Stage 11 — Step 5 Findings

Step 5 redesigned prototypes: multi-scale (narrow, wide), hinge, derivative, and orthogonalized flip_v against flip_h/rotate. We also used per-primitive gating (raw floors, residual ceilings) and stronger consensus. Goal: directly suppress phantom wells, especially for flip_v.

Summary Metrics (50 samples each):

Sweep	Exact Accuracy	Grid Similarity	Precision	Recall	F1	Jaccard	Hallucination	Omission	Margin Mean	Margin Mi
S5-1	0.36	0.481	0.707	1.00	0.792	0.707	0.293	0.00	1.67	0.90
S5-2	0.24	0.387	0.620	1.00	0.732	0.620	0.380	0.00	1.56	0.89
S5-3	0.36	0.477	0.713	1.00	0.796	0.713	0.287	0.00	1.88	1.03

Per-Primitive Breakdown (50 samples):

Sweep	Primitive	True Count	Pred Count	Hallucinations	Halluc Rate
S5-1	flip_h	35	50	15	0.30
S5-1	flip_v	33	50	17	0.34
S5-1	rotate	38	50	12	0.24
S5-2	flip_h	29	50	21	0.42
S5-2	flip_v	30	50	20	0.40
S5-2	rotate	34	50	16	0.32
S5-3	flip_h	34	50	16	0.32
S5-3	flip_v	36	50	14	0.28
S5-3	rotate	37	50	13	0.26

Observations:

• S5-1 and S5-3 improved accuracy to 0.36 (best so far).
• Precision rose to ~0.71 (vs. 0.687 ceiling in Steps 2–4).
• Hallucinations dipped below 0.30 (S5-3 = 0.287).
• Margins strongest yet in S5-3 (mean=1.88, min=1.03).
• flip_v hallucination dropped to 0.28 in S5-3 — first break below the 0.30–0.36 wall.
• S5-2 regressed (hallucinations ↑0.38, accuracy ↓0.24) — over-constraining hurt.

Conclusion:

Step 5 achieved a **breakthrough**: multi-scale/shape prototypes plus consensus and orthogonalization reduced hallucinations and improved precision/accuracy. flip_v finally dropped below the long-stuck 30% hallucination rate. S5-3 in particular delivered the best results so far: accuracy 0.36,

precision 0.713, hallucinations 0.287, strong margins. This validates prototype redesign as the right path. Next steps: scale up to 100+ samples, tune flip_v gates further, and confirm stability.