Task Success Rate per Model (%)

Model	Referring Dot-Pointing	Counting (E)	Counting (H)	Mental Rotation 2D (E)	Mental Rotation 2D (H)	Colorization (E)	Colorization (H)	Matchstick Equation (E)	Matchstick Equation (H)	Matchstick Rotation (E)	Matchstick Rotation (H)	Jigsaw (E)	Jigsaw (H)	Maze 2D (E)	Maze 2D (H)	Zoom-In Puzzle (E)	Zoom-In Puzzle (H)	Fetch Reach	Video Unshuffle (E)	Video Unshuffle (H)	Fetch Pick-Place	Maze 3D (E)	Maze 3D (H)	Sliding Block (E)	Sliding Block (H)	Mental Rotation 3D (Objaverse) (E)	Mental Rotation 3D (Objaverse) (H)	Patch Reassembly (E)	Patch Reassembly (H)	Mental Rotation 3D (Cube) (E)	Mental Rotation 3D (Cube) (H)
Single-Task Fine-Tuning	90.0	54.3	15.7	55.7	75.7	90.0	88.6	50.0	0.0	88.6	70.0	97.1	0.0	97.1	77.1	92.9	57.1	100.0	48.6	2.9	70.0	64.3	28.6	94.3	84.3	15.7	2.9	61.4	0.0	50.0	54.3
Mixed-Task Fine-Tuning	81.4	54.3	8.6	58.6	52.9	35.7	17.1	44.3	0.0	80.0	57.1	90.0	0.0	52.9	31.4	90.0	35.7	98.6	42.9	8.6	67.1	52.9	24.3	91.4	61.4	4.3	0.0	38.6	0.0	24.3	25.7
GPT-5	67.1	40.0	24.3	50.0	30.0	27.1	17.1	77.1	41.4	58.6	52.9	18.6	1.4	18.6	4.3	11.4	4.3	7.1	17.1	0.0	4.3	0.0	0.0	11.4	4.3	5.7	1.4	11.4	1.4	1.4	1.4
Gemini 2.5 Pro	71.4	51.4	20.0	50.0	40.0	24.3	11.4	48.6	1.4	31.4	22.9	27.1	0.0	28.6	10.0	30.0	14.3	10.0	14.3	5.7	2.9	4.3	0.0	32.9	12.9	1.4	1.4	1.4	0.0	1.4	0.0
Grok 4 Fast	42.9	30.0	10.0	17.1	8.6	18.6	7.1	38.6	17.1	5.7	1.4	11.4	1.4	10.0	7.1	2.9	2.9	4.3	4.3	2.9	4.3	7.1	5.7	0.0	0.0	2.9	0.0	0.0	0.0	1.4	0.0
Qwen3 VL 235B Instruct	90.0	40.0	14.3	27.1	10.0	12.9	8.6	0.0	0.0	0.0	0.0	18.6	0.0	8.6	2.9	11.4	2.9	1.4	8.6	0.0	4.3	0.0	1.4	0.0	0.0	1.4	1.4	0.0	0.0	0.0	0.0
Qwen VL Max	47.1	31.4	8.6	22.9	20.0	22.9	11.4	1.4	0.0	2.9		15.7	0.0	8.6	1.4	5.7	2.9	0.0	4.3	1.4	4.3	0.0	2.9	0.0	0.0	2.9	0.0	0.0	0.0	1.4	1.4
GLM 4.5V	70.0	20.0	14.3	12.9	4.3	22.9	4.3	7.1	0.0	8.6	2.9	8.6	0.0	12.9	4.3	1.4	2.9	1.4	4.3	0.0	4.3	1.4	2.9	4.3	1.4	2.9	0.0	0.0	0.0	0.0	0.0
Claude Sonnet 4	47.1	14.3	1.4	15.7	5.7	20.0	17.1	1.4	0.0	2.9	0.0	22.9	0.0	14.3	7.1	4.3	2.9	7.1	7.1	0.0	4.3	10.0	2.9	0.0	0.0	1.4	1.4	0.0	0.0	0.0	0.0
Qwen 2.5 VL 72B Instruct	31.4	31.4	14.3	17.1	5.7	17.1	18.6	1.4	0.0	7.1	1.4	24.3	0.0	4.3	0.0	2.9	2.9	10.0	1.4		4.3	1.4	2.9								
Llama 4 Maverick	44.3	37.1	10.0	20.0	10.0	12.9	8.6	1.4	0.0	2.9	1.4	8.6	0.0	0.0	0.0	5.7	2.9	12.9	5.7	1.4	4.3	0.0	0.0					1.4			
Gemma 3 27B Instruct	31.4	31.4	14.3	17.1	8.6	10.0	7.1	0.0	0.0	1.4	0.0	2.9	0.0	4.3	1.4	2.9	2.9	4.3	7.1	0.0	4.3	17.1	8.6								
UI TARS 1.5 7B	10.0	24.3	5.7	8.6	8.6	14.3	5.7	0.0	0.0	1.4	0.0	1.4	0.0	0.0	0.0	4.3	1.4	1.4	4.3	0.0	1.4	0.0	0.0								1.4