Observed NRMSE (normalized to experiment mean)

Owen2.5 VL 32B - 0.51 0.51 0.43 0.51 0.61 0.51 -0.81 0.81 0.80 0.49 0.80 0.73 -0.56 0.51 0.52 0.45 0.73 0.54 L 1 0 -1.02 0.94 0.83 0.54 0.66 0.65 -0.71 0.71 0.65 0.59 0.61 0.58 Mistral 24B - 0.62 0.61 0.59 0.62 0.61 0.62

Llama 4 Mayerick - 0.63 0.71 0.51 0.63 0.56 0.63 -0.79 0.77 0.50 0.96 0.72 1.00 -0.75 0.72 0.51 0.71 0.54 0.72 -0.33 0.33 0.29 0.44 0.26 0.36 GPT 4.1 mini -0.61 0.53 0.49 0.61 0.53 0.61 -0.31 0.28 0.36 0.32 0.32 0.43

- 0.6 Gemini 2.5 Flash Lite -0.37 0.36 0.39 0.37 0.46 0.37 -0.65 0.68 0.64 0.71 0.65 0.66 -0.66 0.64 0.61 0.64 0.48 0.75

Phi 4 Multimodal - 0.82 0.93 0.81 0.82 0.74 0.82 -1.30 1.40 0.98 1.41 1.33 1.43 -0.82 0.94 0.73 0.85 1.05 0.95 L 0 4

Gemma 3 4B it - 1.04 1.07 0.75 1.04 0.93 1.04 **-1.17 1.01 0.70 1.26 1.06 1.37** -1.04 0.97 0.83 1.18 0.99 1.18

Ablation

Verbal steer Numerical steer Short context Short context Numerical steer Gradual noise Verbal steer Verbal steer short context

Ablation

Ablation