| Model | (Hessel et al., 2023) | | (Zhou et al., 2025) | |
|---|---|---|---|---|
| | Matching | Ranking | 10-vs-1000 | 30-vs-300 |
| Human (Expert) | 100.00 | 100.00 | 60.00 | 40.00 |
| Human (Non-Expert) | 53.03 | 65.61 | 54.70 | 52.27 |
| DeepSeek-R1 | 74.00 | 64.67 | 56.86 | 47.14 |
| o4-mini | <u>75.08</u> | 62.59 | 60.17 | 51.42 |
| o3 | **83.33** | 62.85 | **69.05** | **54.57** |
| GLM-4.1V-9B-Thinking | 59.40 | 55.60 | 52.28 | 49.14 |
| Kimi-VL-A3B-Thinking | 54.00 | 57.45 | 52.30 | 52.01 |
| LlamaV-o1 | 43.67 | 50.90 | 48.57 | 49.42 |
| Qwen2.5-VL-7B-Instruct | 42.67 | 55.06 | 50.57 | 47.99 |
| Qwen2.5-VL-32B-Instruct | 46.67 | 49.87 | 52.00 | 44.57 |
| Qwen2.5-VL-72B-Instruct | 56.00 | 55.58 | 53.71 | 50.29 |
| IRS-7B (Our model) | 59.67 | 64.42 | 56.29 | <u>53.14</u> |
| IRS-32B (Our model) | 62.67 | <u>68.05</u> | <u>62.86</u> | <u>53.14</u> |
| IRS-72B (Our model) | 69.33 | **76.10** | 62.57 | 50.86 |