	Below/ Above	Left/ Right	Big/ Small	Tall/ Short	Wide/ Thin	Behind/ Front	Qualitative Average	Direct Distance	Horizontal Distance	Vertical Distance	Width	Height	Direction
GPT-4V-*Turbo*	66.7	47.6	66.0	64.2	71.1	47.2	60.5	30.4 / 0.87	26.2 / 2.66	33.9 / 0.51	48.8 / 0.35	**69.1** / 1.35	40.1 / 70.0°
SpatialRGPT-7B	**95.8**	**99.0**	**77.4**	**92.9**	**82.7**	**90.9**	**90.0**	**43.2 / 0.32**	**63.9 / 0.27**	**52.8 / 0.26**	**51.1 / 0.31**	54.1 / **1.02**	**95.3 / 15.3°**

Table 1: Augmented SpatialRGPT-Bench results. Numbers represent success rates (↑) and absolute relative error (↓).

	VQA_{v2}	GQA	SQA^I	VQA^T	POPE	MME	MMB	MMB-CN	SEED	SEED^I	MMMU_V	MMMU_T	LLaVA^B	MMVet
VILA-1.5-3B	80.4	61.5	69.0	60.4	**85.9**	**1442**	63.4	52.7	60.9	67.9	**33.3**	30.8	**75.9**	35.4
SpatialRGPT-VILA-1.5-3B	**81.1**	**62.3**	**71.0**	**61.7**	85.5	1424	**65.6**	**53.6**	**61.8**	**69.0**	33.0	**31.3**	71.5	**38.2**
VILA-1.5-8B	80.9	61.9	79.7	66.3	84.4	1577	72.3	66.2	64.2	71.4	36.9	36.0	80.0	38.3
SpatialRGPT-VILA-1.5-8B	**83.3**	**64.1**	**81.6**	**68.3**	**85.5**	**1667**	**75.3**	**74.1**	**67.0**	**74.5**	**41.4**	**37.0**	**84.1**	**42.1**

Table 2: Comparison of SpatialRGPT and base model performance across various model sizes on general VLM benchmarks.

Model	**mAP** (↑)	**Acc.** (%)
CLIP	58.9	-
RegionCLIP	58.3	-
LLaVA-7B	-	40.0
Shikra-7B	-	53.9
GPT4RoI-7B	-	64.0
PVIT-7B	-	64.5
ASM-7B	69.3	-
RegionGPT-7B	70.0	80.6
SpatialRGPT-7B	69.7	79.9
SpatialRGPT-VILA-1.5-3B	72.5	82.5
SpatialRGPT-VILA-1.5-8B	**72.9**	**82.9**

Table 3: Region-level classification results. We follow the evaluation in RegionCLIP, report the results of object classification with ground-truth box on COCO-2017 validation set.

Model	**Qual.** (%)	**Quan.** (%)
SpaceLLaVA-13B	47.2	22.1
GPT-4	57.8	33.5
GPT-4 w/ Cuboids	53.7	35.4
GPT-4V-*Turbo*	58.1	41.9
SpatialRGPT-7B	91.8	60.3
SpatialRGPT-VILA-1.5-3B	91.4	59.7
SpatialRGPT-VILA-1.5-8B	**92.7**	**62.5**
Human	97.0	48.2

Table 4: SpatialRGPT-Bench results. We report the average success rates (↑) for qualitative and quantitative QAs, respectively.

BBox Type	**Width** (↓)	**Height** (↓)
Oriented BBox	17.09	4.83
Axis-aligned BBox	**8.27**	**2.35**

Table 5: Ablation study on axis-aligned vs oriented bounding boxes. Numbers indicate MSE comparing to Omni3D ground truth.

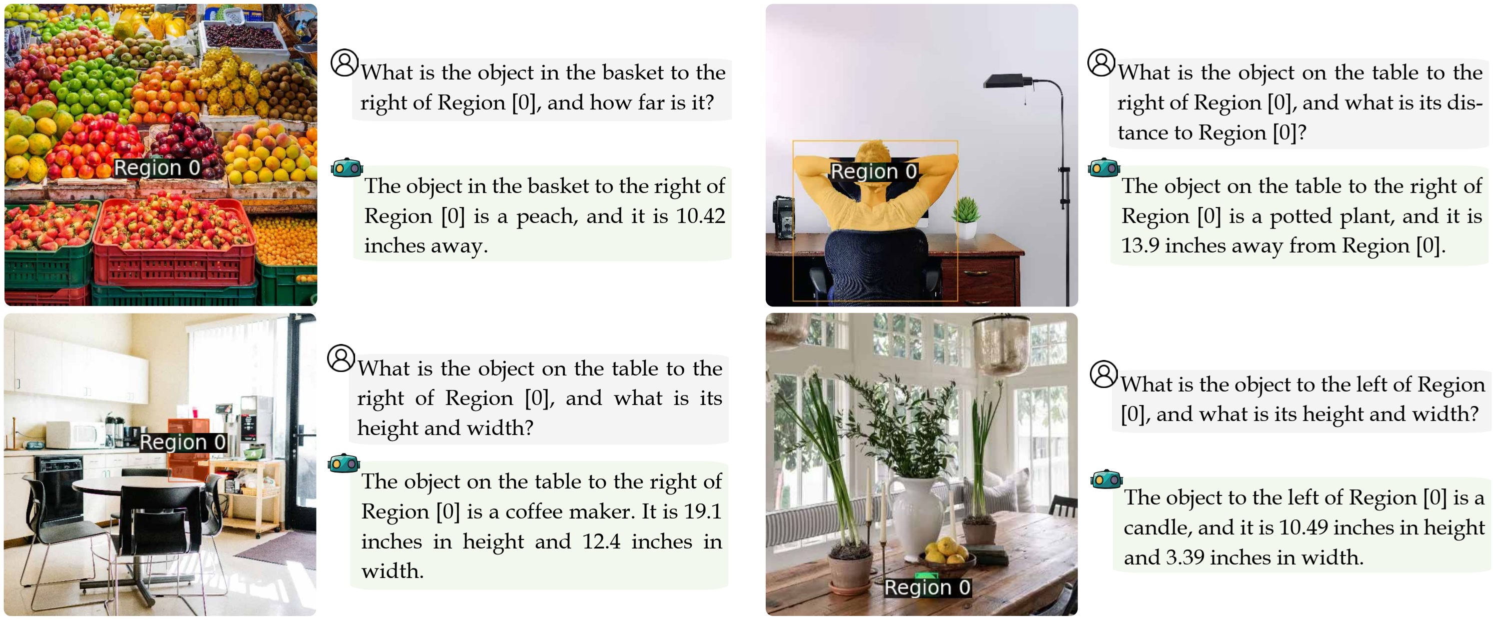
Model	**Acc.** (%)
Qwen-VL-Max	58.9
Gemini Pro	50.0
Claude 3 OPUS	57.3
GPT-4V-*preview*	58.9
GPT-4V-*Turbo*	66.9
GPT-4o	64.5
InstructBLIP-13B	50.0
Yi-VL-34B	53.2
LLaVA-v1.5-13B-xtuner	54.0
LLaVA-v1.5-13B	47.6
LLaVA-v1.6-34B	64.5
MiniGPT-4-v2-7B	49.2
InstructBLIP-7B	50.8
LLaVA-v1.5-7B-xtuner	50.8
CogVLM-7B	50.8
LLaVA-v1.5-7B	51.6
LLaVA-internLM2-7B	52.4
SpatialRGPT-7B	82.3
SpatialRGPT-VILA-1.5-8B	**87.9**

Table 6: $\text{BLINK}_{RelativeDepth}$ results.

	Below / Above		Left / Right		Big / Small		Tall / Short		Wide / Thin		Behind / Front		Avg.	
− width & height	99.1		99.0		75.8		90.8		82.8		92.1		90.5	
+ width & height	99.1	+0	99.0	+0	80.1	+4.3	91.9	+1.1	87.5	+4.7	91.8	-0.3	90.5	+1.2

	Direct Distance		Horizontal Distance		Vertical Distance		Width		Height		Direction	
− width & height	41.2		69.3		54.8		22.8		21.2		95.1	
+ width & height	41.2	+0	65.6	-3.7	51.9	-2.9	49.6	+26.8	57.9	+36.7	95.3	+0.2

Table 7: Ablation study on the impact of width and height data on the performance of other categories. Numbers represent success rates (↑).



What is the object in the basket to the right of Region [0], and how far is it?

The object in the basket to the right of Region [0] is a peach, and it is 10.42 inches away.

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 on the table to the right of Region [0], and what is its height and width?

The object on the table to the right of Region [0] is a coffee maker. It is 19.1 inches in height and 12.4 inches in width.

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.

Figure 1: Examples of SpatialRGPT multi-hop reasoning.