SpatialRGPT: Grounded Spatial Reasoning in Vision Language Models

An-Chieh Cheng[1], Hongxu Yin[2], Yang Fu[1], Qiushan Guo[2], Ruihan Yang[1], Jan Kautz[2], Xiaolong Wang[1,2], Sifei Liu[2]

[1] UC San Diego | [2] NVIDIA

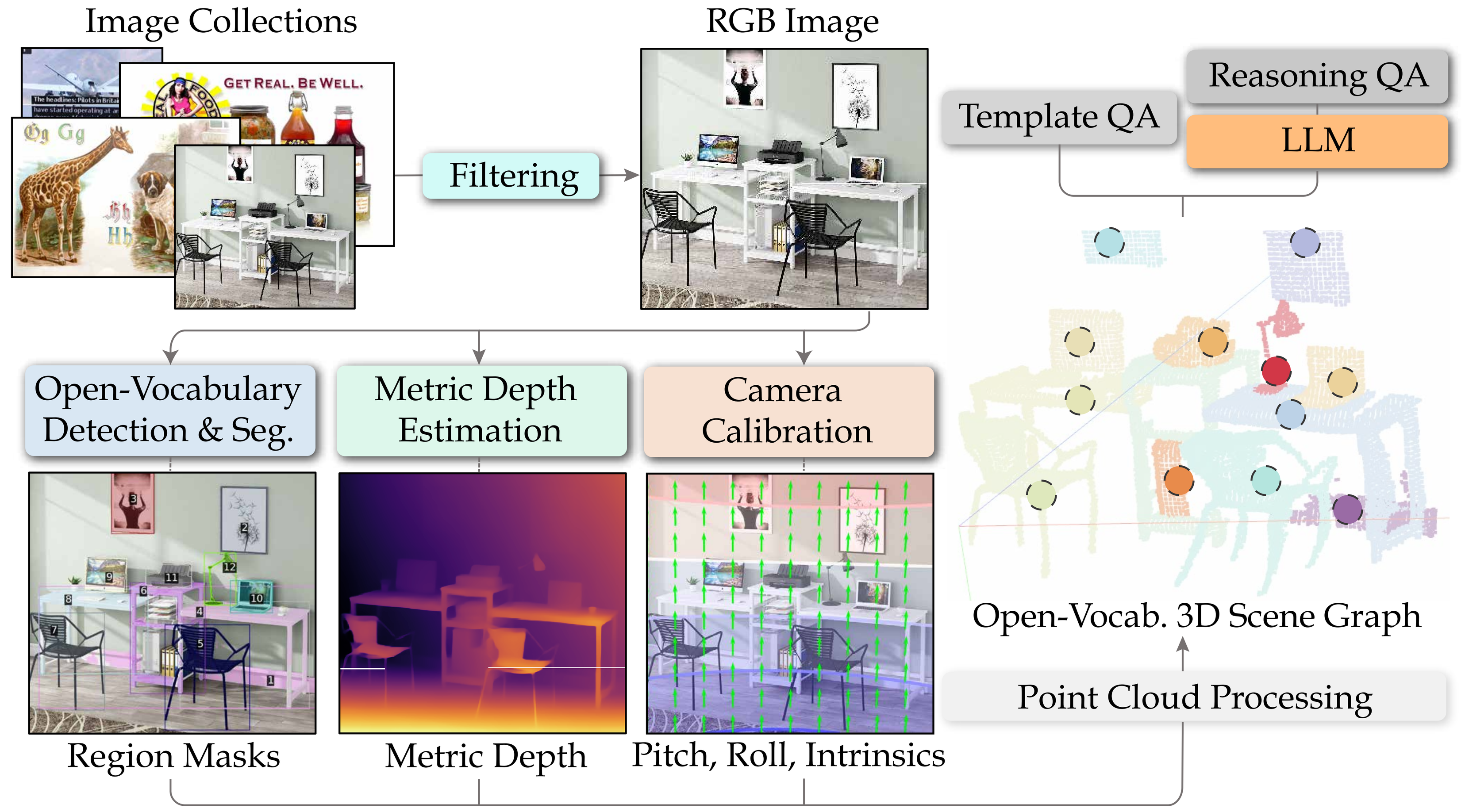
NEURAL INFORMATION PROCESSING SYSTEMS

Code | Weight
Data | Benchmark

Overview

- Current VLMs lack spatial knowledge and spatial reasoning
- SpatialRGPT excels in 2D/3D spatial understanding, using region proposals to answer complex spatial questions
- SOTA results on our new benchmark and public benchmarks

Open Spatial Dataset: Pipeline

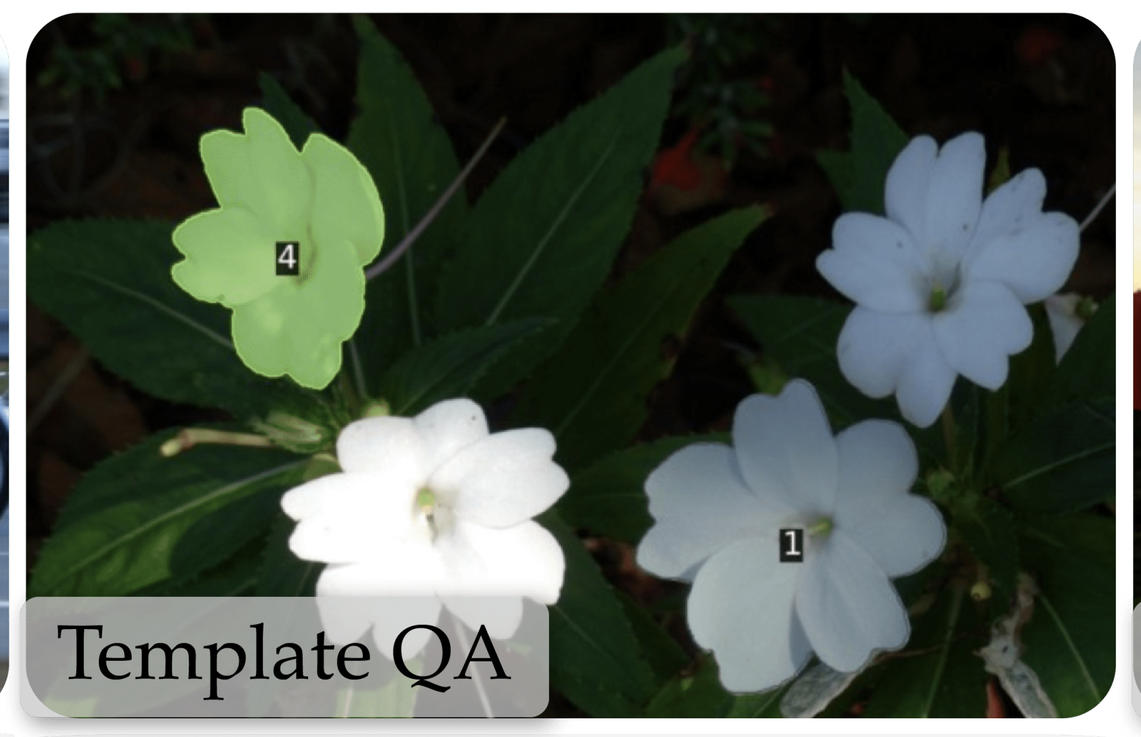


Open Spatial Dataset: Samples

- 8.7M spatial concepts grounded in 5M regions of 1M images

How wide is Region [1]?

The width of Region [1] is 7.73 feet.

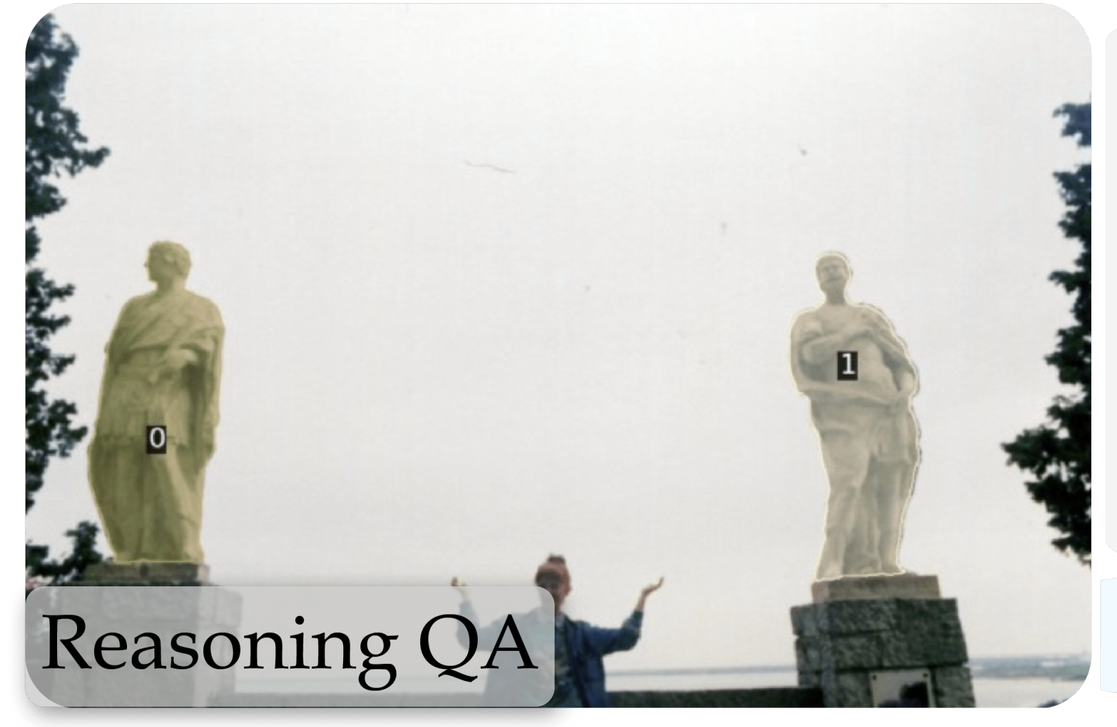


Is Region [1] behind Region [4]?

No, it is in front of Region [4].

Between Region [0] and Region [2], which one has more height?

Region [2] is taller.



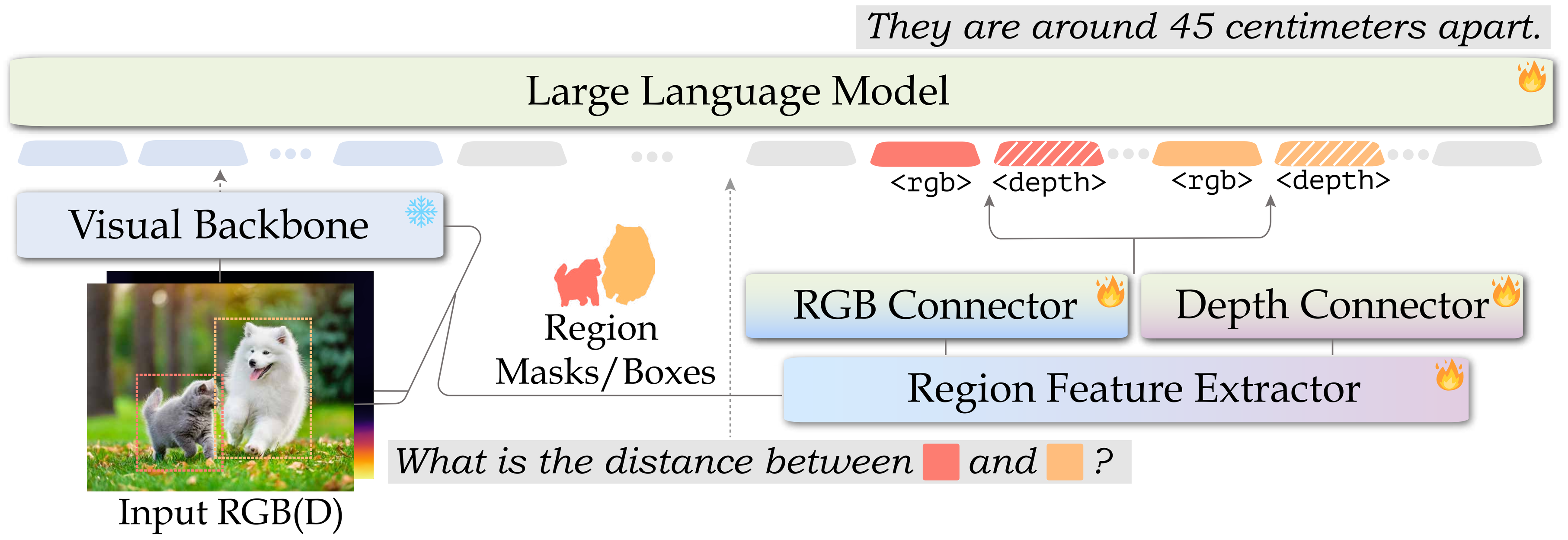
You are a visitor in a museum and see two sculptures, one in Region [0] and the other in Region [1]. If you walk from one sculpture to the other, how far will you have walked?

You will have walked 4.85 meters.

You are a helicopter pilot flying over the city and you see Region [1] and Region [7]. Which one is higher?

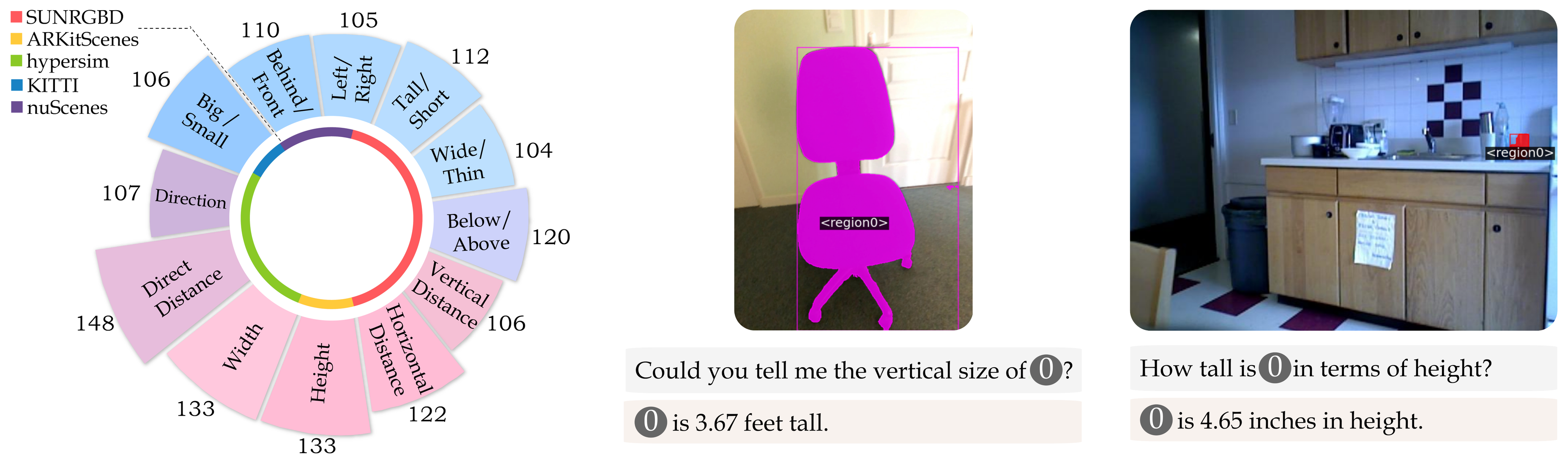
The tower in Region [1] is higher than the skyscraper in Region [7].

Model Architecture



SpatialRGPT-Bench

- 1406 VQA Pairs / 88 Classes / Ground-truth 3D Annotations



Results: SpatialRGPT-Bench

	Below/ Above	Left/ Right	Big/ Small	Tall/ Short	Wide/ Thin	Behind/ Front	Avg.
GPT-4 [55]	64.16	42.85	42.85	61.60	61.60	49.09	57.83
GPT-4V [55]	63.34	46.67	64.15	60.71	68.26	45.45	58.14
LLaVA-v1.6-34B [56]	44.16	45.71	36.79	53.57	37.50	45.45	43.98
SpatialRGPT-8B	**99.17**	**100.0**	**84.90**	**89.28**	**91.34**	**90.90**	**92.69**

	Direct Distance	Horizontal Distance	Vertical Distance	Width	Height	Direction
GPT-4 [55]	21.6 / 1.29	11.5 / 2.08	33.0 / 0.65	**52.3** / 0.52	48.1 / 1.40	34.6 / 83.7°
GPT-4V [55]	29.7 / 0.92	25.4 / 2.75	33.0 / 0.48	51.1 / 0.37	**68.4** / 1.57	43.9 / 69.9°
LLaVA-v1.6-34B [56]	24.3 / 0.76	24.5 / 1.59	30.1 / 0.62	30.8 / 0.40	42.8 / 1.96	33.6 / 78.2°
SpatialRGPT-8B	**45.9** / **0.31**	**68.0** / **0.22**	**56.6** / **0.28**	48.9 / **0.28**	61.7 / **0.41**	**95.3** / **9.7°**

Grounded Spatial Reasoning

(A) Relative Relation

Can you confirm if 4 has a greater width compared to 6?

In fact, 4 might be narrower than 6.

Which of these four, 3, 4, 5, 6 is the tallest?

6 is the tallest.

From the viewer's perspective, is the position of 6 closer than that of 3?

No, 3 is closer.

(B) Metric Measurement

If you are at 3, where will you find 7?

7 is at 1 o'clock from 3.

What is the height of 4?

4 is 1.38 meters tall.

How far apart are 2 and 3 horizontally?

They are 15.33 meters apart horizontally.

(C) Spatial Reasoning

Can you try to estimate how many floors are in 1? Think step by step.

The height of 1 is 204.54 feet. Assuming each floor is about 10 feet high, the total number of floors would be 20.454. Since you can't have a fraction of a floor, the total number of floors would be approximately 20.

I'm planning to build a bridge from 2 to 1, is it possible?

The distance between 2 and 1 is 14.82 meters, so it is possible to build a bridge between them.

If you are riding a motorcycle with 36 inches wide, do you think you can pass through the area between 3 and 4?

The distance between 3 and 4 is 38.95 inches, so yes, you can pass through the area between 3 and 4 since the motorcycle is narrower than the distance between them.

Multi-hop Reasoning

What is the object on the table to the right of Region [0], and what is its distance to Region [0]?

The object on the table to the right of Region [0] is a potted plant, and it is 13.9 inches away from Region [0].

What is the object to the left of Region [0], and what is its height and width?

The object to the left of Region [0] is a candle, and it is 10.49 inches in height and 3.39 inches in width.

BLINK$_{Depth}$

Model	Acc. (%)
Qwen-VL-Max	58.9
Gemini Pro	50.0
Claude 3 OPUS	57.3
GPT-4V-*Turbo*	66.9
GPT-4o	64.5
SpatialRGPT-8B	**87.9**

Region-aware Dense Reward Annotator