See and Think: Embodied Agent in Virtual Environment

Zhonghan Zhao[1*], Wenhao Chai[2*], Xuan Wang[1], Boyi Li[1], Shengyu Hao[1], Shidong Cao[1], Tian Ye[3], and Gaoang Wang[1†]

1 Zhejiang University 2 University of Washington 3 Hong Kong University of Science and Technology (GZ)

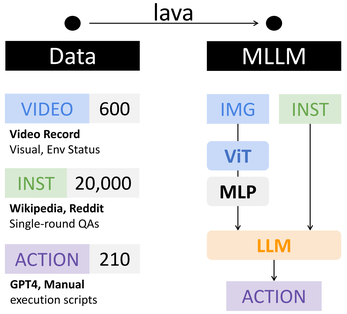
Motivation & Contribution

Motivation

Large language models (LLMs) have made significant strides in handling open-world tasks. However, prior to our work, they lacked natural perceptual abilities for effective instruction interpretation, with scant research focusing on how visual perception impacts performance in open-ended tasks. This deficiency in natural perception has limited the developmental potential of embodied LLM agents.

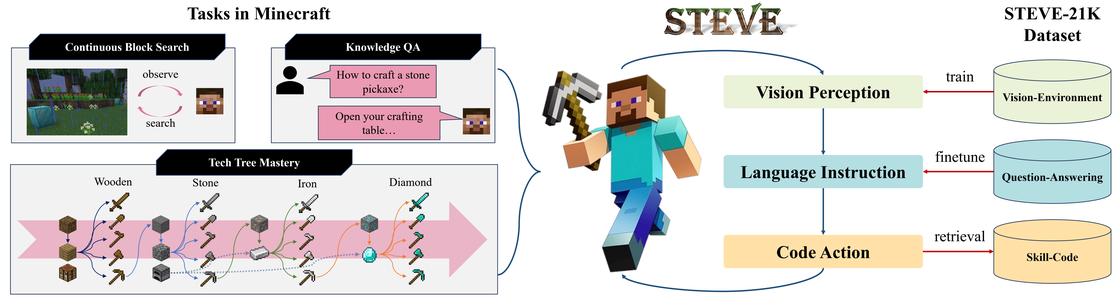
Contribution

- STEVE is the first pure LLM-based embodied agent with vision perception and comprehensive control capabilities. STEVE includes vision perception, language instruction, and code action, achieving 1.5× faster unlocking of key tech trees and is 2.3× quicker in block search tasks compared to previous SOTA methods.

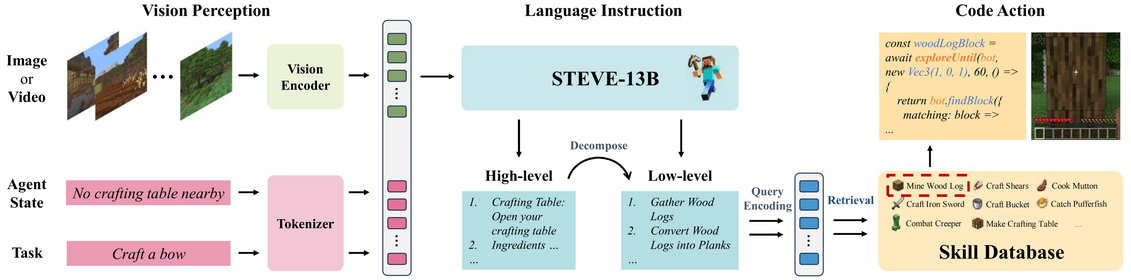


- STEVE-7B/13B is a series of multi-modal large language model obtained by fine-tuning with Minecraft knowledge question-answering pairs from Llama2-7B/13B.
- STEVE-21K dataset, includes 600+ vision-environment pairs, 20K knowledge question-answering pairs, and 200+ skill-code pairs, for justifying the effective performance of STEVE

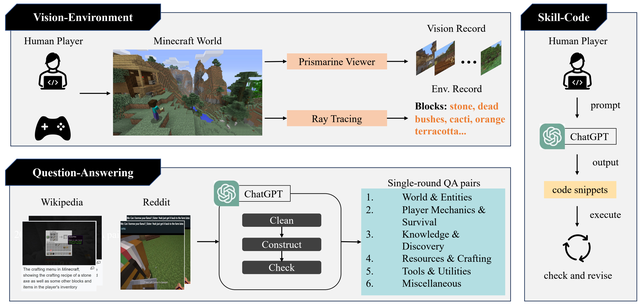
Overview



Method



Dataset: STEVE-21K



Environment [Blocks] stone, dead bushes, cacti, orange terracotta, ...
[Equipment] iron sword
[Entities] creeper, Rabbit
[Health] 15/20
[Hunger] 18/20
[Inventory] Carrots, torches

Context
[Question] How to survive, according to the current situation?
[Answer] Dealing with the Creeper and Health Management...

Chat Log
A Creeper is Nearby, please combat Creeper with the iron sword first.

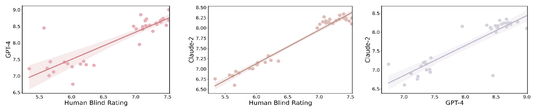
Plan
Use '**combatOneCreeper**' function to combat the Creeper until the Creeper is eliminated. Once the Creeper is eliminated, call the '**eatCarrot**' ...

Case Study

Experiment

Knowledge QA

Method	Wld. & Ent.	Mech. & Surv.	Know. & Disc.	Res. & Craft.	Tl. & Util.	Miscell.	Overall
Llama2-7B	6.44	6.68	6.58	6.42	6.80	6.96	6.56
Llama2-13B	6.93	6.95	6.77	6.77	6.98	6.64	6.89
STEVE-7B	7.99	7.88	7.84	7.95	7.93	7.82	7.94
STEVE-13B	**8.14**	**8.13**	8.03	**8.15**	**8.12**	7.72	**8.12**
GPT-4	8.06	8.07	**8.07**	7.92	8.09	**8.21**	8.04

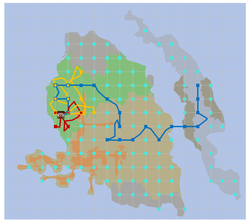


Tech Tree Mastery

Method	Wooden Tool	Stone Tool	Iron Tool	Diamond Tool
AutoGPT [49]	92 ± 72 (3/3)	94 ± 72 (3/3)	135 ± 103 (3/3)	N/A (0/3)
Voyager [57]	6 ± 2 (3/3)	11 ± 2 (3/3)	21 ± 7 (3/3)	**102** (1/3)
STEVE	**4 ± 1** (3/3)	**8 ± 1** (3/3)	**15 ± 2** (3/3)	106 ± 12 (3/3)

Method	Wooden Tool	Stone Tool	Iron Tool	Diamond Tool
w/o vision unit	11 ± 5 (3/3)	27 ± 5 (3/3)	46 ± 11 (3/3)	158 (1/3)
STEVE (GPT-4)	6 ± 2 (3/3)	10 ± 1 (3/3)	**14 ± 3** (3/3)	**89 ± 9** (3/3)
STEVE (Ours)	**4 ± 1** (3/3)	**8 ± 1** (3/3)	15 ± 2 (3/3)	106 ± 12 (3/3)

Continuous Block Search



Component Comparison

	VPT [1]	DreamerV3 [5]	DECKARD [7]	DEPS [11]	Plan4MC [13]	Voyager [9]	STEVE (ours)
Demos	Videos	None	Videos	None	None	None	Videos
Rewards	Sparse	Dense	Sparse	None	Dense	None	None
Observations	Pixels Only	Pixels & Meta	Pixels & Inventory	Feedback & Inventory	Pixels & Meta	Feedback & Meta & Inventory	Pixels & Feedback & Meta & Inventory
Actions	Keyboard & Mouse	Discrete	Keyboard & Mouse	Keyboard & Mouse	Discrete	Code	Code
Iterative Planning				✓		✓	✓
Skill Database					9	172	210
Gradient-Free						✓	✓