VISUAL AGENTIC AI FOR SPATIAL REASONING WITH A DYNAMIC API

TEAM VISIONCORE

- Amarathunga D.N 210037G
- Jayathilaka D.E.U 210254T

OVERVIEW

"VADAR is a **multi-agent system** designed for **complex visual reasoning**, where large language model agents dynamically generate and execute python programs to solve **3D spatial queries**."

Query: If the black armchair is 1 cm away in 3D, how far away in 3D is the dining table?

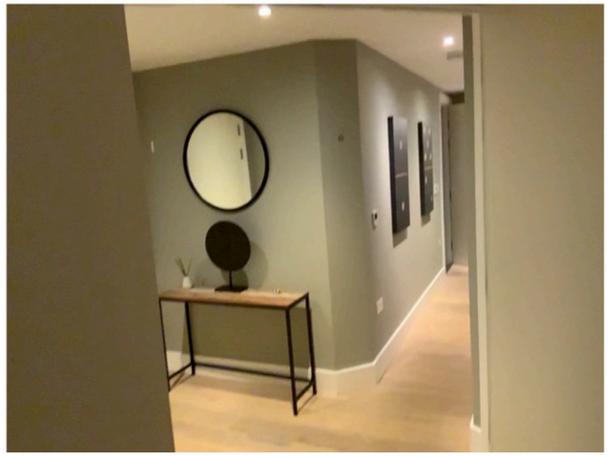
Answer: 3.5cm

Query: How many chairs are there further away from the camera than the bowl?

Answer: 3

Query: If the table is 20 meters tall in 3D, what is the radius of the mirror? Respond with a number in meters.

Answer: 8.0m



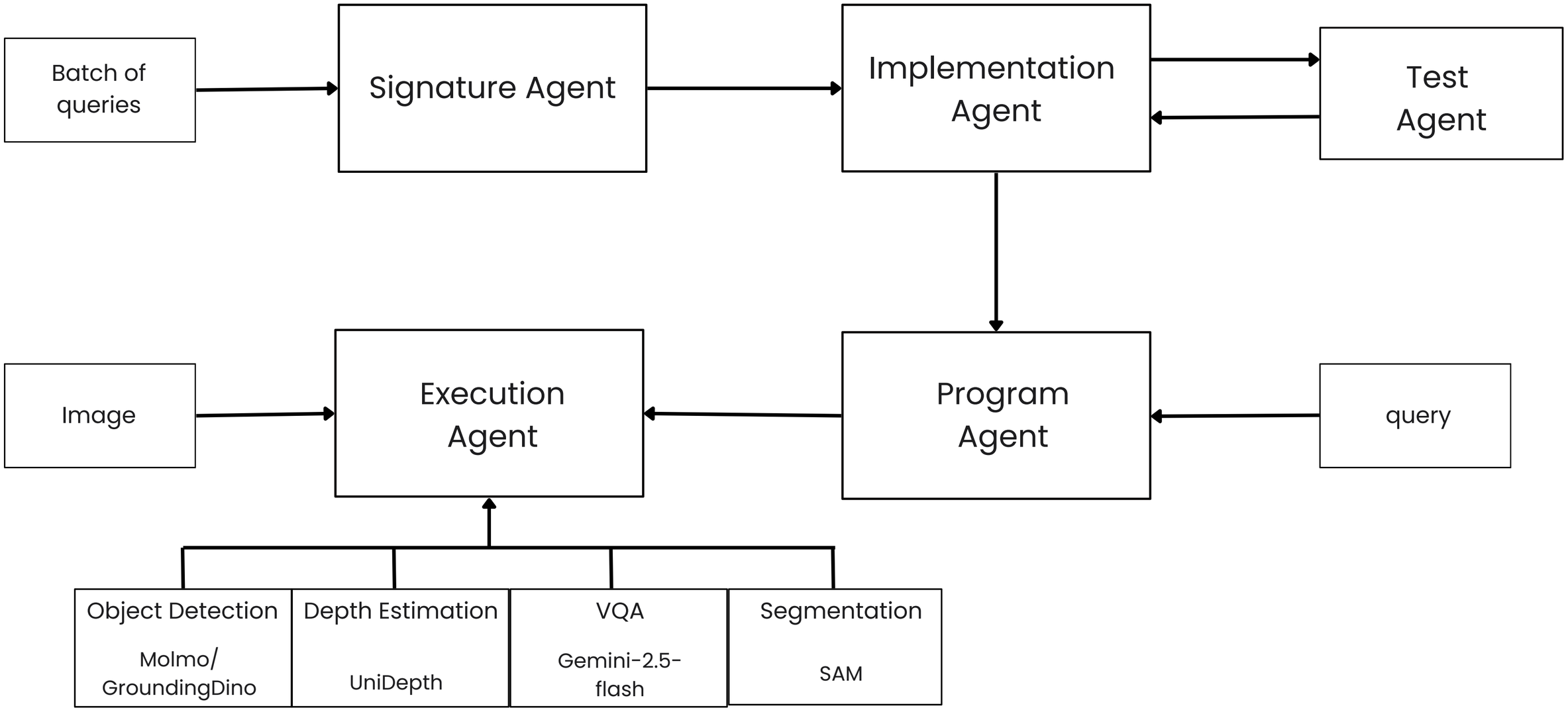
IMPORTANTANCE OF VADAR COMPARED TO TRADITIONAL VLMS

- Stronger 3D spatial reasoning.
- Outperforms prior zero-shot systems.
- Dynamic API resulting in more expressive reasoning.

		CLEVR				OMNI3D-BENCH				
		numeric	y/n	multi-choice	Total	numeric (ct)	numeric (other)	y/n	multi-choice	Total
VLMs	GPT4o [1]	52.3	63.0	60.0	58.4	**28.1**	**35.5**	**66.7**	57.2	**42.9**
	Claude3.5-Sonnet [2]	44.7	61.4	**72.2**	**58.9**	22.4	20.6	62.2	50.6	32.2
	Llama3.2 [9]	34.6	45.6	49.0	42.8	24.3	19.3	47.5	27.4	25.6
	Gemini1.5-Pro [36]	44.9	59.7	67.0	56.9	25.2	28.1	46.2	37.6	32.0
	Gemini1.5-Flash [36]	43.1	58.8	56.8	52.8	24.3	27.6	51.1	52.9	35.0
	Molmo [8]	11.0	42.6	51.4	34.4	21.4	21.7	29.3	41.2	26.1
	SpaceMantis [6, 17]	14.5	52.9	32.3	33.2	20.0	21.7	50.6	48.2	30.3
Program Synthesis	ViperGPT [35]	20.5	43.4	13.4	26.2	20.0	15.4	56.0	42.4	26.7
	VisProg [12]	16.7	48.4	28.3	31.2	2.9	0.9	54.7	25.9	13.5
	VADAR (ours)	**53.3**	**65.3**	40.8	53.6	21.7	**35.5**	56.0	**57.6**	40.4

	CLEVR				OMNI3D-BENCH				
	numeric	y/n	multi-choice	Total	numeric (ct)	numeric (other)	y/n	multi-choice	Total
ViperGPT [35]	38.5	57.8	30.2	42.6	50.0	17.8	66.7	49.3	54.9
VisProg [12]	25.3	52.5	41.8	39.9	**100.0**	23.5	68.5	66.7	66.0
VADAR (ours)	**82.4**	**85.4**	**81.0**	**83.0**	**100.0**	**82.3**	**100.0**	**94.1**	**94.4**
GPT4o	52.3	63.0	66.0	58.4	30.0	29.4	77.8	44.0	53.7
Claude3.5-Sonnet	44.7	61.4	72.2	58.9	30.0	35.3	83.3	56.0	59.3

METHODOLOGY



INTERMIDIATE RESULTS

Query: Which object has the largest height-to-width ratio: the tv or the leftmost painting?

APIs Used

- loc(image, object)
- get_2D_object_size(image, bounding_box)

Generated Program

```
tv_bboxes = loc(image, "tv")
tv_bbox = tv_bboxes[0] # Assuming "the tv" refers to the first detected tv

painting_bboxes = loc(image, "painting")

leftmost_painting_bbox = None
min_xmin = float('inf')

for bbox in painting_bboxes:
    xmin = bbox[0]
    if xmin < min_xmin:
        min_xmin = xmin
        leftmost_painting_bbox = bbox

tv_width, tv_height = get_2D_object_size(image, tv_bbox)
tv_ratio = tv_height / tv_width

painting_width, painting_height = get_2D_object_size(image, leftmost_painting_bbox)
painting_ratio = painting_height / painting_width

if tv_ratio > painting_ratio:
    final_result = "the tv"
else:
    final_result = "the leftmost painting"
```

EXECUTION TRACE

```
[solution_program] Line 4: tv_bboxes = loc(image, "tv")
```

Locate: tv

1.

1 tv found

Boxes: [[334, 276, 448, 316]]

Locate: painting

2.

4 paintings found

Boxes: [[165, 250, 242, 308], [48, 305, 105, 344], [333, 275, 448, 317], [48, 261, 105, 301]]

EXECUTION TRACE

```
[solution_program] Line 18: tv_width, tv_height = get_2D_object_size(image, tv_bbox)
```

2D Object Size

3.

Width: 114, Height: 40

2D Object Size

4.

Width: 57, Height: 39

KEY CHALLENGES AND MITIGATIONS

- **Large Language Model Selection**

Our Design

Research Paper

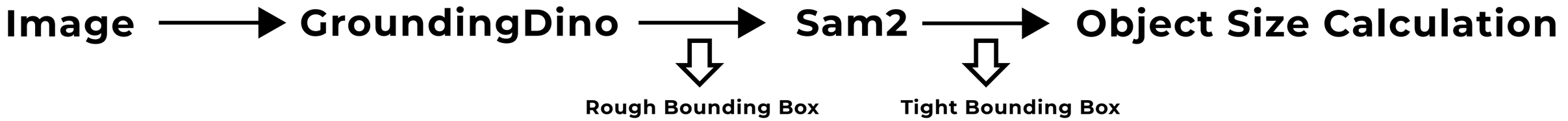
- **Molmo-7B caused CUDA Out-of-Memory errors due to limited GPU VRAM.**

FP16 precision and 4-bit NF4 quantization for weights, with dequantization to FP16 during computation.

- **Low numeric accuracy for Omni3D-Bench compared to the original research paper.**

Object Depth Extraction

Object Size Extraction



RESULT COMAPARISON

	Original Paper Accuracy		Implementation Accuracy	
Category	Omni-3D	CLEVR	Omni-3D	CLEVR
Numeric (count)	21.7	53.3	19	81.8
Numeric (float)	35.5	NA	31.2	NA
Yes or No	56	65.3	47	68.2
Multiple Choice	57.6	40.8	58.8	70.6
Total	40.4	53.6	38.9	72

VADAR + Scenes Data for CLEVR Dataset	
Original Paper Accuracy	**83.0**
Implemetation Accuracy	**88.5**

THANK YOU!