Stage 11 — Step 4 Findings

Step 4 introduced **per-primitive thresholds** (e.g. higher raw floors for flip_v), residual ceilings, and **robust prototype aggregation** (soft-min, median, consensus). The aim was to directly collapse phantom wells rather than just sharpen the landscape.

Summary Metrics (50 samples each):

Sweep	Exact Accuracy	Grid Similarity	Precision	Recall	F1	Jaccard	Hallucination	Omission	Margin Mean	Margin Min
S4-1	0.24	0.392	0.687	1.00	0.78	0.687	0.313	0.00	1.71	0.97
S4-2	0.18	0.344	0.687	1.00	0.78	0.687	0.313	0.00	2.09	1.15
S4-3	0.28	0.424	0.687	1.00	0.78	0.687	0.313	0.00	1.13	0.46

Per-Primitive Breakdown (50 samples):

Sweep	Primitive	True Count	Pred Count	Hallucinations	Halluc Rate
S4-1	flip_h	33	50	17	0.34
S4-1	flip_v	35	50	15	0.30
S4-1	rotate	35	50	15	0.30
S4-2	flip_h	33	50	17	0.34
S4-2	flip_v	35	50	15	0.30
S4-2	rotate	35	50	15	0.30
S4-3	flip_h	33	50	17	0.34
S4-3	flip_v	35	50	15	0.30
S4-3	rotate	35	50	15	0.30

Observations:

- Accuracy varied: 0.18 (S4-2) to 0.28 (S4-3), similar to prior steps.
- Precision stuck at ~0.687, hallucinations fixed at ~0.31.
- Recall remained perfect at 1.0 (no omissions).
- Margins improved in S4-2 (mean ≈ 2.09, min ≈ 1.15) — sharpest wells so far.
- Per-primitive hallucination rates remained flat: flip_v ≈ 30%, flip_h ≈ 34%, rotate ≈ 30%.

Conclusion:

Step 4 sharpened wells and enforced stricter conditions, but **hallucinations persisted at ~30%**. This confirms that thresholds and aggregation rules alone are insufficient. The structural issue lies in **prototype design and alignment** — phantom bumps still find acceptable matches. Next steps: redesign prototypes (e.g., asymmetry, multi-scale) and calibrate per-primitive acceptance regions.