Stage 11 — Step 2 Findings

In Step 2, we tested three sweeps (S2-1, S2-2, S2-3) using Sweep C as baseline, but reweighted energy terms and strengthened inhibition. Goal: reduce hallucinations, especially flip_v, while preserving recall and improving margins.

Summary Metrics (50 samples each):

Sweep	Exact Accuracy	Grid Similarity	Precision	Recall	F1	Jaccard	Hallucination	Omission	Margin Mean	Margin Min
S2-1	0.30	0.433	0.72	1.00	0.81	0.72	0.28	0.00	1.81	0.87
S2-2	0.30	0.433	0.72	1.00	0.81	0.72	0.28	0.00	2.07	1.06
S2-3	0.28	0.418	0.72	1.00	0.81	0.72	0.28	0.00	1.69	0.93

Per-Primitive Breakdown:

Sweep	Primitive	True Count	Pred Count	Hallucinations	Halluc Rate
S2-1	flip_h	38	50	12	0.24
S2-1	flip_v	32	50	18	0.36
S2-1	rotate	38	50	12	0.24
S2-2	flip_h	38	50	12	0.24
S2-2	flip_v	32	50	18	0.36
S2-2	rotate	38	50	12	0.24
S2-3	flip_h	38	50	12	0.24
S2-3	flip_v	32	50	18	0.36
S2-3	rotate	38	50	12	0.24

Observations:

• Accuracy nudged up to 0.30 in S2-1 and S2-2 compared to 0.28 in Step 1 Sweep C.
• Precision and hallucination rates remained flat (Precision ≈ 0.72, Hallucination ≈ 0.28).
• Recall remained perfect at 1.0 across all sweeps.
• Margins improved, especially in S2-2 (mean margin ≈ 2.07).
• flip_v hallucinations remained fixed at 36%, indicating channel-specific bias.

Conclusion:

Step 2 improved **margins** (well sharpness) and modestly nudged accuracy, but did not reduce hallucinations. The persistent flip_v over-prediction suggests that the problem is structural (prototype alignment / residual artifacts). Next steps should focus on prototype diversity or per-primitive calibration to target flip_v specifically.