MineAnyBuild: Benchmarking Spatial Planning for Open-world AI Agents

Ziming Wei[1*], Bingqian Lin[2*], Zijian Jiao[1*], Yunshuang Nie[1], Liang Ma[3], Yuecheng Liu[4], Yuzheng Zhuang[4], Xiaodan Liang[1†]

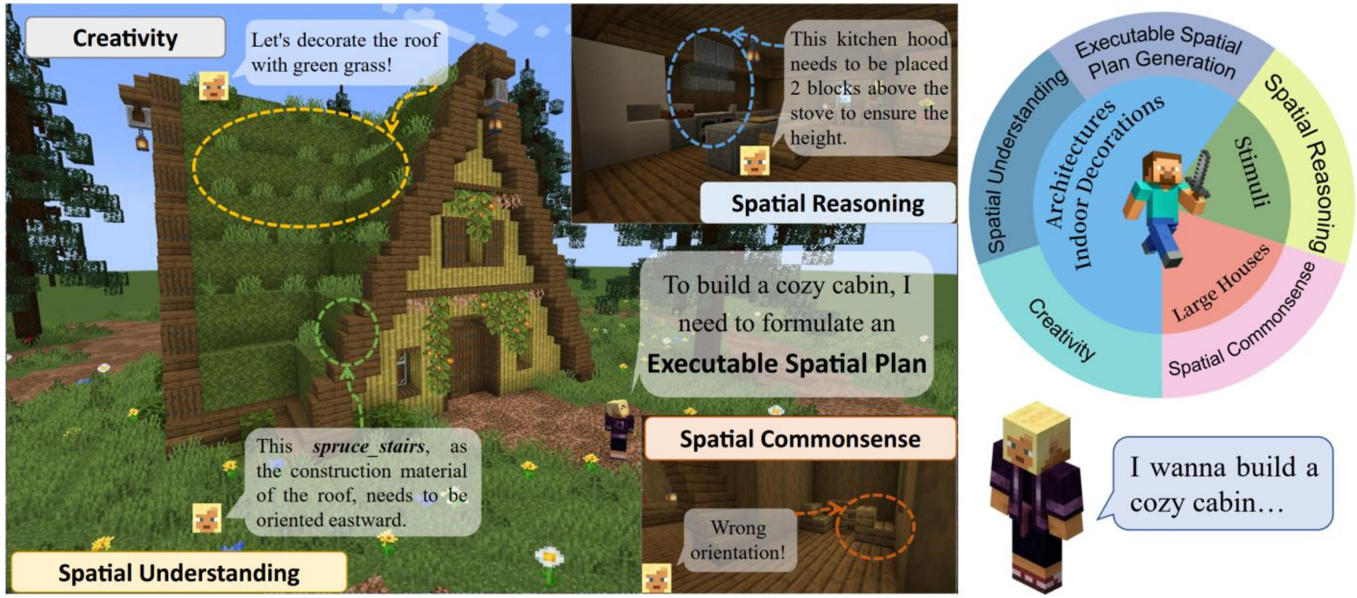
[1]Shenzhen Campus of Sun Yat-sen University, [2]Shanghai Jiao Tong University,

[3]Mohamed bin Zayed University of Artificial Intelligence, [4]Huawei Noah's Ark Lab

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

- Spatial Planning is a crucial part in the field of spatial intelligence, which requires the **understanding and planning about object arrangements in space perspective**.
- AI agents with the spatial planning ability can better adapt to **various real-world applications**, including robotic manipulation, automatic assembly, urban planning, etc.
- Recent works have attempted to construct benchmarks for evaluating the spatial intelligence of Multimodal Large Language Models (MLLMs). Nevertheless, these benchmarks primarily focus on spatial reasoning based on typical Visual Question-Answering (VQA) forms, which suffers from **the gap between abstract spatial understanding and concrete task execution**.



Contributions

- We propose **MineAnyBuild**, which benchmarks the **spatial planning evaluation** for **open-world AI agents** in the Minecraft game. MineAnyBuild covers diverse evaluation dimensions, including **spatial reasoning, creativity, spatial commonsense**, etc. Through requiring the agent to generate **executable architecture building plans**, our MineAnyBuild significantly mitigate the gap between abstract spatial understanding and concrete task execution.

- We test various existing MLLM-based AI agents for spatial planning in multiple perspectives and difficulties, which exposes the **insufficiency** of the existing AI agents' capabilities in spatial planning. We provide the **visualization results on executable planning outputs and failure cases**, revealing that current AI agents are still facing tough issues such as **spatial misunderstanding and implementation gap** to be handled.

- We propose an **infinitely expandable data curation pipeline** to scale our benchmark and datasets, where we can collect **endless player-generated content** on the Internet and automatically convert it into **processable data**. Our pipeline well utilize the **abundant creations** made by millions of players to benefit the **training and evaluation of open-world AI agents**.

Benchmark and Task Suite

NEURAL INFORMATION PROCESSING SYSTEMS

Executable Spatial Plan Generation

Input:
Let's build a desert small house.
Begin by laying a foundation of sand in a rectangular shape, with some smooth sandstone and terracotta blocks forming a patterned interior... Once the walls reach their full height, cap the entire structure with a flat roof made of cut sandstone, ensuring it covers the full footprint of the house...

Output:
Here's my plan for this small house in desert.
We firstly use sand blocks to build the floor…
Then, we build the structure with sandstone…
The blueprint matrix is …

Creativity

Input:
Construct an Olympic Rings.

Output:
Here's my plan for building an Olympic Rings. I will use 5 colors of wools. The colors are blue, black, red, yellow, lime…

Spatial Understanding

Input:
Let's build a piano with a potted flower.
Layer 1: quartz_pillar: [(1,1), (2,1), (3,1)],
smooth_quartz_stairs: [(1,2), (2,2), (3,2)].
Layer 2: ...

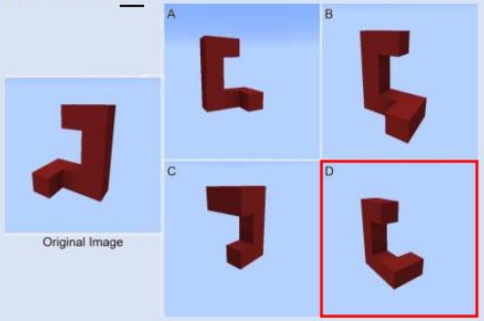
Output: Here's a piano with a potted flower…
The blueprint matrix is:

```
[
  [[1, 2], ..., [1, 2]],
  ...,
  [[5, -1], ..., [6, -1]]
]
```

Spatial Reasoning

Question:
Which option is the same as the original image, aside from its orientation?

Answer: **D**



Spatial Commonsense

Question:
Where should I go if my bedroom is on the second floor?

Answer:
You should go up the stairs on the right front to reach your bedroom.

Our MineAnyBuild involves approximately 4,000 spatial planning tasks with 500+ buildings/indoor decoration assets.

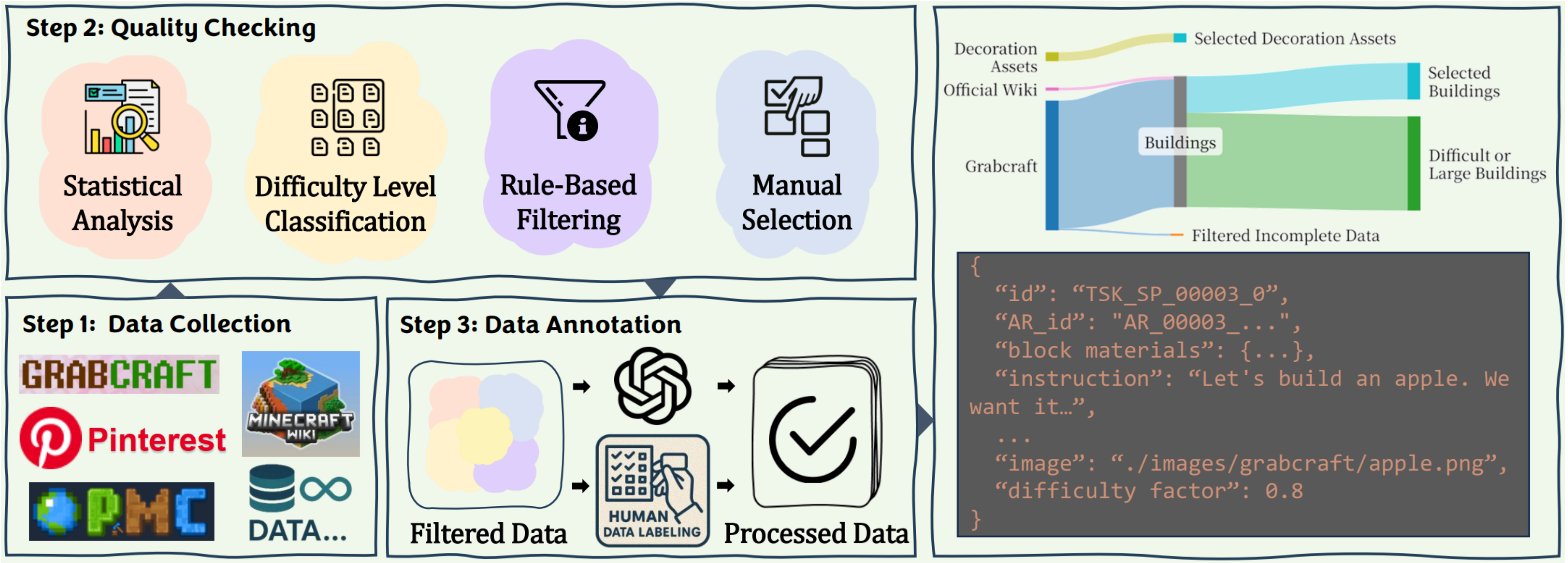
These tasks correspond to diverse evaluation dimensions, thereby conducting a comprehensive assessment of AI agents' spatial planning capabilities.

Data Curation Pipeline

NEURAL INFORMATION PROCESSING SYSTEMS

MineAnyBuild is a comprehensive benchmark with diverse architectures and indoor decorations, aligned with various instructions and visual reference images.

We build our benchmark based on the following steps: 1) data collection, 2) quality checking, and 3) data annotation.



Experimental Results

We mainly conduct our evaluation on MLLM-based agents that suitable to address the spatial planning task in our benchmark. We evaluated 13 SOTA MLLM-based agents, including proprietary models and open-source models.

Findings:

- Even the best proprietary model, GPT-4o, scores only 41 out of 100, showing that current agents are still far from real spatial planning ability.
- Among all dimensions, spatial reasoning and plan execution are the weakest, while commonsense and understanding remain the relatively stronger areas.
- Open-source models perform much worse, with most failing to produce valid 3D plans.

Models	Executable Spatial Plan Generation	Spatial Understanding	Spatial Reasoning	Creativity	Spatial Commonsense	Overall
	Score ↑	Score ↑	Accuracy ↑	Score ↑	Score ↑	
Proprietary						
Claude-3.5-Sonnet	3.21	4.63	19.8	3.24	6.90	39.92
Claude-3.7-Sonnet	3.48	5.07	17.6	3.10	6.94	40.70
Gemini-1.5-Flash	2.87	2.49	25.8	2.71	7.12	35.54
Gemini-1.5-Pro	3.53	4.80	16.9	2.73	7.52	40.54
Gemini-2.0-Flash	2.63	4.19	16.0	2.44	6.82	35.36
GPT-4o	3.27	4.75	24.4	2.73	7.32	41.02
GPT-4o-mini	2.08	2.52	26.7	2.38	7.14	33.58
Open-source						
InternVL2.5-2B	0.24	0.34	19.8	0.28	4.94	15.56
InternVL2.5-4B	0.32	0.42	20.0	0.63	5.66	18.06
InternVL2.5-8B	0.68	0.62	20.4	0.66	5.62	19.24
Qwen2.5VL-3B	0.63	0.61	17.0	0.54	5.46	17.88
Qwen2.5VL-7B	1.29	1.12	16.0	1.34	6.30	23.30
LLava-Onevision-7B	0.73	0.92	19.6	0.98	5.54	20.26

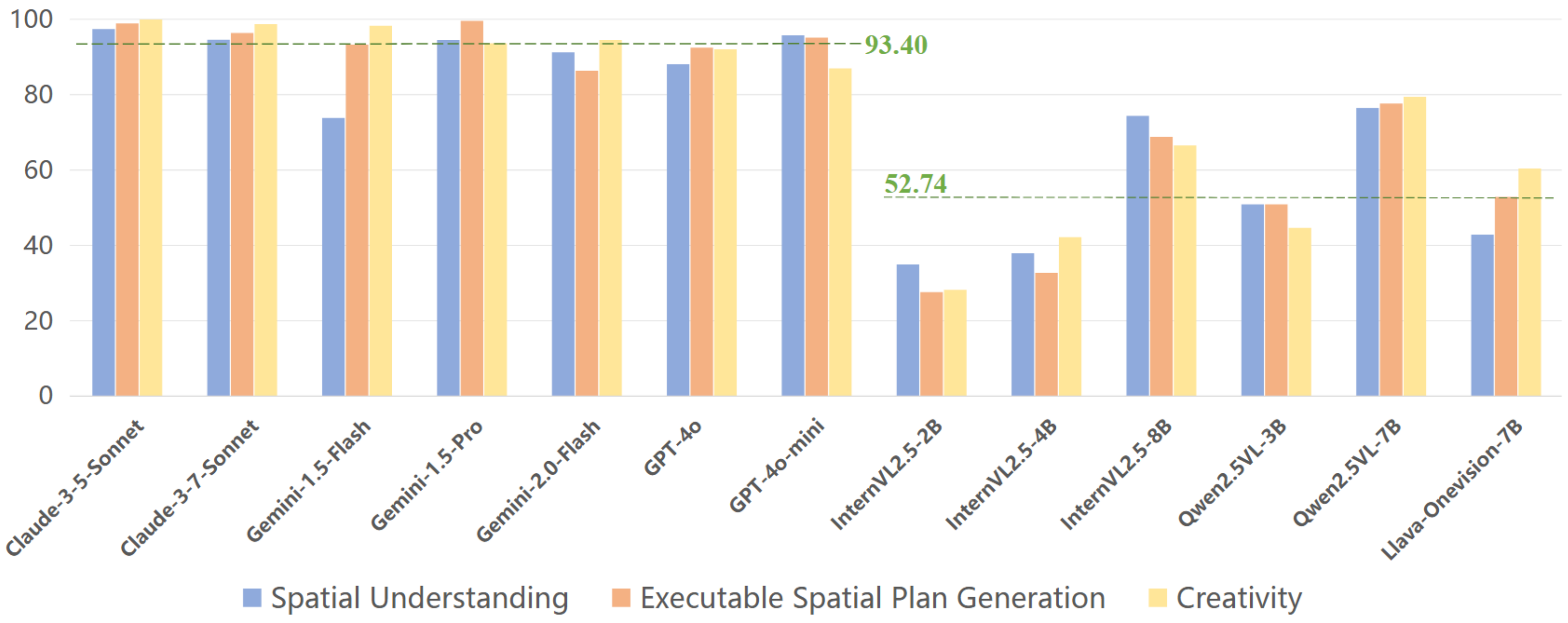
Experimental Results

NEURAL INFORMATION PROCESSING SYSTEMS

We mainly conduct our evaluation on MLLM-based agents that suitable to address the spatial planning task in our benchmark. We evaluated 13 SOTA MLLM-based agents, including proprietary models and open-source models.

Findings:

- Even the best proprietary model, GPT-4o, scores only 41 out of 100, showing that current agents are still far from real spatial planning ability.
- Among all dimensions, spatial reasoning and plan execution are the weakest, while commonsense and understanding remain the relatively stronger areas.
- Open-source models perform much worse, with most failing to produce valid 3D plans.

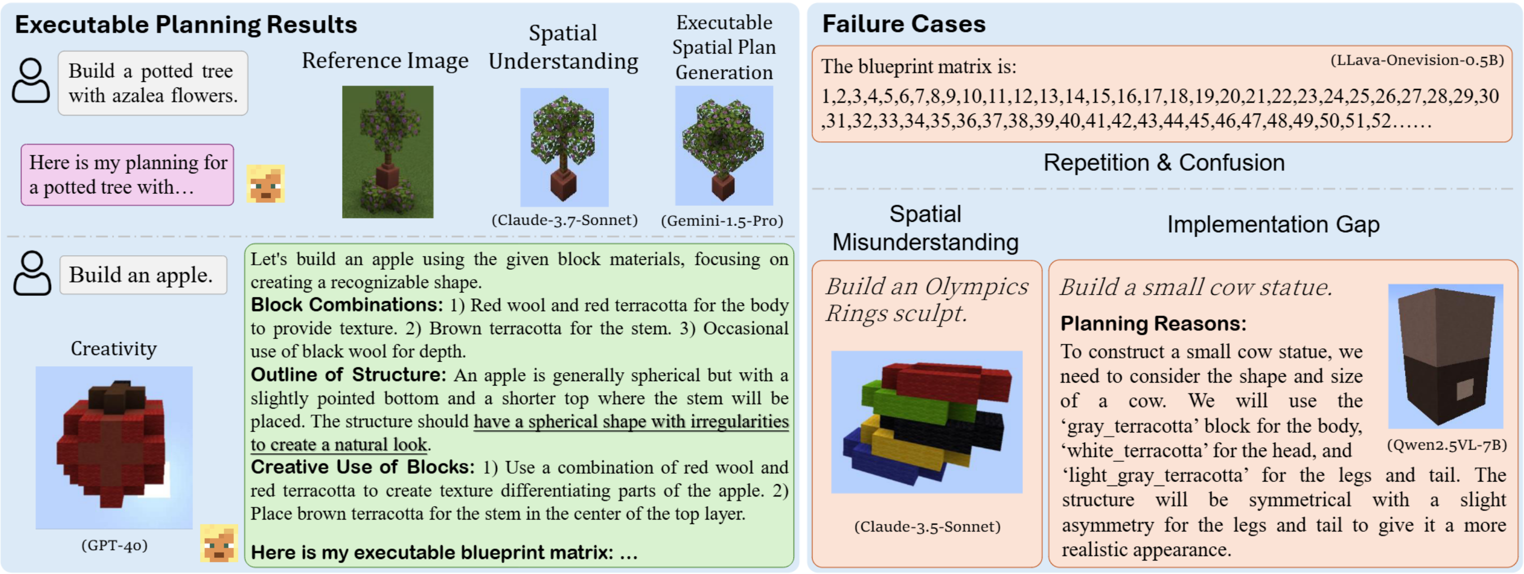


Results Analysis

There are some weaknesses.

- **Spatial Misunderstanding**: incorrect spatial arrangements;
- **Implementation Gap**: incorrect block indexing, orientation errors or inconsistent spatial logic;
- **Structural Degeneration under Complexity**: limited ability to scale from basic patterns to more abstract and complex architectural concepts.

These failure modes reflect deeper limitations in MLLM's capabilities to perform hierarchical spatial planning, maintain geometric consistency and ground language into manipulable 3D structures. They also provide more research directions for MLLMs, e.g., to improve multi-modal spatial understanding, align linguistic abstraction with executable plans or enhance agent's ability for structural composition in open-ended 3D environments.



Thank you!

Please read our paper for more discussions and analyses!

Code	Project	arXiv	Ziming's Homepage	Bingqian's Homepage