Weekly Report: Patch Size Analysis, Auto-Regressive Rollout, and Dynamics Training

Zhang Yiran

10 December 2025

1 Patch Size Comparison (3×3, 5×5, 7×7 Models)

1.1 Experimental Setup

For each local neighborhood size (3×3, 5×5, 7×7), we train an MLP classifier to predict the center cell state using weighted loss

$$w_{\text{pos}} = \min(5, \sqrt{r}), \qquad r = \frac{N_{\text{dead}}}{N_{\text{alive}}}.$$

All spatial patches are extracted using **wrap-around (toroidal) padding**, meaning patches crossing the boundary of the grid are sampled cyclically. This ensures spatial homogeneity and avoids boundary artifacts.

Two types of datasets are used:

- **General dataset**: $p = 0.5$, burn-in = 50, 60 evaluation steps.
- **Matched datasets**: densities $p \in \{0.2, 0.4, 0.6\}$ and regimes *early* (burn-in = 10), *mid* (burn-in = 60), *late* (burn-in = 160), each with 40 evaluation steps.

All models share the same architecture (two-layer MLP with hidden size 128) and training schedule. Performance is measured using Accuracy, Precision, Recall, and F1-score.

1.2 Overall Performance on General Dataset

Table 1: Performance on the general dataset (p = 0.5). Larger patches consistently improve F1.

Patch Size	Accuracy	Precision	Recall	F1
3×3	0.8721	0.4462	**0.8231**	0.5787
5×5	0.8930	0.4991	0.8009	0.6150
7×7	**0.8985**	**0.5163**	0.7741	**0.6195**

Observation. F1 improves monotonically as patch size increases. Although recall decreases slightly from 5×5 to 7×7, the corresponding gain in precision leads to the strongest overall F1 for 7×7.

1.3 Performance Across Densities and Regimes

Since reporting all 27 configurations (3 densities × 3 regimes × 3 patch sizes) would be redundant, we summarize the key outcome using **regime-averaged F1**. This aggregates across densities within each regime (useful for identifying dynamical trends).

Table 2: Regime-averaged F1-score across patch sizes (averaged over p = 0.2, 0.4, 0.6).

Regime	3×3 F1	5×5 F1	7×7 F1
Early	0.5787	0.6065	**0.6119**
Mid	0.5960	0.6268	**0.6375**
Late	0.6219	0.6601	**0.6736**

Key trends:

- In all regimes, F1 improves as patch size increases.
- Improvements from 3×3 → 5×5 are substantial in all regimes.
- Improvements from 5×5 → 7×7 are smaller but highly consistent.
- The advantage of 7×7 is clearest in the *late* regime, where local patterns become more structured and larger spatial context becomes beneficial.

1.4 Interpretation

Three main observations emerge from the comparison across patch sizes:

1. **Limited spatial context strongly restricts model performance.** The 3×3 model systematically underperforms across all densities and regimes. Although its recall is relatively high, its precision and F1 are consistently lower, indicating that the restricted neighborhood does not provide enough information for reliable discrimination of subtle local configurations.
2. **5×5 patches offer a substantial improvement and form a strong baseline.** Expanding the receptive field to 5×5 yields a significant increase in F1-score across all settings. This suggests that many cell-state patterns in the Game of Life require contextual cues beyond immediate neighbors but still within a moderately sized spatial footprint.
3. **7×7 patches provide the most stable and highest overall F1.** While the performance gain from 5×5 to 7×7 is smaller than from 3×3 to 5×5, the improvement is remarkably consistent across all densities and regimes. Larger spatial context helps resolve ambiguous or multi-cell interactions, especially in the *mid* and *late* regimes where structured patterns (oscillators, still lifes, moving fronts) dominate.

Overall, the results indicate that expanding the spatial neighborhood improves single-step predictive fidelity in a stable and monotonic manner. Although precision and recall individually fluctuate across settings, the F1-score—which best captures the precision–recall trade-off—is consistently highest for the 7×7 model.

2 Autoregressive Rollouts

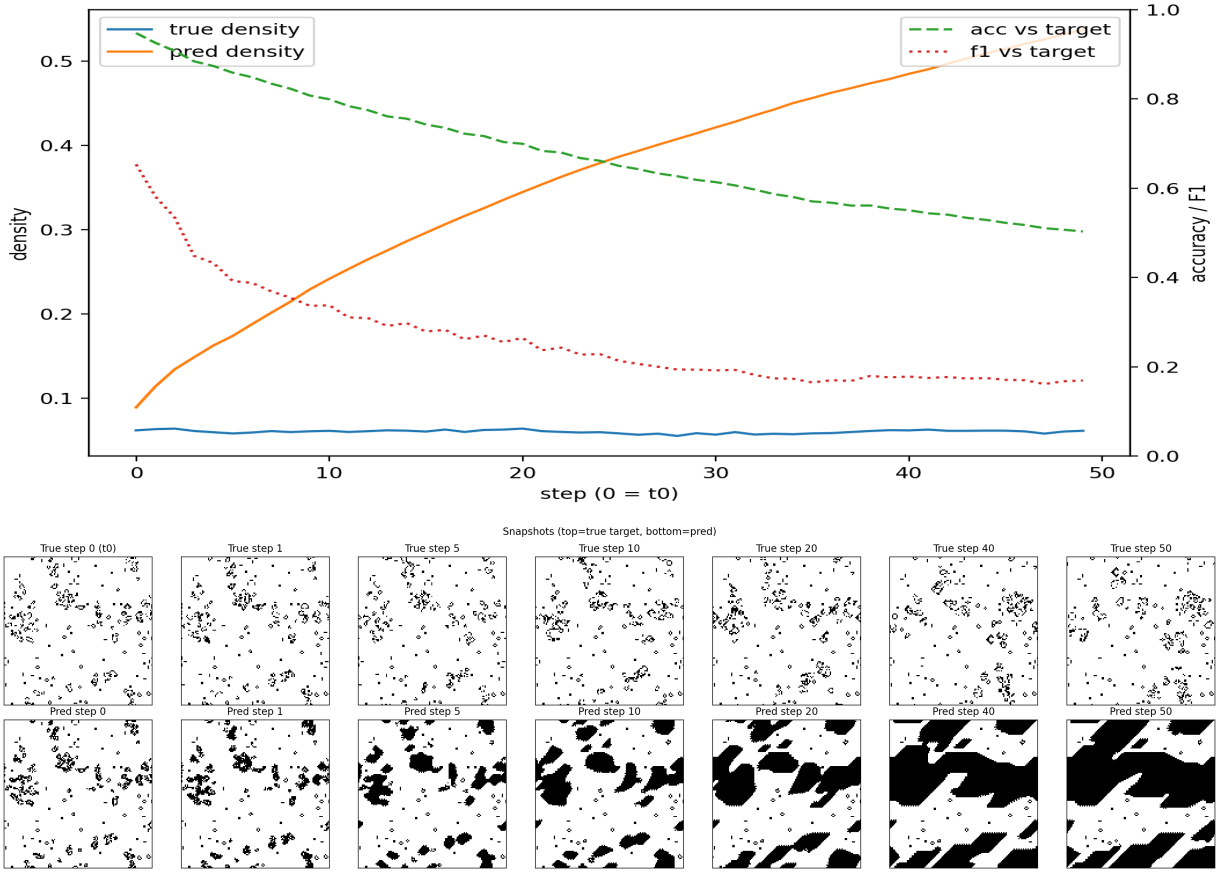
2.1 Rollout without Calibration (alignment = current, temp = 1.0, thr = 0.5, 50 steps)

2.1.1 General Models

Without calibration the rollouts drift/bloom; avg_f1 is low. Larger patches help somewhat (late, p=0.6) but degradation remains.

Table 3: Non-calibrated, general models (avg over 50 steps).

Setting	patch3 (avg_acc / avg_f1)	patch5	patch7
early, p=0.2	**0.367** / 0.230	0.199 / 0.245	0.255 / **0.254**
mid, p=0.4	0.378 / 0.207	0.253 / 0.232	**0.397** / **0.251**
late, p=0.6	0.392 / 0.154	0.384 / 0.196	**0.677** / **0.262**

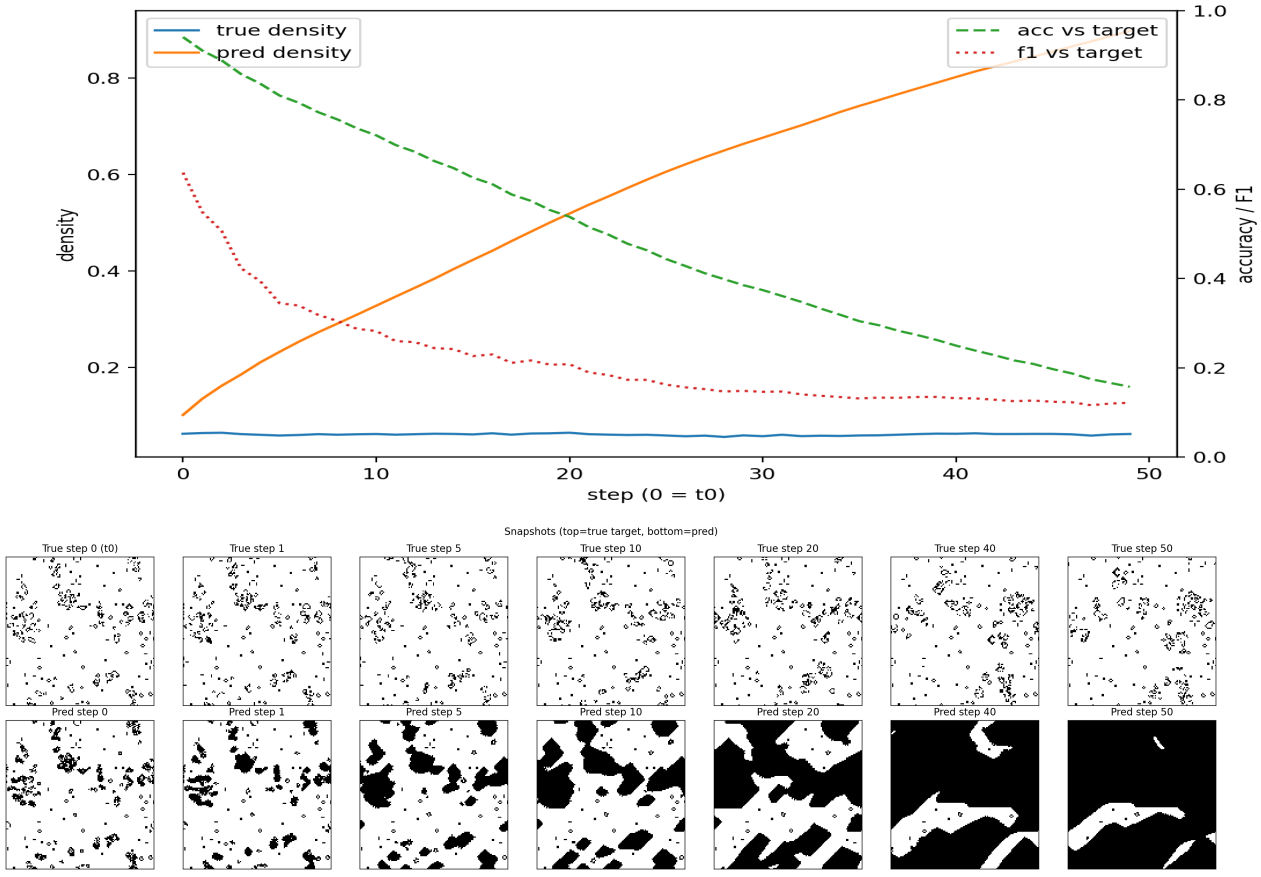


2.1.2 Matched Models

Matched ckpts do not materially fix the drift; f1 remains low, though patch7 is consistently best.

Table 4: Non-calibrated, matched models (avg over 50 steps).

Setting	patch3 (avg_acc / avg_f1)	patch5	patch7
early, p=0.2	0.191 / 0.244	0.200 / 0.245	**0.222** / **0.249**
mid, p=0.4	0.233 / 0.225	0.253 / 0.230	**0.328** / **0.246**
late, p=0.6	0.244 / 0.162	0.301 / 0.178	**0.488** / **0.216**



Takeaways (non-calibrated) Explosive drift dominates; patch7 is least bad (especially late/p=0.6) but f1 stays low. Matched models offer only limited gains over general.

2.2 Rollout with Calibration (alignment = current; tuned temp/thr; adaptive thr on)

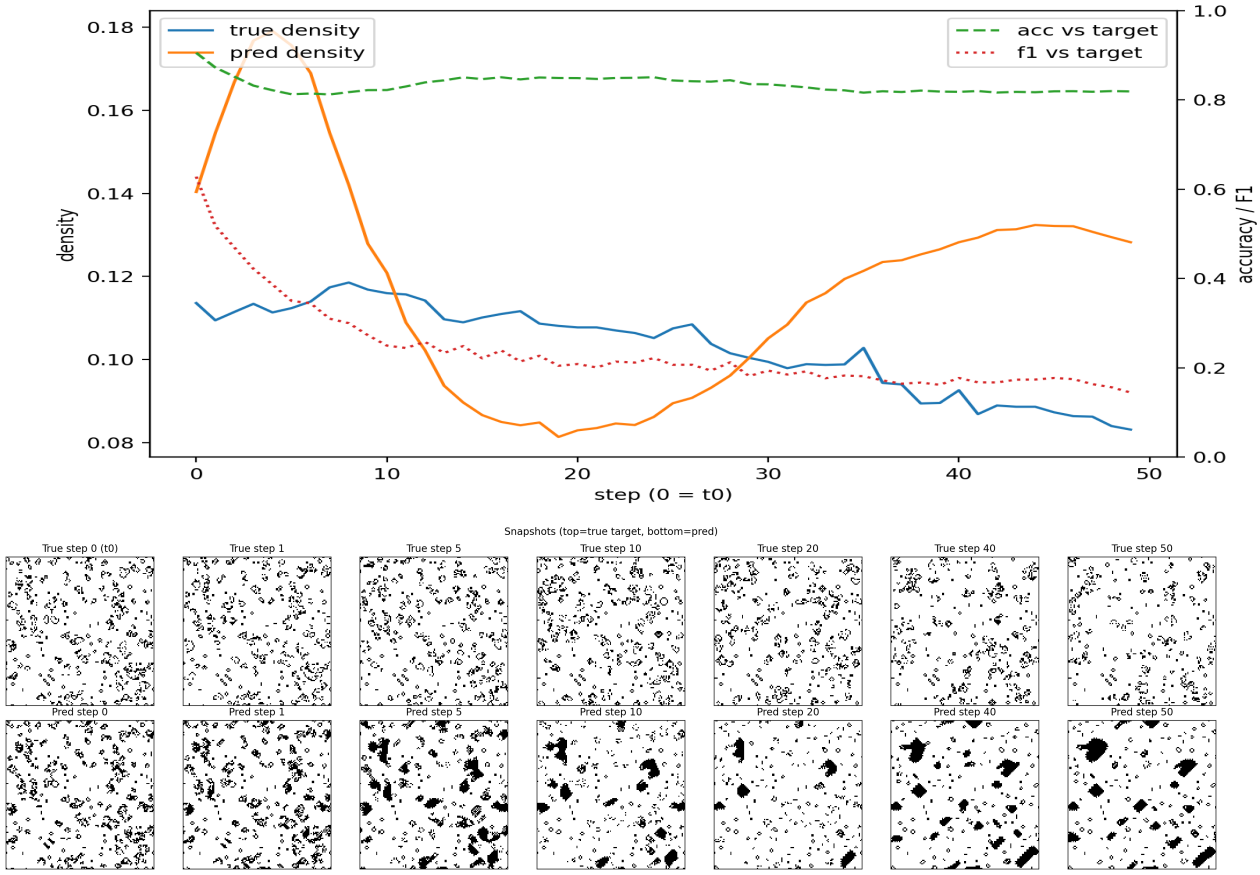
Settings: temps {1.0, 1.5, 2.0}, thresholds {0.6, 0.7}, short-horizon tuning 10 steps, adaptive threshold (gain=0.5, thr_min=0.2, thr_max=0.85).

2.2.1 General Models

Calibration suppresses explosion and raises avg_acc; avg_f1 improves modestly. (Late p=0.6 not run.)

Table 5: Calibrated, general (avg over 50 steps; temp/thr in parentheses).

Setting	patch3 (avg_acc / avg_f1)	patch5	patch7
early, p=0.2	0.758 / 0.190 (1.00, 0.60)	0.752 / 0.176 (1.00, 0.60)	**0.763** / **0.197** (1.00, 0.60)
mid, p=0.4	0.827 / 0.209 (1.00, 0.60)	0.818 / 0.185 (1.00, 0.60)	**0.833** / **0.237** (1.00, 0.60)

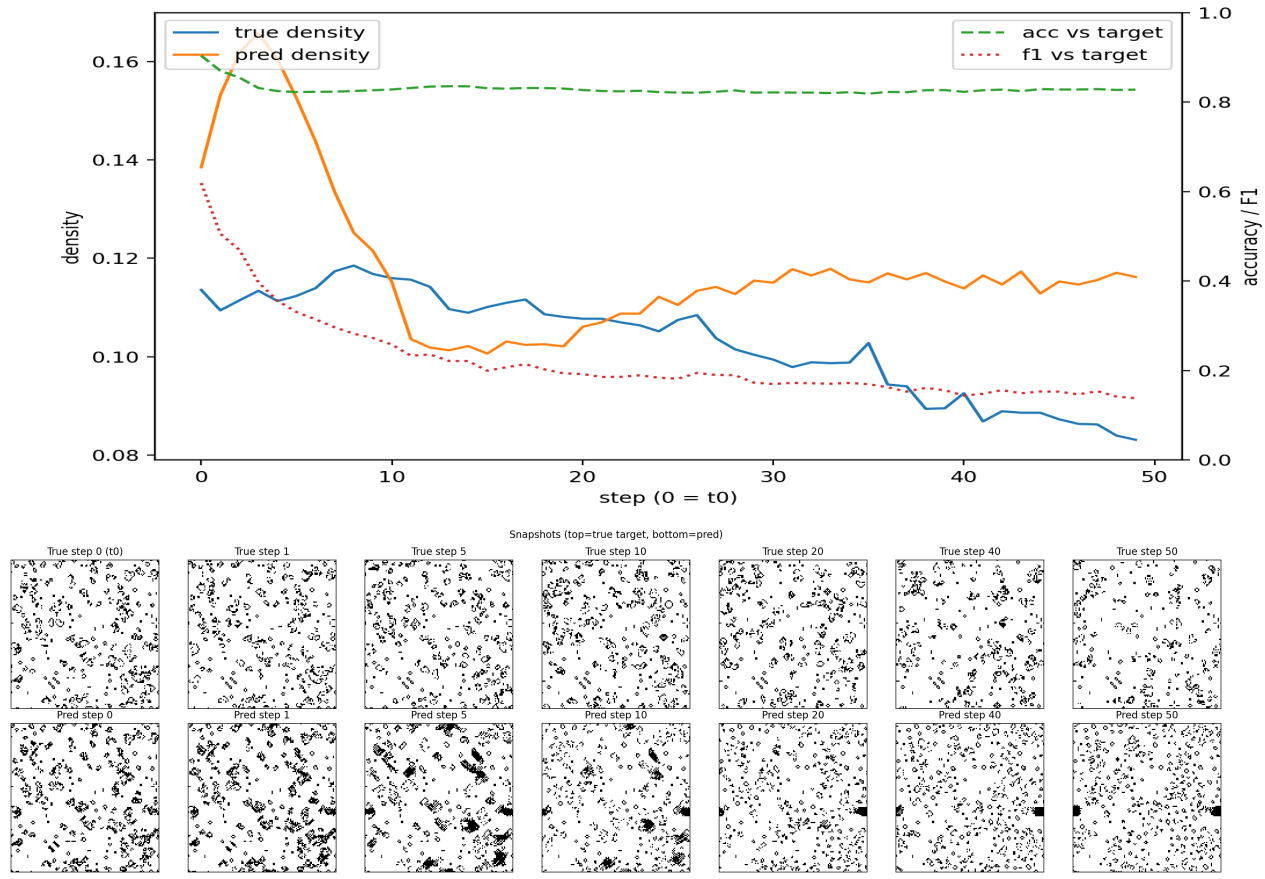


2.2.2 Matched Models

Calibration (tuned temp/thr + adaptive) on matched ckpts shows similar patterns; late p=0.6 not run.

Table 6: Calibrated, matched (avg over 50 steps; temp/thr in parentheses).

Setting	patch3 (avg_acc / avg_f1)	patch5	patch7
early, p=0.2	**0.761** / **0.192** (1.50, 0.60)	0.758 / 0.185 (1.50, 0.60)	0.758 / 0.193 (1.00, 0.60)
mid, p=0.4	0.825 / 0.199 (1.00, 0.60)	0.818 / 0.177 (1.00, 0.60)	**0.829** / **0.221** (1.00, 0.60)
late, p=0.6		not run	



Takeaways (calibrated) Calibration (tuned temp/thr + adaptive) boosts avg_acc and stabilizes density; avg_f1 rises slightly. Patch7 remains the strongest among sizes. Late regime remains to be evaluated.

2.3 Overall Observations

- **Non-calibrated rollouts:** All models drift rapidly; avg_f1 remains low. Patch7 is the least bad (notably in late p=0.6), but matched checkpoints offer only marginal gains over general; early/mid cases remain unstable.
- **Calibrated rollouts:** Tuning temp/thr plus adaptive thresholding markedly increases avg_acc and stabilizes density. Avg_f1 improves slightly but remains modest. Patch7 remains the strongest. Calibrated late p=0.6 is still missing.
- **Patch-size trend:** Across calibrated and non-calibrated settings, larger patches (5/7) outperform patch3; patch7 consistently yields the best avg_f1/stability.
- **Model source:** Matched vs general yields limited benefit; even after calibration, improvements are small. Priority should be on patch size and calibration, and on completing the calibrated late regime to firm up conclusions.

3 Dynamics Model: 3×3 → Next-Center (9→1)

3.1 Experimental Setup

- Data generation: wrap/toroidal GoL, board 128×128, p=0.5, burn-in 50, 60 steps; 16/4 boards (train/test), 200 patches/step, seed=42. Stored at `data/dynamics_patch3.npz`.
- Model: MLP (9→64→64→1), BCEWithLogitsLoss, Adam (lr=1e-3), batch 2048, max 10 epochs, patience 2 (early stopping).
- Label: center cell at t+1; input: full 3×3 patch at t (including center).
- Checkpoint: `checkpoints_dynamics/best_model_dynamics_patch3.pth`.

3.2 Test-Set Performance

Table 7: Patch-level test (p=0.5 distribution).

Accuracy	Precision	Recall	F1
1.0000	**1.0000**	**1.0000**	**1.0000**

The task is deterministic (2^9=512 possible inputs); the model perfectly captures the GoL update rule on this distribution.

3.3 Board-Level Rollout Checks (10 steps, alignment to true GoL)

All runs: acc@1 = 1.0000, acc@10 = 1.0000, avg_acc = 1.0000; only densities change due to GoL dynamics (not model drift).

Table 8: Board rollout accuracy vs. true GoL; dens_last = predicted density at step 10.

Setting	acc@1	acc@10	avg_acc	dens_last
early, p=0.2 (burn=10)	1.0000	1.0000	1.0000	0.1372
early, p=0.4 (burn=10)	1.0000	1.0000	1.0000	0.1733
mid, p=0.2 (burn=60)	1.0000	1.0000	1.0000	0.0919
mid, p=0.4 (burn=60)	1.0000	1.0000	1.0000	0.1160
late, p=0.2 (burn=160)	1.0000	1.0000	1.0000	0.0712
late, p=0.4 (burn=160)	1.0000	1.0000	1.0000	0.0738

Interpretation With full 3×3 input (including center), the model exactly reproduces the GoL transition; multi-step rollouts match true GoL with no drift. Residual density changes reflect GoL's own evolution, not model error. This can serve as an upper bound/reference for approximate local predictors.