Fine-Tuned (Ours)

	Easy	Hard

Avg. Success Rate (%)

- Single-Task Fine-Tuning: 71.88 (Easy), 41.81 (Hard)
- Mixed-Task Fine-Tuning: 58.75 (Easy), 26.00 (Hard)
- GPT-5: 26.43 (Easy), 12.57 (Hard)
- Gemini 2.5 Pro: 26.79 (Easy), 9.52 (Hard)
- Grok 4 Fast: 12.32 (Easy), 4.57 (Hard)
- Qwen3 VL 235B Instruct: 13.75 (Easy), 3.05 (Hard)
- Qwen VL Max: 10.27 (Easy), 3.81 (Hard)
- GLM 4.5V: 11.16 (Easy), 2.76 (Hard)
- Claude Sonnet 4: 10.54 (Easy), 2.86 (Hard)
- Qwen 2.5 VL 72B Instruct: 9.38 (Easy), 3.33 (Hard)
- Llama 4 Maverick: 9.55 (Easy), 2.57 (Hard)
- Gemma 3 27B Instruct: 8.12 (Easy), 3.14 (Hard)
- UI TARS 1.5 7B: 4.38 (Easy), 1.62 (Hard)