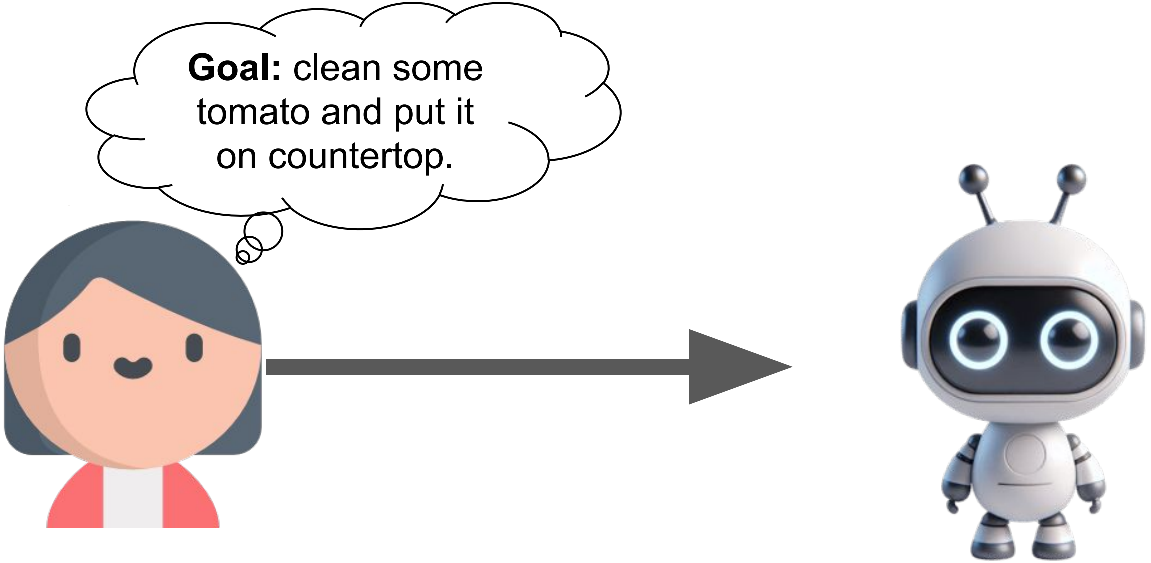
Instruction-based planning for embodied AI with LLMs and VLMs

Salim Aissi
Laure Soulier, Nicolas Thome, Olivier Sigaud

Instruction based planning



- The robot must:
  - Understand the task goal
  - Perceive and interpret the environment

Textual instruction based planning

Leveraging Large Language Models (LLMs), internal Common sense knowledge

<u>Goal:</u> clean some tomato and put it on countertop.

Textual Description
You open the fridge 1. The fridge 1 is open. In it, you see a egg 1, a mug 1, a tomato 1,.....

Planner (LLM)

Plan

- Go to fridge
- Open fridge
- Take tomato from fridge
- Close fridge
- Go to sinkbasin
- Clean tomato with sinkbasin
- Go to countertop
- Put tomato 1 in/on countertop

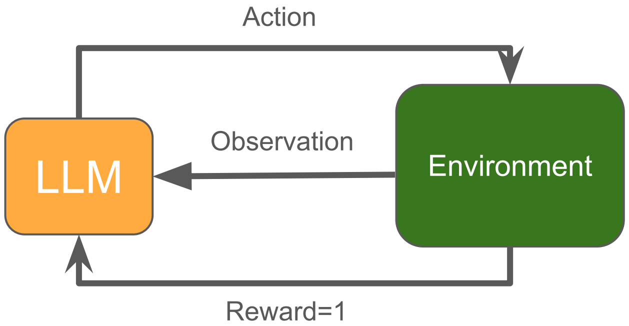
Textual instruction based planning

- **<u>BUT:</u> grounding issues:**
  - Affordance misunderstanding
  - Physics misunderstanding

LLM possible answer:
"You can use the egg to hammer the nail into the wall."

LLM possible answer:
"Yes, you can carefully balance the bowling ball on the soda can."

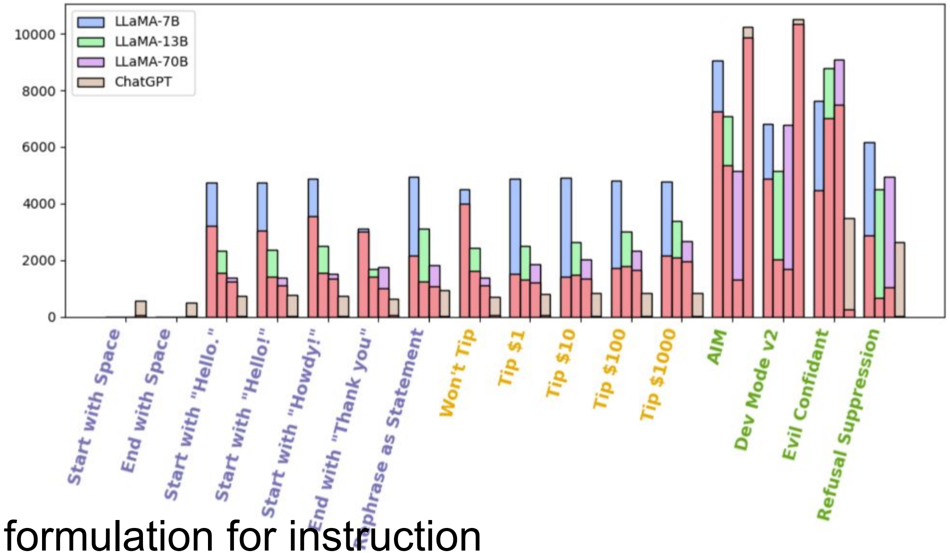
- Recent works on fine-tuning embodied agents with RL, e.g., PPO [1,2]
- **Open questions:**
  - Generalization
  - Sensitivity



[1] T. Carta, C. Romac, T. Wolf, S. Lamprier, O. Sigaud, and P.Y. Oudeyer. Grounding large language models in interactive environments with online reinforcement learning. ICML 2023
[2] W. Tan, W. Zhang, S. Liu, L. Zheng, X. Wang, and B. An. True knowledge comes from practice: Aligning llms with embodied environments via reinforcement learning. ICLR, 2024.

LLMs Prompt sensitivity

- Multiple studies have highlighted the sensitivity of LLMs to minor perturbations in the prompt, leading to substantially different outputs.



Our work:

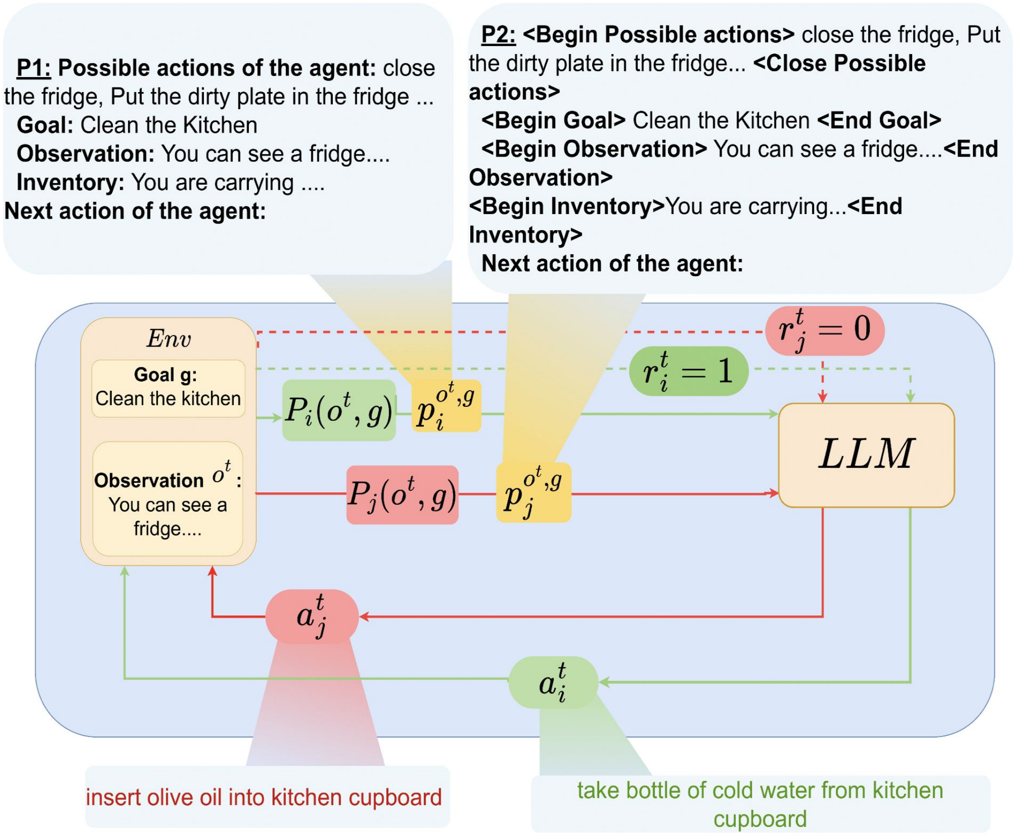
- Analyzing LLMs' performance wrt prompt formulation for instruction based planning.
- Detailed analysis of how the LLM processes each change
- Propose a solution to mitigate prompt overfitting.

[3] A. Salinas and F. Morstatter. The butterfly effect of altering prompts: How small changes and jailbreaks affect large language model performance. arXi, 2024.
[4] T. Z. Zhao, E. Wallace, S. Feng, D. Klein, S. Singh. Calibrate before use: Improv ing few-shot performance of language models. arXiv, 2021

Problem statement

- Partially Observable Markov Decision Process (POMDP) **M = (S, V, A, T , R, G, O, γ)**

- S is the state space.
- V is the language vocabulary
- A is the action space
- G is the goal space
- T is the transition function
- R the goal-conditioned reward function
- O is the observation function mapping a state to a textual description
- γ is the discount factor



Goal and observation formatted using a prompt formulation P_i

[5] Reinforcement Learning for Aligning LLM Agents: Quantifying and Mitigating Prompt Overfitting. S. Aissi, C. Romac, T. Carta, S. Lamprier, P.Y. Oudeyer, O. Sigaud, L. Soulier, N. Thome. NAACL 2025 Findings.

Prompt overfitting protocol

Prompt Strategy

P$_0$:
Possible actions of the agent: close the fridge, Put the dirty plate in the fridge ...
Goal: Clean the Kitchen
Observation: You can see a fridge. Empty! You can see ...
Inventory: You are carrying
Next action of the agent:

P$_1$:
Goal: Clean the Kitchen
Inventory: You are carrying...
Observation: You can see a fridge. Empty! You can see ...
Possible actions of the agent: close the fridge and put the dirty
Next action of the agent:

P$_2$:
<Begin Possible actions> close the fridge, Put the dirty plate in the fridge...
<Close Possible actions>
<Begin Goal> Clean the Kitchen **<End Goal>**
<Begin Observation> You can see a fridge. Empty! You can see ...
<End Observation>
<Begin Inventory: > You are carrying... **<End Inventory: >**
Next action of the agent:

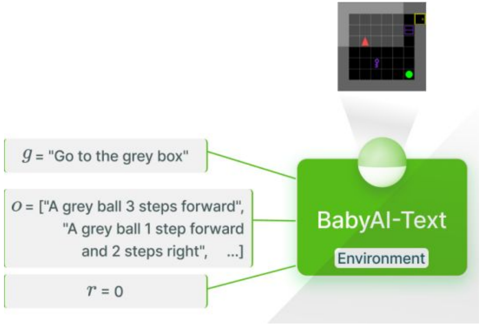
P$_3$: Welcome to TextWorld! You find yourself in a messy house...What you can do is to close the fridge, Put the dirty plate in the fridge...Your goal is to clean the Kitchen. You can see a fridge....now, You are carrying nothing., and your next action is to :

switch in order

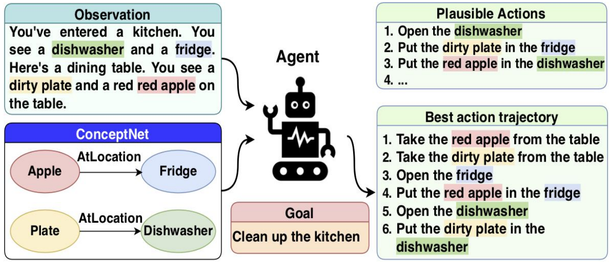
Rigid syntaxe

Paraphrase Natural Language

Environments



Baby Ai Text: An environment that requires exploration and understanding of the positions of objects



Text World Common sense: An environment that requires commonsense knowledge about the world.

Prompt overfitting protocol

Prompt Strategy

P_0:
Possible actions of the agent: close the fridge, Put the dirty plate in the fridge ...
Goal: Clean the Kitchen
Observation: You can see a fridge. Empty! You can see ...
Inventory: You are carrying
Next action of the agent:

P_1:
Goal: Clean the Kitchen
Inventory: You are carrying...
Observation: You can see a fridge. Empty! You can see ...
Possible actions of the agent: close the fridge and put the dirty
Next action of the agent:

P_2:
<Begin Possible actions> close the fridge, Put the dirty plate in the fridge...
<Close Possible actions>
<Begin Goal> Clean the Kitchen **<End Goal>**
<Begin Observation> You can see a fridge. Empty! You can see ... **<End Observation>**
<Begin Inventory: > You are carrying... **<End Inventory: >**
Next action of the agent:

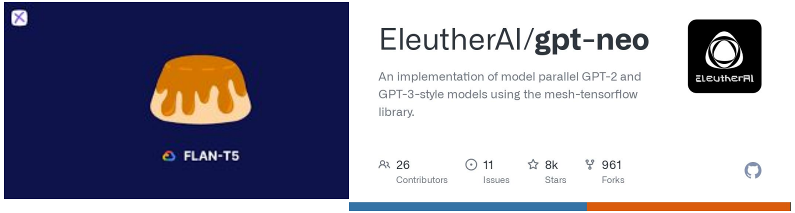
P_3 : Welcome to TextWorld! You find yourself in a messy house...What you can do is to close the fridge, Put the dirty plate in the fridge...Your goal is to clean the Kitchen. You can see a fridge....now, You are carrying nothing., and your next action is to :

switch in order

Rigid syntaxe

Paraphrase Natural Language

Training & Evaluation Scenarios



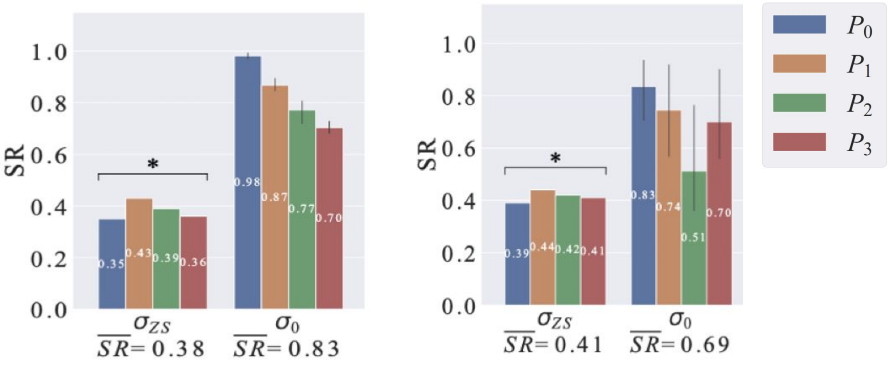
- -Zero shot evaluation
- -Train with one Strategy and Evaluate with others
- -Train on All Strategy

Experimental results: prompt sensitivity

Metrics:

- Success Rate (SR): $ESR = \frac{n_e}{N}$
- Mean Episode Length

Results for FLAN T5 780M



TWC Medium

BabyAI Text

- RL boost performances
- **But strong overfitting to the prompt**
- Similar trend with GPT-neo

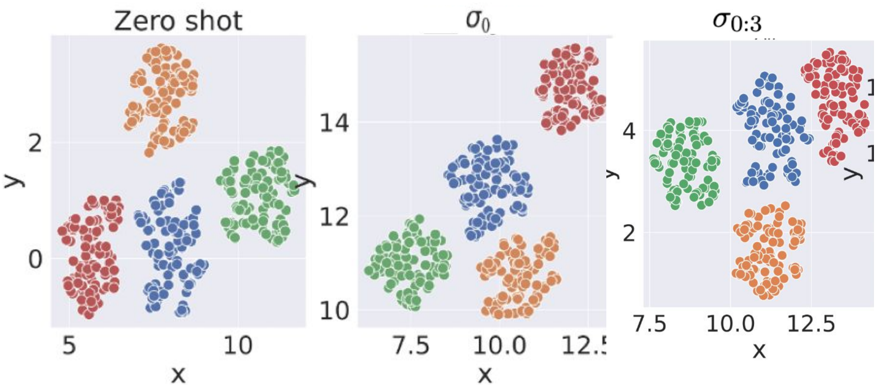
Experimental results: state representation in LLMs

$$Intra(P_i) = \frac{1}{|\Gamma|^2 - |\Gamma|} \sum_{\substack{(o,g)\in\Gamma \\ (o',g')\in\Gamma\setminus\{o,g\}}} cos\left(z_i^{o,g}, z_i^{o',g'}\right)$$

$$Inter(P_i, P_j) = \frac{1}{|\Gamma|} \sum_{(o,s)\in\Gamma} cos\left(z_i^{o,g}, z_j^{o,g}\right)$$

- Intra(P_i) ≃ 1 => different state (goal and observation) but same prompt
- Inter(P_i , P_j) < 0.5 => same states with different prompt

Models	78M		780M	
	$Intra(P_i)$	$Inter(P_i, P_j)$	$Intra(P_i)$	$Inter(P_i, P_j)$
Zero-shot	0.992 $\pm$ 0.003	0.376 $\pm$ 0.019	0.998 $\pm$ 0.001	0.469 $\pm$ 0.462
σ_0	0.991 $\pm$ 0.003	0.382 $\pm$ 0.020	0.997 $\pm$ 0.001	0.458 $\pm$ 0.449
$\sigma_{0:3}$	0.991 $\pm$ 0.003	0.371 $\pm$ 0.020	0.998 $\pm$ 0.001	0.47 $\pm$ 0.461

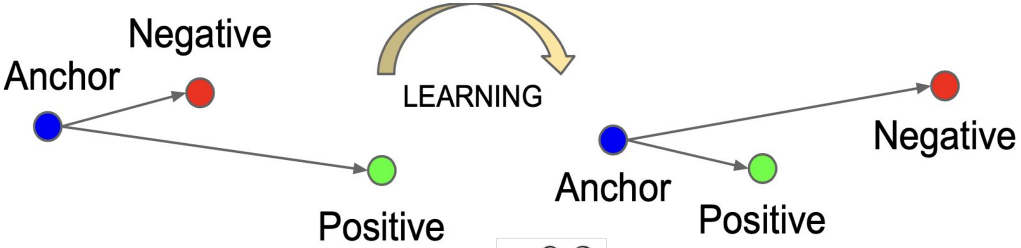


=> Learned representations: encode prompt but not useful information (state)!

Experimental results: state representation in LLMs

contrastive loss enforcing:

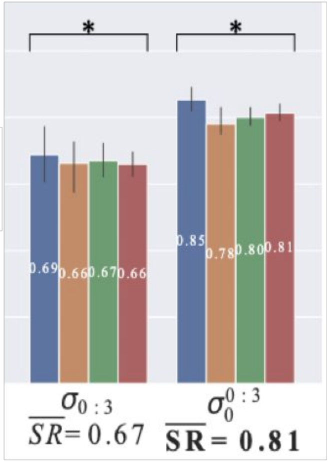
$$\underbrace{\Delta(z_\theta(p_i^{o,g}), z_\theta(p_j^{o,g}))}_{\text{Same state, different prompts}} \leq \underbrace{\Delta(z_\theta(p_i^{o,g}), z_\theta(p_i^{o',g'}))}_{\text{Same prompt, different states}} + 1$$



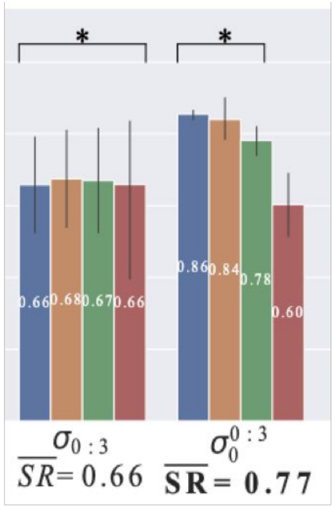
- Trained with PPO on a single prompt $\sigma_0^{0:3}$
- Compa to models trained on all prompts $\sigma_{0:3}$

Mitigating prompt overfitting: results

TWC Medium



BabyAI Text



Contrastive $\sigma_0^{0:3}$ **:** same homogeneity, better performance than $\sigma_{0:3}$

	$\sigma_{0:3}$	$\sigma_0^{0:3}$
78 M	0.77 ± 0.11 (3%)	0.92 ± 0.02 (97%)
780 M	0.80 ± 0.06 (4.7%)	0.86 ± 0.05 (91%)
1.3 B	0.66 ± 0.02 (99%)	0.76 ± 0.03 (98%)

Much better results in zero-shot prompt transfer!

- Prompt P_4 unseen during training

x

Visual instruction based planning

<u>Goal:</u> clean some tomato and put it on countertop.

Planner

Plan

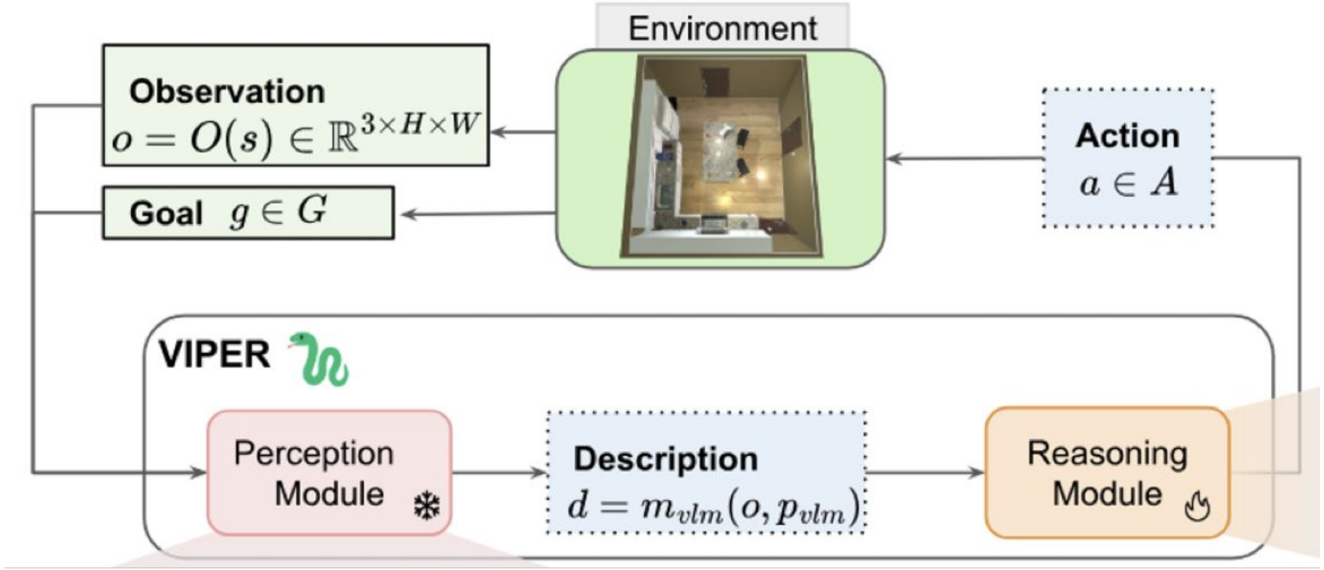
- Go to fridge
- Open fridge
- Take tomato from fridge
- Close fridge
- Go to sinkbasin
- Clean tomato with sinkbasin
- Go to countertop
- Put tomato 1 in/on countertop

Visual instruction based planning: Alfworld

- Both image and detailed textual description of each image
- Successful attempts for using LLMs, either with RL or filtering relevant actions
- However, performances with VLMs from visual inputs are way less successful, e.g., limited performances in recent works RL4VLM [6]
  - EMMA [7] uses textual guidance, cumbersome and unrealistic requirement

[6] Y. Zhai, H. Bai, Z. Lin, J. Pan, S. Tong, Y. Zhou, A. Suhr, S. Xie, Y.LeCun, Y. Ma, S. Levine. Fine-tuning large vision-language models as decision-making agents via reinforcement learning. NeurIPS 2024.
[7] Y. Yang, T. Zhou, K. Li, D. Tao, L. Li, L. Shen, X. He, J. Jiang, Y. Shi. Embodied multi-modal agent trained by an llm from a parallel textworld. CVPR 2024.

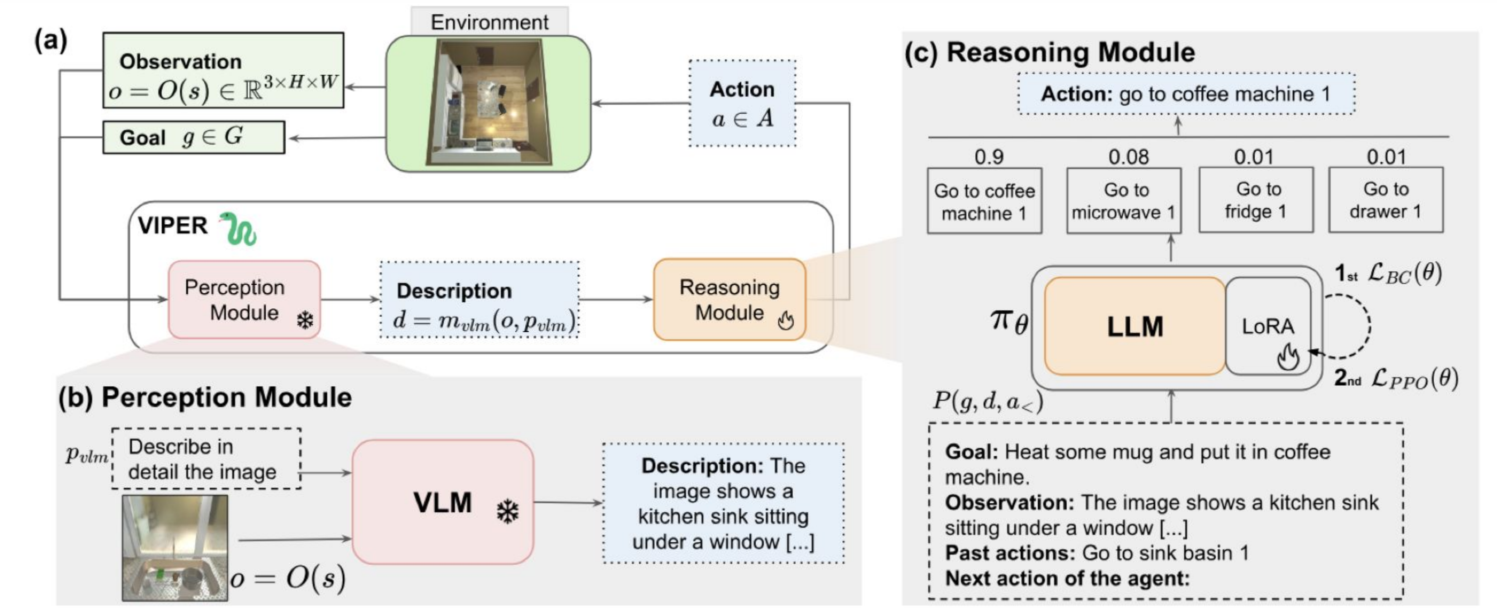
VIPER: Visual Perception and Explainable Reasoning for Sequential Decision-Making



- Use text as intermediate for visual instruction-based planning
- Reasoning module fine tuned with Behavioral Coining (BC) and RL
- Intermediate text => rich monitoring potential

[9] VIPER: Visual Perception and Explainable Reasoning for Sequential Decision-Making S. Aissi, C. Grislain, M. Chetouani, O. Sigaud, L. Soulier, N. Thome.

VIPER architecture

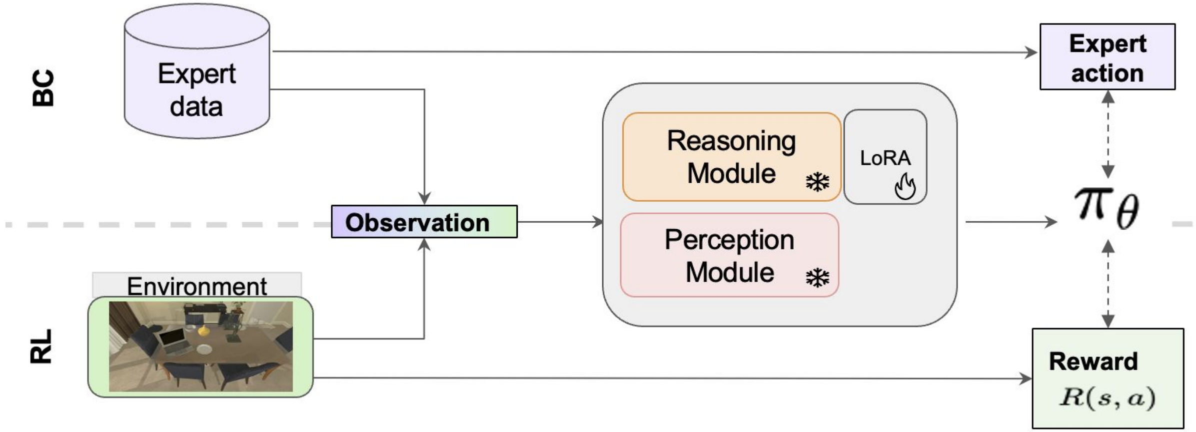


- Perception: frozen VLM
- Reasoning: prediction in support of possible actions, as in GLAM [8] but ≠ RL4VLM [12]

VIPER training

Fine-tune the reasoning module (LLM) with sequential training strategy

- **1. Supervised fine-tuning** : behavioural cloning
  - Using a rule-based expert
- **2. Online fine-tuning** : reinforcement learning (PPO)
  - Interaction with the environment and reward feedback



Experiments: AlfWorld

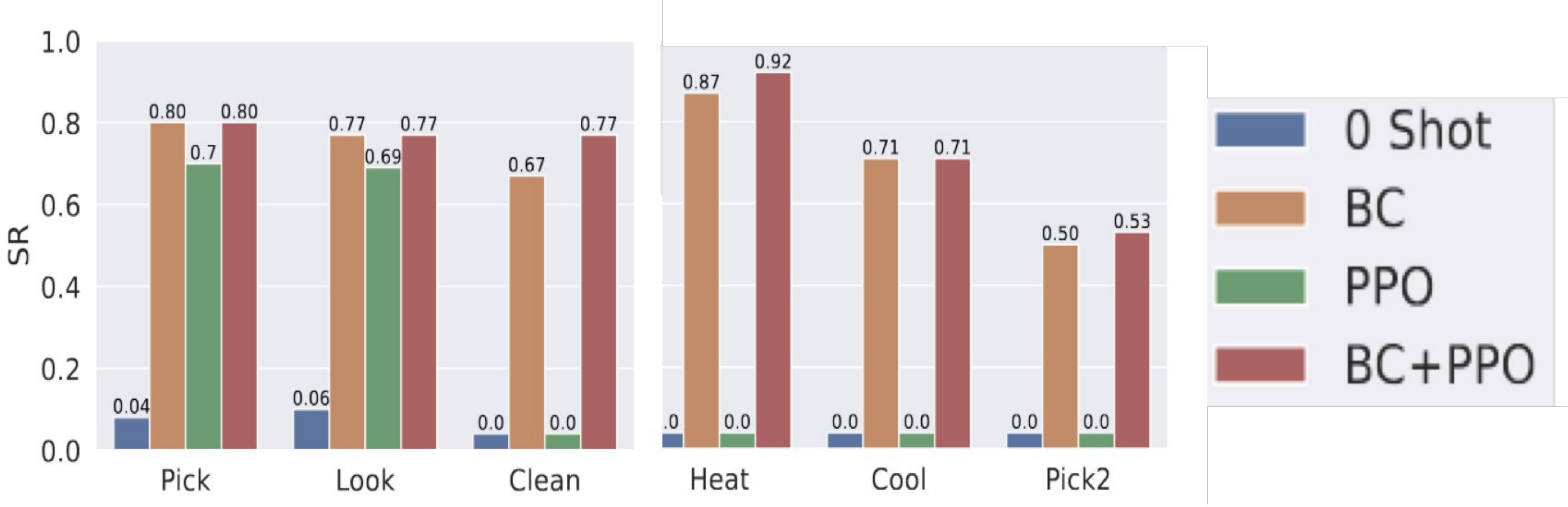
- Baselines
  - Oracle methods use textual observation (training and/or inference)
  - Zero-shot or fine-tuned VLM agents
- VIPER: +50 pts and reduces the gap with oracle methods

	Observation	Pick	Look	Clean	Heat	Cool	Pick2	Avg
AutoGen* [27]	📝	0.92 (-)	0.83 (-)	0.74 (-)	0.78 (-)	0.86 (-)	0.41 (-)	0.77 (-)
ReAct* [30]	📝	0.71 (18.1)	0.28 (23.7)	0.65 (18.8)	0.62 (18.2)	0.44 (23.2)	0.35 (25.5)	0.54 (20.6)
DEPS* [25]	📝	0.93 (-)	1.00 (-)	0.50 (-)	0.80 (-)	1.00 (-)	0.00 (-)	0.76 (-)
Reflexion* [19]	📝	0.96 (17.4)	0.94 (16.9)	1.00 (17.0)	0.81 (19.4)	0.83 (21.6)	0.88 (21.6)	0.91 (18.7)
EMMA* [29]	📝→👁	0.71 (19.3)	0.88 (19.6)	0.94 (17.5)	0.85 (19.6)	0.83 (19.9)	0.67 (22.4)	0.82 (19.5)
Florence-2 [28]	👁	0.00 (30.0)	0.06 (28.5)	0.0 (30.0)	0.0 (30.0)	0.0 (30.0)	0.0 (30.0)	0.01 (29.7)
Idefics-2 [13]	👁	0.04 (29.2)	0.06 (28.2)	0.0 (30.0)	0.0 (30.0)	0.0 (30.0)	0.0 (30.0)	0.02 (29.5)
MiniGPT-4* [36]	👁	0.04 (29.0)	0.17 (17.7)	0.0 (30.0)	0.19 (26.3)	0.17 (26.7)	0.06(28.9)	0.16 (26.9)
InstructBLIP* [4]	👁	0.50 (21.5)	0.17 (26.8)	0.26 (25.0)	0.23 (27.2)	0.06 (28.9)	0 (30.0)	0.22 (26.2)
RL4VLM* [33]	👁	0.47 (-)	0.14 (-)	0.10 (-)	0.14 (-)	0.18 (-)	0.18 (-)	0.21 (-)
VIPER	👁	0.80 (13.1)	0.77 (16.7)	0.77 (19.5)	0.92 (14.3)	0.71 (20.4)	0.53 (24.0)	0.75 (18.0)

Demo: AlfWorld

Experiments: impact of BC/RL training

RL works for simple tasks, but need BC guidance on more complex ones



Failure analysis and monitoring

Goal: Heat some egg and put it on countertop

Description: The image shows a table with a variety of items on it, including an **apple** and other **objects**. In the [...]

Action: Take apple 1 from dining table 2

Perception error: missed egg

Failure analysis and monitoring

Goal: Your task is to: put some box on dresser

t=1

Current Observation: This is an animated image. In this image we can see a sofa. On the sofa there is a box and a paper. In the background there is wall.

Next Action: go to sofa1

t=5

Current Observation: In this image we can see a chair on the floor. On the right side of the image there is a table on which a lamp is placed. We can also see a cardboard box placed on the table. In the background there is wall.

Next Action: take box 1 from garbagecan 1

t=10

Current Observation: This is an animated image. In this image we can see a monitor and a keyboard on the table.

Next Action: go to garbagecan 1

t=16

Current Observation: This is an animated image. In this image we can see a chair, bed, table, lamp and wall. At the bottom of the image there is a floor

Next Action: go to dresser 1

t=40

Current Observation: This is an animated image. In this image we can see a monitor and a keyboard on the table.

Next Action: go to garbagecan 1

t=50

Current Observation: This is an animated image. In this image we can see a chair, bed, table, lamp and wall. At the bottom of the image there is a floor

Next Action: go to dresser 1

Reasoning error: box detected

Perspective & future works

- Text as intermediate representation in instruction-based planning
  - Enhancing perception through the interaction with reasoning (example VQA)
  - Applying continual learning methodologies across diverse tasks to enable incremental knowledge acquisition