



The following can be seen in the image: tall tree directly right of the orange truck bed and over the tree over the back end of the brown truck bed. Open overcast sky over everything on the ground. Right group of tree leaves over the orange truck bed and between the cranes that are suspending it. Window area of the driver's side door of the white truck on the left of the brown truck bed... crane holding up the orange truck bed to the right of another crane... Please answer the following question based on the provided information.

Q: How many cranes are involved in lifting the orange truck bed? Available options:
A. Two B. One
C. Three D. Four

Figure 1: Illustration of the **Spatial-Real** dataset. **Spatial-Real** is a newly curated dataset that builds on the Densely Captioned Images (DCI) dataset, which features detailed captions averaging over 1000 words per image. We create multiple-choice questions with annotated answers focused on spatial reasoning, including object counting, relation, and position understanding.

Input Format	Model	Average Accuracy
Text-only (TQA (LLM))	Vicuna-13B-1.5	0.845
	Vicuna-7B-1.5	0.716
	Mistral-7B	0.800
	LLaMA-3-8B	0.884
Vision-only (VQA)	CogVLM	0.419
	Qwen-VL	0.329
	LLaVA-1.6-Mistral-7B	0.368
	CogAgent	0.400
	LLaVA-1.6-Vicuna-7B	0.432
	LLaVA-1.6-Vicuna-13B	0.445
Vision-text (VTQA)	CogVLM	0.594
	Qwen-VL	0.729
	LLaVA-1.6-Mistral-7B	0.761
	CogAgent	0.471
	LLaVA-1.6-Vicuna-13B	0.832
	LLaVA-1.6-Vicuna-7B	0.806

Table 1: Performance on the **Spatial-Real** task. The same trends still hold on real images: VQA vs. VTQA, TQA (LLM) vs. VTQA, and TQA (LLM) vs. VQA.

Input Format	Model Name	Q1 Acc	Q2 Acc	Q3 Acc	Q4 Acc	Q5 Acc	Q6 Acc	Avg Acc
Vision-only	LLaVA-1.6-34B	0.32	0.29	0.21	0.40	0.38	0.34	0.32
	LLaVA-1.6-Mistral-7B	0.28	0.69	0.45	0.67	0.37	0.48	0.49
	LLaVA-1.6-Vicuna-13B	0.30	0.58	0.30	0.59	0.48	0.45	0.45
Vision-text	LLaVA-1.6-34B	0.44	0.40	0.33	0.62	0.43	0.36	0.43
	LLaVA-1.6-Mistral-7B	0.32	0.79	0.65	0.88	0.40	0.56	0.60
	LLaVA-1.6-Vicuna-13B	0.29	0.78	0.36	0.75	0.45	0.44	0.51

Table 2: **[R1]** Detailed results for the **Spatial-Grid** task. We added three more questions (Q4-Q6). LLaVA-1.6-34B excels in counting (Q1) but struggles to process dense and fine-grained image information (Q2-Q6) compared to other LLaVA models.