NRMSE vs Bayesian-ness (one dot per model × modality) Modality 3.5 Gemma 3 4B it text ↓ better 3.0 -Qwen2.5 VL 32B Observed NRMSE (mean across ablations) 2.5 -Phi 4 Multimodal 2.0 1.5 -Gemini 2.5 Flash Lite Llama 4 Maverick 1.0 Mistral 24B 0.5 Claude 3.7 Sonnet 0.5 0.7 0.3 0.4 0.6 0.2 8.0 0.9

Bayesian-ness (mean probability across ablations)