Supplement Table 1 Performance of five multimodal large language models on the sensitivity analysis set (*n* = 109) using the majority vote standard

Model	Accuracy	Macro F1	Macro sensitivity
GPT-5	0.569	0.529	0.581
GPT-4o	0.312	0.298	0.386
Gemini-2.5-pro	0.367	0.342	0.401
Grok-4	0.312	0.245	0.295
Qwen-VL-Max	0.294	0.265	0.326