Reasoning Paths with Reference Objects Elicit Quantitative Spatial Reasoning in Large Vision-Language Models

Yuan-Hong Liao[1], Rafid Mahmood[2,3], Sanja Fidler[1,2], David Acuna[2]

[1] University of Toronto & Vector Institute, [2] NVIDIA, [3] University of Ottawa

Scan Me!

Key Contributions

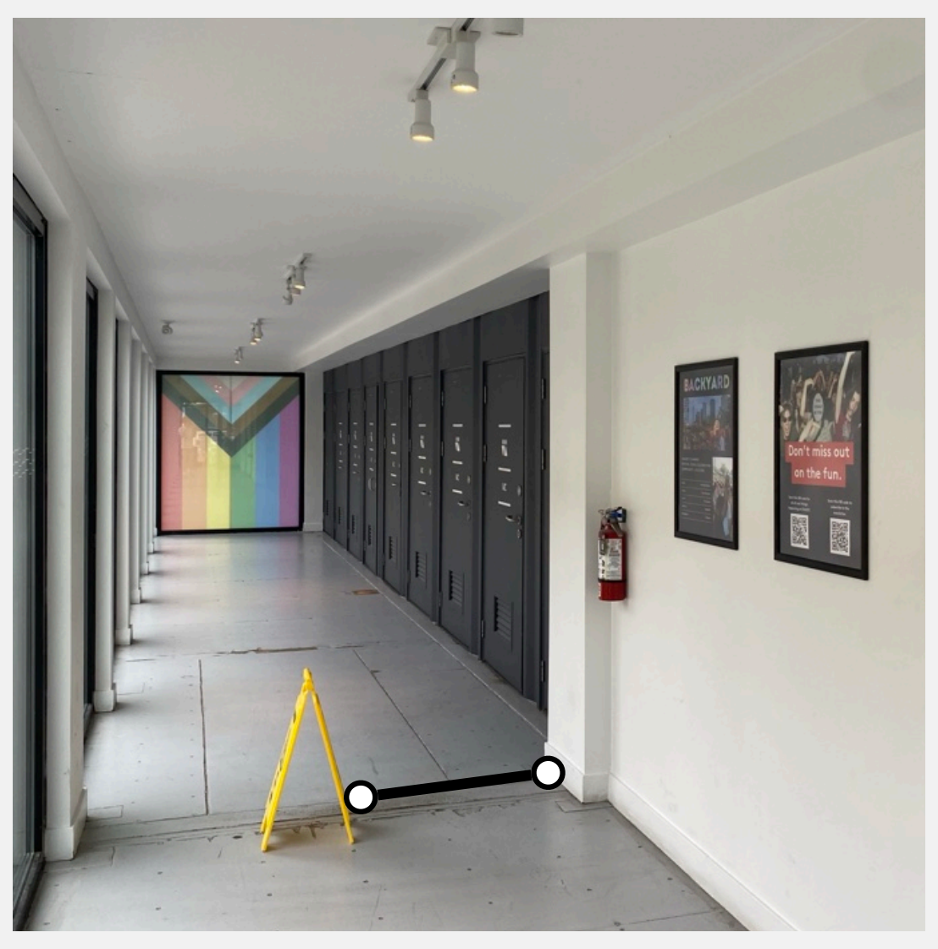
Keywords: Vision-Language Models, Spatial Reasoning

1. **Q-Spatial Bench**
   - Quantitative spatial reasoning
   - Features high-precision answers
2. **SpatialPrompt**
   - Reference objects
   - Improves Gemini-1.5-Pro and GPT-4V by 40 and 30, respectively

Quantitative Spatial Reasoning?

VQA formulation

Question: What is the gap between the yellow sign and the right wall?
Answer: 94 cms



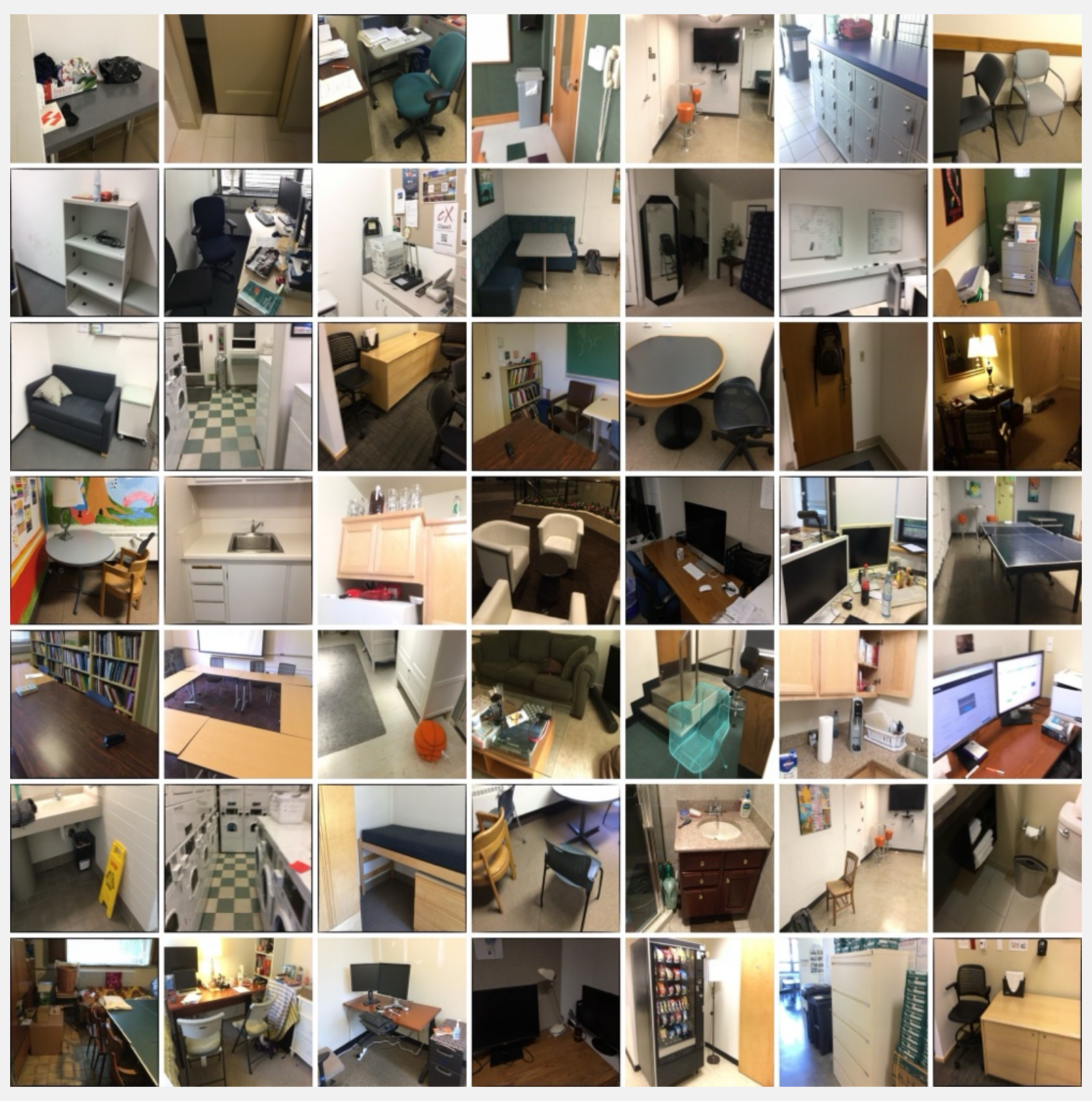
Why "Quantitative"?

Modelling spatial relationships are limited

Can I maneuver through the aisle?
Can I move an office chair to the end of aisle?

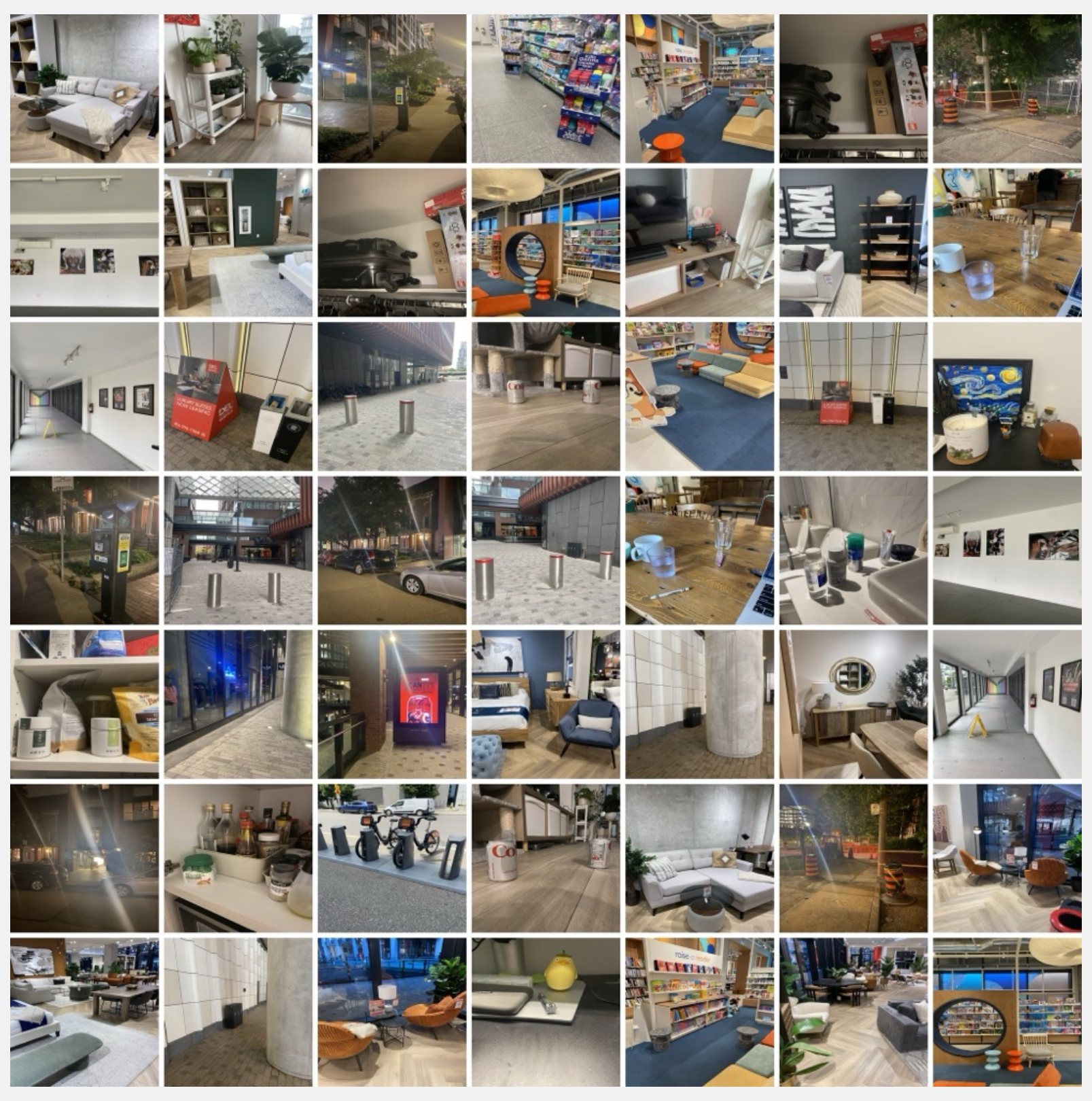
Q-Spatial Bench

Q-Spatial-ScanNet



- 170 VQAs
- 99 images
- 5 question types

Q-Spatial++



- 101 VQAs
- 87 images (freshly captured)
- 2 question types

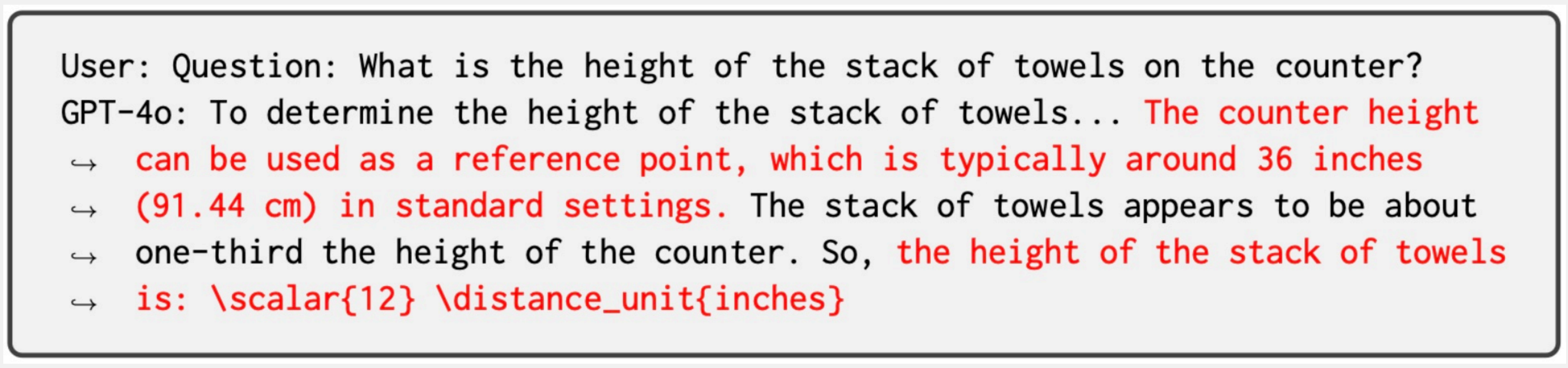
Evaluation metric: $\max\left(\frac{\text{Ground truth}}{\text{Prediction}}, \frac{\text{Prediction}}{\text{Ground truth}}\right)$

Benchmark Results

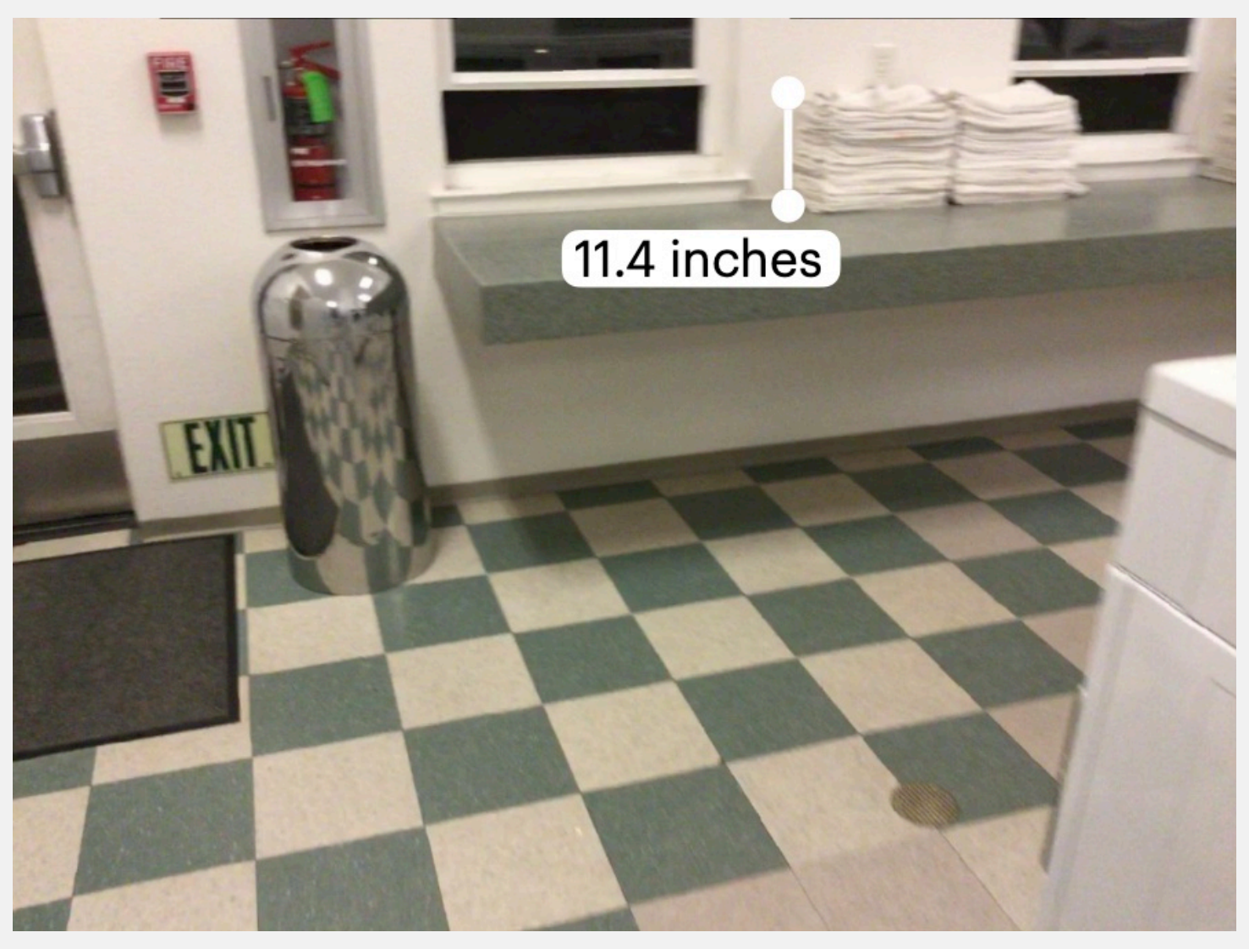
- Easy: Objects width/height
- Difficult: Distances between objects
- GPT-4o >> GPT-4V, Gemini-Pro, Gemini-Flash

	Gemini 1.5 Pro*	Gemini 1.5 Flash	GPT-4V	👑 GPT-4o
Object width	0.87	43.48	44.93	**81.16**
Object height	0	31.82	30.3	**93.94**
Horizontal distance	0.33	13.33	10	**49.44**
Vertical distance	0.69	20.69	33.33	**71.26**
Direct distance	1.11	8.33	42.59	**78.7**

Analyzing GPT-4o Responses

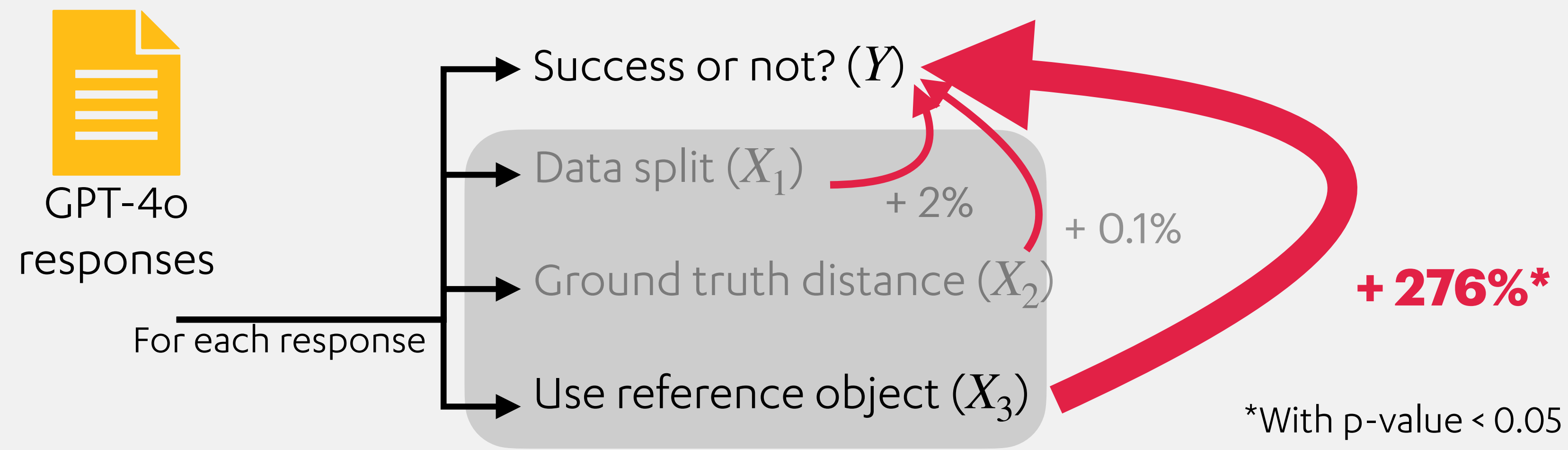


```
User: Question: What is the height of the stack of towels on the counter?
GPT-4o: To determine the height of the stack of towels... The counter height
↪   can be used as a reference point, which is typically around 36 inches
↪   (91.44 cm) in standard settings. The stack of towels appears to be about
↪   one-third the height of the counter. So, the height of the stack of towels
↪   is: \scalar{12} \distance_unit{inches}
```



Hypothesis: *Using reference objects for quantitative spatial reasoning leads to strong performances.*

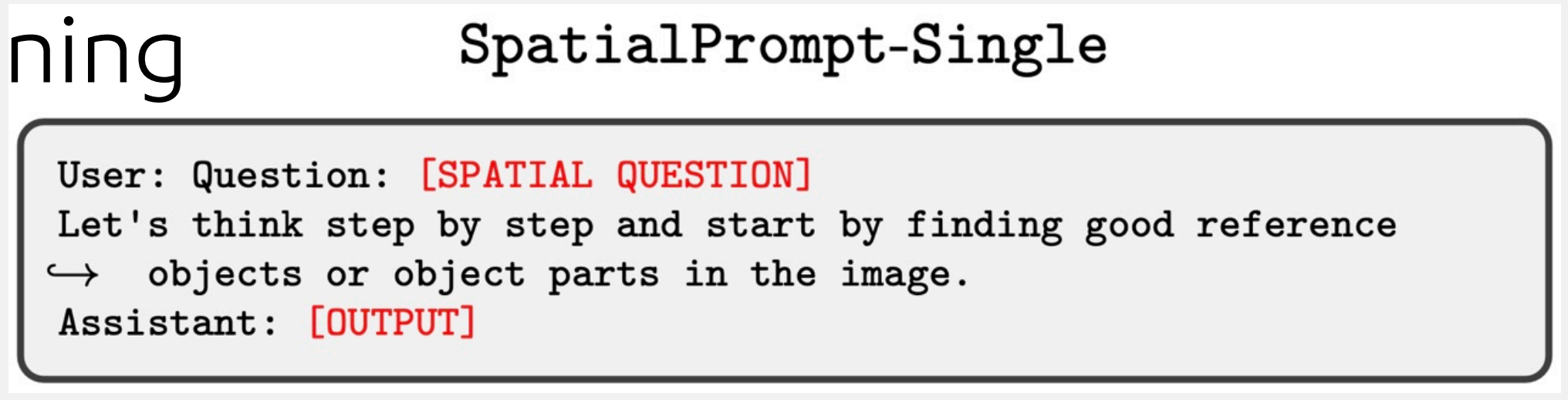
Tool: *Logistic Regression Model*



*With p-value < 0.05

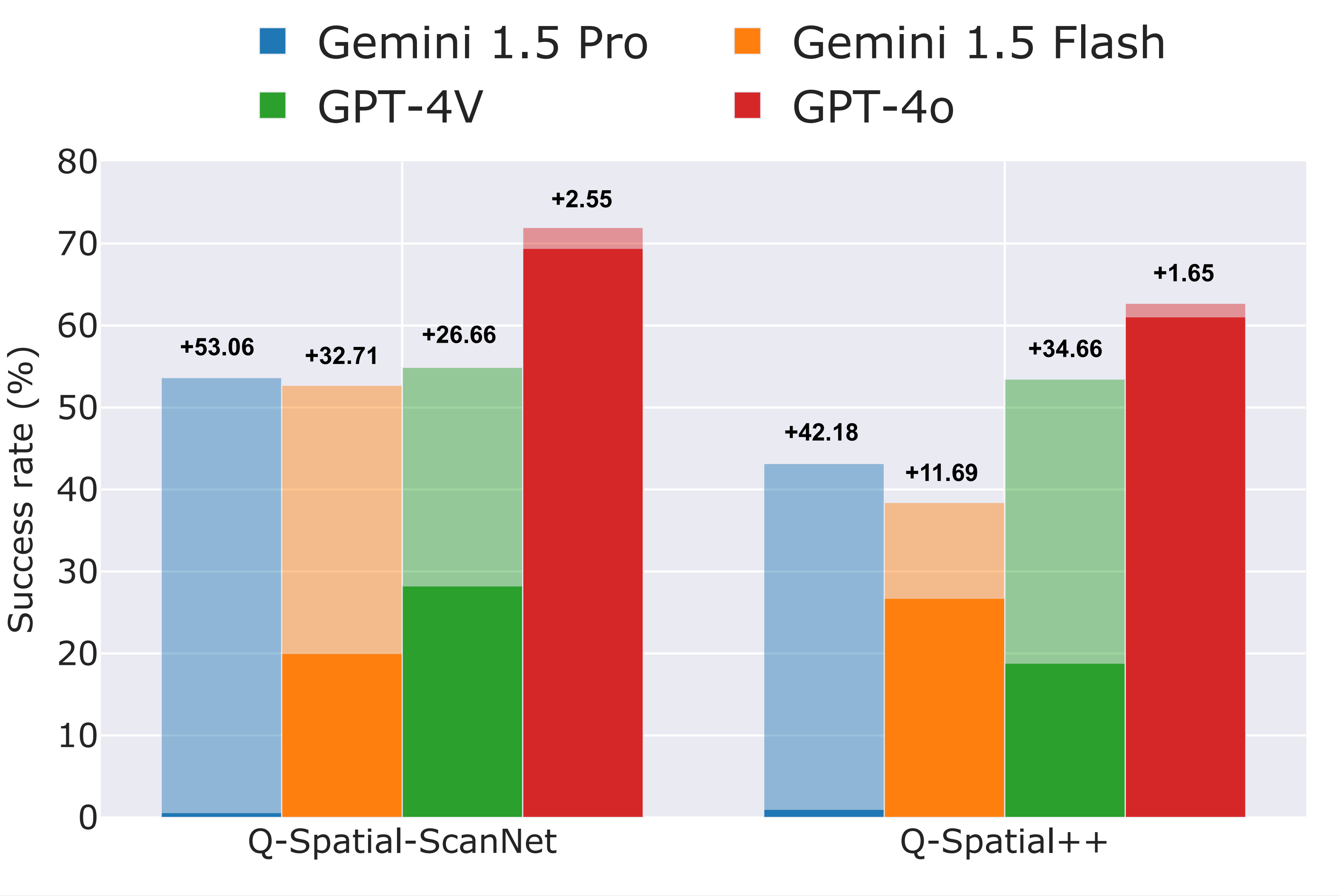
SpatialPrompt

Idea: Prompt to encourage VLMs to use reference objects for quantitative spatial reasoning

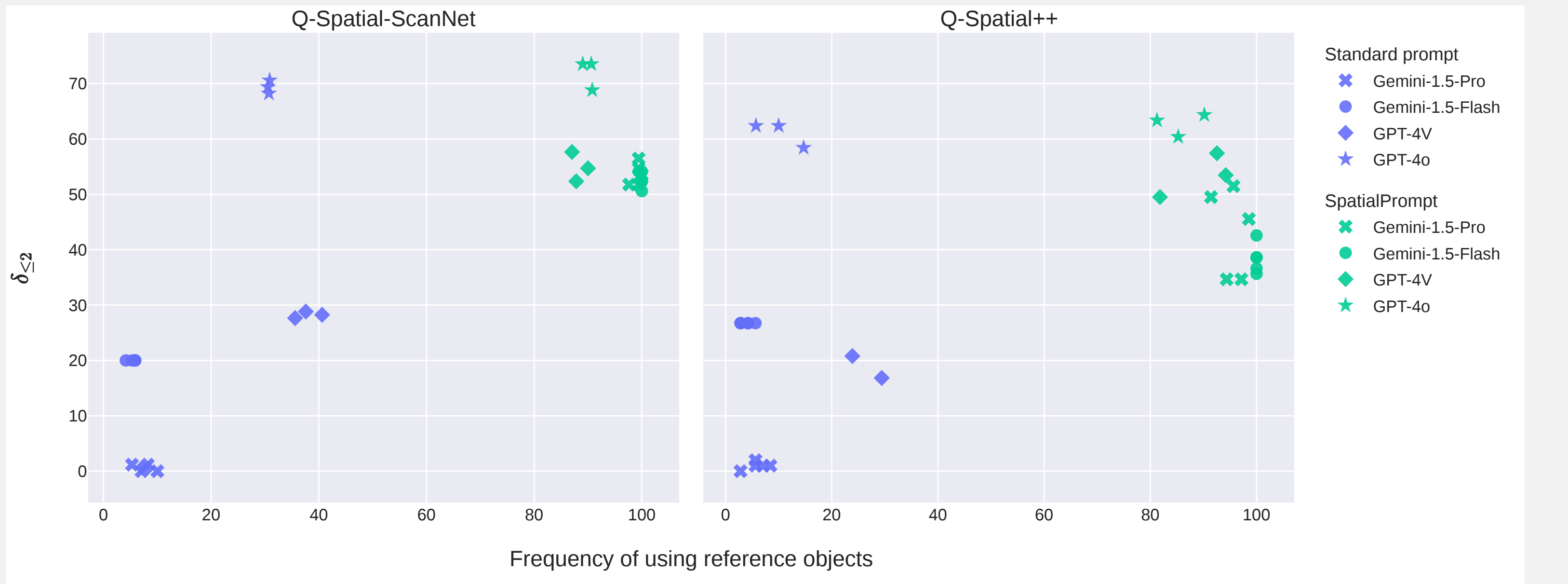


SpatialPrompt-Single

```
User: Question: [SPATIAL QUESTION]
Let's think step by step and start by finding good reference
↪   objects or object parts in the image.
Assistant: [OUTPUT]
```



SpatialPrompt encourages VLMs to use reference objects


Q-Spatial-ScanNet
Q-Spatial++
$\delta_{\leq 2}$
Frequency of using reference objects
Standard prompt
Gemini-1.5-Pro
Gemini-1.5-Flash
GPT-4V
GPT-4o
SpatialPrompt
Gemini-1.5-Pro
Gemini-1.5-Flash
GPT-4V
GPT-4o