

$s_0$  $a_0$  $r_0$ 

...

 $s_k$  $a_k$  $r_k$  $s_{k+1}$ 

...

 $r_{K-1}$ 

Trajectory Buffer

Reward

Environment

+

Visual Reasoning

+

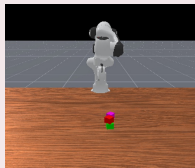
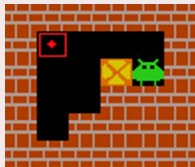
Format

State



You are a home robot and perform navigation tasks according to instructions.  
...

Environment



Rollout Loop

VLM-Agent

Action

&lt;think&gt;

<observation>There is a plate on the dining table to the right.</observation>

<reasoning> First, I should move forward to get closer to the table. Then, I can move to the right to be in front of the table.</reasoning>

<prediction>I will move closer to the table.</prediction>

&lt;/think&gt;

<answer>moveahead, moveright</answer>

Training Loop

Policy Update

PPO

+

Bi-Level  
Advantage  
Estimation