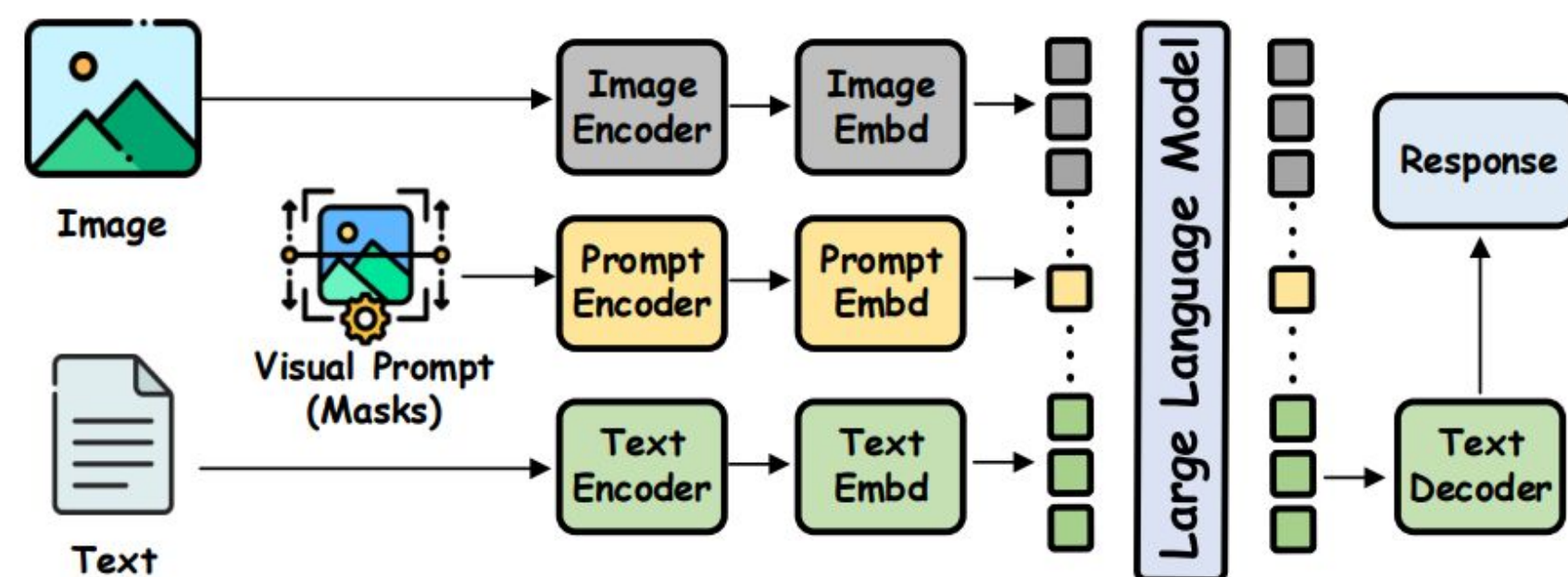




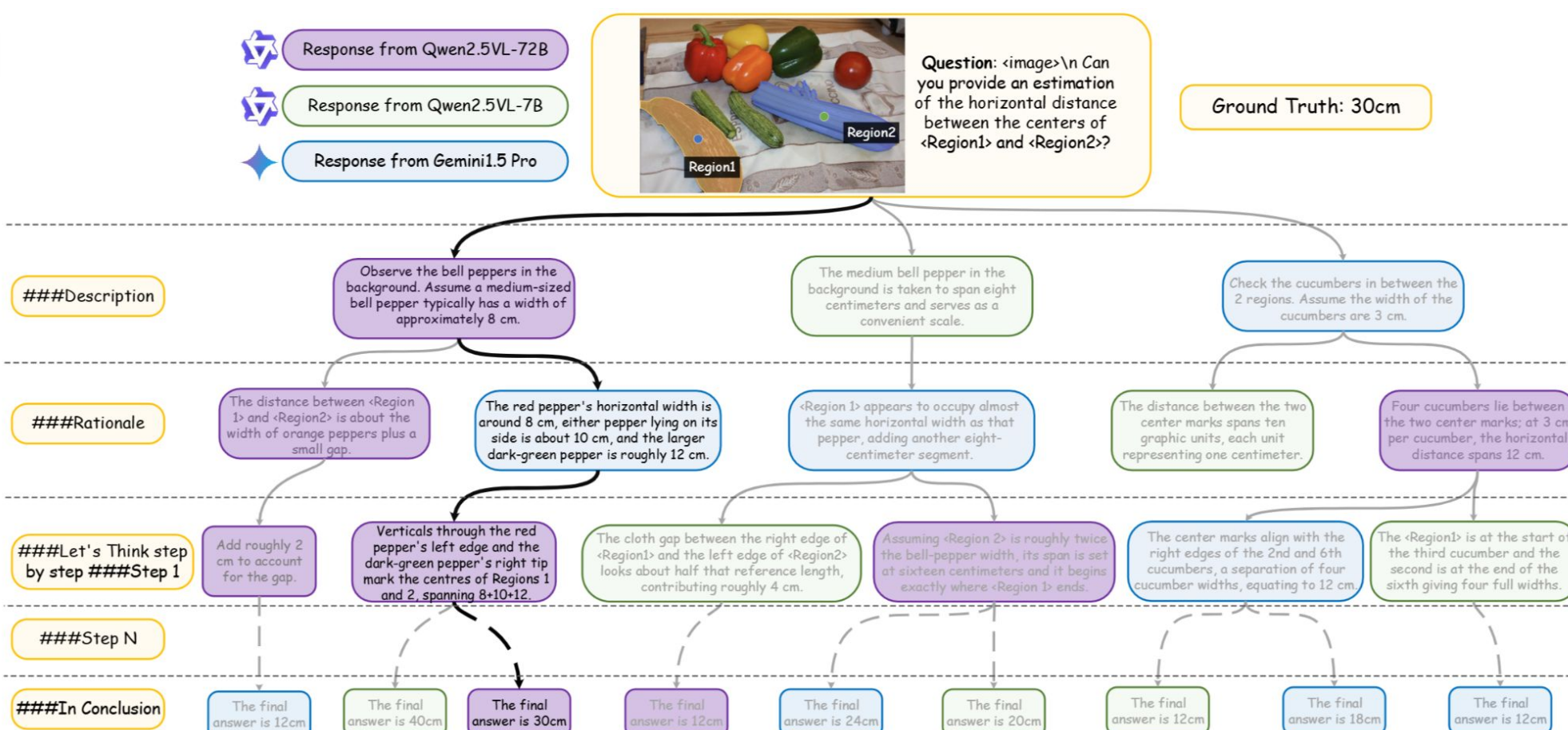
Overview

- Current MLLMs struggle with fine-grained visual understanding and coherent step-by-step reasoning.
- We propose a three-stage pipeline: generate LongCoT with **M3CTS**, construct preference pairs, apply a novel **fine-grained Preference Optimization (fDPO)** method that employs segment-specific learning updates tailored for descriptive grounding and logical reasoning

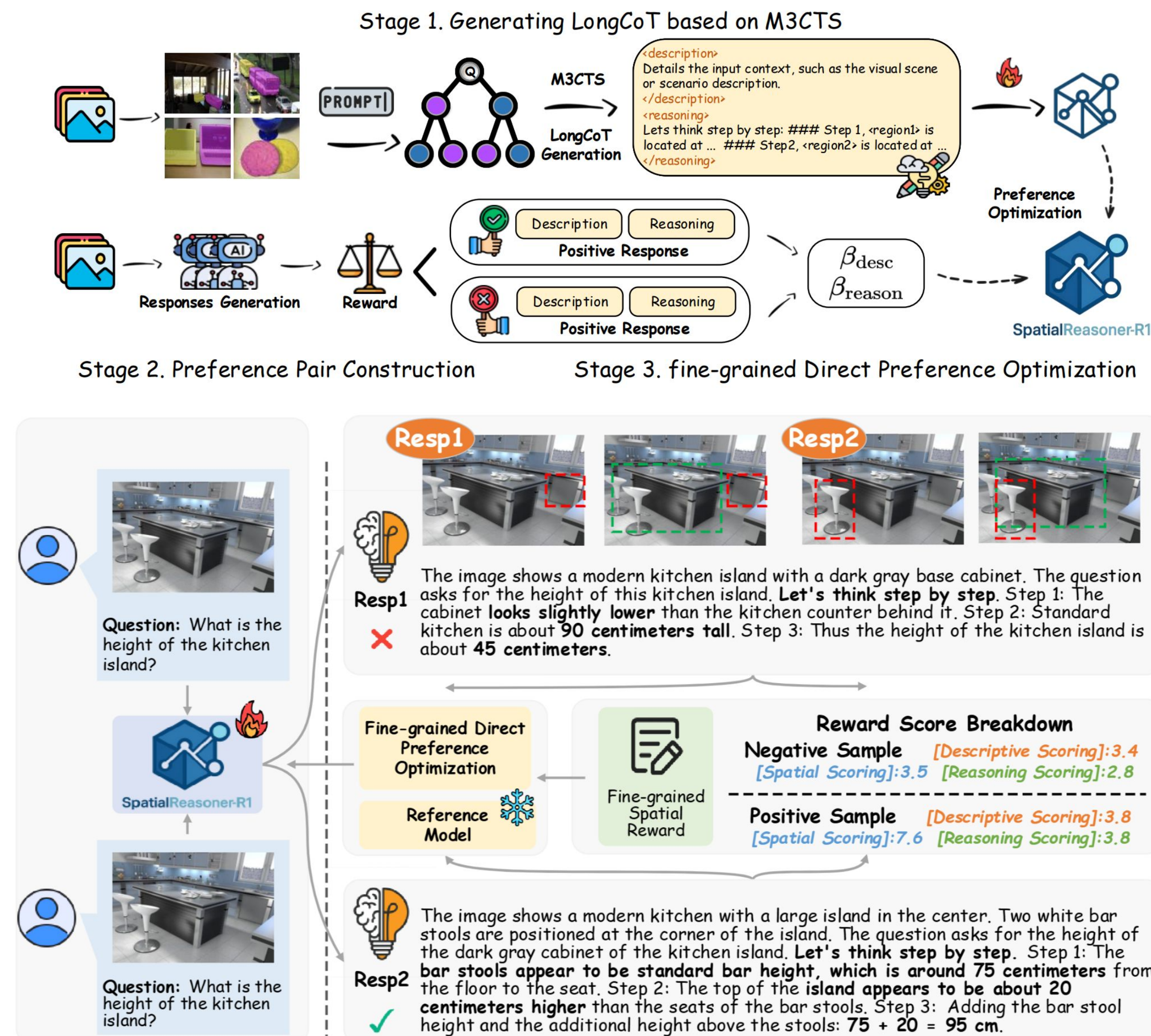
Architecture



Dataset Pipeline



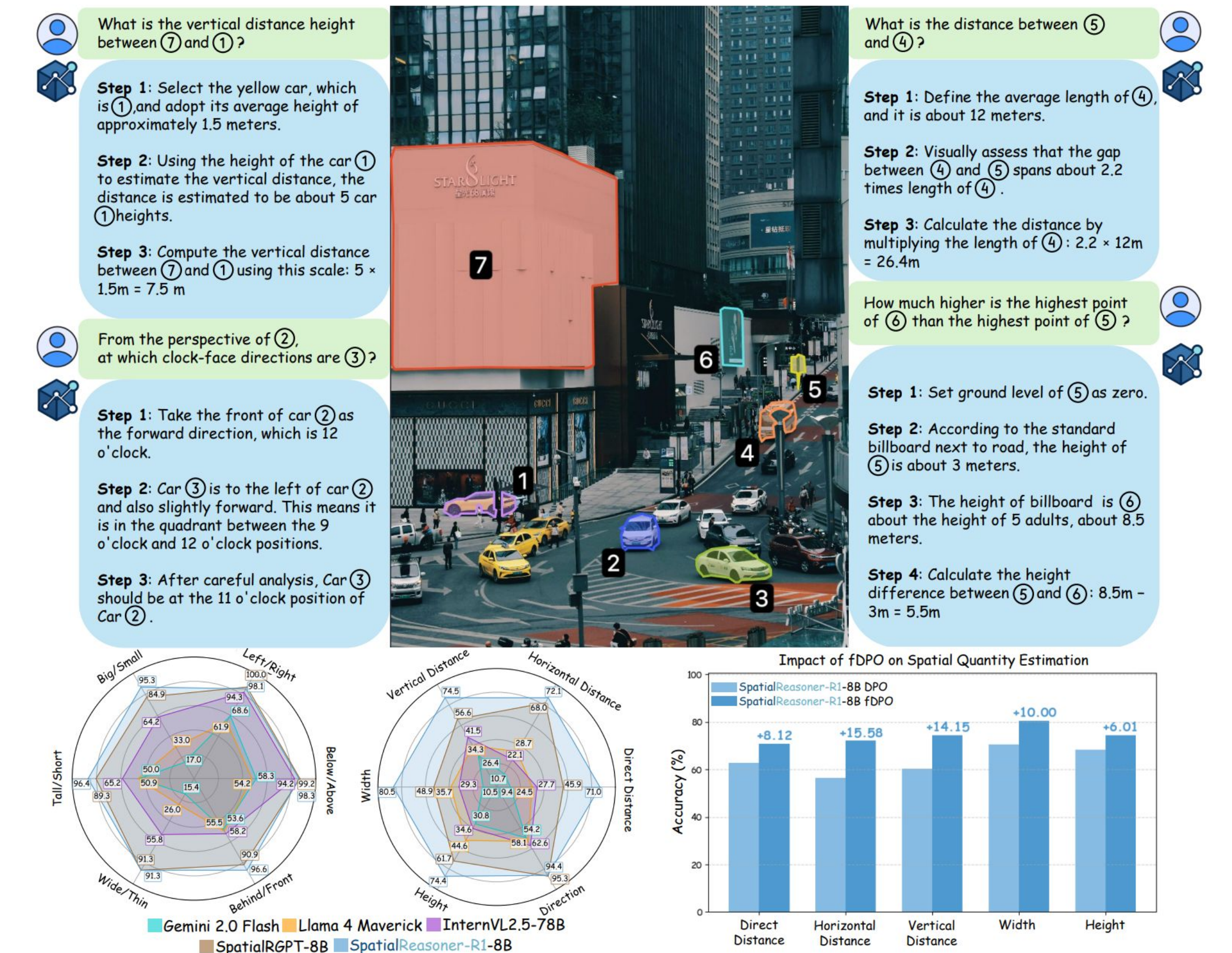
Method



Experiment

	Below/ Above	Left/ Right	Big/ Small	Tall/ Short	Wide/ Thin	Behind/ Front	Qual Acc
Gemini 2.0 Flash [21]	58.33	68.57	16.98	50.00	15.38	53.63	44.29
Llama 4 Maverick [47]	54.17	61.90	33.02	50.89	25.96	55.45	47.18
Gemini 1.5 Pro [63]	85.83	56.19	58.49	71.42	55.76	60.00	65.14
ChatGPT-4o [49]	87.50	80.00	53.77	63.39	51.92	60.90	66.67
SpatialBot-3B [6]	52.50	62.86	57.54	49.11	49.04	62.73	55.56
SpaceThinker Qwen2.5VL-3B [4]	89.16	63.81	76.41	56.25	56.73	70.91	69.25
InternVL2.5-78B [12]	94.16	94.28	64.15	65.17	55.76	58.18	72.29
SpatialReasoner-R1 fDPO 8B	98.33	98.10	95.28	96.43	91.34	93.64	95.59
	Direct Distance	Horizontal Distance	Vertical Distance	Width	Height	Direction	Quan Acc
Gemini 2.0 Flash [21]	9.45	10.65	26.41	10.52	30.82	54.20	22.43
Llama 4 Maverick [47]	24.48	28.68	34.28	35.71	44.61	58.09	36.72
Gemini 1.5 Pro [63]	14.18	17.21	14.15	19.54	36.09	30.84	21.90
ChatGPT-4o [49]	/	/	/	/	/	60.75	/
SpatialBot-3B [6]	6.00	15.51	8.00	10.52	18.75	39.00	15.62
SpaceThinker Qwen2.5VL-3B [4]	24.32	17.21	59.43	23.27	23.62	32.35	28.97
InternVL2.5-78B [12]	27.70	22.13	41.50	29.32	34.58	62.61	35.25
SpatialReasoner-R1 fDPO 8B	70.95	72.13	74.52	80.45	74.43	94.39	77.30

Spatial Reasoning



Q: Can you give me an estimation of the horizontal distance between center of <Region1> and <Region 2>?

Ground Truth

Center of <Region1> and <Region2> are 1.6 meters apart horizontally

Response from SpatialRGP 8B

The horizontal distance of Region 1 from Region 2 is 1.08 meters.

Response from InternVL2.5-78B

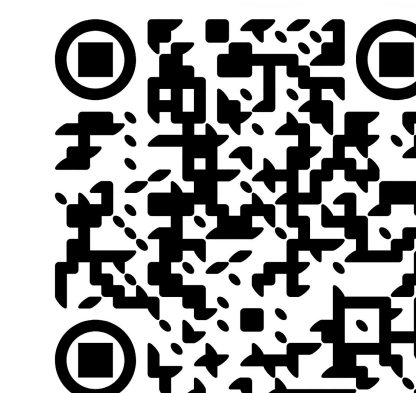
Step 1: Estimate the width of Region 1 (fireplace) as 1.2 meter.
 Step 2: Estimate the width of Region 2 (TV cabinet) as 1.5 meters.
 Step 3: the distance between the fireplace and the TV stand seems to be about the width of one object (either the fireplace or the TV stand)
 Step 4: We can estimate that the total horizontal distance between the centers of Region1 and Region2 is approximately the sum of half the width of the fireplace, the distance between the objects, and half the width of the TV stand.
 Step 5: Distance = $1.2/2 + 1.5 + 1.5/2 = 2.85$ meters

Response from Gemini 1.5 Pro

Step1: The right edge of the fireplace seems aligned with the left edge of the TV, and the TV is centered above the TV stand, which is slightly wider than the TV.
 Step2: Let's assume the width of fireplace is 1 meter. The distance between the center of the fireplace and the left edge of the TV can be estimated as half the fireplace's width, which is 0.5 meters.
 Step3: Let's assume the width of the TV stand is about 1 meter. The distance from the left edge of the TV to the center of the TV stand can be estimated as 1/4 of the stand's width, which is 0.25 meters.
 Step4: Combining steps 2 and 3, the total horizontal distance is roughly 0.75 meters.

Response from SpatialReasoner-R1 8B

Step 1: Estimate the width of the fireplace. The fireplace looks like a standard size, about 1.2 meters wide.
 Step 2: Estimate the width of the TV stand. The TV stand looks slightly wider than the TV, about 1.4 meters wide.
 Step 3: Estimate the gap between the fireplace and the TV stand. There's a small gap between the fireplace and the TV stand, which appears to be about 0.2 meters.
 Step 4: Add the estimated widths and the gap. $1.2\text{m}/2$ (fireplace) + 0.2m (gap) + $1.4\text{m}/2$ (TV) = 1.5 meters. So, the region1 and region2 are 1.5 meters apart horizontally.



<https://plan-lab.github.io/projects/spatialreasoner>

- Learn more about me
- Connect with WeChat

