① Input

3D Encoder

Visual Features V_1

2D Encoder

Selection

Query Feature F_1

Text Encoder

Text Feature F_1

Point	Cloud	Depth	Map	of	a	[CLASS]

Refrigerator

② Multimodal RAG

a_1 a_2 a_3

Retrieval
Rerank Img+Text

Goal

{ Refrigerator (14, 6)}

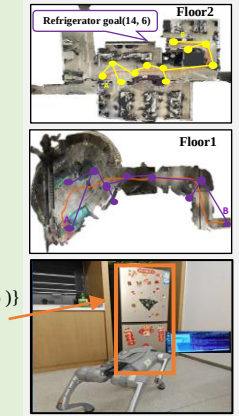
Date Vector

Graph

Multimodal Knowledge DateBase

Update

③ Embodied Planning



Real World

3D Localization

Object Goal Nevigation

Embodied Action Selection

EmbodiedVLM(1.7B)

Embodied Action Selection

SFT-InternLM

PPO-Clip

Action Instruction:

- ☑ **Climb Stairs**
- ○ **Turn Right**
- ○ **Turn Left**
- ○ **Move Forward**
- ○ **Down Stairs**
- ○ **StandDown**

……

I. Text Input: I want to go to the refrigerator

II. Output Result: This is a decorated refrigerator, located at coordinates (14, 6)