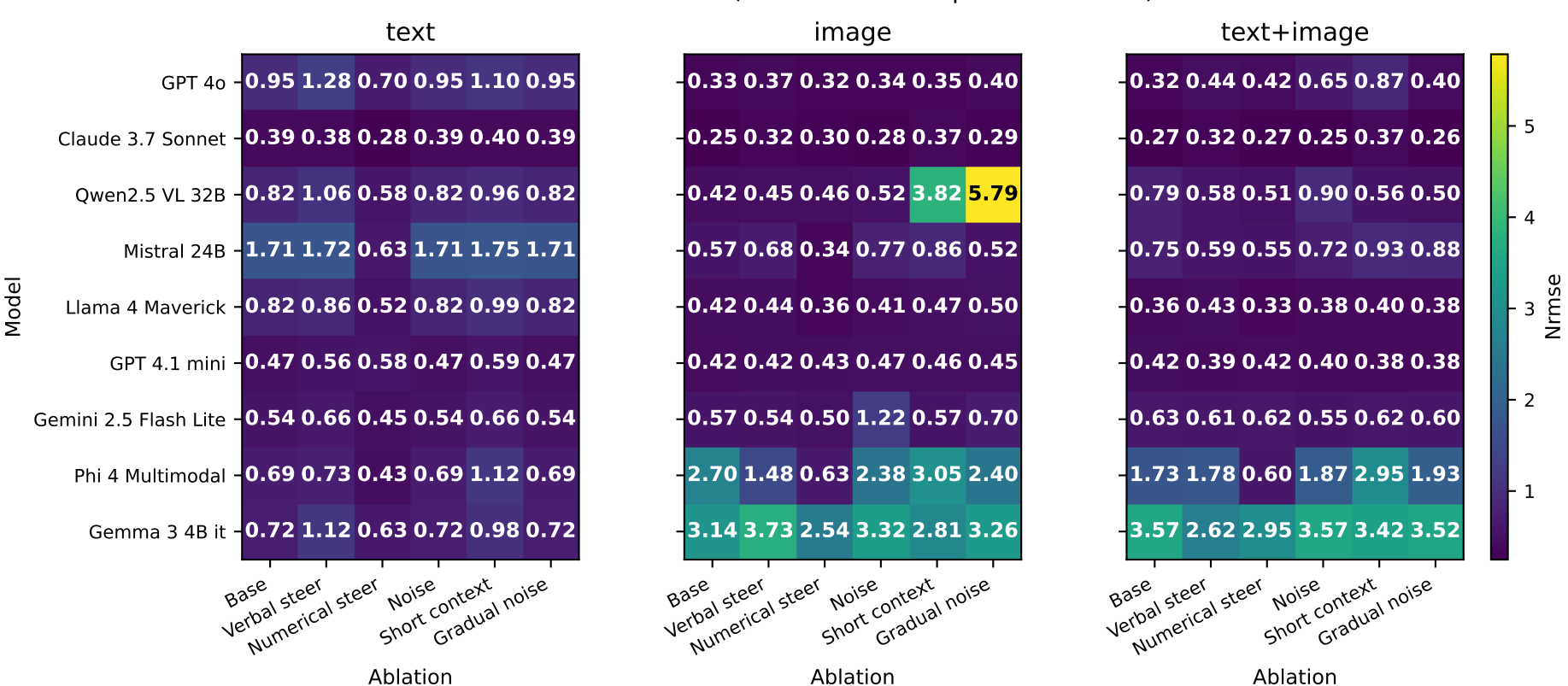
Observed NRMSE (normalized to experiment mean)
text
image
text+image
Model
GPT 4o
Claude 3.7 Sonnet
Qwen2.5 VL 32B
Mistral 24B
Llama 4 Maverick
GPT 4.1 mini
Gemini 2.5 Flash Lite
Phi 4 Multimodal
Gemma 3 4B it
0.95 1.28 0.70 0.95 1.10 0.95
0.39 0.38 0.28 0.39 0.40 0.39
0.82 1.06 0.58 0.82 0.96 0.82
1.71 1.72 0.63 1.71 1.75 1.71
0.82 0.86 0.52 0.82 0.99 0.82
0.47 0.56 0.58 0.47 0.59 0.47
0.54 0.66 0.45 0.54 0.66 0.54
0.69 0.73 0.43 0.69 1.12 0.69
0.72 1.12 0.63 0.72 0.98 0.72
0.33 0.37 0.32 0.34 0.35 0.40
0.25 0.32 0.30 0.28 0.37 0.29
0.42 0.45 0.46 0.52 3.82 5.79
0.57 0.68 0.34 0.77 0.86 0.52
0.42 0.44 0.36 0.41 0.47 0.50
0.42 0.42 0.43 0.47 0.46 0.45
0.57 0.54 0.50 1.22 0.57 0.70
2.70 1.48 0.63 2.38 3.05 2.40
3.14 3.73 2.54 3.32 2.81 3.26
0.32 0.44 0.42 0.65 0.87 0.40
0.27 0.32 0.27 0.25 0.37 0.26
0.79 0.58 0.51 0.90 0.56 0.50
0.75 0.59 0.55 0.72 0.93 0.88
0.36 0.43 0.33 0.38 0.40 0.38
0.42 0.39 0.42 0.40 0.38 0.38
0.63 0.61 0.62 0.55 0.62 0.60
1.73 1.78 0.60 1.87 2.95 1.93
3.57 2.62 2.95 3.57 3.42 3.52
Base
Verbal steer
Numerical steer
Noise
Short context
Gradual noise
Ablation
Nrmse
1
2
3
4
5