = {"log":3},
          tool = None)
    # step 3: craft 4 sticks from 2 planks
    craft(obj = {"stick":4}, materials = {"planks"
         :2}, tool = None)
    # step 4: craft 1 crafting_table from 4 planks
    craft(obj = {"crafting_table":1}, materials = {"
         planks":4}, tool = None)
    # step 5: craft 1 wooden_axe from 3 planks and 2
         sticks on crafting table
    craft(obj = {"wooden_axe":1}, {"planks": 3, "
         stick": 2}, "crafting_table")
    return "wooden_axe"
```

Lots of errors!

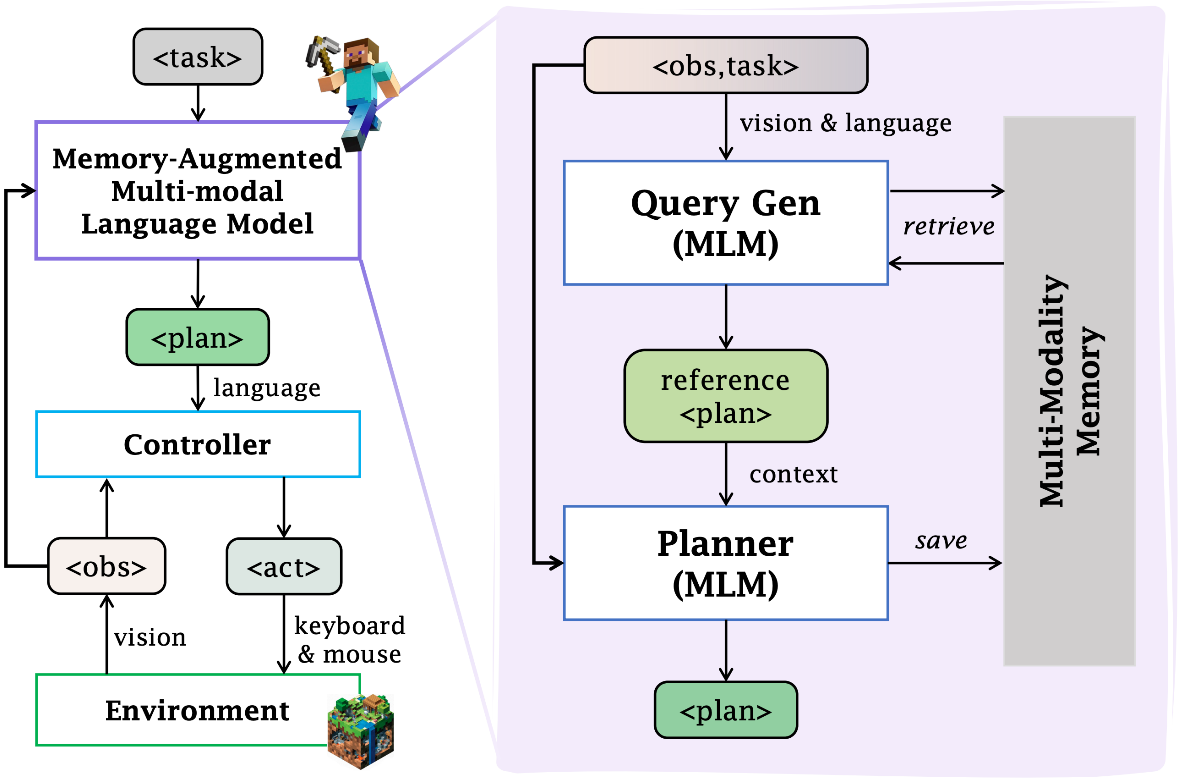
Self-correction

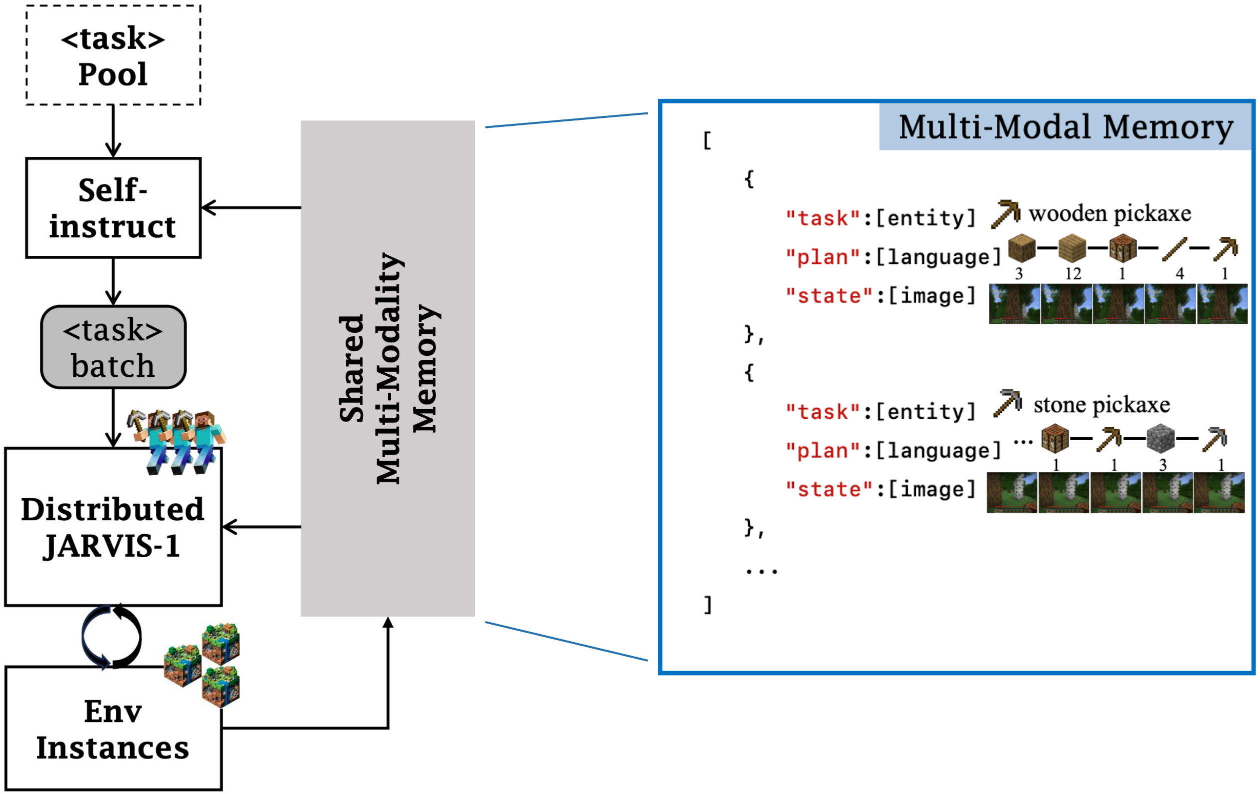


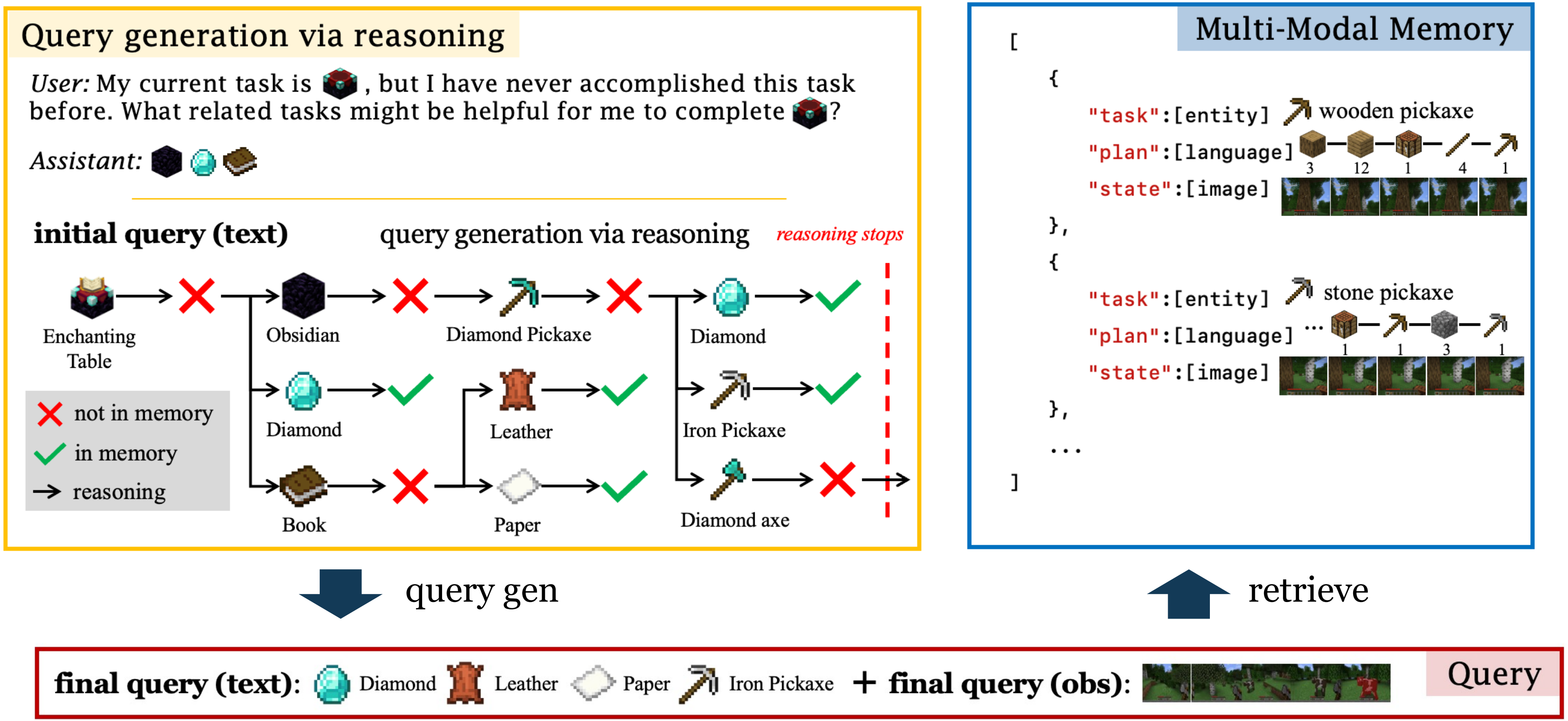


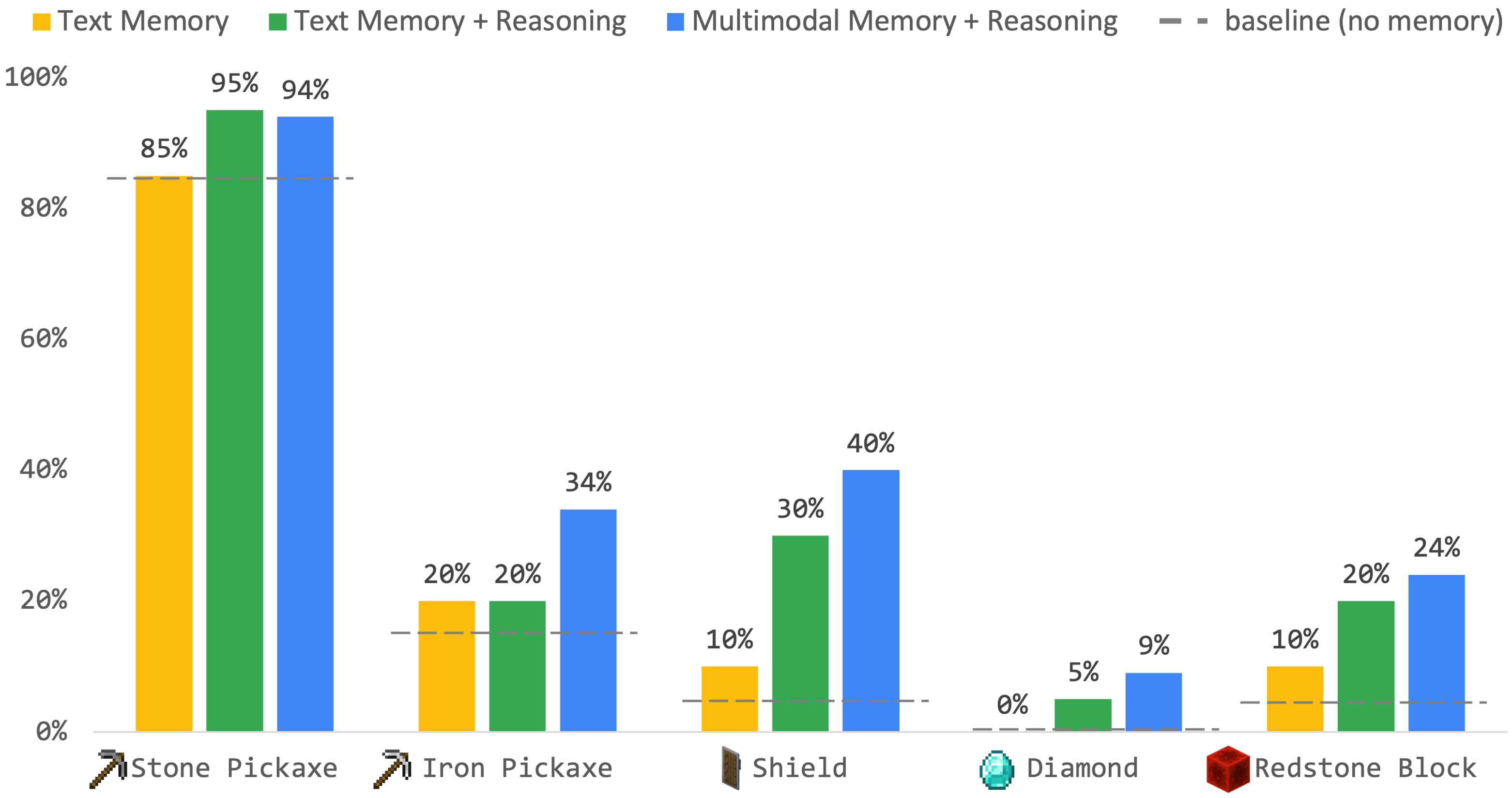
<task>: Obtain a diamond in *Minecraft* step-by-step?; <obs>:
original <plan>:
3 12 1 4 1 11 1 1 3 1 3 1 1
Self-check: When simulating on the goal , I find are not enough (lack of 2). So I need craft more from . More require more . So I need to mine more .
refined <plan> :
4 16 1 8 1 11 1 1 3 1 3 1 1
multi-modal <feedback> : I **failed** on . My current state is:
is broken; I still have in the inventory. My position is …
Self-explain: Because mining needs , which I do not have in the inventory. Crafting needs . So I need to smelt into first.
new <plan> by re-planning:
3 1 1

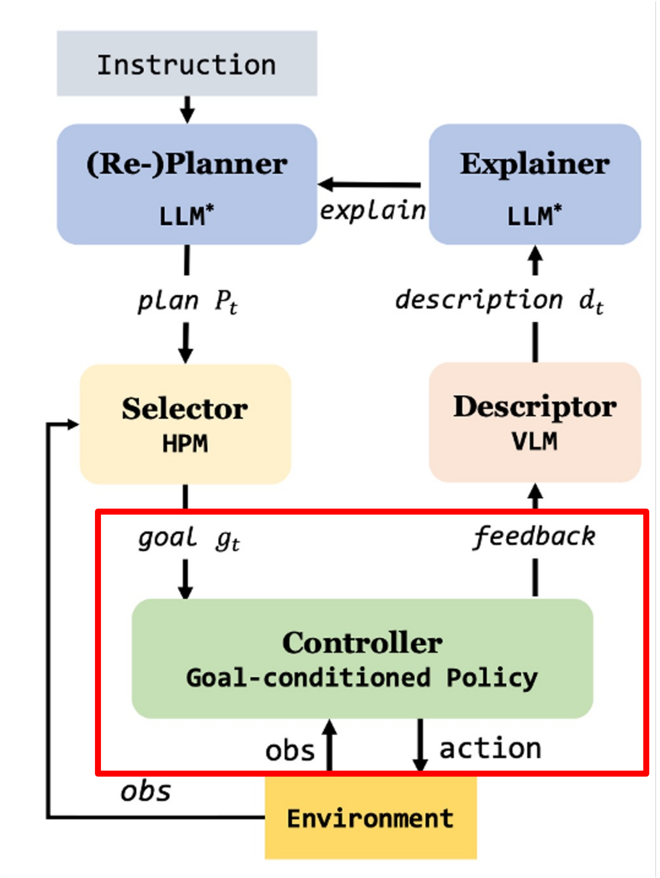
Embodied RAG (retrieval-augmented generation)



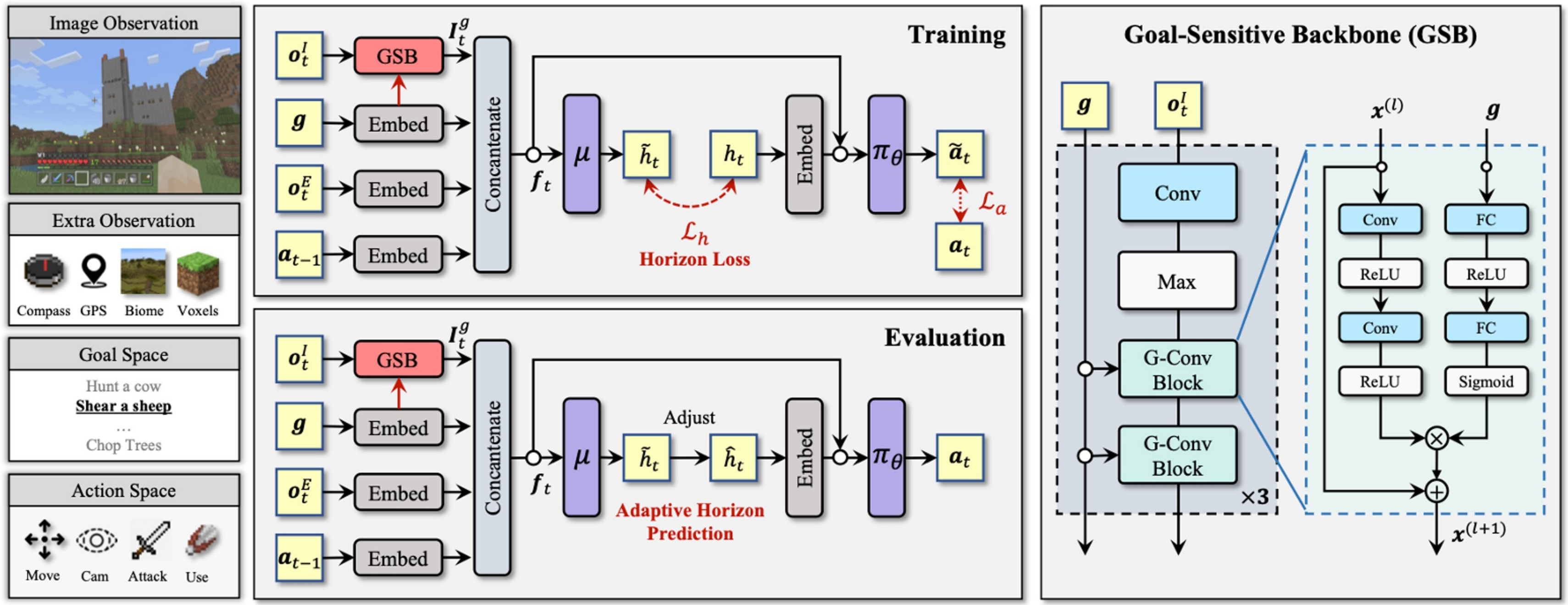

<task>
Pool
Self-
instruct
<task>
batch
Distributed
JARVIS-1
Env
Instances
Shared
Multi-Modality
Memory
Multi-Modal Memory
[
{
"task":[entity]
wooden pickaxe
"plan":[language]
3
12
1
4
1
"state":[image]
},
{
"task":[entity]
stone pickaxe
"plan":[language]
1
1
3
1
"state":[image]
},
...
]


Query generation via reasoning
User: My current task is , but I have never accomplished this task before. What related tasks might be helpful for me to complete ?
Assistant:
initial query (text)
query generation via reasoning
reasoning stops
Enchanting Table
Obsidian
Diamond Pickaxe
Diamond
Diamond
Leather
Iron Pickaxe
Book
Paper
Diamond axe
not in memory
in memory
reasoning
Multi-Modal Memory
[
{
"task":[entity] wooden pickaxe
"plan":[language]
3
12
1
4
1
"state":[image]
},
{
"task":[entity] stone pickaxe
"plan":[language] ...
1
1
3
1
"state":[image]
},
...
]
query gen
retrieve
final query (text): Diamond Leather Paper Iron Pickaxe + final query (obs):
Query

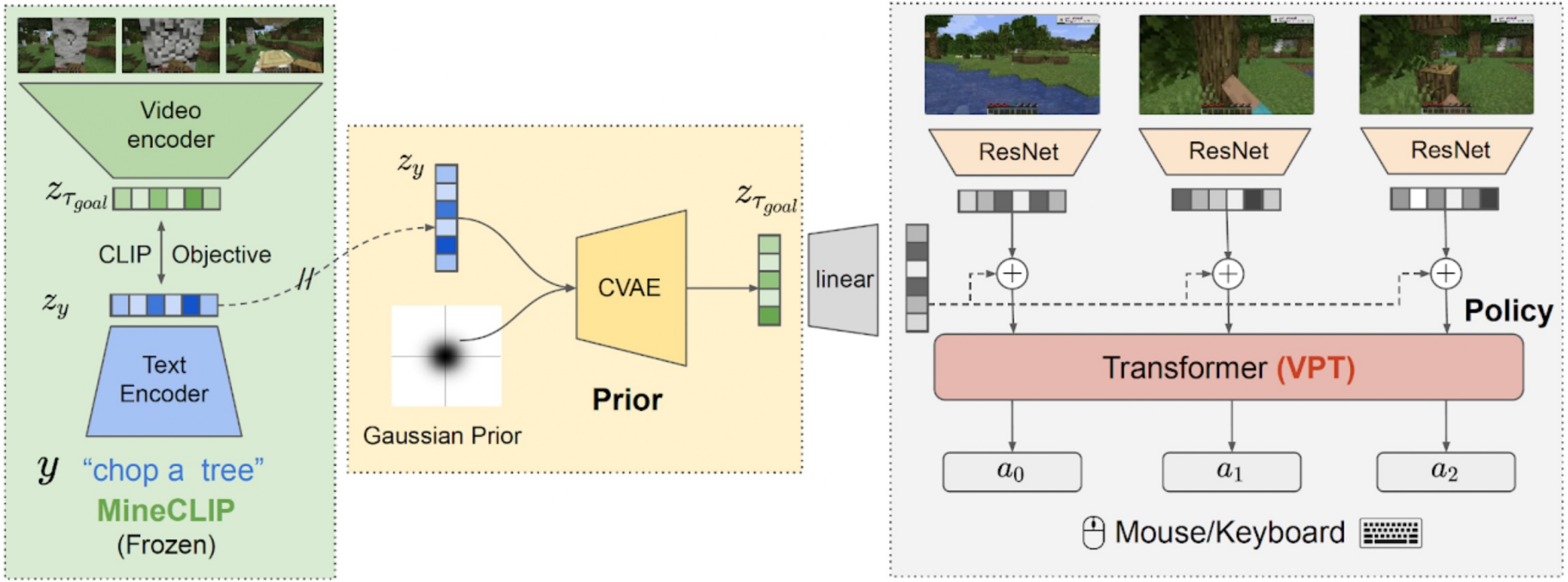

Text Memory
Text Memory + Reasoning
Multimodal Memory + Reasoning
baseline (no memory)
100%
80%
60%
40%
20%
0%
85%
95%
94%
20%
20%
34%
10%
30%
40%
0%
5%
9%
10%
20%
24%
Stone Pickaxe
Iron Pickaxe
Shield
Diamond
Redstone Block


Instruction
(Re-)Planner
LLM*
explain
Explainer
LLM*
plan P_t
description d_t
Selector
HPM
Descriptor
VLM
goal g_t
feedback
Controller
Goal-conditioned Policy
obs
action
obs
Environment

Open world embodied control: goal-aware representation learning and horizon prediction



Open world embodied control: pretraining and alignment

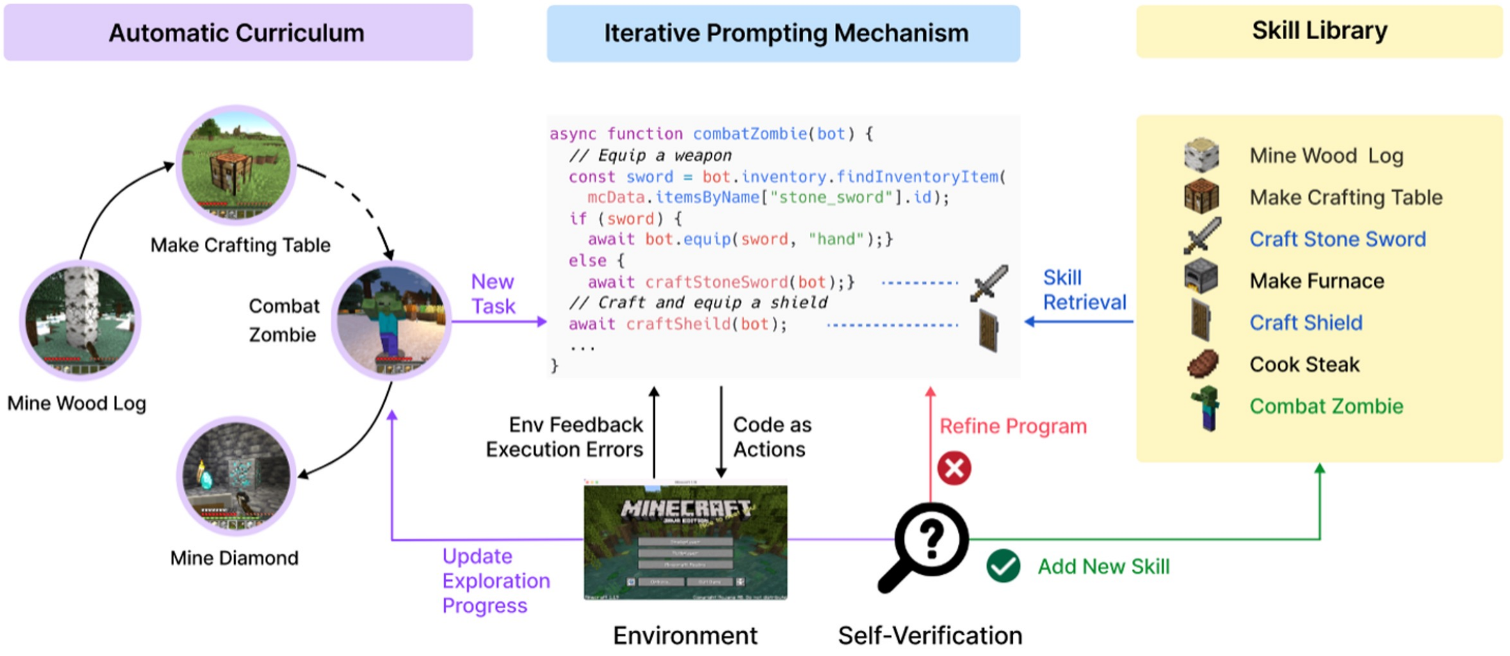


STEVE-1: A Generative Model for Text-to-Behavior in Minecraft


.26 .08 .18 .24 .22 .23 .02
.02 .06 .08 .05 .05 .05
.04 .14 .14 .02 .02 .05 .02 .06
.09
.06 .11 .10 .04 .11
.05 .02 .02 .04 .11
.37 .45 .50 .32 .50 .31 .32
.55
.34 .33 .32 .33 .40 .25
.38 .36 .32 .28 .36 .28 .36
.37 .30 .38 .33 .05 .25
.90 .67
.57
.97 .90 .87 .95 .94 .87 .91 .95 .88 .79 .84 .94
.91 .75 .97 .79 .92 .97 .96 .90 .89 .93 .92
.97

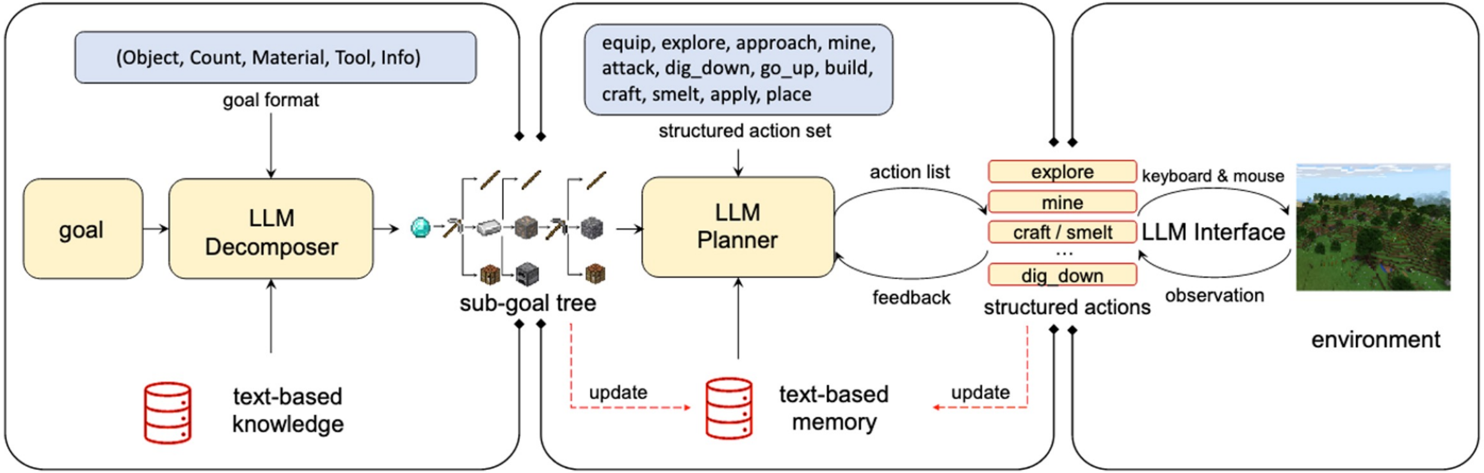
Some follow-up research projects built upon CraftJarvis

Voyager: GPT-4 based language agent



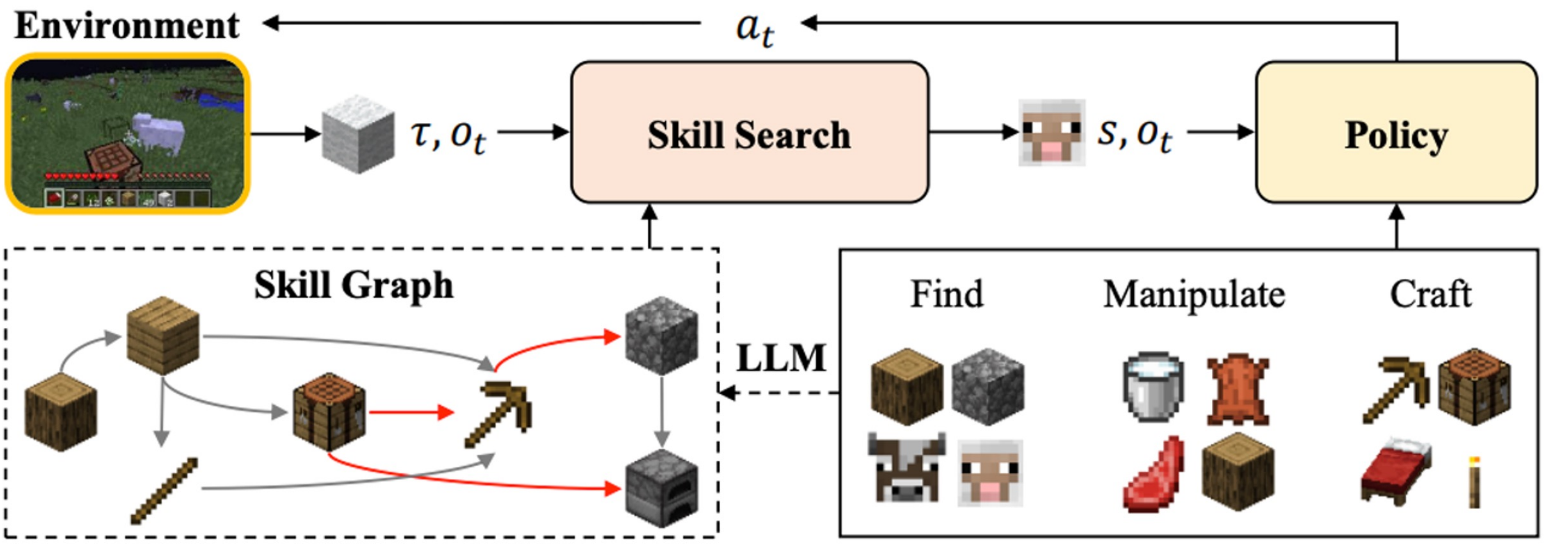
Voyager: An Open-Ended Embodied Agent with Large Language Models

GITM: language agent with structured knowledge library



Ghost in the Minecraft: Generally Capable Agents for Open-World Environments via Large Language Models with Text-based Knowledge and Memory

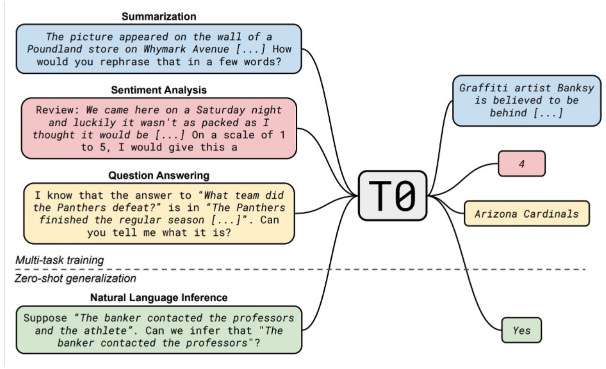
Plan4MC: language model + RL skills



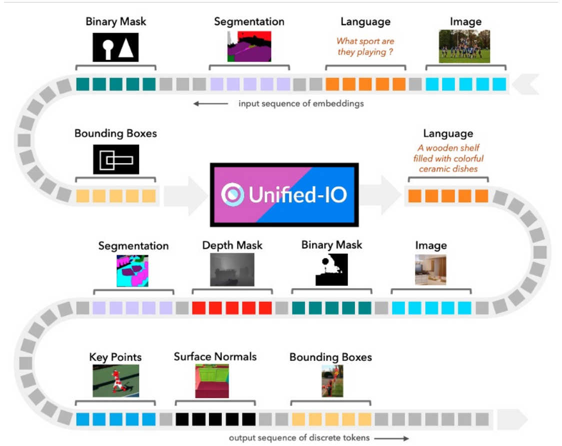
Plan4MC: Skill Reinforcement Learning and Planning for Open-World Minecraft Tasks

What’s next?

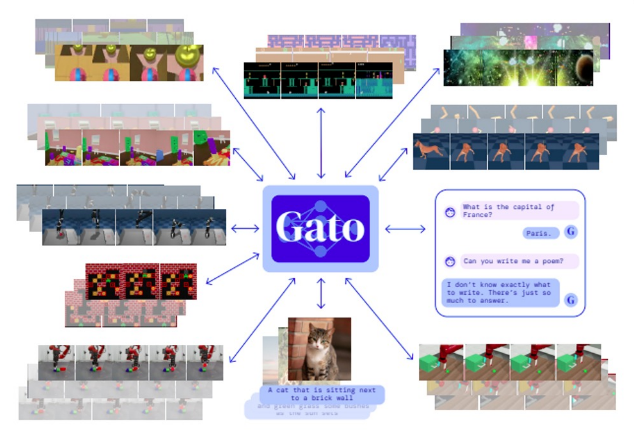
From unified models to unified agents



unified text models

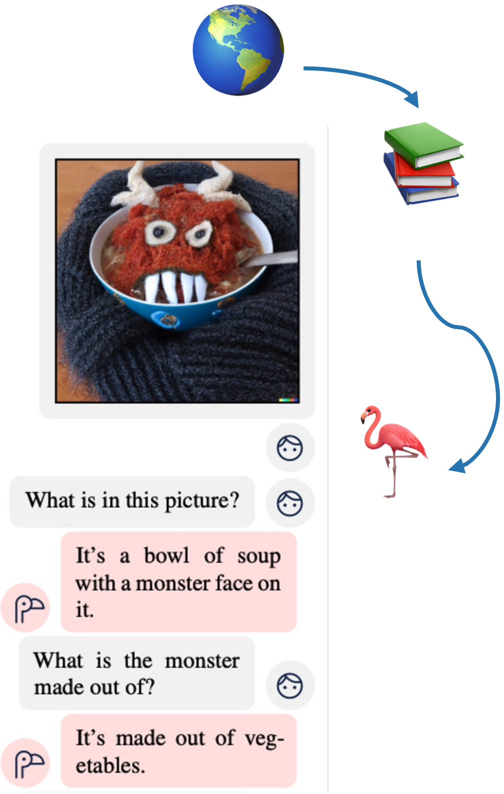


unified multimodal models



unified agents

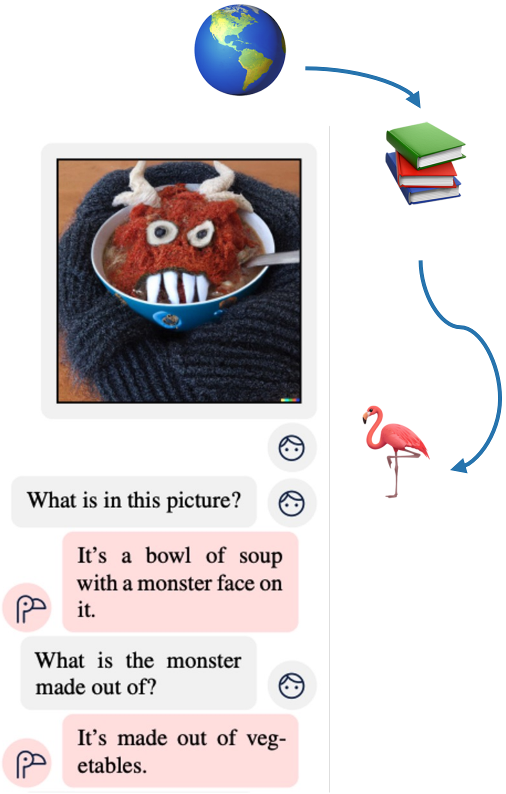
From unified models to unified agents



The following facts seems to be true:

1. Learning from massive web-scale data 📚
2. Large scale architecture O(10B) 🐘
3. Multi-tasking 🐙
4. (optional) Multimodal understanding 📖👀👂

From unified models to unified agents



The following facts seems to be true:

1. Learning from massive web-scale data 📚
2. Large scale architecture O(10B) 🐘
3. Multi-tasking 🐙
4. (optional) Multimodal understanding 📖👀👂

Definition

A language-piloted, large-scale agent that can fulfill arbitrary goals from multimodal input in embodied environments.

From unified models to unified agents

Few-shot learning

Enable few-shot (in-context) learning in these models?

More modalities

Unified models for other modalities (3D, egocentric videos, proprioception, high-res structured input, etc)?

Beyond zero-shot

Fine-tuning:
-Flexibility to new tasks/domains
-Preventing the acquired knowledge from being wiped out

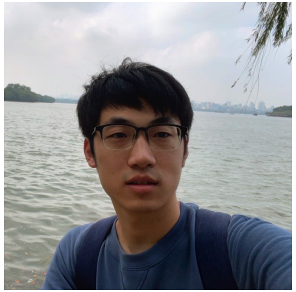
Road to agents

We need better learning algorithms for:
-episodic memory & situation awareness
-learning from interactions

unified text models

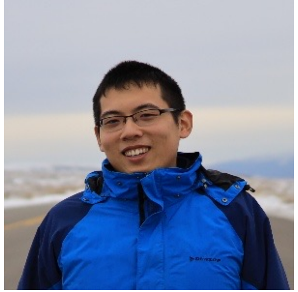
unified agents









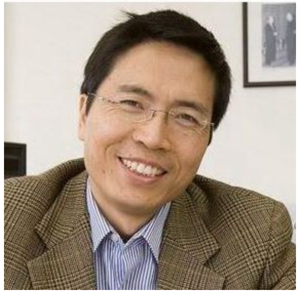



PEKING UNIVERSITY 1898

TSINGHUA UNIVERSITY 1911

UCLA



Thank you