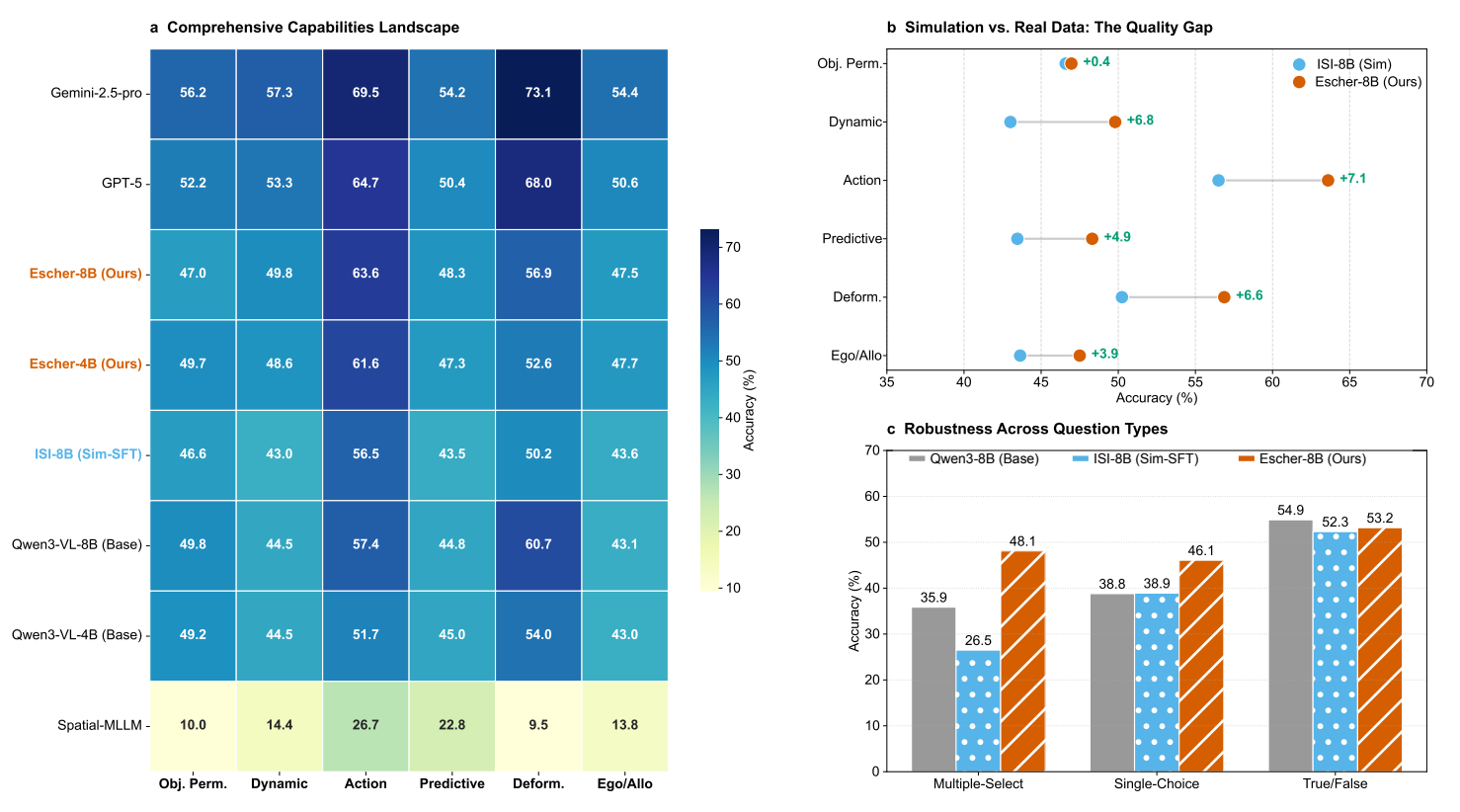

a Comprehensive Capabilities Landscape
Gemini-2.5-pro
GPT-5
Escher-8B (Ours)
Escher-4B (Ours)
ISI-8B (Sim-SFT)
Qwen3-VL-8B (Base)
Qwen3-VL-4B (Base)
Spatial-MLLM
56.2 57.3 69.5 54.2 73.1 54.4
52.2 53.3 64.7 50.4 68.0 50.6
47.0 49.8 63.6 48.3 56.9 47.5
49.7 48.6 61.6 47.3 52.6 47.7
46.6 43.0 56.5 43.5 50.2 43.6
49.8 44.5 57.4 44.8 60.7 43.1
49.2 44.5 51.7 45.0 54.0 43.0
10.0 14.4 26.7 22.8 9.5 13.8
Obj. Perm.
Dynamic
Action
Predictive
Deform.
Ego/Allo
Accuracy (%)
70
60
50
40
30
20
10
b Simulation vs. Real Data: The Quality Gap
ISI-8B (Sim)
Escher-8B (Ours)
Obj. Perm.
Dynamic
Action
Predictive
Deform.
Ego/Allo
+0.4
+6.8
+7.1
+4.9
+6.6
+3.9
35
40
45
50
55
60
65
70
Accuracy (%)
c Robustness Across Question Types
Qwen3-8B (Base)
ISI-8B (Sim-SFT)
Escher-8B (Ours)
Accuracy (%)
0
10
20
30
40
50
60
70
35.9
26.5
48.1
38.8
38.9
46.1
54.9
52.3
53.2
Multiple-Select
Single-Choice
True/False