

User Instruction
I feel like tomato soup tonight

Current observation



I_t

I_t^m



Large Multimodal Model (GPT-4v)



\mathcal{F}^{ground}

CoT: Looking for a soup, potentially a canned item.
Target object: [6]

\mathcal{F}^{plan}

CoT: The marker [1] is blocking the can [6].
Plan:

1. remove [1]
2. pick [6]

\tilde{n} Region of Interest



$c_{\tilde{n}}$

Grasp Markers

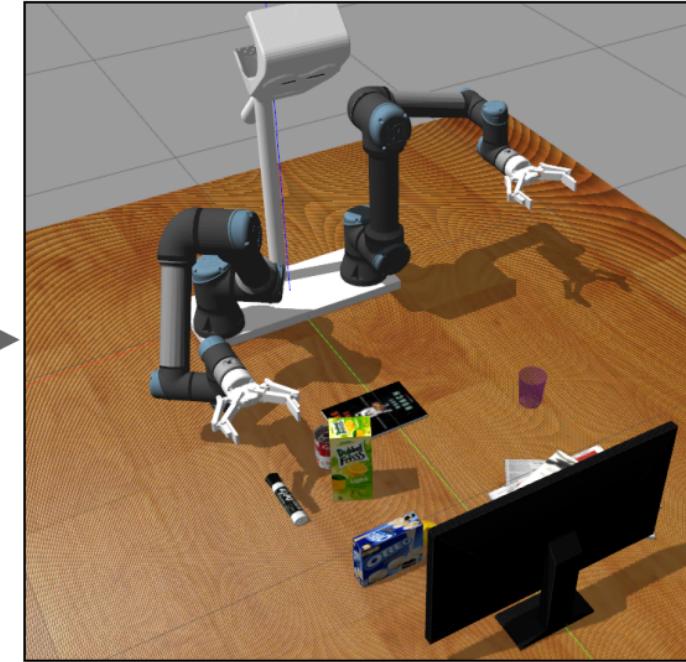


$c'_{\tilde{n}}$

\mathcal{F}^{rank}

CoT: Grasp near the marker body, avoid the can,
Ranked grasps: [3, 4, 2, 1, 5, 6]

Update observation



World state after robot execution