COMPARING MULTI-MODAL AND LANGUAGE MODEL SPATIAL REASONING IN DISCRETE ENVIRONMENTS

Michael Sheroubi, Alessandro Pranzo, Emiliano Pizana-Vela, Sandro Mikautadze, Mattia Martino
Ecole Polytechnique, France
{firstname.lastname}@polytechnique.edu

Felix Zieger
ai@felixzieger.de

ABSTRACT

This study compares the performance of a Multimodal Model (Pixtral Large) and a Large Language Model (Mistral Large) in navigating identical game states. The Multimodal Model processes image-based inputs, while the LLM relies on textual descriptions derived from the same images. The evaluation focuses on their decision-making, reasoning, and adaptability to provide insights into their respective strengths and limitations.

***K*eywords** Multimodal Model · Large Language Model · Spatial Reasoning

1 Introduction

Recent advancements in AI lead to the development of powerful Multimodal Models and Large Language Models (LLMs). These models demonstrate remarkable capabilities in various tasks. This paper compares their performance in navigating identical game states, exploring their strengths and limitations in decision-making tasks.

2 Experimental Setup

Drawing inspiration from Singh et al. [1], the agent operates within a discrete grid-based environment where it moves in four directions: up, down, left, or right. The objective of the agent is to maximize its score by eating an apple and reaching the goal while avoiding penalties for stepping on a knife or moving out of bounds.

Game states are randomly generated as matrices adhering to specific constraints to ensure that each state is solvable. These matrices are then converted into visual representations as images for use with the Multimodal Model. For the Large Language Model (LLM), textual descriptions of the matrices are generated to serve as input.

The scoring system rewards favorable behavior while penalizing risky or incorrect moves. The scoring system is detailed in Table 1.

Table 1: Scoring System for Agent Actions

Action	Score
Eating an apple	+1
Reaching the goal	+2
Moving out of bounds	-1
Stepping on a knife	-2

The agent must also perform an action at the apple and the goal (*eat*, *exit*). So it is not sufficient to only wander onto the target. This setup allows for a direct comparison of decision-making capabilities between the Multimodal Model and the LLM. A sample game state is shown in Figure 1.

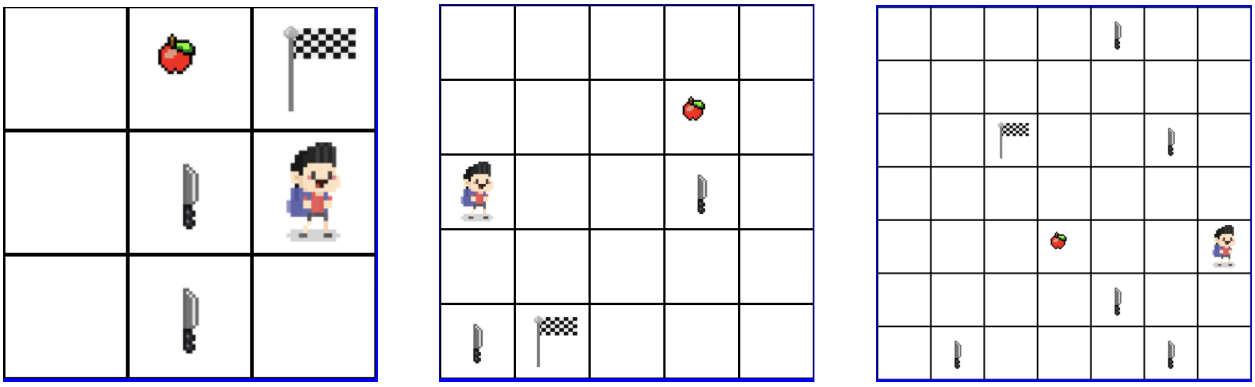

Figure 1: Example States (3x3, 5x5, 7x7)

3 Experimental Results

Success rates measure how often each model achieves the goal state without incurring significant penalties. Efficiency is evaluated based on the number of steps taken to achieve objectives.

Pixtral Large demonstrates the best performance when working with only textual input, while Mistral Large also performs reasonably well with textual data. However, Pixtral faces challenges when relying solely on visual input. This highlights how visual language models tend to struggle with spatial reasoning tasks [2].

Model	Successes_3x3	Successes_5x5	Successes_7x7	Successes_9x9
Pixtral Large (Visual)	40	9	0	0
Mistral Large	70	**62**	50	41
Pixtral Large (Text Only)	**73**	54	**62**	**46**
Ministral 3b	1	0	1	1
Ministral 8b	2	1	4	5
Ministral Small	30	29	28	29
Mistral Nemo	7	4	2	2
Pixtral 12b (Text Only)	23	4	8	5

Table 2: Success (%) across different grid sizes (3x3, 5x5, 7x7, and 9x9).

Model	Rewards_3x3	Rewards_5x5	Rewards_7x7	Rewards_9x9
Pixtral Large (Visual)	0.64	-0.39	-1.13	-1.09
Mistral Large	1.61	**1.26**	0.96	0.80
Pixtral Large (Text Only)	**1.77**	1.02	**1.60**	**1.01**
Ministral 3b	-1.20	-0.97	-0.73	-0.67
Ministral 8b	-0.73	-0.86	-0.90	-0.75
Ministral Small	-0.19	-0.15	-0.19	0.12
Mistral Nemo	-0.89	-0.95	-0.99	-0.86
Pixtral 12b (Text Only)	0.00	-0.73	-0.48	-0.60

Table 3: Rewards across different grid sizes (3x3, 5x5, 7x7, and 9x9).

4 Future Work

While Pixtral Large provides the best results, we believe that we can achieve measurable performance by fine-tuning a smaller model on state & action sequence pairs. We would like to explore using a structure similar Group Relative Policy Optimization [3] to generate high-quality training data, then fine-tuning a small model on the output.

References

[1] Satinder Singh, R. Lewis, and A. Barto. Where do rewards come from? 01 2009.

[2] Kanchana Ranasinghe, Satya Narayan Shukla, Omid Poursaeed, Michael S. Ryoo, and Tsung-Yu Lin. Learning to localize objects improves spatial reasoning in visual-llms, 2024.

[3] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025.

A Appendix

Model	Num Parameters
Pixtral Large	124B
Mistral Large	123B
Mistral Small	22B
Mistral Nemo	12B
Pixtral 12b	12B
Ministral 8b	8B
Ministral 3b	3B

Table 4: Number of parameters for Mistral Models

Model	NumSteps_3x3	NumSteps_5x5	NumSteps_7x7	NumSteps_9x9
Pixtral Large (Visual)	6.68	**8.61**	15.63	18.73
Mistral Large	6.79	10.18	14.10	17.30
Pixtral Large (Text Only)	7.01	10.19	13.69	17.69
Ministral 3b	6.34	8.64	14.12	16.76
Ministral 8b	**4.29**	9.40	14.95	18.30
Ministral Small	5.51	10.22	14.67	**15.10**
Mistral Nemo	8.81	15.39	26.59	19.30
Pixtral 12b (Text Only)	5.46	8.68	**12.65**	20.66

Table 5: Average Num Steps across different grid sizes (3x3, 5x5, 7x7, and 9x9).

B Prompts

B.1 Main Prompt

```
Analyze the text and environment to generate a sequence of actions using ONLY:
{Available Classes}.
Print your initial position, and the final exit position.
Explain the path taken and the values encountered in each position.
"At last print ALL the sequence of actions used in the explained part, including
the EAT and EXIT"

Follow these rules:

1. Movement Constraints:
- Directions alter your position in the matrix as follows:
* ‘up‘: Move to previous row (row_index - 1), same column.
* ‘down‘: Move to next row (row_index + 1), same column.
* ‘left‘: Move to previous column (column_index - 1), same row.
* ‘right‘: Move to next column (column_index + 1), same row.

- **Knife Avoidance**:
- You CANNOT **enter, step into, or pass through** positions with knives (‘2‘).
- You MUST reroute around knives, DO NOT PASS THROUGH KNIVES.

- **Boundary Constraints**:
- You CANNOT move outside the grid (row/column indices must be 0-{rows-1}).

- **Apple and Exit Interaction**:
- ‘eat‘ MUST be executed **immediately** when reaching the apple’s cell (‘3‘).
- After eating, navigate to the exit (‘4‘). The exit is your final destination.

Key:
0 = Empty, 1 = YOU (start), 2 = KNIFE (blocked), 3 = APPLE, 4 = EXIT (goal)

Example (Different 5x5 Environment):
Text: "Go to the apple and then to the exit"
Environment Matrix:
Row 0: |1|0|0|0|0|  ← YOU start at [0][0]
Row 1: |0|2|0|0|0|  ← KNIFE at [1][1]
Row 2: |0|0|2|0|0|  ← KNIFE at [2][2]
Row 3: |0|0|0|2|0|  ← KNIFE at [3][3]
Row 4: |3|0|0|0|4|  ← APPLE at [4][0], EXIT at [4][4]

Explanation:
1. Start at [0][0] (YOU).
2. Move down → [1][0] (0).
3. Move down → [2][0] (0).
4. Move down → [3][0] (0).
5. Move down → [4][0] (APPLE) → eat.
6. Move right → [4][1] (0).
7. Move right → [4][2] (0).
8. Move right → [4][3] (0).
9. Move right → [4][4] (EXIT).

Valid Output: down, down, down, down, eat, right, right, right, right, exit


2. Environment Layout for THIS TASK:
{’-’*40}
```

```
{environment_layout}
{'-'*40}
Columns: {columns_header}

Your Task: "Take the actions to go eat the apple and then go to the exit,
IMPORTANT: "do NOT step on any knife and do NOT go out of boundary""
"THE LAST LINE HAS TO BE JUST THE ACTIONS BETWEEN COMAS "
```

B.2 Multimodal Prompt

This prompt was used to initially test the multi-modal model. For the test results, all models used the Main prompt.

```
You are situated in the environment depicted in the provided image, you are represented by the
pixeled kid icon. Your goal is to reach the apple and then the exit. You must avoid the knives.

First, describe what you observe in the environment. Then, outline a strategy to accomplish
the following tasks:
1. Locate and consume the apple.
2. Reach the exit safely.
Be cautious of the knives in the environment; treat them as obstacles to avoid them at all
stages of the movement. DO NOT ENTER THEIR CELLS.

In order to help you with the process, predict where you end up after each movement,
check what you will find in such cell and make sure that it is an allowed move,
otherwise go back and change action until possible then you move.
Generate a sequence of actions using ONLY: {", ".join(CLASSES)}.
These actions must be in the correct order in order to complete the tasks.

After displaying your process, as a last line, respond STRICTLY with comma-separated actions.

Follow these rules:

1. Movement Constraints:
- Directions alter your position in the matrix as follows:
* up: Move to upper cell.
* down: Move to lower cell.
* left: Move to left cell.
* right: Move to right cell.

- *Knife Avoidance*:
- You CANNOT *enter, step into, or pass through* cells with knives.
- You MUST reroute around knives, even if adjacent.

- *Boundary Constraints*:
- You CANNOT move outside the grid.

- *Apple and Exit Interaction*:
- eat MUST be executed *immediately* when reaching the apple's cell.
- After eating, navigate to the exit. The exit is your final destination.

The icons in the image represent:
- The pixel art kid: Your starting position.
- The pixel art knife: Obstacle to avoid.
- The pixel art apple: Food to eat.
- The pixel art flag: Your final destination.

Your Task:
Text: "Go and eat the apple, then reach the exit"
Actions:
```