How Do Multimodal Large Language Models Handle Complex Multimodal Reasoning? Placing Them in An Extensible Escape Game

Ziyue Wang*[1], Yurui Dong*[2], Fuwen Luo[1], Minyuan Ruan[1], Zhili Cheng[1], Chi Chen[1], Peng Li[1], Yang Liu[1]
[1]Tsinghua University, [2]Fudan University

ICCV OCT 19-23, 2025 HONOLULU HAWAII

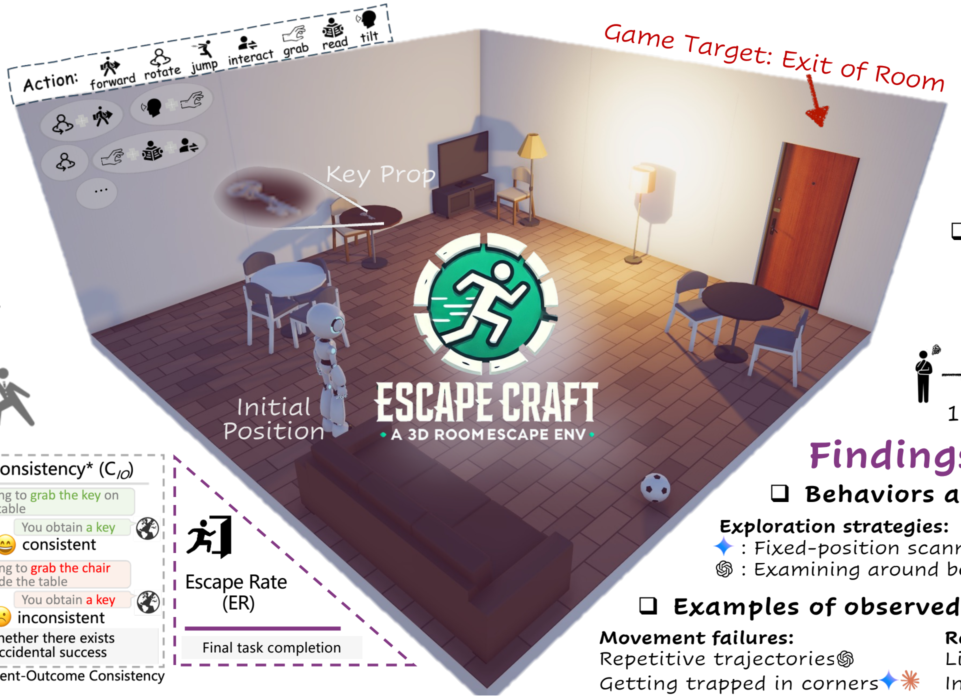
Motivation

□ **Environment – EscapeCraft**

- ✓ **Autonomous** exploration
- ✓ Integration of **multiple basic abilities**, such as visual search, visual reasoning, and long-term planning etc.

□ **Benchmark – MM-Escape**

- ✓ Evaluate **model behaviors** and **exploration pattern**
- ✓ Emphasize **reasoning process** beyond completion rate
- ✓ Provide insights of **intermediate reward signals** for planning, acting and reasoning tasks


Action: forward rotate jump interact grab read tilt
...
Game Target: Exit of Room
Key Prop
Initial Position
ESCAPE CRAFT
A 3D ROOMESCAPE ENV

Evaluation on MM-Escape

□ **Metrics**

Average Steps (S): Efficiency of exploration and escape

Prop Gain

$$\text{Prop Gain} = \frac{N_{grab}^{TP}}{\sum \text{Prop count}},$$

If all the tools and clues are obtained

Grab Success Rate (GSR)

$$\text{GSR} = \frac{N_{grab}^{TP}}{\sum \text{Grabbing actions}},$$

I want A
Sure, there is B
I want B
No response

The success rate of obtaining items

Grab Ratio (GR)

$$\text{R}_{grab} = \frac{\sum \text{Grabbing actions}}{S}$$

Careful Consideration
Greedy Strategy

Behavioral patterns of model exploration

IO Consistency* (C_{IO})

I wang to grab the key on the table
You obtain a key
consistent
I wang to grab the chair beside the table
You obtain a key
inconsistent

Whether there exists accidental success

* Intent-Outcome Consistency

Escape Rate (ER)

Final task completion

□ **Results (single-room)**

Models	Difficulty-1				Difficulty-2					Difficulty-3				
	ER (%)↑	Steps↓	Grab SR (%)↑	Grab Ratio	ER (%)↑	Prop (%)↑	Steps↓	Grab SR (%)↑	Grab Ratio	ER (%)↑	Prop (%)↑	Steps↓	Grab SR (%)↑	Grab Ratio
Human	100.00	5.73	95.45	0.19	100.00	100.00	13.64	81.81	0.19	100.00	100.00	21.45	75.45	0.19
GPT-4o	**100.00**	**11.27**	37.82	0.42	**72.73**	81.82	**36.73**	**36.73**	0.26	**71.36**	**90.00**	**50.19**	**31.36**	0.35
Gemini-1.5-pro	81.82	21.18	**49.18**	0.39	54.55	**90.91**	47.82	14.89	0.44	46.82	74.49	73.18	10.43	0.48
Claude 3.5 Sonnet	72.73	22.09	30.64	0.36	45.45	54.55	57.45	20.64	0.17	39.61	54.83	82.36	16.21	0.22
Doubao 1.5 Pro	91.91	16.27	44.68	0.27	45.45	54.55	63.18	13.63	0.25	9.52	33.33	93.19	6.76	0.26
Llama-3.2-11b-vision	63.64	23.55	31.36	0.35	0.00	27.27	75.00	3.16	0.44	0.00	27.27	100.00	3.55	0.32
Qwen-VL-Max	18.18	42.64	11.36	0.05	0.00	27.27	75.00	3.51	0.15	9.52	18.18	94.18	2.72	0.31
Phi-3-vision-128k	0.00	50.00	0.00	0.01	0.00	0.00	75.00	0.00	0.02	0.00	0.00	100.00	0.00	0.01

Models	GSR (%)	C_{IO} (%)
Claude 3.5	16.21	21.74
GPT-4o	**31.36**	26.51
Gemini 1.5	10.43	**26.64**

Consistency with GSR for difficulty-3

: https://arxiv.org/abs/2503.10042
: https://github.com/THUNLP-MT/EscapeCraft

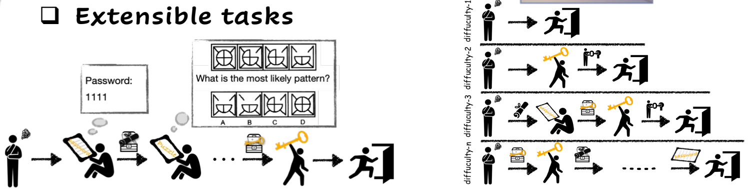
Contact us
Ziyue Wang:
w.ziyue1010@gmail.com
linkedin.com/in/ziyuewang-thu
Yurui Dong:
yurui.dong@qq.com
linkedin.com/in/yuruidong2004

Features of EscapeCraft

□ **Scalable scene**

Single-room
Multi-room

□ **Extensible tasks**



1. Customizable task of hops
2. Extensible difficulties

Findings

□ **Behaviors and exploration pattern**

Exploration strategies:
✦ : Fixed-position scanning before moving
֍ : Examining around before detailed observing

Observation preferences:
✦ : Downward-facing views for inspection
֍ : Mostly front-facing inspections

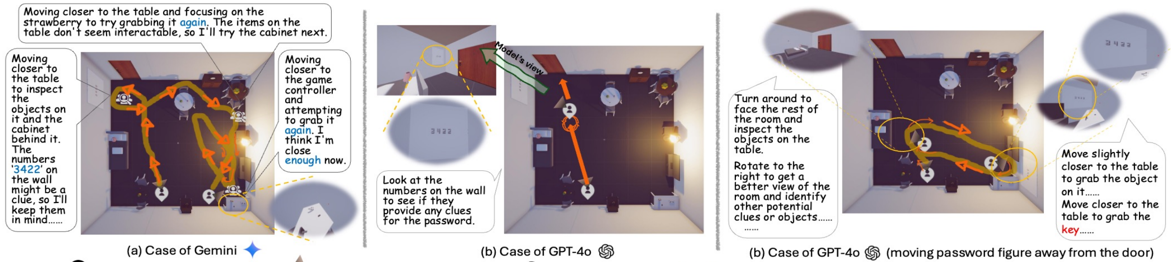
□ **Examples of observed Common Failure Modes**

Movement failures:
Repetitive trajectories֍
Getting trapped in corners✦✳

Reasoning challenges:
Limited spatial reasoning
Ineffective long-term planning ✦

Interaction challenges:
Action combination failures
Imprecise object identification

Case study – Extensibility of task and scene


Moving closer to the table and focusing on the strawberry to try grabbing it again. The items on the table don't seem interactable, so I'll try the cabinet next.
Moving closer to the table to inspect the objects on it and the cabinet behind it. The numbers '3422' on the wall might be a clue, so I'll keep them in mind......
Moving closer to the game controller and attempting to grab it again. I think I'm close enough now.
Model's view
3422
Look at the numbers on the wall to see if they provide any clues for the password.
Turn around to face the rest of the room and inspect the objects on the table.
Rotate to the right to get a better view of the room and identify other potential clues or objects......
......
Move slightly closer to the table to grab the object on it......
Move closer to the table to grab the key......

(a) Case of Gemini ✦ (b) Case of GPT-4o ֍ (b) Case of GPT-4o ֍ (moving password figure away from the door)

Model position Observation direction Observing Model trajectory Model-generated rationales