NeurIPS #4053

SpatialPIN: Enhancing Spatial Reasoning Capabilities of Vision-Language Models through Prompting and Interacting 3D Priors

Table T1. **Qualitative spatial VQA on ScanQA dataset (similar to qualitative IaOR-VQA).** We select 20 3D scenes and 100 QA pairs. Ours and SpatialVLM take 2D screenshots of 3D scenes as input. Chat-3D and Chat-3D-v2 take 3D scenes as direct input.

	GPT-4o		SpatialVLM	Chat-3D	Chat-3D-v2
	w/o ours	w ours			
Accuracy %	61	**76**	64	68	71

Table T2. **Qualitative IaOR-VQA.** We include the comparison to Llama 3.1-70B in addition to the models included in our main paper Table 1.

	GPT-4o		Llama 3.1-70B		SpatialVLM
	w/o ours	w ours	w/o ours	w ours	
Accuracy %	69.0	**87.3**	72.7	82.3	76.7

Table T3. **Quantitative IaOR-VQA.** Same as our main paper Table 2, we measure the accuracy by the percentage of answers that fall within 0.5x to 2.0x, 0.75x to 1.33x, and 0.9x to 1.11x of the ground truth value. "Output number" means VLMs produce number in the response instead of vague descriptions.

	GPT-4o		Llama 3.1-70B		SpatialVLM
	w/o ours	w ours	w/o ours	w ours	
Output numbers %	31.5	**99.5**	37.5	99.0	91.0
In range [50, 200] %	14.0	**74.5**	18.0	53.5	33.5
In range [75, 133] %	8.5	**70.5**	9.5	38.0	20.5
In range [90, 111] %	3.0	**55.0**	2.5	20.0	7.5

Table T4. **Detailed inference speed in seconds.** Since coarse and fine-grained 3D understanding runs in parallel, we take their maximum. For VLM understanding inference time, the first and the second numbers are LLaVA-1.5 and GPT-4o respectively. For VLMs, their inference times are insensible to input image sizes since LLaVA-1.5 resizes them to 224×224. For GPTs, time of API call is the main overhead. Thus, here we assume the inference speed of VLMs is identical across different input image sizes (experiment conducted on 640×480 resolution).

	2D Understanding			Coarse 3D Understanding		Fine-Grained 3D	Answering	Total
	VLM	Seg	Inpaint	VLM	Depth + Camera	Reconstruction	VLM	
320×240						8.0		[8.42, 16.62]
640×480	[0.17, 4.4]	0.12	∼0	[0.17, 4.2]	∼0	8.2	[0.13, 4.1]	[8.62, 16.82]
1280×960						16.1		[16.52, 24.72]

Table T5. **Peg insertion.** We classify the success rates into: 1) successfully picked, 2) successfully picked and contacted the small hole on the cube to be inserted, and 3) successfully picked and inserted.

	GPT-4o + ours	Our Direct 3D Info	SpatialVLM + RRT*	VoxPoser
Picked %	**63.3**	56.7	16.7	53.3
Picked & contacted %	**26.7**	6.7	3.3	16.7
Picked & inserted %	**20.0**	6.7	3.3	13.3

Figure F1. **Accurately, partially reconstructed 3D scenes.** For each two-image column, the left image is the input image, and the right image is the reconstructed scene. The reconstructions show high visual alignment with the input images.

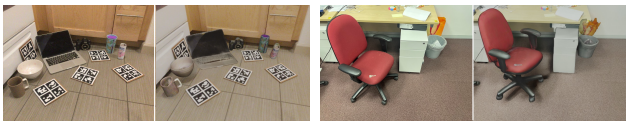

Figure F2. **Failure cases of single-view object 3D reconstruction.** In the left and right scenes, the silver laptop and the orange plastic bag have imperfect appearances. Nevertheless, their general shapes are acceptable.

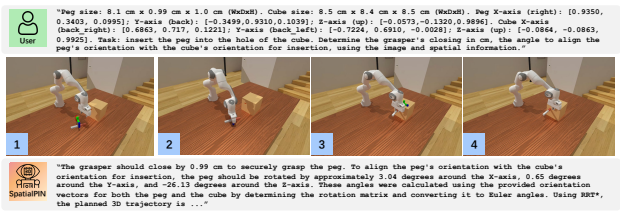

Figure F3. **Peg insertion.** SpatialPIN successfully outputs the insertion policies by reasoning about the size of the peg and how to rotate the peg to align its orientation with the cube.