Model	V^*Bench	MMVP	Depth	Spatial	Jigsaw	Vis. Corr.	Sem. Corr.
Prior multimodal LLMs							
LLaVA-1.5-7B [26]	48.7	-	52.4	61.5	11.3	25.6	23.0
LLaVA-1.5-13B [26]	-	24.7	53.2	67.8	58.0	29.1	32.4
LLaVA-NeXT-34B [27]	-	-	67.7	74.8	54.7	30.8	23.7
Claude 3 OPUS [2]	-	-	47.6	58.0	32.7	36.6	25.2
Gemini-Pro [41]	48.2	40.7	40.3	74.8	57.3	42.4	26.6
GPT-4V-preview [35]	55.0	38.7	59.7	72.7	70.0	33.7	28.8
Previous state of the art	75.4 [50]	49.3 [10]	67.7 [27]	76.2 [42]	70.0 [35]	42.4 [41]	33.1 [48]
Latest multimodal LLMs + Visual Sketchpad							
GPT-4 Turbo	52.5	71.0	66.1	68.5	64.7	48.8	30.9
+ Sketchpad	71.0	73.3	68.5	80.4	68.5	52.3	42.4
	+18.5	+2.3	+2.4	+11.9	+3.8	+3.5	+11.5
GPT-4o	66.0	85.3	71.8	72.0	64.0	73.3	48.6
+ Sketchpad	**80.3**	**86.3**	**83.9**	**81.1**	**70.7**	**80.8**	**58.3**
	+14.3	+1.0	+12.1	+9.1	+6.7	+7.5	+9.7

Table 2: Accuracy on complex visual reasoning tasks. **SKETCHPAD enhances both GPT-4 Turbo and GPT-4o performance, establishing new SOTA performance levels on all the tasks.**