VisionGraph: Leveraging Large Multimodal Models for Graph Theory Problems in Visual Context

Yunxin Li[1], Baotian Hu[1], Haoyuan Shi[1], Wei Wang[2], Longyue Wang, Min Zhang[1]

[1]Harbin Institute of Technology, Shenzhen

[2]Shenzhen Compus of Sun Yat-sen University

Paper Link: https://arxiv.org/abs/2405.04950

Github Code: https://github.com/HITsz-TMG/VisionGraph/

Core Issue: Complex Logical Reasoning Ability

Logical Reasoning of Multimodal Large Models

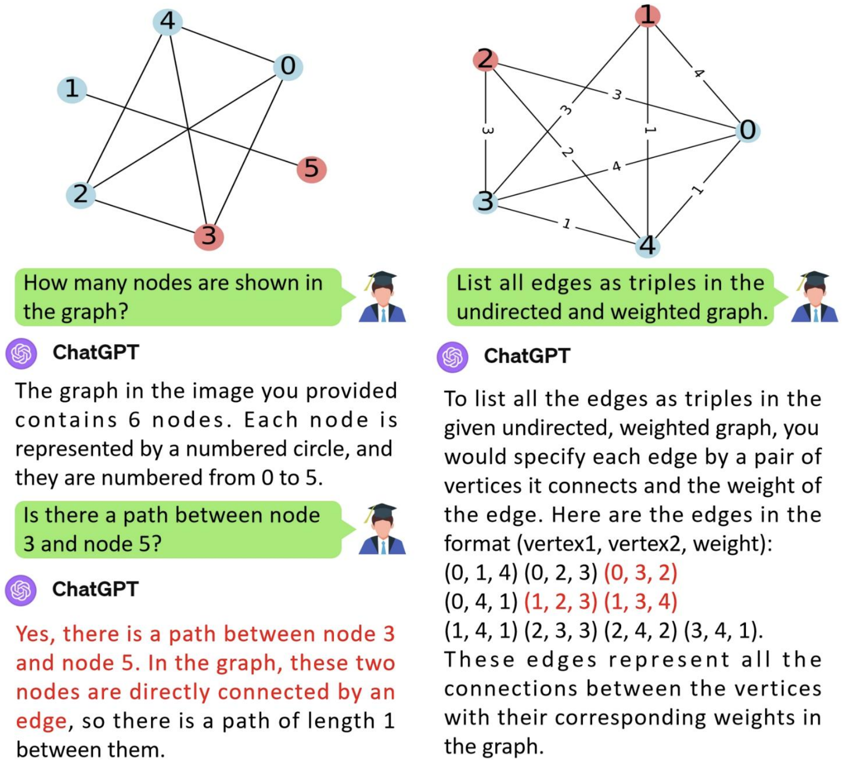
- Mathematical reasoning is one of the core aspects for evaluating logical reasoning ability of large models.
- Our task: inputting a visual image and answering the corresponding graph theory questions.
- The graph theory questions require MLMs to: 1) Accurately understand the graph structure. 2) Utilize knowledge to perform multi-step reasoning in the visual graph.
- Graph theory problems are involved in various agent scenarios:

    1) AI for Science, e.g., Molecular Structure

    2) Visual language navigation,

    3) Robot planning and control

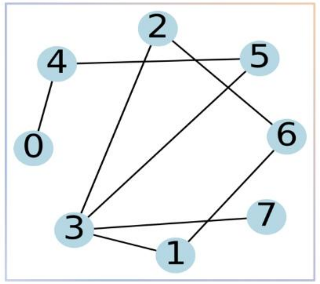
- Most agent scenarios focus on Web UI and OS environments.
- VisionGraph provides a new testboard for assessing multimodal agent.



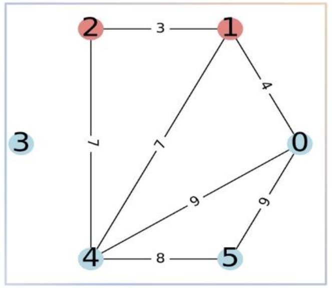
Yunxin Li, Baotian Hu, Haoyuan Shi, Wei Wang, Longyue Wang, Min Zhang. VisionGraph: Leveraging Large Multimodal Models for Graph Theory Problems in Visual Context. *ICML 2024*

VisionGraph: Introduction

- Propose a benchmark set (6,000): covering 8 types of graph theory problems and 3 levels of difficulty.
- Evaluate spatial understanding and reasoning ability (2+1 questions):
  - Node Recognition: *How many nodes are shown in the graph?*
  - Edge Recognition: *List all edges as triples in the undirected and weighted graph.*
  - Eight Specific Graph Theory problems
- Perception-Enhanced Data: 200k edge-relevant VQA



Q: Which edges are connected to node 1?
A: The edges connected to node 1 are: (1, 3), (1, 6).
Q: Please use tuples to represent the edges in the graph. Each tuple should consist of two nodes that are connected by an undirected edge.
A: The edges are represented by the tuples: (0, 4), (1, 3), (1, 6), (2, 3), (2, 6), (3, 5), (3, 7), (4, 5).



Q: Which edges are connected to node 2?
A: The edges connected to node 2 are: (1, 2, 3), (2, 4, 7).
Q: Please use tuples to represent the edges in the graph. Each tuple should consist of three elements: (node1, node2, weight).
A: The edges are represented by the tuples: (0, 1, 4), (0, 4, 9), (0, 5, 9), (1, 2, 3), (1, 4, 7), (2, 4, 7), (4, 5, 8).

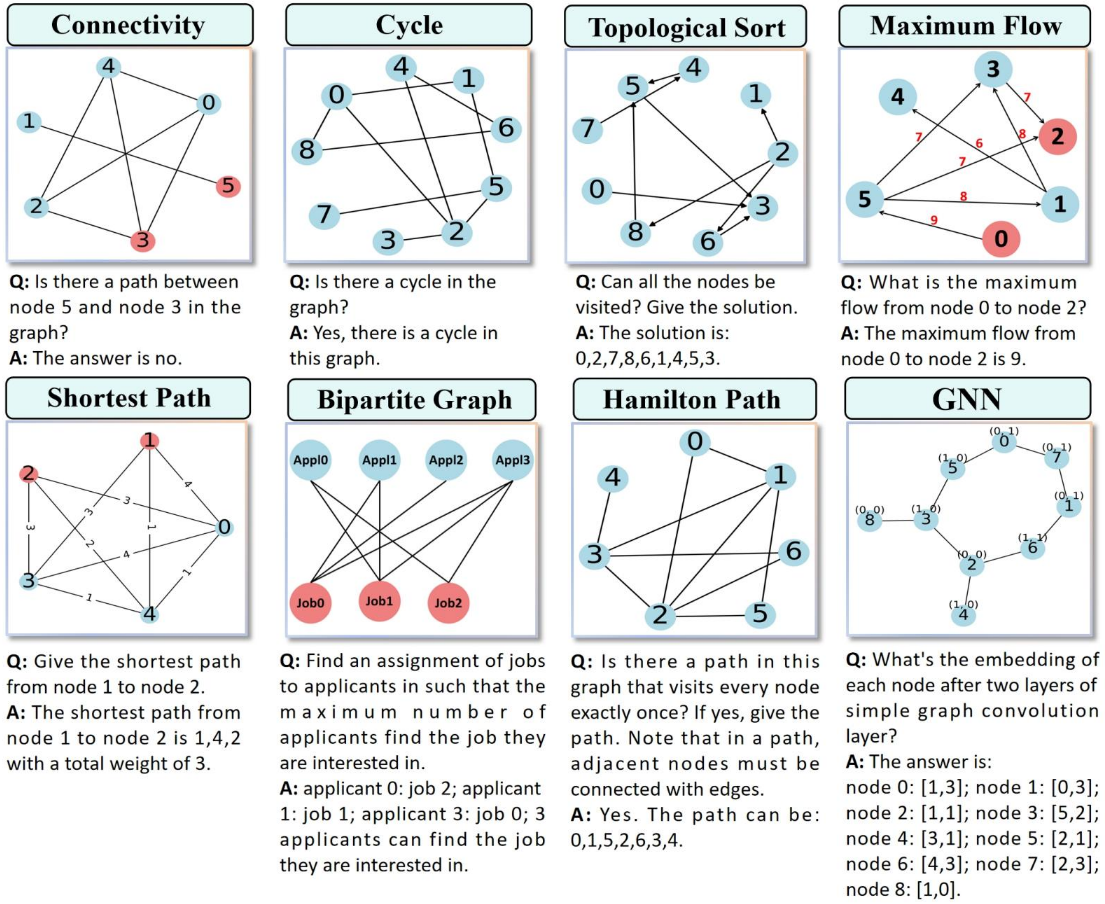


Table 1. Overview of the VisionGraph Benchmark. 'SPEC.' represents the level of difficulty, indicated by the number of nodes in each graph. Key metrics include the number of samples (S), images (I), questions (Q), and answers (A). Each visual graph is accompanied by three questions: two focus on general graph comprehension, and one addresses specific graph theory problems.

Subset	Connect.	Cycle	Topo. Sort	Shortest Path	Max. Flow	Bipartite Graph	Hamilton Path	GNNs
# EASY	352	150	180	180	150	300	150	100
SPEC.	$n:5-10$	$n:5-10$	$n:5-10$	$n:5-10$	$n:5-10$	$n:6-20$	$n:5-10$	$n:5-8$
# MEDIUM	1,200	600	150	-	-	-	-	-
SPEC.	$n:11-25$	$n:11-25$	$n:11-25$	-	-	-	-	-
# HARD	680	400	200	200	200	210	200	140
SPEC.	$n:26-35$	$n:26-35$	$n:26-35$	$n:11-20$	$n:11-20$	$n:17-33$	$n:11-20$	$n:9-15$
# Total S/I	2,232	1,150	530	380	350	510	350	240
#Q/A	6,696	3,450	1,590	1,140	1,050	1,530	1,050	720
#Len_Q	45.0	39.0	52.0	60.0	61.0	54.0	62.0	76.0
#Len_A	162.26	63.94	194.06	95.10	141.47	126.01	101.03	61.79

VisionGraph: Evaluation

Comprehensive Evaluation

- ➢ Close: GPT-4V, Gemini, Qwen-plus/max
- ➢ Open: MiniGPT-4, InstructBLIP, LLaVA

Understanding of Graph Structures

- ➢ Closed-source GPT-4V outperforms Gemini.
- ➢ Open-source multimodal large models have very poor zero-shot spatial understanding capability.
- ➢ High error rates and error propagation affect reasoning ability.

Impact of Supervised Fine-tuning

- ➢ Using 200K edge VQA data for enhancement, edge recognition reduces error rates, especially for Cycle and Connectivity.
- ➢ The improved accuracy of graph problem shows that improving spatial reasoning requires more underlying perception capabilities.

Table 3. Overall results in the VisionGraph benchmark. ♣ refers to that the corresponding model is trained using the training set of VisionGraph. The results in parentheses for Gemini and GPT-4V are the accuracy of the detailed path, yet other LMMs can not follow the instructions to provide specific paths. Bold words refer to the best results.

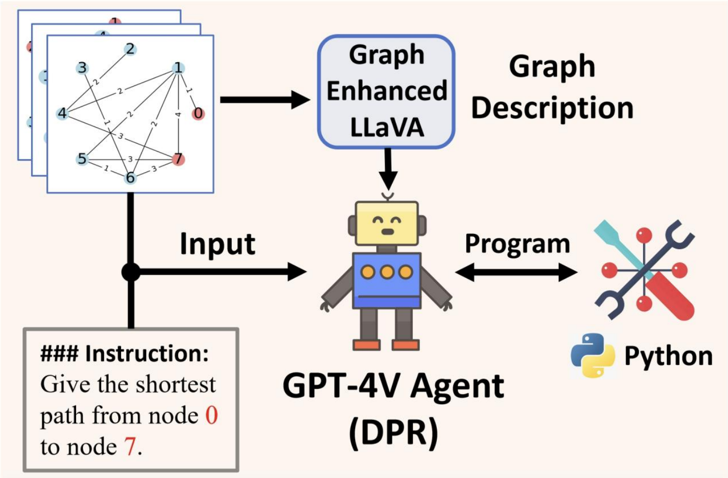
Model↓ Task Types →	Connect	Cycle	Topo. Sort	Shortest Path	Max. Flow	Bipartite Graph	Hamilton Path	GNNs
				Node Recognition ↑				
MiniGPT-4 ♣(Vicuna-7b)	19.14	12.04	42.96	42.19	32.76	8.33	60.34	53.85
BLIP-2 ♣(FlanT5-xxl)	37.74	52.88	47.41	81.25	67.24	22.62	62.07	61.54
InstructBLIP♣ (FlanT5-xl)	36.12	47.64	46.67	75.00	56.90	36.90	53.45	74.36
InstructBLIP♣ (FlanT5-xxl)	35.31	52.88	61.48	85.94	77.59	17.86	65.52	61.54
Sphinx♣	61.99	98.95	94.07	100.0	91.38	55.95	100.00	97.44
Internlm♣	67.92	100.0	97.78	100.0	98.25	77.38	100.0	100.0
Llava-v1.5-7b♣	64.15	96.86	92.59	100.00	93.10	13.10	100.00	94.87
Llava-v1.5-13b♣	62.26	97.91	91.11	100.00	96.55	11.9	100.00	97.44
Qwen-Plus (0-shot)	2.96	0.00	0.00	0.00	5.17	0.00	0.00	56.41
Qwen-max (0-shot)	29.11	31.94	30.37	12.50	3.45	14.29	29.31	46.15
Gemini (0-shot)	40.97	42.93	47.41	67.19	72.41	10.71	65.52	35.90
GPT-4V (0-shot)	46.49	81.15	81.48	89.06	58.62	20.24	100.00	97.44
				Edge Recognition (Correct ↑ / Error ↓)				
MiniGPT-4 ♣(Vicuna-7b)	11.78/31.78	0.68/1.59	12.54/58.89	4.78/87.20	0.61/61.15	14.45/47.53	28.48/34.69	37.48/55.05
BLIP-2 ♣(FlanT5-xxl)	12.49/84.03	15.11/84.69	0.08/2.14	1.75/96.84	0.00/0.00	9.92/75.89	11.73/45.55	17.26/88.84
Sphinx ♣	44.76/66.69	22.13/79.69	37.84/73.07	39.88/70.62	20.68/86.57	83.93/53.51	66.26/71.15	60.66/61.43
Internlm ♣	53.08/35.01	40.78/60.05	55.70/50.85	57.82/45.02	23.45/80.27	71.21/42.34	73.98/36.00	83.00/19.69
InstructBLIP♣ (FlanT5-xl)	17.24/87.62	26.02/88.06	0.00/0.00	5.70/93.93	0.00/0.00	12.72/83.13	37.07/82.85	49.18/81.28
InstructBLIP♣ (FlanT5-xxl)	16.34/81.50	16.04/85.54	0.00/0.00	3.58/98.31	0.00/0.00	13.26/76.86	32.05/65.84	37.70/67.57
Llava-v1.5-7b♣	46.81/58.13	23.23/77.63	36.56/72.97	38.76/66.47	9.80/91.56	63.10/54.70	80.14/48.06	69.85/32.92
w/ Graph Understanding Data	54.87/38.55	49.86/42.36	30.37/64.41	49.86/40.49	8.50/90.45	35.44/53.50	71.90/14.77	58.73/24.07
Llava-v1.5-13b♣	51.18/53.41	22.60/76.91	38.80/70.26	41.93/63.50	9.89/91.72	67.88/54.21	76.26/45.21	67.40/33.59
w/ Graph Understanding Data	55.76/36.09	47.57/38.91	31.47/61.66	50.81/35.17	9.77/86.36	54.45/56.46	72.07/11.80	60.54/14.60
Qwen-Plus	30.46/64.78	27.42/82.37	10.59/68.46	6.16/81.60	1.32/64.62	75.93/58.65	48.63/50.41	33.71/60.56
Qwen-max	25.71/63.21	20.92/83.50	16.70/76.00	1.63/95.70	1.12/96.58	42.59/55.55	40.47/51.61	35.17/55.81
Gemini (0-shot)	23.26/52.35	21.65/80.09	19.11/66.94	16.18/83.09	4.79/94.78	66.01/53.90	39.40/37.80	40.83/52.60
GPT-4V (0-shot)	14.10/23.09	17.50/72.97	9.64/30.58	23.01/66.85	5.31/43.62	24.13/32.33	29.22/38.03	46.14/42.74
GPT-4V (4-shot)	20.63/34.52	26.25/69.95	13.19/51.75	23.40/61.90	6.12/84.94	46.33/51.69	58.49/49.79	48.06/35.01
				Accuracy ↑ on Specific Graph Theory Problems				
MiniGPT-4 ♣(Vicuna-7b)	50.67	48.69	0.00	0.00	0.00	5.95	0.00	0.00
BLIP-2 ♣(FlanT5-xxl)	46.63	61.26	0.00	0.00	13.79	0.00	0.00	0.00
InstructBLIP♣ (FlanT5-xl)	48.79	47.12	0.00	0.00	6.90	0.00	0.00	0.00
InstructBLIP♣ (FlanT5-xxl)	48.25	52.88	0.00	0.00	12.07	0.00	0.00	0.00
Llava-v1.5-7b♣	53.37	47.12	0.00	3.12	1.72	0.00	0.00	0.00
w/ Graph Understanding Data	63.61 ↑	56.02 ↑	0.00	0.00	1.72	0.00	0.00	0.00
Llava-v1.5-13b♣	52.83	47.12	0.00	4.69	3.45	0.00	0.00	0.00
w/ Graph Understanding Data	60.38 ↑	53.93 ↑	0.00	0.00	0.00	4.76 ↑	3.45 ↑	0.00
Gemini (0-shot)	55.52(14.01)	48.69(**6.80**)	0.00	0.00	3.45	1.72	0.00	0.00
GPT-4V (0-shot)	38.81(13.74)	49.21(0.52)	-	3.12	-	-	0.00	-
GPT-4V (2-shot)	54.98(19.13)	52.35(0.52)	-	6.25	-	-	0.00	-
GPT-4V (0-COT)	30.45(13.20)	50.26(0.00)	-	**7.69**	-	-	0.00	-
GPT-4V (2-COT)	54.71(**19.40**)	52.87(0.52)	-	6.25	-	-	0.00	-

VisionGraph: Improving the spatial reasoning of MLMs

Multi-Step Graph Reasoning – Multimodal Agents

- Propose DPR to enhance graph structure perception and multistep reasoning.
- GPT-4V+DPR can be considered as a multimodal agent, integrating complex task decomposition, small model enhancement, code generation, and tool invocation.

1) Using graph understanding-enhanced Llava-7b to generate graph explanations, inputting them into GPT-4V to enhance graph structure understanding.
2) Generate graph representation: adjacency matrices.
3) Select specific algorithms and generate relevant codes.
4) Invoke tools to execute the produced codes.



Description-Programming-Reasoning (DPR)

Question: *Is there a path between node 3 and node 1? If yes, give the path.*

Gemini (0-shot): Yes, there is a path between node 3 and node 1. The path is 3, 2, 0, 1.

GPT-4V (0-shot): Yes, there is a path between node 3 and node 1. The path is 3, 7, 0, 4, 1.

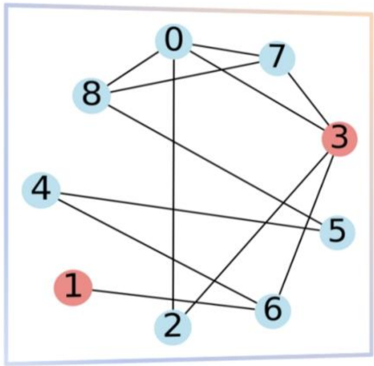
GPT-4V (2-shot+COT): Yes, there is a path between node 3 and node 1. The path is 3, 5, 2, 1.

GPT-4V (DPR):

1. *The edges of the graph are represented as tuples*: (0, 2), (0, 3), (0, 7), (0, 8), (1, 6), (2, 6), (3, 6), (4, 5), (4, 6), (5, 8), (6, 7).

2. *The algorithm and corresponding codes are*: To determine if there is a path between node 3 and node 1, we can use a simple graph traversal algorithm such as Depth First Search (DFS) or Breadth First Search (BFS). Here is an example of Python code using BFS:

```python
from collections import deque
def bfs(graph, start, end):
    visited = set()
    queue = deque([start])
    while queue:
        node = queue.popleft()
        if node == end:
            return True
        if node not in visited:
            visited.add(node)
            queue.extend(graph[node] - visited)
    return False
graph = {0: {2, 3, 7, 8}, 1: {6}, 2: {0, 6}, 3: {0, 6}, 4: {5, 6}, 5: {4, 8}, 6: {1, 2, 3, 4, 7}, 7: {0, 6}, 8: {0, 5}}
print(bfs(graph, 3, 1))
```



3. *The process of multi-step reasoning is as follows*:
The neighboring nodes to node 3 are nodes 0 and 6.
The neighboring nodes to node 0 are nodes 2, 7, and 8.
The neighboring nodes to node 6 are nodes 1, 2, 4, and 7.
Since node 6 is a neighbor of node 3 and node 1 is a neighbor of node 6, there is a path from node 3 to node 1 through node 6.

4. Yes, there is a path between node 3 and node 1. The path is 3-6-1.

VisionGraph: Improving the spatial reasoning of MLMs

Multistep Graph Reasoning Capability

- The interleaved logical reasoning chain of natural language and code enhances the complex problem-solving abilities.
- GPT-4V excels in selecting and understanding algorithms and DPR can enhance strengths and mitigate weaknesses.
- The DPR practice results on open-source models are also promising.
- After invoking the Python Tool, GPT-4V's performance is further improved.
- However, the limited visual perception capability of LMM leads to hallucinations (nonexistent paths), ultimately resulting in poor performance on complex problems, such as the shortest path.

Table 4. Model performance on three common graph theory problems in VisionGraph. ♣ refers to that the corresponding model is trained using the training set of VisionGraph. "w/ Python" shows using external Python interpretation to run algorithms and return final answers.

Task Types →	Connectivity ↑				Cycle ↑				Shortest Path ↑		
Model↓	Easy	Medium	Hard	Avg.	Easy	Medium	Hard	Avg.	Easy	Hard	Avg.
MiniGPT-4 ♣ (Vicuna-7b)	60.71	53.57	52.94	54.45	36.00	51.40	59.32	51.83	0.00	0.00	0.00
BLIP-2 ♣ (FlanT5-xxl)	37.50	43.37	56.30	46.63	88.00	63.55	45.76	61.26	0.00	0.00	0.00
InstructBLIP♣ (FlanT5-xl)	46.43	46.43	53.78	48.79	36.00	50.47	45.76	47.12	0.00	0.00	0.00
Sphinx	39.29	45.41	52.1	46.63	64.00	49.53	54.24	52.88	6.90	0.00	3.12
Sphinx♣ w/ DPR	67.86	59.69	52.94	58.76	64.00	49.53	54.24	52.88	13.78	0.00	6.25
Internlm♣	78.57	66.33	52.10	52.94	52.00	55.14	59.32	56.02	0.00	0.00	0.00
Internlm♣ w/ DPR	89.29	72.96	56.30	70.08	44.00	57.01	64.41	57.59	0.00	0.00	0.00
Llava-v1.5-7b♣	64.29	50.00	53.78	53.27	36.00	50.47	45.76	47.12	6.90	0.00	3.12
w/ Graph Understanding Data	89.29	64.80	49.58	63.61	68.00	53.27	55.93	56.02	0.00	0.00	0.00
w/ DPR	80.36	68.37	48.74	63.88 ↑	68.00	51.40	55.93	54.97	0.00	0.00	0.00
Llava-v1.5-13b♣	71.43	49.49	49.58	52.83	36.00	50.47	45.76	47.12	10.34	0.00	4.69
w/ Graph Understanding Data	78.57	62.76	47.90	60.38	64.00	51.40	54.24	53.93	0.00	0.00	0.00
w/ DPR	83.93	70.92	50.42	66.31 ↑	60.00	64.49	55.93	61.26 ↑	0.00	0.00	0.00
Gemini (0-shot)	69.64(39.29)	56.63(13.78)	47.06(2.52)	55.52(14.01)	60.00(0.00)	47.66(4.67)	45.76(13.56)	48.69(6.80)	0.00	0.00	0.00
Gemini (DPR)	66.07(57.14)	52.04(27.04)	36.97(5.88)	49.32(24.79)	76.00(16.00)	27.10(5.61)	22.03(0.00)	31.93(5.23)	0.00	0.00	0.00
Qwen-plus	62.50(19.64)	56.63(9.69)	47.06(3.36)	54.45(9.16)	64.00(0.00)	49.53(0.00)	54.24(0.00)	52.88(0.00)	0.00	0.00	0.00
Qwen-plus w/ DPR	57.14(1.79)	46.43(4.08)	35.29(5.88)	44.47(4.31)	64.00(16.00)	56.07(22.43)	52.54(20.34)	56.02(20.94)	6.90	0.00	3.12
Qwen-max	62.50(16.07)	56.63(3.06)	46.22(0.84)	54.18(4.31)	64.00(16.00)	49.53(0.00)	54.24(0.00)	52.88(0.00)	0.00	0.00	0.00
Qwen-max w/ DPR	60.71(12.50)	51.02(12.24)	27.73(6.72)	45.01(10.51)	64.00(16.00)	52.34(10.28)	50.85(1.69)	53.40(8.38)	20.69	2.86	10.93
GPT-4V (0-shot)	69.64(46.43)	42.86(12.76)	17.65(0.00)	38.81(13.74)	60.00(4.00)	48.60(0.00)	45.76(0.00)	49.21(0.52)	6.90	0.00	3.12
GPT-4V (2-shot)	67.86(42.86)	56.12(18.88)	47.06(8.40)	54.98(19.13)	64.00(4.00)	48.60(0.00)	54.24(0.00)	52.35(0.52)	13.79	0.00	6.25
GPT-4V (0-COT)	64.29(37.50)	34.69(13.78)	7.56(0.84)	30.45(13.20)	64.00(0.00)	47.66(0.00)	49.15(0.00)	50.26(0.00)	17.24	0.00	7.69
GPT-4V (2-COT)	67.86(44.64)	56.63(22.96)	45.38(1.68)	54.71(19.40)	64.00(4.00)	49.53(0.00)	54.24(0.00)	52.87(0.52)	13.79	0.00	6.25
GPT-4V (DPR)	92.86(89.29)	58.67(44.90)	36.97(16.81)	56.87(42.58)	76.00(52.00)	48.60(15.89)	45.76(1.69)	51.30(16.23)	24.14	2.86	12.50
w/ Python	92.86(91.07)	61.73(53.06)	51.26(35.29)	**63.07(53.09)** ↑	88.00(72.00)	61.68(34.58)	55.93(20.34)	**63.35(35.07)** ↑	31.03	11.43	**20.31** ↑

Take Away

- The intertwined thinking chain of natural language and code facilitates tool invocation, maintains rigorous reasoning, and is easy to trace.
- Breakthroughs in spatial perception (Feifei Li) will significantly enhance the spatial planning capabilities of visual-language large models, benefiting robotic planning and vision-language navigation.
- Combining advanced visual-language descriptions with fundamental visual perception may help improve the overall capabilities of multimodal large language models to solve complex visual reasoning problems.

Paper Link: https://arxiv.org/abs/2405.04950

Github Code: https://github.com/HITsz-TMG/VisionGraph/tree/main