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

Ziyue Wang

2025.04.07

Agenda

- Motivation
- Introduction Video
- Introduction of our benchmark
- Introduction the constructed environment
- Evaluation
- Analysis and conclusion

Agenda

- **Motivation**
- Introduction Video
- Introduction of our benchmark
- Introduction the constructed environment
- Evaluation
- Analysis and conclusion

Motivation

- What are MLLMs capable of?

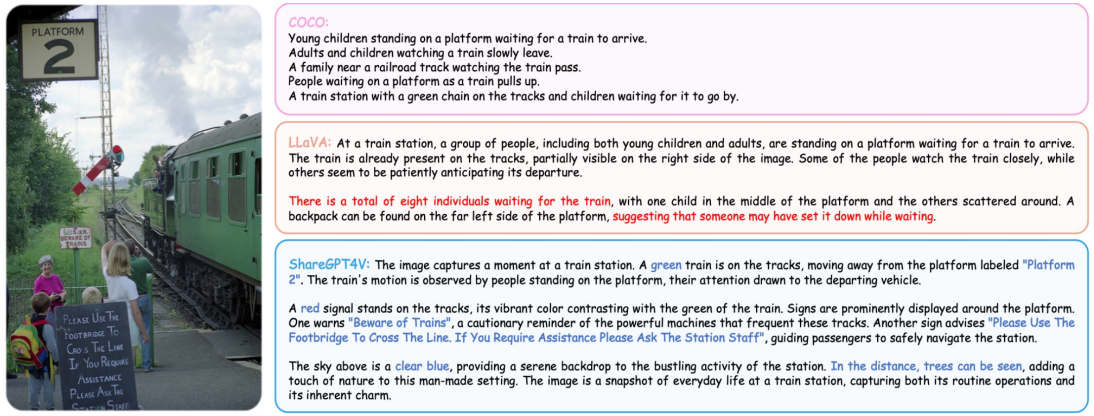
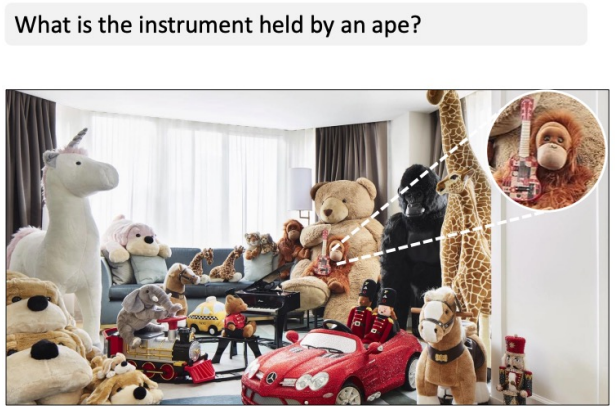
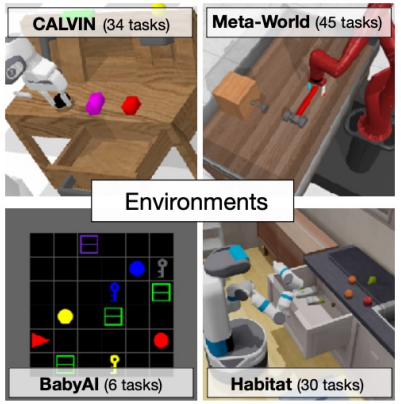


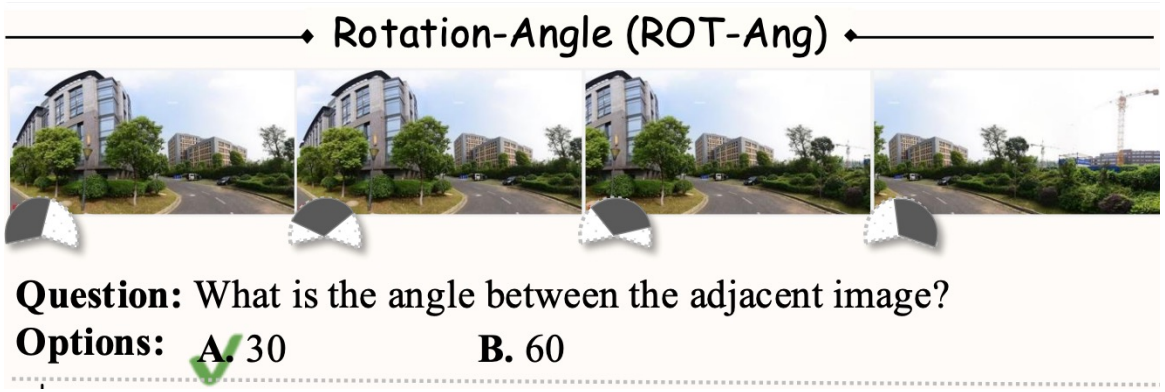
Image captioning



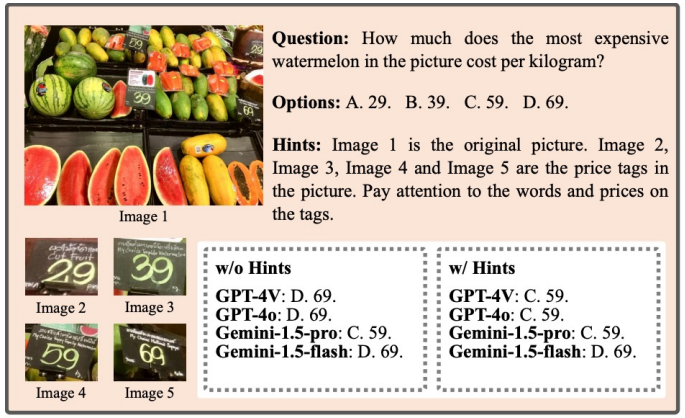
Visual search



Action & planning



Spatial reasoning



Visual reasoning

...

etc.

Motivation

- What are MLLMs capable of?

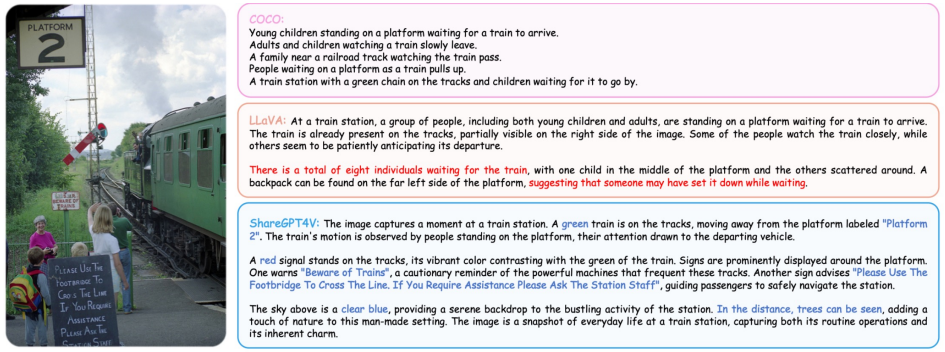
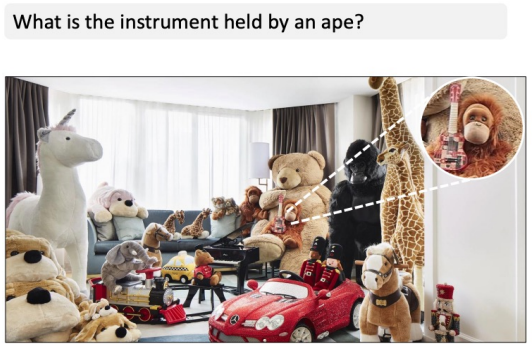
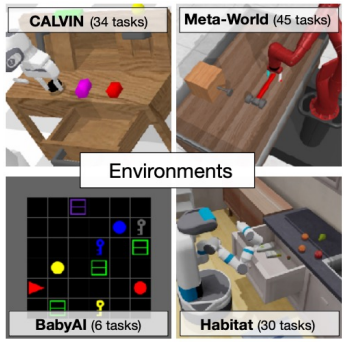


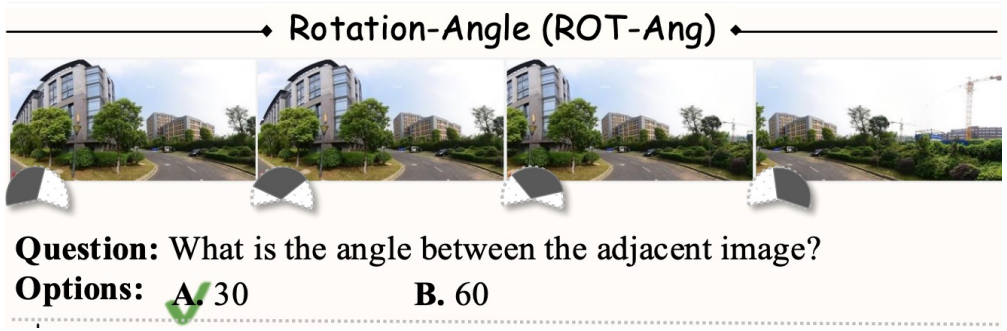
Image captioning



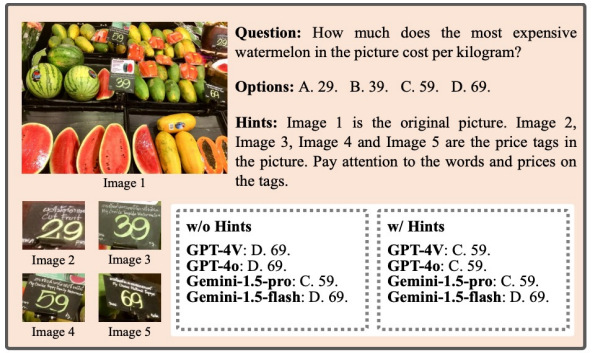
Visual search



Action & planning



Spatial reasoning



Visual reasoning

...

etc.

Are they able to accomplish **complex tasks requiring integrations of these capabilities?**

Agenda

- Motivation
- **Introduction Video**
- Introduction of our benchmark
- Introduction the constructed environment
- Evaluation
- Analysis and conclusion

Introduction

- Let's watch a video first 

Agenda

- Motivation
- Introduction Video
- **Introduction of our benchmark**
- **Introduction the constructed environment**
- Evaluation
- Analysis and conclusion

Introduction of our benchmark

- Target: To explore how MLLMs handle **complex multimodal reasoning**

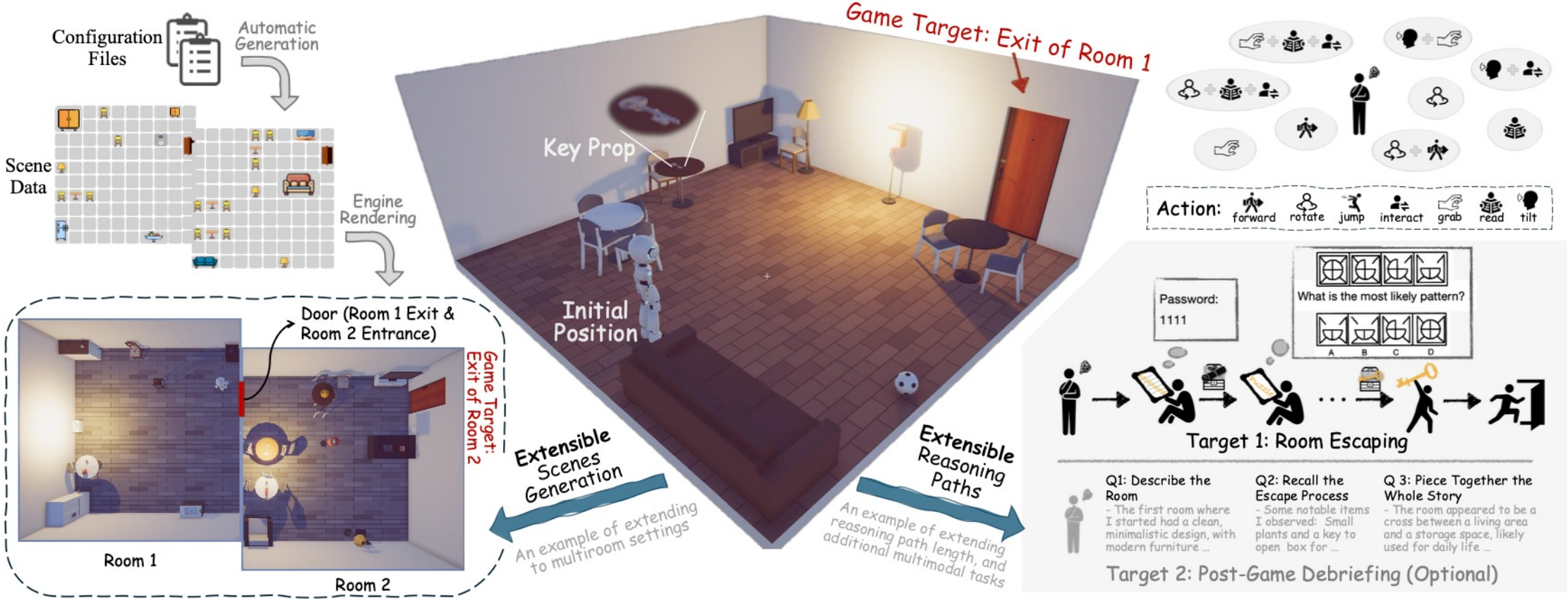


Figure 1. Illustration of our proposed room escape environment EscapeCraft, which allows us to generate customized room scenes (left) and define groundtruth reasoning path of tasks (right). Based on EscapeCraft, we create MM-Escape benchmark, targeting at evaluating both the task completion performance and the entire multimodal reasoning process of MLLMs.

Introduction of our benchmark

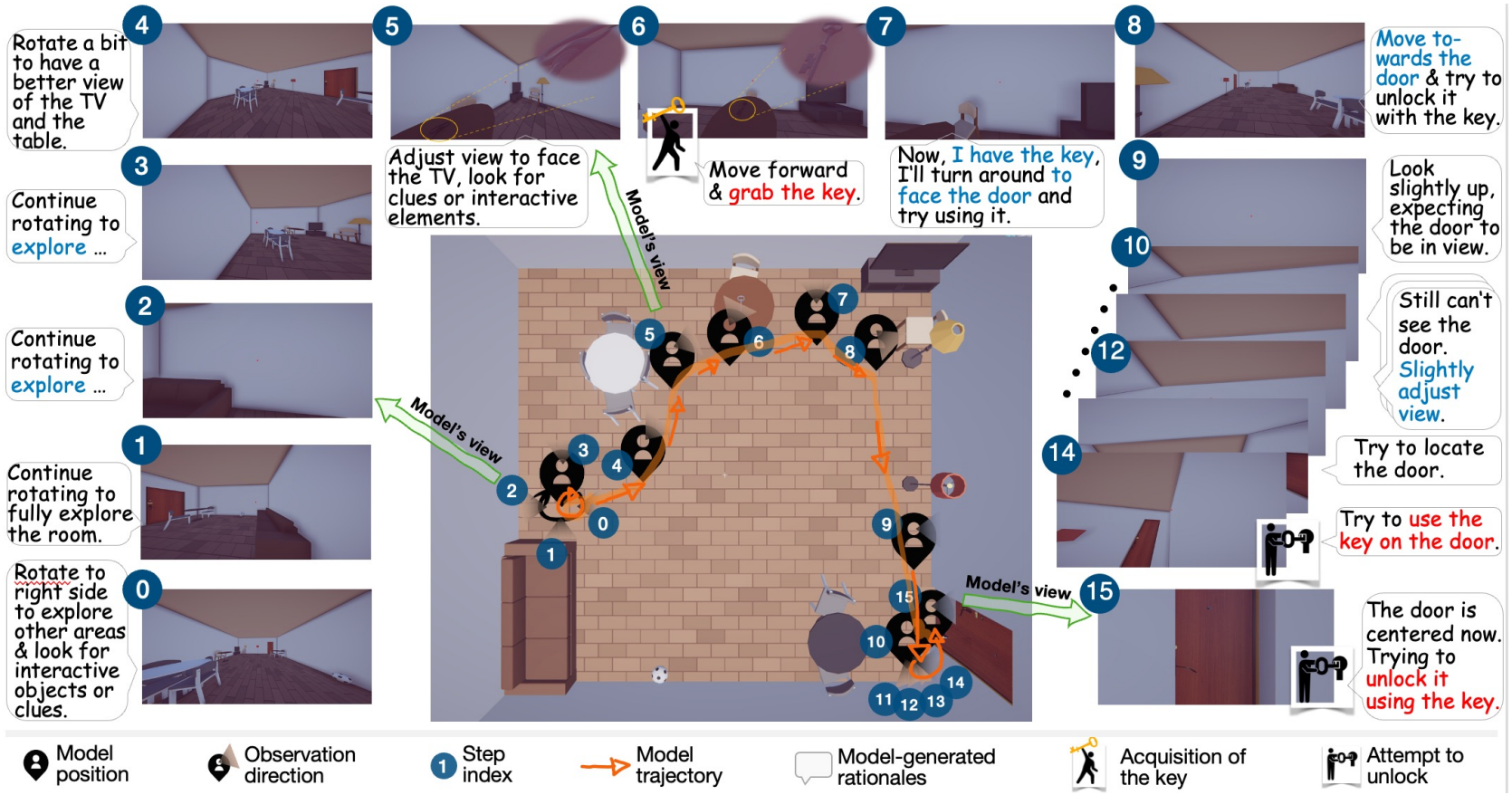
- **What** is complex multimodal reasoning?
  - Multimodal reasoning in a virtual/real environment
  - Requiring integration of multiple basic abilities, such as visual search, visual-spatial reasoning, long-term reasoning etc.

- **Why** complex multimodal reasoning?
  - Limitations of the primary focus of current evaluations:
    - Performance task completion rather than the reasoning process
    - Isolated abilities
    - Limited metrics and evidence for analysis
  - Limitations of environment:
    - Simplified the autonomous reasoning process
    - With provided structured knowledge libraries

Introduction of our benchmark

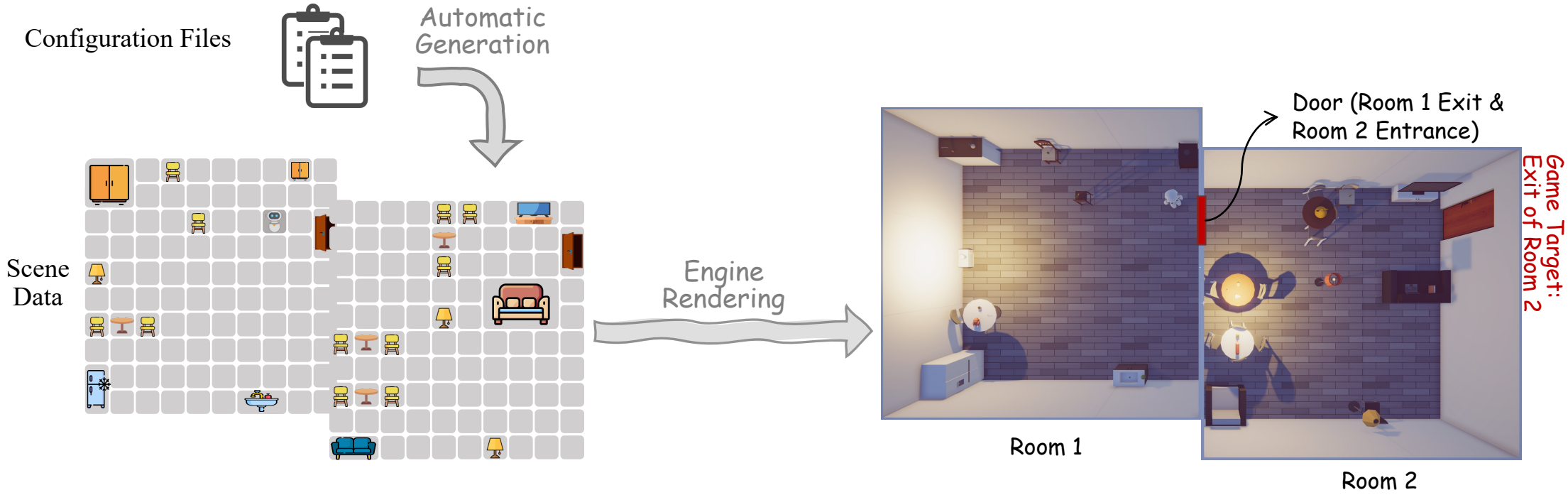
- How to evaluate the entire multimodal reasoning process?

- ✓ **MM-Escape**:
  An extensible benchmark inspired by real-world escape games
- ✓ Built on **EscapeCraft**:
  A customizable open environment
- ✓ Requires free-form exploration
- ✓ Requires coordination of multiple abilities
- ✓ Metrics emphasizing intermediate model behaviors alongside task completion



Introduction of our environment

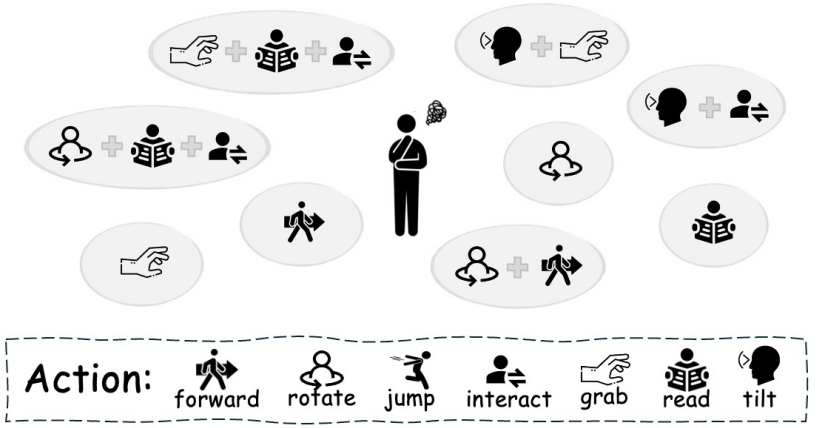
- EscapeCraft Environment Features
  - Automatic room scene generation
  - Customizable room styles (living room, kitchen, bathroom, bedroom)



Introduction of our environment

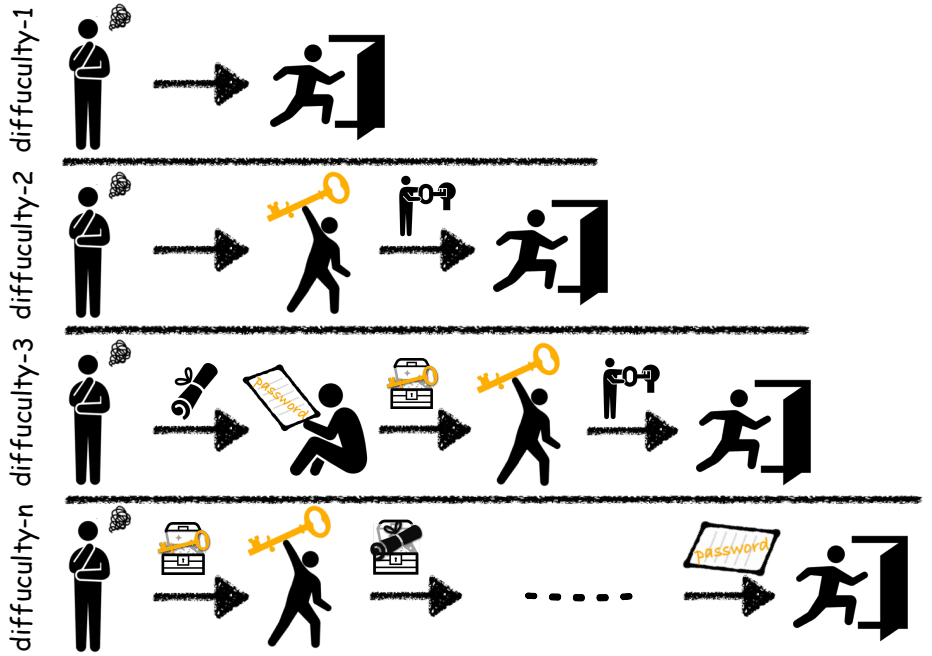
- EscapeCraft Environment Features
  - Automatic room scene generation
  - Customizable room styles (living room, kitchen, bathroom, bedroom)

  - Comprehensive action space:
    - Movement (forward)
    - View adjustment (rotation, tilting)
    - Interactions (grab, use, read, input)



Introduction of our environment

- EscapeCraft Environment Features
  - Automatic room scene generation
  - Customizable room styles (living room, kitchen, bathroom, bedroom)
  - Comprehensive action space:
    - Movement (forward)
    - View adjustment (rotation, tilting)
    - Interactions (grab, use, read, input)
  - Extensible escaping path and difficulty

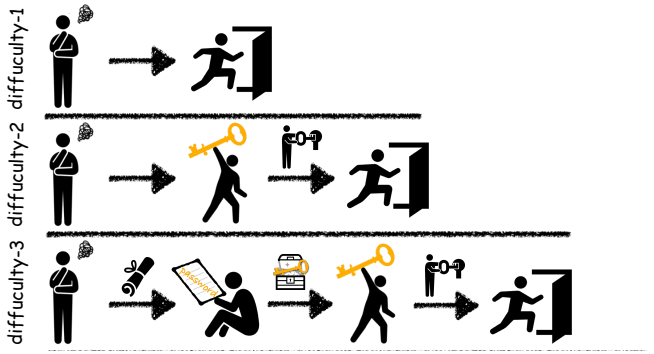


Agenda

- Motivation
- Introduction Video
- Introduction of our benchmark
- Introduction the constructed environment
- **Evaluation**
- Analysis and conclusion

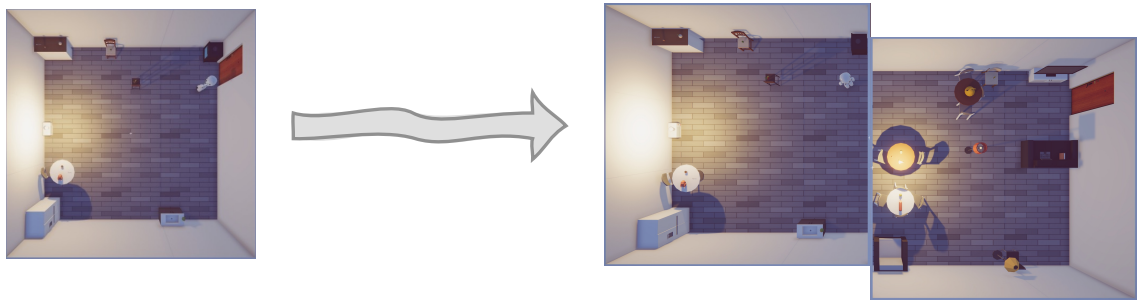
Evaluation

- Benchmark construction
  - Variable difficulty levels based on reasoning path length:
    - Difficulty-1: (one-hop)
      - Direct door interaction
    - Difficulty-2: (two-hop)
      - Find key/password + unlock door
    - Difficulty-3: (three-hop)
      - Find password + find key + unlock door



  - Extensible to more complex chains and multi-room settings

Single room



Multi-room

  - Statistics
    - 63 scenes, 73 evaluation games (including combinations of multi-room settings)

Evaluation

- Evaluation metrics design:
  - **Final** task completion:
    - Escape Rate (**ER**)

  - **Intermediate** process metrics:
    - **Prop Gain:** Successful acquisition of necessary items)

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

    - **Average Steps:** Efficiency of the whole process
    - **Grab Success Rate:** Precision and reasonability of model-env interactions

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

    - **Grab Ratio:** Interaction strategy

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

Evaluation

- Results
  - Models perform **significantly below human level** across all metrics
  - Performance **sharply declines as game difficulty increases**
  - GPT-4o does not always lead the performance for the easiest setting

Table 2. Results of standard single room setting. Prop: Prop Gain; Steps: average steps used to complete the game; Grab SR: the precision of grabbing; Grab Ratio: the portion of grabbing actions regarding the total consumed steps. Note that Difficulty-1 requires no prop, and the prop gain is therefore omitted for this setting. The max allowed steps are 50, 75, 100 for Difficulty-1, -2, -3 respectively. The best score of each metrics is **bolded** and the second is underlined.

Models	Difficulty-1				Difficulty-2					Difficulty-3					AVG
	ER (%)↑	Steps↓	Grab SR (%)↑	Grab Ratio	ER (%)↑	Prop (%)↑	Steps↓	Grab SR (%)↑	Grab Ratio	ER (%)↑	Prop (%)↑	Steps↓	Grab SR (%)↑	Grab Ratio	ER (%)↑
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	100.00
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	81.36
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	61.06
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	52.60
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	48.96
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	21.21
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	9.23
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	0.00

Evaluation

- Results
  - For models fail to escape, we can still assess the performance via prop gain and interactive behaviors
    - Llama 3.2 and Qwen present several successful interactions and manage to obtain some props
    - Phi-3 attempts to interact with environment
  - Grab Ratio does not directly contribute to the other metrics, it is a neutral indicator of interaction behavior and strategy

Models	Difficulty-1				Difficulty-2					Difficulty-3					AVG
	ER (%)↑	Steps↓	Grab SR (%)↑	Grab Ratio	ER (%)↑	Prop (%)↑	Steps↓	Grab SR (%)↑	Grab Ratio	ER (%)↑	Prop (%)↑	Steps↓	Grab SR (%)↑	Grab Ratio	ER (%)↑
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	100.00
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	81.36
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	61.06
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	52.60
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	48.96
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	21.21
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	9.23
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	0.00

Evaluation

- Extended Scenarios
  - Are models able to detect **a potential prop when it is blended within the environment** without hint?

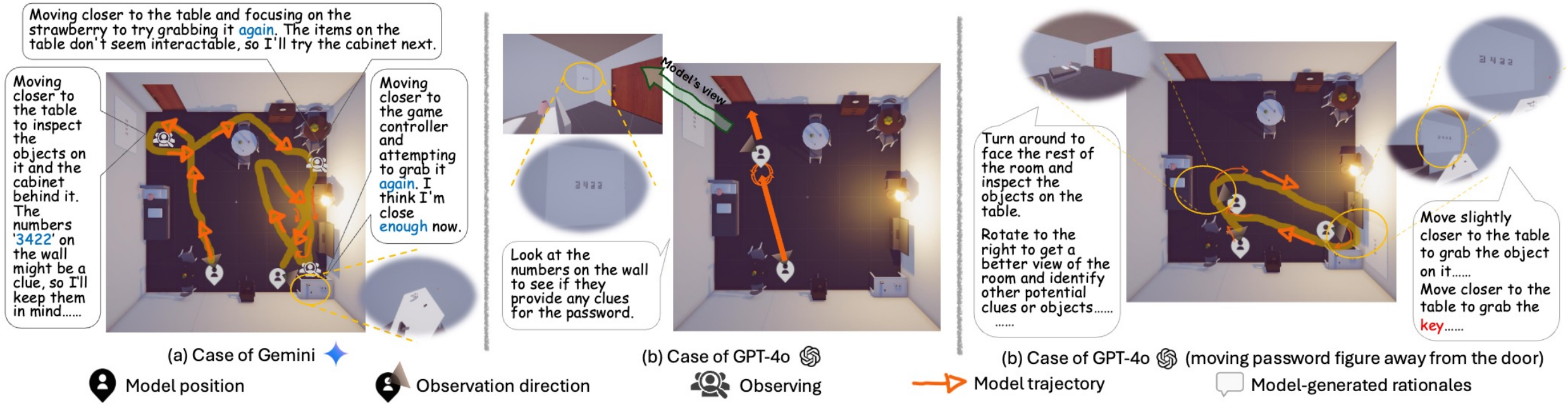
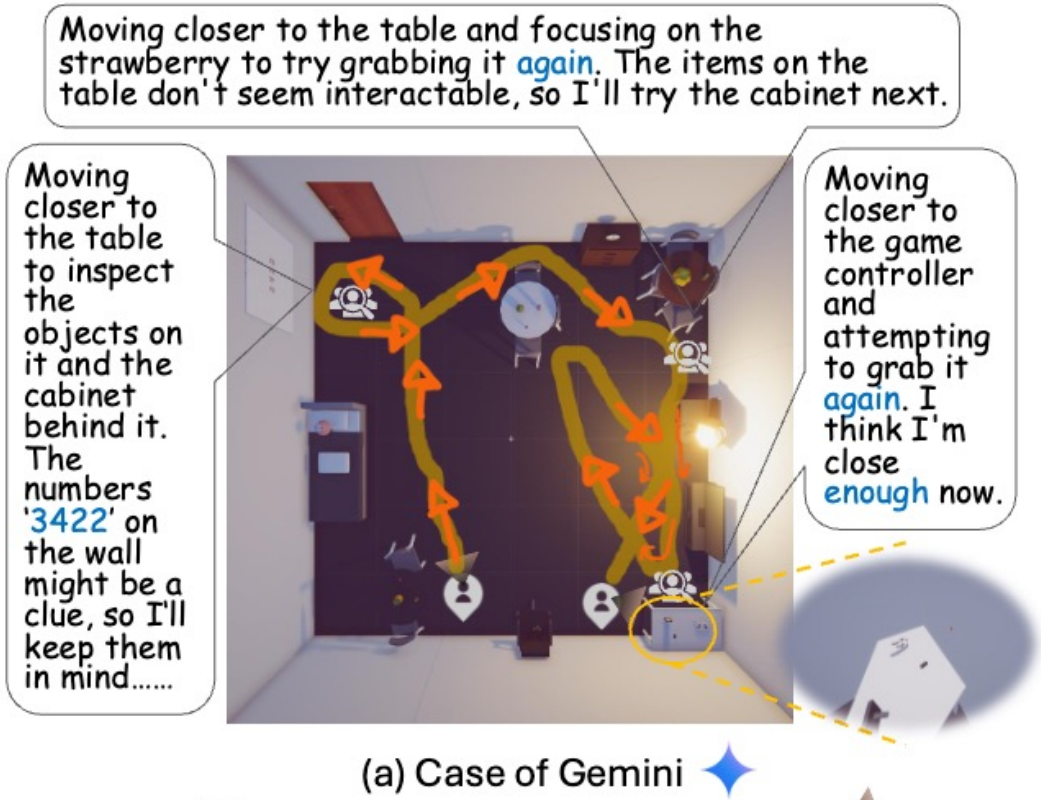


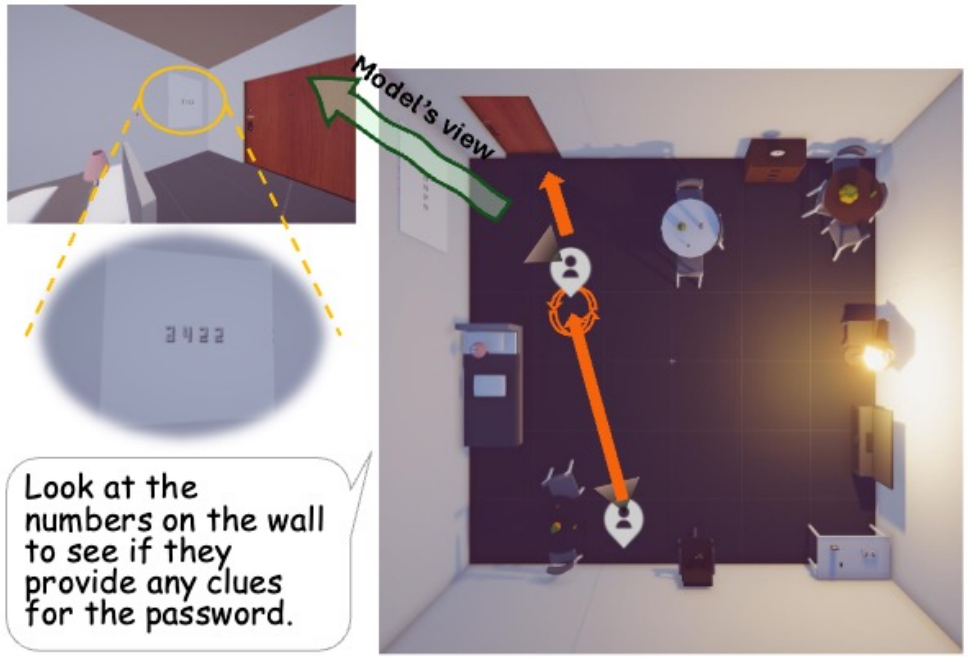
Figure 5. The extended scenario where the required password is displayed via a numerical pattern on the wall, rather than explicitly written on notes. GPT-4o completes reading it at once and exits within five steps, while Gemini struggles to repetitively search the room. Moving the pattern away from the door further challenges GPT-4o, leading to a failure of escaping.

Evaluation

- Extended Scenarios
  - Are models able to detect **a potential prop when it is blended within the environment** without hint?
    - Placed near the exit



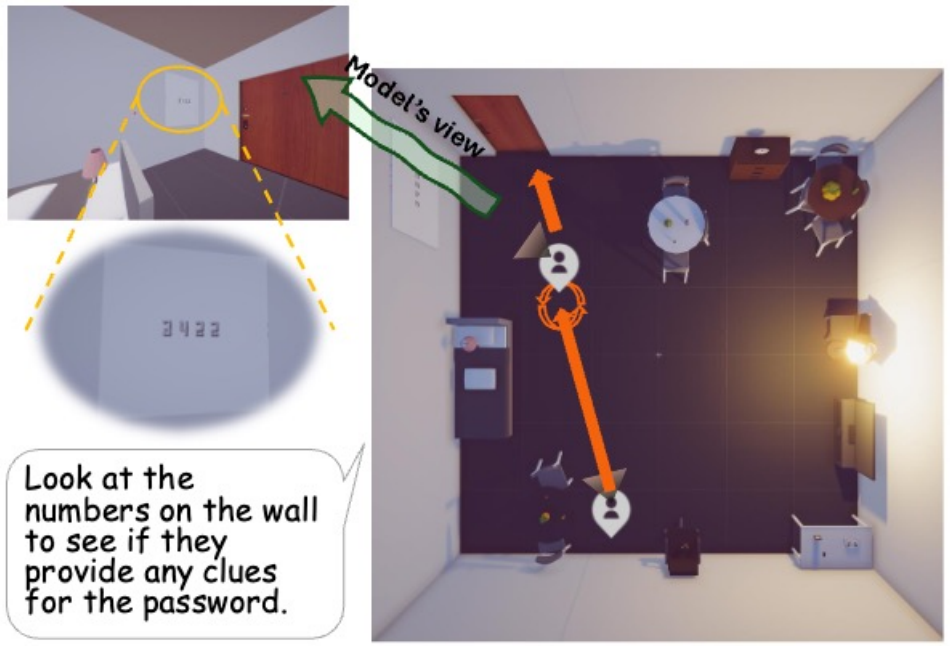
(a) Case of Gemini



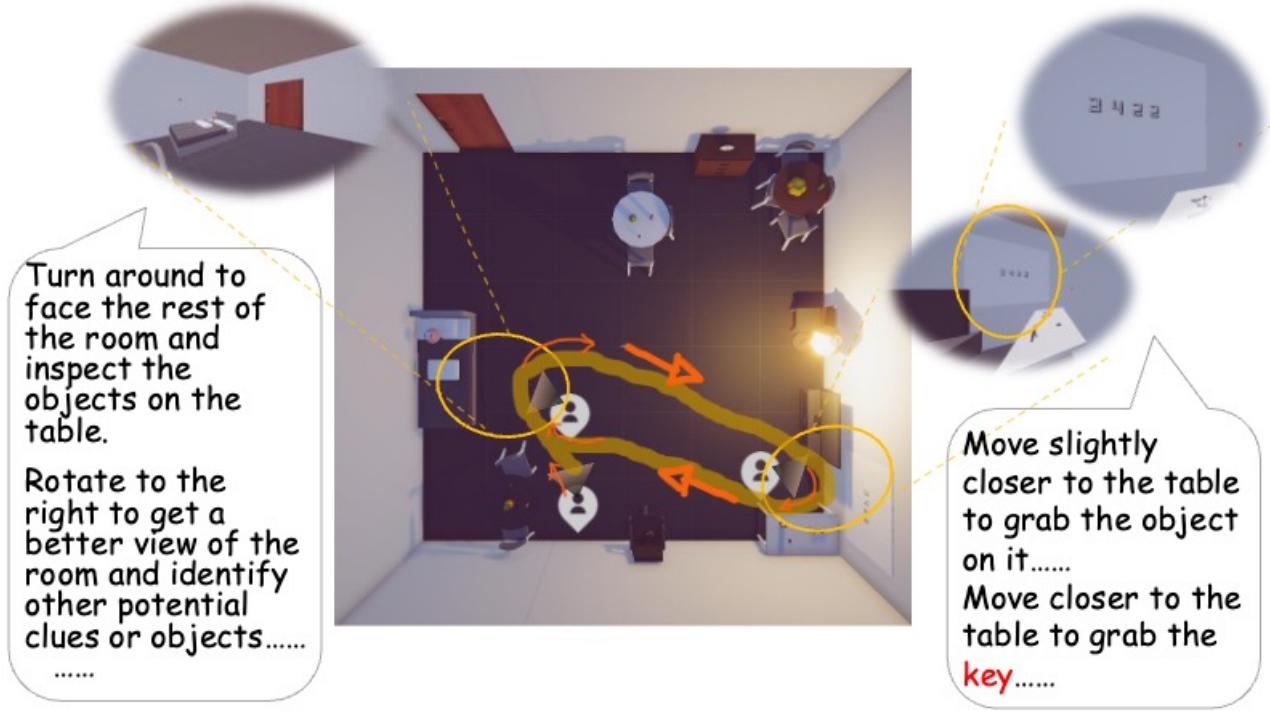
(b) Case of GPT-4o

Evaluation

- Extended Scenarios
  - Are models able to detect **a potential prop when it is blended within the environment** without hint?
    - Move away from the exit



(b) Case of GPT-4o



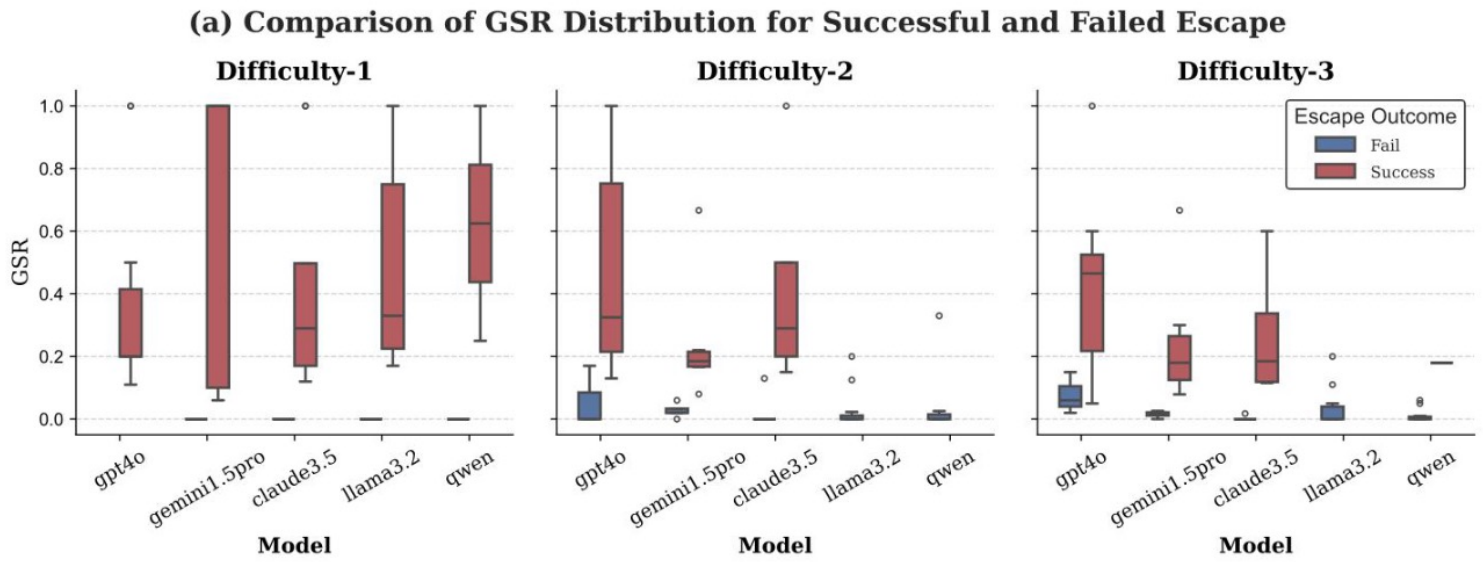
(b) Case of GPT-4o (moving password figure away from the door)

Agenda

- Motivation
- Introduction Video
- Introduction of our benchmark
- Introduction the constructed environment
- Evaluation
- **Analysis and conclusion**

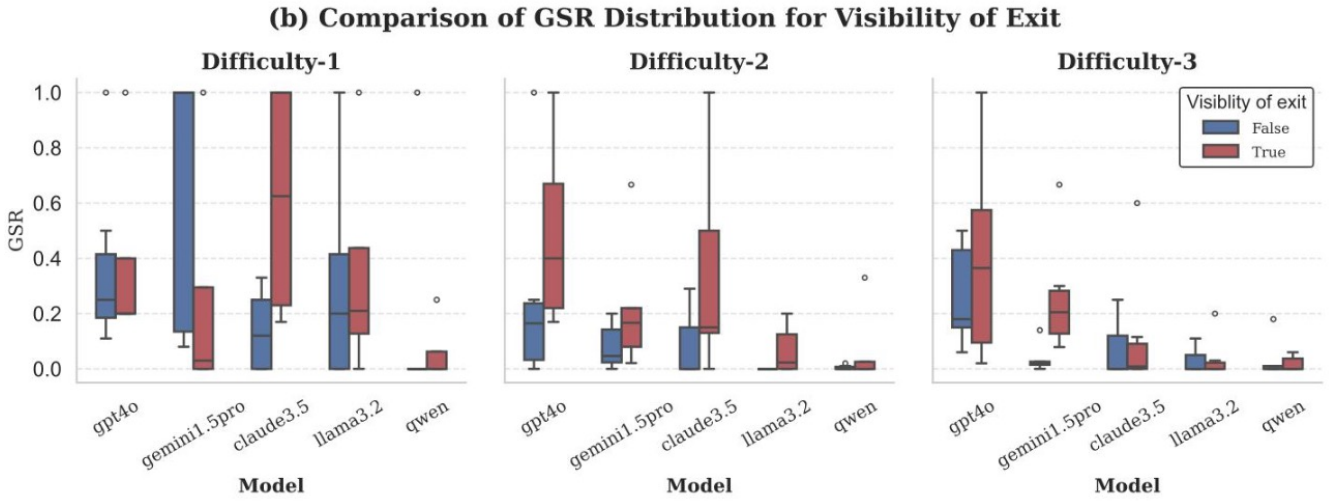
Analysis

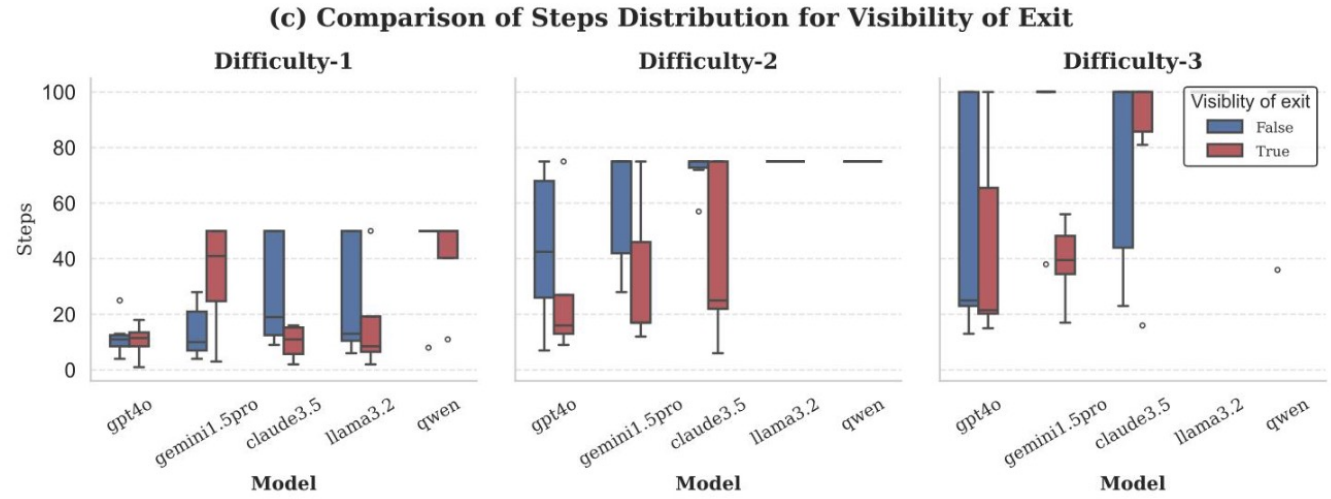
- Is grab success rate corelated to escape rate?
  - Higher grab success rate largely contributes to successful escape



Analysis

- Does the “visibility of exit at initial locations” affect the grab success rate and steps?
  - For games whose exits are visible at initial locations, models tend to exhibit **higher GSR**
  - For games whose exits are visible at initial locations, models often **use less steps to escape**





Conclusion

- Some Distinct Model Behaviors
  - Different exploration strategies:
    - Gemini: fixed-position scanning before moving
    - GPT-4o: global understanding before detailed inspection
  - Observation preferences:
    - Gemini: *downward-facing view* to inspect objects
    - GPT-4o: predominantly *front-facing* view
- Observed Common Failure Modes
  - Movement failures:
    - Repetitive trajectories (GPT-4o)
    - Getting trapped in corners (Gemini, Claude)
  - Interaction challenges:
    - Action combination failures (Phi-3, Qwen-VL)
    - Imprecise object identification
  - Limited spatial awareness and long-term planning

Thanks for listening

Q & A

MM-Escape

- Results
  - Multi-room settings expose limitations in spatial reasoning
  - A successful experience helps model to escape

Models	Difficulty-1 & Difficulty-1				Difficulty-1 & Difficulty-2					Difficulty-2 & Difficulty-2				
	ER(%)↑	Steps↓	Grab SR (%)↑	Grab Ratio	ER(%)↑	Prop (%)↑	Steps↓	Grab SR (%)↑	Grab Ratio	ER(%)↑	Prop (%)↑	Steps↓	Grab SR (%)↑	Grab Ratio
GPT-4o	**75.00**	35.50	34.25	0.32	**90.00**	**100.00**	**34.90**	35.52	0.31	**70.00**	**80.00**	**39.50**	**42.32**	0.37
Gemini-1.5-pro	22.22	40.22	22.89	0.38	40.00	50.00	56.60	16.79	0.05	60.00	**80.00**	60.00	22.71	0.34
Llama-3.2-11b-vision	55.56	**31.00**	**36.25**	0.36	10.00	60.00	66.40	4.40	0.40	10.00	40.00	76.80	27.00	0.19
Claude 3.5 Sonnet	22.22	45.22	10.62	0.08	20.00	20.00	71.90	6.75	0.09	10.00	10.00	80.00	23.20	0.06
Qwen-VL-max	22.22	40.33	12.96	0.16	30.00	50.00	57.70	**42.30**	0.28	0.00	10.00	80.00	23.66	0.32

Table 3. Performance on multi-room setting for different room scenes. To assist in the more challenging setting, we provide models with a full successful escape path from Room 1 (9 steps) for self-reflection when they try to unlock Room 2. Hence, the Prop Gain (Prop (%)) in the results refers only to Room 2. Further challenges of escaping from the very beginning of multi-room setting are discussed in Supplementary Material F.