

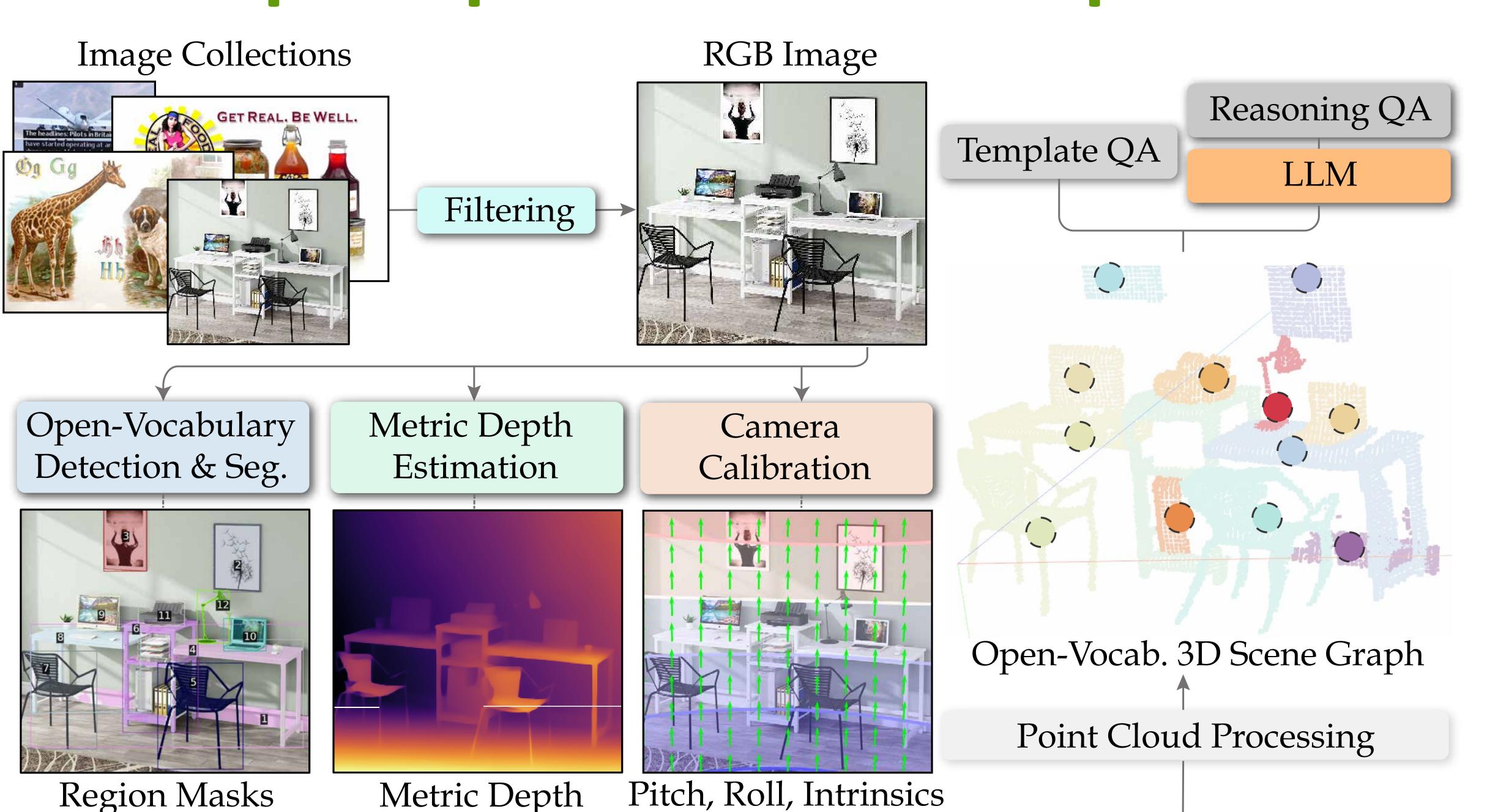
# SpatialRGPT: Grounded Spatial Reasoning in Vision Language Models

An-Chieh Cheng¹, Hongxu Yin², Yang Fu¹, Qiushan Guo², Ruihan Yang¹, Jan Kautz², Xiaolong Wang¹,², Sifei Liu²

#### Overview

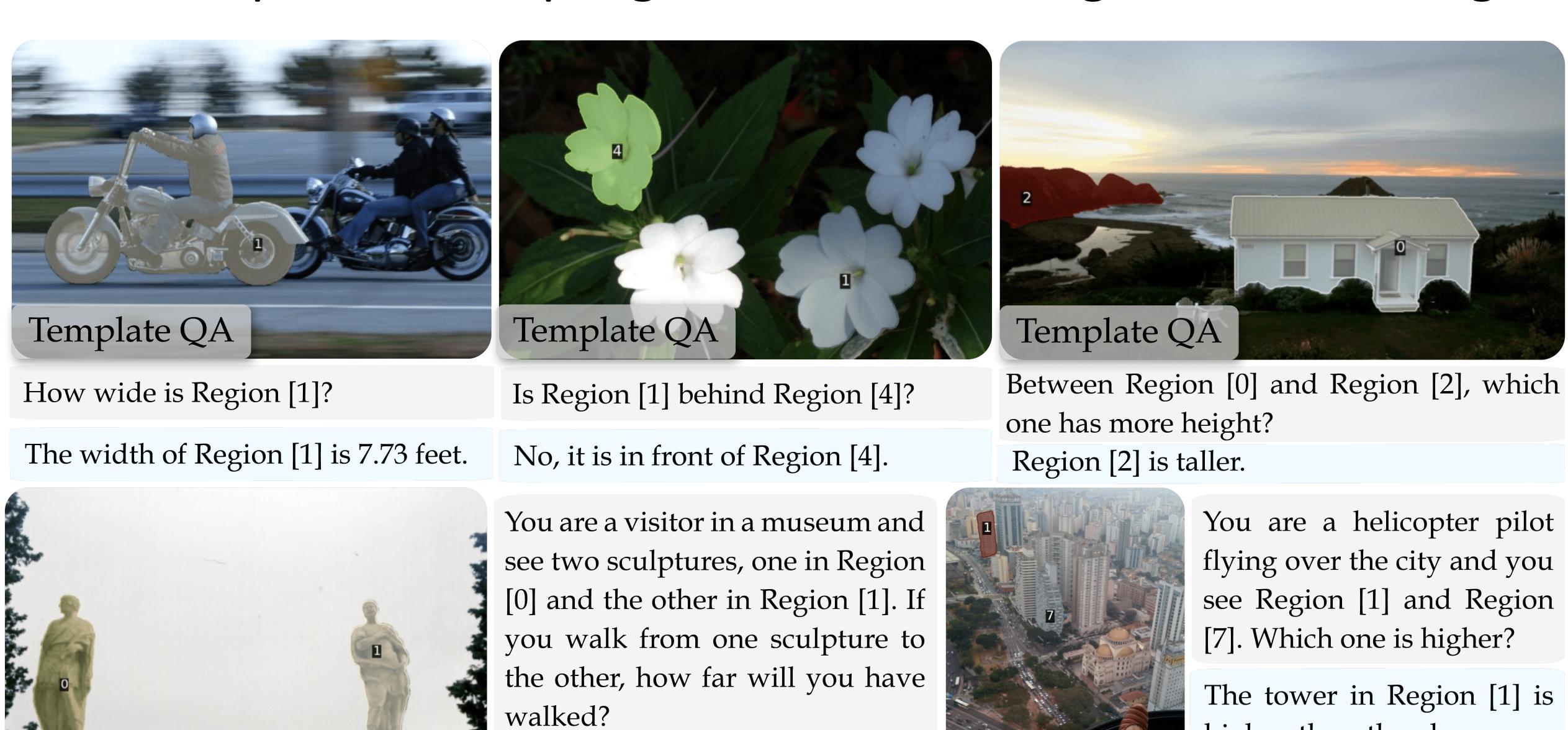
- Current VLMs lack spatial knowledge and spatial reasoning
- SpatialRGPT excels in 2D/3D spatial understanding, using region proposals to answer complex spatial questions
- SOTA results on our new benchmark and public benchmarks

## Open Spatial Dataset: Pipeline



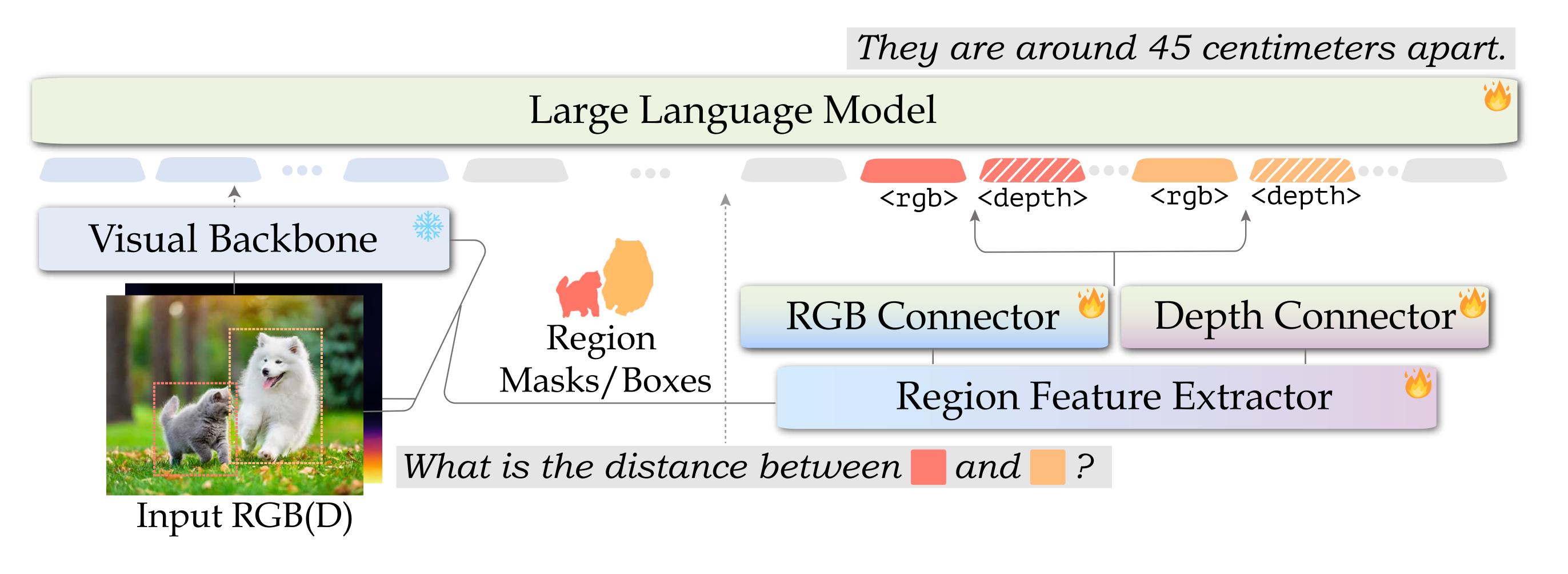
# Open Spatial Dataset: Samples

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



You will have walked 4.85 meters. Reasoning QA in Region [7].

#### Model Architecture

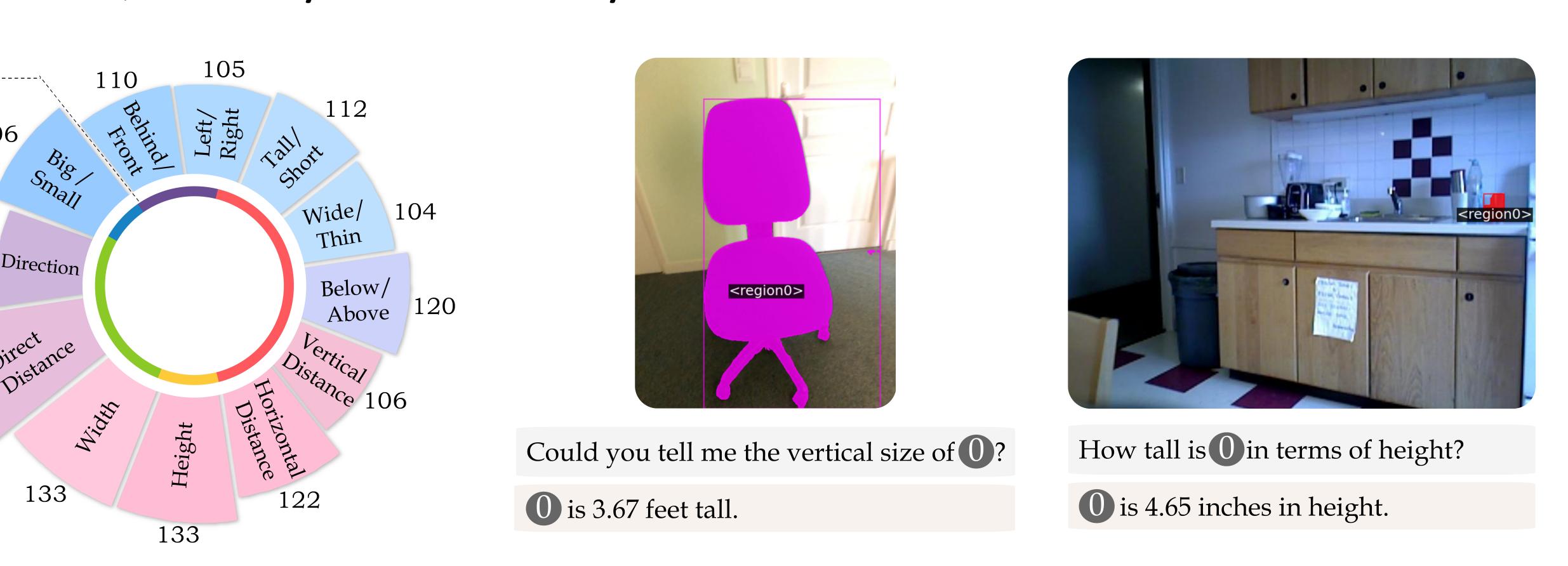


#### SpatialRGPT-Bench

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

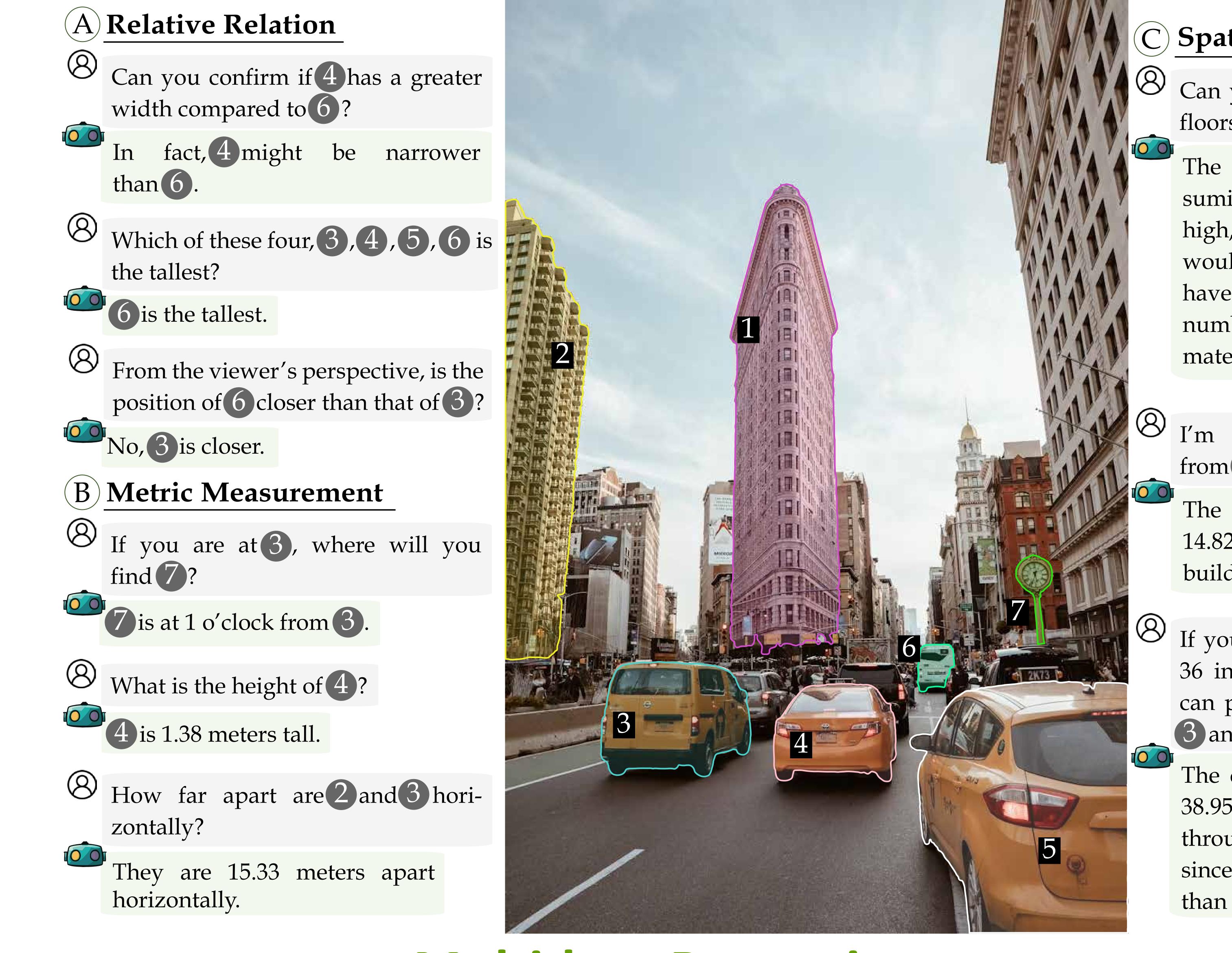
hypersinKITTInuScenes

nigher than the skyscraper

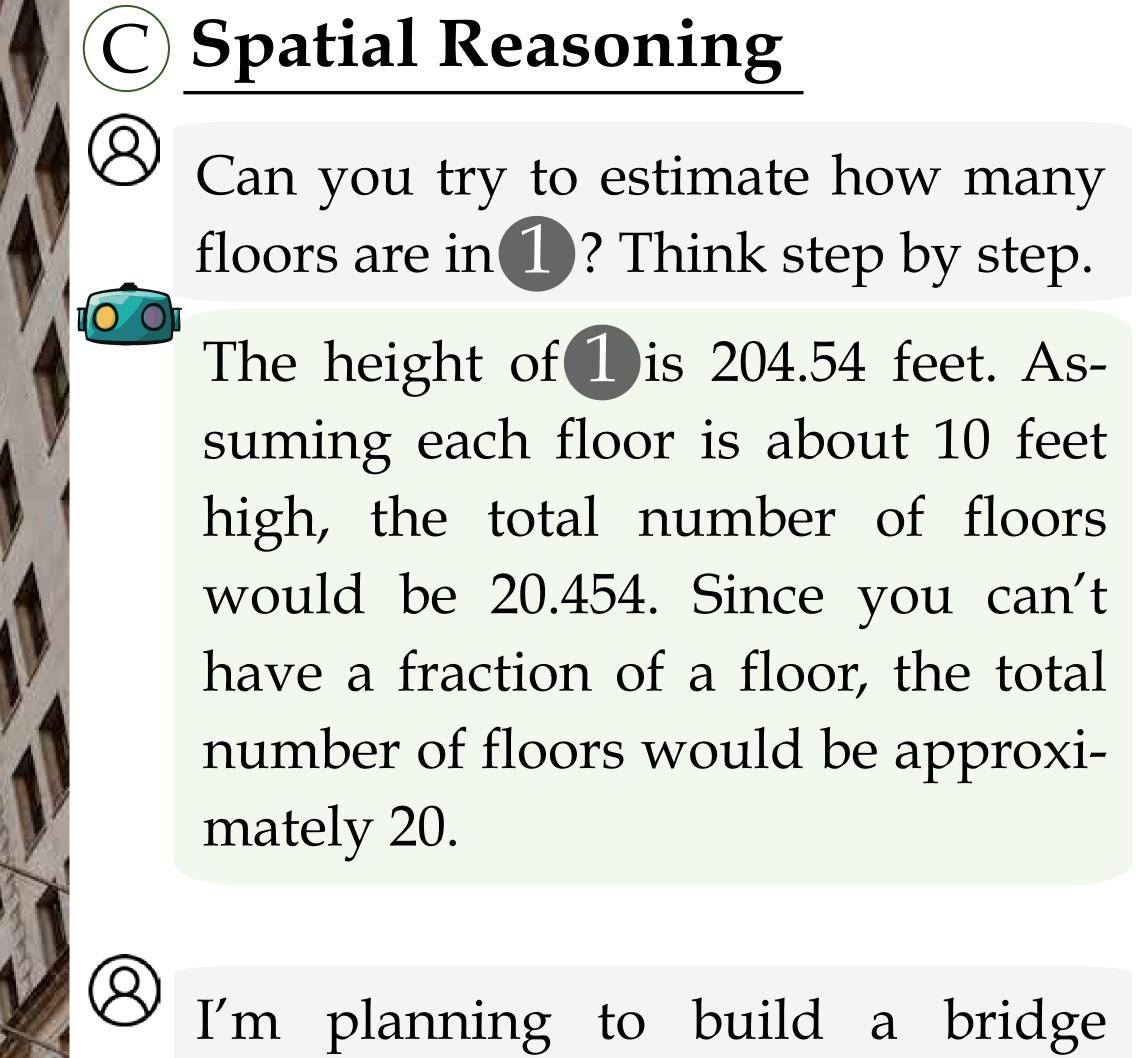


### Results: SpatialRGPT-Bench

	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] LLaVA-v1.6-34B [56]	63.34 44.16	46.67 45.71	64.15 36.79	60.71 53.57	68.26 37.50	45.45 45.45	58.14 43.98
SpatialRGPT-8B	99.17	100.0	84.90	89.28	91.34	90.90	92.69
	Direct Distance	Horizonta Distance		rtical tance	Width	Height	Direction
GPT-4 [55]	21.6 / 1.29	11.5 / 2.08	33.0	/ 0.65	<b>52.3</b> / 0.52	48.1 / 1.40	34.6 / 83.7°
GPT-4V [55] LLaVA-v1.6-34B [56]	29.7 / 0.92 24.3 / 0.76	25.4 / 2.75 24.5 / 1.59		/ 0.48 / 0.62	51.1 / 0.37 30.8 / 0.40	<b>68.4</b> / 1.57 42.8 / 1.96	43.9 / 69.9° 33.6 / 78.2°
SpatialRGPT-8B	<b>45.9</b> / 0.31	68.0 / 0.22	56.6	/ 0.28	48.9 / <b>0.28</b>	61.7 / <b>0.41</b>	95.3 / 9.7°



# **Grounded Spatial Reasoning**



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]?

> Region [0] is a potted plant, and it is 13.9 inches away from Region [0].



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.

[0], and what is its height and width?

Model	Acc. (%)
Qwen-VL-Max	58.9
Gemini Pro	50.0
Claude 3 OPUS	57.3

GPT-4V-Turbo

BLINKDepth

#### Region-aware Dense Reward Annotator

