Observed vs Bayes(range) — NRMSE (Base, Text+Image) Observed LLM (text+image — base) Bayes(range) combo (base) 1.04 1.03 1.0 0.82 NRMSE (↓ better) 0.75 8.0 0.73 0.71 0.66 0.62 0.56 0.6 0.55 0.51 0.43 0.43 0.4 0.33 0.33 0.31 0.23 0.21 0.2 0.0 Gemini 2.5 Flash Lite Qwen2.5 VL 32B Claude 3.7 Sonnet Gemma 3 4B it _{Phi} 4 Multimodal Llama 4 Maverick GPT 4.1 mini Mistral 24B GPT 40 Ranking by Bayes(range) / Observed (Base, Text+Image) 120% 120 Bayes as % of Observed 99% 100 89% 82% 77% 76% 80 70% 66% 60 46% 40 20 #1 #6 #8 0 Gemini 2.5 Flash Lite Claude 3.7 Sonnet GPT 4.1 mini Qwen2.5 VL 32B Mistral 24B Llama 4 Maverick Phi 4 Multimodal Gemma 3 4B it GPT 40