Agent with the Big Picture: Perceiving Surroundings for Interactive Instruction Following

Byeonghwi Kim[1] Suvaansh Bhambri[1] Kunal Pratap Singh[2] Roozbeh Mottaghi[2] Jonghyun Choi[1]

GIST[1] AI2[2]

CVPR VIRTUAL JUNE 19-25

Problem Setup

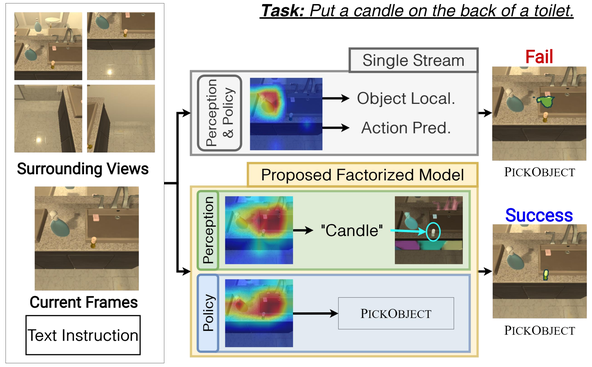
- Given **natural language instructions** and **egocentric RGB observations**, complete tasks by predicting a sequence of **actions** and **masks** for object interaction.

[Goal statement]
Put the heated mug in the coffee maker.

[Step-by-step instructions]
Turn around and walk through the kitchen, turning right to face the kitchen island opposite the sink. Pick up the white mug from the kitchen island. Turn right and walk over to the microwave above the oven. Heat the mug in the microwave then remove it. Turn right and walk over the coffee maker on the counter. Put the mug in the coffee maker.

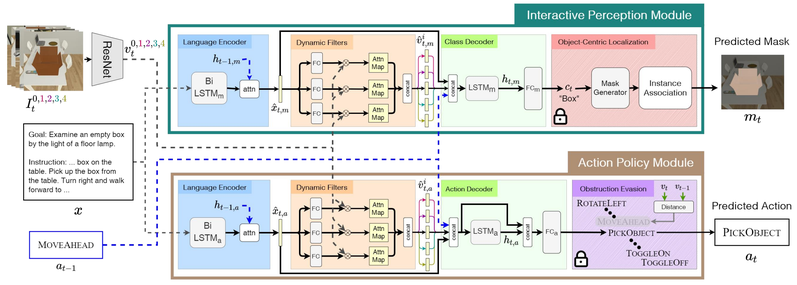
Main Idea

- A small FOV often limits the understanding of an environment.
  -> Perceive **surrounding views** to acquire rich information.
- The agent address tasks that require different semantics.
  -> Factorize the tasks into **separate streams**.

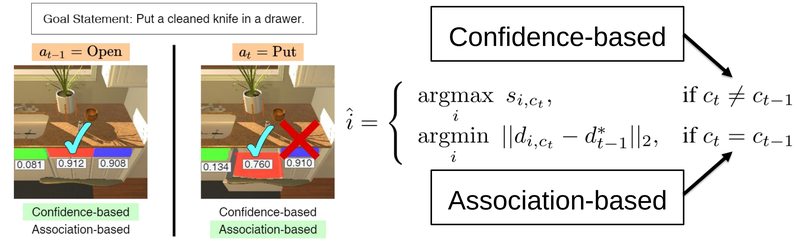


Method

Model Architecture

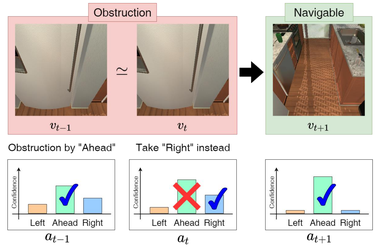


Instance Association for Object-Centric Localization



$$\hat{i} = \begin{cases} \underset{i}{\text{argmax}}\ s_{i,c_t}, & \text{if } c_t \neq c_{t-1} \\ \underset{i}{\text{argmin}}\ ||d_{i,c_t} - d^*_{t-1}||_2, & \text{if } c_t = c_{t-1} \end{cases}$$

Obstruction Evasion



If $d(v_{t-1}, v_t) < \epsilon$, then

$$a_t = \underset{k}{\text{argmax}}\, \text{FC}_a([u_a; h_{t,a}]), \quad k \in [1, N_{action}] - \{k'\}$$

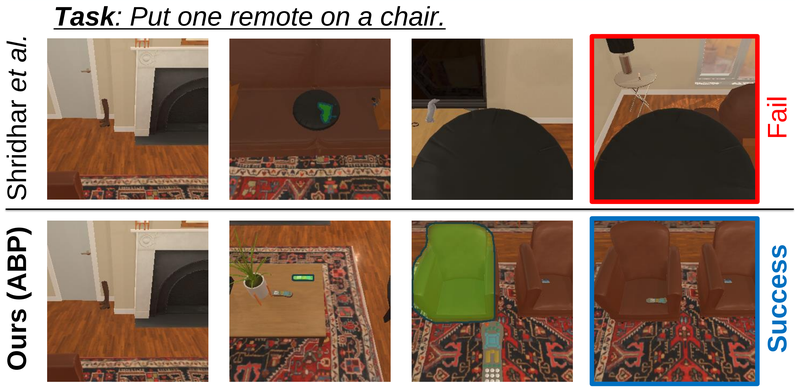
Action that causes **obstruction**

Experiments

Quantitative Analysis

Split	Model	Seen		Unseen	
		Task	Goal-Cond	Task	Goal-Cond
Val.	Shridhar *et al.* [8]	4.00 (2.10)	10.50 (7.20)	0.20 (0.10)	7.50 (5.10)
	LWIT [7]	33.70 (28.40)	43.10 (38.00)	9.70 (7.30)	23.10 (18.10)
	Ours (ABP)	**42.93** (3.84)	**50.45** (4.76)	**12.55** (1.05)	**25.19** (2.25)
Test	Shridhar *et al.* [8]	3.98 (2.02)	9.42 (6.27)	0.39 (0.08)	7.03 (4.26)
	LWIT [7]	29.16 (24.67)	38.82 (34.85)	8.37 (5.06)	19.13 (14.81)
	LWIT [7]*	30.92 (25.90)	40.53 (36.76)	9.42 (5.60)	20.91 (16.34)
	Ours (ABP)	**44.55** (3.88)	**51.13** (4.92)	**15.43** (1.08)	**24.76** (2.22)

Qualitative Analysis



Summary

- We explore the problem of interactive instruction following.
- Our model exploits surrounding views for rich information.
- We factorize the task into two streams, interactive perception and action policy with the improved components.