# WHEN DEPARTURE RATES SATURATE: ENDPOINT DIAGNOSTICS FOR LEARNED DYNAMICS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

The any-time departure rate is the fraction of predicted trajectories that leave a prescribed region at least once. At long horizons, this rate can saturate at one even when models differ in their return behavior. We study the additional information obtained by recording whether the same predicted trajectories end inside or outside the region. An exact decomposition distinguishes terminal exclusion from departure followed by return. A two-state calculation shows how identical departure rates can coexist with different return behavior while endpoint and later-occupancy rankings still agree. For fixed candidate models, we give finite-sample endpoint-risk bounds that allow diagnostic eligibility screens and randomized ties, while isolating the additional condition needed for temporal transfer. In a complete 30-model mushroom-billiard cohort, every departure rate equals one, yet terminal exclusion predicts later outside occupancy on disjoint starts with Spearman correlation $0.905$. The association survives a diagnostic concentration screen and appears in a separate study of both Koopman autoencoders and residual predictors. Exploratory selection analyses improve on the saturated rule, including after accuracy screening; the residual-family selection gain remains unresolved. These results support a low-cost evaluation practice: report endpoint status alongside departure and predictive fidelity, and evaluate selection performance against a separately specified regional target.

## 1 INTRODUCTION

Long-horizon learned simulators are often evaluated by whether their rollouts remain in a prescribed region. Such a region may encode an invariant component, an admissible state range, or a task constraint. The any-time departure rate answers an essential question: did the predicted trajectory leave at least once? However, the score accumulates failures over the entire horizon. Once every evaluated trajectory has departed, it assigns the same value to models with very different subsequent behavior. A saturated alarm is still a valid record of a violation, but it no longer orders regional outcomes.

The membership traces $(0, 0, 1, 1)$ and $(0, 0, 1, 0)$ illustrate the distinction, with $1$ denoting exclusion. Both paths depart; one ends outside and the other returns. Their common any-time label discards this difference. We retain the final membership label using the same predictor, starts, horizon, and oracle (Figure 1). When the full rollout has already been evaluated for departure, retaining the endpoint label requires no additional model or oracle evaluation. If the region is a hard safety constraint, both paths still fail that constraint. Endpoint status describes the state after a violation; it does not erase the violation.

The substantive question is whether this inexpensive label contains information about behavior beyond the diagnostic rollout. It need not: endpoint rankings can reverse later. We separate three claims. An exact identity specifies what the label recovers. A simple departure–return model exhibits a regime in which departure rates lose all ranking information while endpoint and later-occupancy rankings still agree. A finite-sample result then bounds selection within a declared eligible candidate set and identifies the additional membership-change term needed to relate endpoint selection to later-window occupancy. Our contribution is to identify an evaluation blind spot caused by departure-rate saturation, assess the value of endpoint status for distinguishing models, and establish the conditions and limits of its use in model selection.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

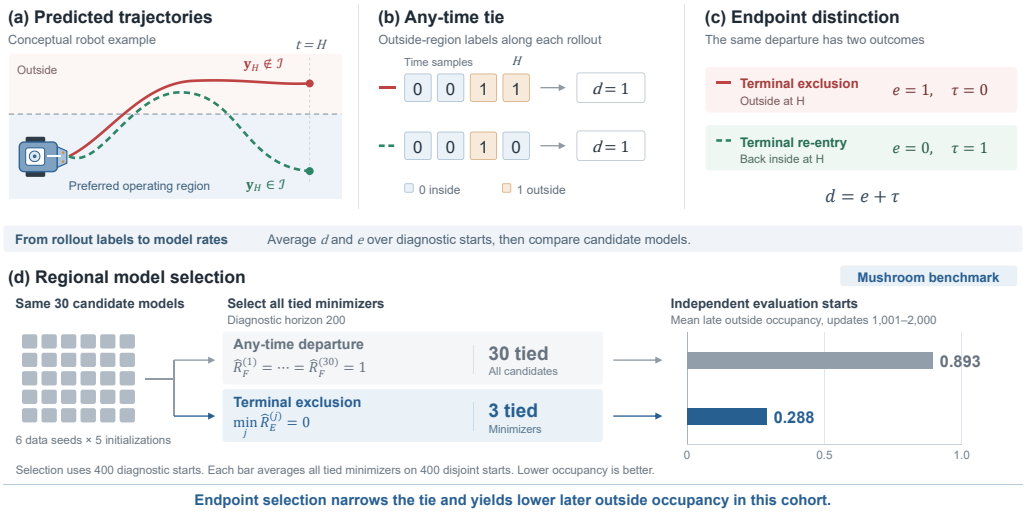



Figure 1: **Endpoint information for regional model selection.** (a–c) Both conceptual robot rollouts leave the region but differ in endpoint status. (d) In the mushroom cohort, minimizing any-time departure leaves all 30 models tied, whereas minimizing terminal exclusion yields three tied minimizers. Mean late outside occupancy is 0.893 and 0.288, respectively, averaged over all tied minimizers on independent evaluation starts.

The experiments provide evidence at two levels. A complete crossed cohort of 30 mushroom-billiard Koopman autoencoders has identical any-time rates but nine distinct endpoint rates. Endpoint exclusion predicts later outside occupancy on disjoint starts, and the association persists after a diagnostic concentration screen. A separate 60-model architecture study supports the same endpoint–target association beyond latent-linear predictors. Averaging endpoint status over longer terminal windows yields no statistically resolved gain in association with the later target. Exploratory selection comparisons quantify the value and the risks of using this signal, including accuracy screens and heterogeneous tied minimizers. Standard-map and smooth-flow controls establish that both terminal outcomes occur under other independently specified membership rules. These results support retaining endpoint status as a complementary regional diagnostic.

## 2 Related Work

**Learning and validating long-horizon dynamics.** DMD and EDMD approximate evolution operators (Schmid, 2010; Williams et al., 2015); neural Koopman models learn coordinates and latent evolution jointly (Lusch et al., 2018). Re-encoding and action-conditioned representations improve rollouts and planning (Fathi et al., 2024; Mondal et al., 2024). Hamiltonian networks, SympNets, conservation corrections, invariant-measure objectives, and stable-set objectives alter the predictor, update, or training loss (Greydanus et al., 2019; Jin et al., 2020; McGreivy & Hakim, 2023; Jiang et al., 2023; Schiff et al., 2024; Takeishi & Kawahara, 2021). Our diagnostic holds the fitted predictor fixed. It complements symplectic validation (Higgins et al., 2021), residual spectral checks (Colbrook & Townsend, 2024), and feature-space invariance diagnostics (Conradie et al., 2026) by measuring the regional status of decoded trajectories.

**Regional and temporal information.** Transfer-operator methods identify almost-invariant or metastable sets (Dellnitz & Junge, 1999; Froyland & Dellnitz, 2003; Deuflhard & Weber, 2005; Mardt et al., 2018); ergodic partitions and trajectory averages organize state space (Mezić & Wiggins, 1999; Budišić & Mezić, 2012; Duignan & Meiss, 2023). Set invariance, recurrent-event analysis, and temporal logic distinguish path constraints from state or event summaries (Blanchini, 1999; Aalen et al., 2008; Donzé & Maler, 2010). Learned safe sets can support constrained control and return-to-safe-set planning (Wilcox et al., 2022). We use a prescribed membership oracle, without

learning a safe set or guaranteeing a controller. Endpoint membership itself is elementary. Our focus is its measurable value after departure-score saturation, and the distinction between endpoint estimation, temporal transfer, and selection of a faithful simulator.

**Controlled systems and statistical scope.** Mushroom billiards offer a sharp regular–chaotic division (Bunimovich, 2001; Dettmann & Georgiou, 2011), while the standard map provides mixed phase space (Meiss, 1992). These systems permit explicit membership semantics and reference controls. The selection result uses standard bounded-variable concentration (Hoeffding, 1963); its role is to state exactly which sampling and transfer assumptions support a regional decision. The empirical contribution is the crossed-model evaluation under those fixed semantics, including negative and unresolved outcomes.

## 3 Endpoint Information and Its Limits

### 3.1 A common rollout and a fixed membership rule

Let $T : \mathcal{X} \to \mathcal{X}$ be the reference map and $m : \mathcal{X} \to \{0, 1\}$ a fixed oracle for an evaluated region $\mathcal{I}$. Write $\ell = 1 - m$ for exclusion. Each predictor produces a path $\mathbf{y}^F_{i,0:H}$ from an inside start:

$$\mathbf{y}^F_{i,0} = \mathbf{x}_i, \qquad m(\mathbf{x}_i) = 1, \qquad i = 1, \ldots, n. \tag{1}$$

Overflowed outputs receive $\ell = 1$. The empirical any-time, endpoint, and return labels are

$$d_i = \max_{0 \le t \le H} \ell(\mathbf{y}^F_{i,t}), \qquad e_i = \ell(\mathbf{y}^F_{i,H}), \qquad \tau_i = d_i - e_i. \tag{2}$$

Their sample means are $\widehat{R}_F$, $\widehat{R}_E$, and $\widehat{R}_{\mathrm{tr}}$, respectively.

**Proposition 1** (Terminal-status accounting)**.** *Let $H \ge 1$, let $m$ be binary, and let the paths satisfy Equation* (1)*. Then $e_i \le d_i$, $\tau_i \in \{0, 1\}$, and*

$$\widehat{R}_F = \widehat{R}_E + \widehat{R}_{\mathrm{tr}}. \tag{3}$$

This identity makes no claim about future persistence or predictive accuracy. It separates paths ending outside from paths that have returned by $H$; it does not require uninterrupted exclusion. Appendix A gives the proof and corresponding label-disagreement bounds.

### 3.2 Why saturation can coexist with informative endpoints

A two-state model isolates the roles of departure and return. We use this two-state model only to illustrate the roles of departure and return. Its parameters are not estimated from the Hamiltonian experiments. Let $Z_t = 0$ denote inside and $Z_t = 1$ outside, with transition probabilities $a$ from inside to outside and $b$ from outside to inside.

**Proposition 2** (Common departure, different return)**.** *Let $Z_0 = 0$ and let $(Z_t)$ be a time-homogeneous Markov chain on $\{0, 1\}$ with $a \in (0, 1)$ and $b \in [0, 1 - a]$. For every integer $H \ge 1$,*

$$\Pr\left(\max_{0 \le t \le H} Z_t = 1\right) = 1 - (1 - a)^H, \qquad \Pr(Z_H = 1) = \frac{a}{a + b}\left[1 - (1 - a - b)^H\right]. \tag{4}$$

*For fixed $a$ and $H \ge 2$, the endpoint probability is strictly decreasing in $b$, as is $|\mathcal{W}|^{-1} \sum_{t \in \mathcal{W}} \Pr(Z_t = 1)$ for every finite nonempty $\mathcal{W} \subset \{H + 1, H + 2, \ldots\}$.*

Thus candidates can have identical departure probabilities at every horizon, approaching one, while their endpoint and later-occupancy rankings agree. The any-time event depends only on the first departure; the endpoint also reflects return. This calculation explains a possible source of diagnostic value without asserting endpoint superiority to every trajectory summary. Its assumptions matter: without a temporal relation, later rankings can reverse completely (Appendix C).

### 3.3 SELECTION WITHIN AN ELIGIBLE CANDIDATE SET

Let $F_1, \dots, F_M$ be fixed candidates. Stratum $s$ has start distribution $\mu_s$, size $n_s$, and fixed weight $w_s \geq 0$, with $\sum_s w_s = 1$. Let $\mu = \sum_s w_s \mu_s$ and $R_E^{(j)} = \mathbb{E}_\mu[e^{(j)}]$; estimate it by $\widehat{R}_E^{(j)} = \sum_s w_s n_s^{-1} \sum_i e_{s,i}^{(j)}$. For a fixed finite nonempty later window $\mathcal{W}$, define

$$Q^{(j)} = \mathbb{E}_\mu\left[\frac{1}{|\mathcal{W}|}\sum_{t\in\mathcal{W}} \ell(\mathbf{y}_t^{(j)})\right], \quad B^{(j)} = \mathbb{E}_\mu\left[\frac{1}{|\mathcal{W}|}\sum_{t\in\mathcal{W}} |\ell(\mathbf{y}_t^{(j)}) - e^{(j)}|\right]. \tag{5}$$

$B^{(j)}$ measures disagreement with the diagnostic endpoint, not the number of boundary crossings. An eligibility rule returns a nonempty set $A \subseteq \{1, \dots, M\}$ using diagnostic information. It can enforce numerical validity, a fidelity threshold, or a dispersion threshold. For tolerance $\eta \geq 0$, define

$$I_\eta(A) = \left\{ j \in A : \widehat{R}_E^{(j)} \leq \min_{k\in A} \widehat{R}_E^{(k)} + \eta \right\}. \tag{6}$$

A selector uses a probability vector $\pi$ supported on this set; uniform $\pi$ at $\eta = 0$ specifies the exact-tie policy used below.

**Proposition 3** (Screened selection, ties, and temporal transfer)**.** *Let the $n_s \geq 1$ starts be mutually independent within and across strata and independent of the $M$ fixed candidates. Let fixed weights $w_s \geq 0$ sum to one, and let $\mathcal{W} \subset \{H+1, H+2, \dots\}$ be fixed and finite nonempty. For $\delta \in (0,1)$, set $\varepsilon_\delta = \sqrt{\frac{1}{2}(\sum_s w_s^2/n_s)\log(2M/\delta)}$. With probability at least $1-\delta$, for every nonempty $A$, every $\eta \geq 0$, and every probability vector $\pi$ supported on $I_\eta(A)$,*

$$\begin{aligned} &\sum_j \pi_j R_E^{(j)} - \min_{j\in A} R_E^{(j)} \leq 2\varepsilon_\delta + \eta, \\ &\sum_j \pi_j Q^{(j)} - \min_{j\in A} Q^{(j)} \leq 2\varepsilon_\delta + \eta + \sum_j \pi_j B^{(j)} + B^{(j_Q^\star)}, \qquad j_Q^\star \in \arg\min_{j\in A} Q^{(j)}. \end{aligned} \tag{7}$$

The simultaneous event permits a diagnostic-dependent screen; the comparison is with the best member of that same eligible set. The result does not certify that the screen itself enforces physical fidelity. It also holds for each exact minimizer by taking $\pi$ to be a point mass. Proofs, the sharp departure-only information limit, and margin-based sufficient conditions for transfer appear in Appendices A and C. For $M = 30$, $n_s = 200$ in two equally weighted strata, and $\delta = 0.05$, the endpoint term $2\varepsilon_\delta$ is $0.188$. The empirical later-window study is not a numerical transfer certificate: the saved aggregate records do not identify the paired membership disagreements $B^{(j)}$.

### 3.4 IMPLEMENTATION AND COST

The criterion applies to any predicted path. Our Koopman model uses the periodic embedding $\boldsymbol{\phi}_\sigma(\theta, p) = (\cos\theta, \sin\theta, p/\sigma)$ and unembedding $\boldsymbol{\psi}_\sigma(\mathbf{r}) = (\operatorname{atan2}(r_2, r_1) \bmod 2\pi, \sigma r_3)$:

$$\mathbf{z}_{i,0} = E(\boldsymbol{\phi}_\sigma(\mathbf{x}_i)), \quad \mathbf{z}_{i,t+1} = \mathbf{K}\mathbf{z}_{i,t}, \quad \mathbf{y}_{i,t}^F = \boldsymbol{\psi}_\sigma(D(\mathbf{z}_{i,t})) \quad (t \geq 1). \tag{8}$$

Appendix D specifies the reconstruction, one-step prediction, and latent-linearity losses. Reference and persistence controls are

$$\mathbf{y}_{i,t}^T = T^t(\mathbf{x}_i), \qquad \mathbf{y}_{i,t}^P = \mathbf{x}_i. \tag{9}$$

They verify the reference region and time-zero convention. Retaining $e_i$ needs no additional predictor or oracle call if endpoint membership was already evaluated for $d_i$. Computing the departure score can still be costly when the membership oracle is expensive; the standard-map oracle uses 5,000 reference steps. For a signed Lipschitz membership margin, endpoint labels are stable away from the boundary; Appendix A gives the corresponding stability bound. Optional circular-mean prototypes test a different, spatial aggregation question (Appendix B).

## 4 EXPERIMENTAL DESIGN

**Primary association cohort.** We use all 30 models obtained by crossing six mushroom-billiard training-data seeds with five initialization seeds. Diagnostic evaluation uses 200 starts in each of two

Table 1: **Membership semantics fixed independently of the predictor.** Exact formulas, reference controls, and numerical tolerances are in Appendix E.

| System | Evaluated region and oracle |
|---|---|
| Mushroom billiard | Cap/shoulder regular region: correct boundary piece and angular momentum $\lvert L\rvert \geq r$. |
| Standard map | Finite-time boundedness: $\lvert p_t\rvert < 2$ for reference iterates $t = 0, \ldots, 5000$. |
| Smooth flow | Exact invariant momentum band $\lvert p\rvert \leq p_0$ for the declared Hamiltonian. |

boundary-depth strata and horizon $H = 200$. A disjoint panel has another 200 starts per stratum and horizon 2,000. The target $Q$ is outside occupancy over updates 1001:2000. The diagnostic-only cohort rule requires $\widehat{R}_F \geq 0.975$; all 30 models pass with $\widehat{R}_F = 1$. Exact reference trajectories remain confined on both panels. Seeds, training details, and the exact geometric oracle are given in Appendix F.1.

**Architecture and terminal-window study.** A separate $6 \times 5$ grid per architecture compares Koopman autoencoders (11,155 parameters) with direct residual autoregressive MLPs (10,803 parameters). Both families share training-data realizations, initialization labels, evaluation panels, and horizons. All 30 Koopman models and 25 of 30 residual models meet the same departure threshold. We compare final windows of $1, 5, 10, 20$, and $50$ updates. This is an additional architecture study on the same system and evaluation panels, not a new-system or new-panel replication.

**Estimands, inference, and provenance.** The saved primary analysis plan specifies the model-level Spearman association between diagnostic endpoint exclusion and the later target. Crossed-factor bootstrap intervals independently resample data and initialization labels; they condition on the fixed start panels. The primary interval uses the original 2,000 attempted resamples. Secondary displays use 10,000 draws, and paired comparisons use common draws. The paired correlation contrast, selection bootstrap, accuracy screens, and concentration screen are exploratory reanalyses of the retained records. These analyses remain exploratory. An internal plan hash or replay of saved records does not constitute independent prospective validation.

For selection, we specify uniform random choice among exact diagnostic minimizers (numerical tolerance $10^{-12}$). Its expected target is the mean over the full tie set. Regret subtracts the smallest held-out target within the same eligible cohort. The held-out target is used for evaluation, never to choose a tied candidate. Comparators are any-time departure, whole-trajectory diagnostic occupancy, and one-step RMSE. RMSE uses generic full-boundary transitions, whereas regional scores use exact-island starts, so the RMSE comparison includes a distribution difference. Secondary screens retain either the recorded lowest-RMSE fractions or models with diagnostic endpoint spread above $10^{-5}$; they do not redefine the primary cohort.

**Other systems.** Fixed-checkpoint, fresh-training, and smooth-flow controls test terminal outcomes under the three independently specified oracles in Table 1. Each comparison fixes starts, horizon, and oracle. These controls test the departure/return distinction; only the mushroom studies include a disjoint longer-horizon selection target.

# 5 Results

## 5.1 Endpoint information predicts later regional behavior

All 30 primary models have any-time departure one, but endpoint exclusion takes nine values spanning $[0, 1]$ (Figure 2a). Thus saturation hides substantial variation even on the same starts and horizon. Endpoint exclusion is strongly associated with later outside occupancy on disjoint starts: $\rho = 0.905$, with original 95% interval $[0.719, 1.000]$. Whole-trajectory occupancy is also informative, with $\rho = 0.758$. The exploratory paired correlation contrast favors the endpoint by $0.148$ ($[0.011, 0.407]$). First-departure time, one-step RMSE, and endpoint spread have intervals including zero (Tables 4–5). These results support a predictive relationship between endpoint exclusion

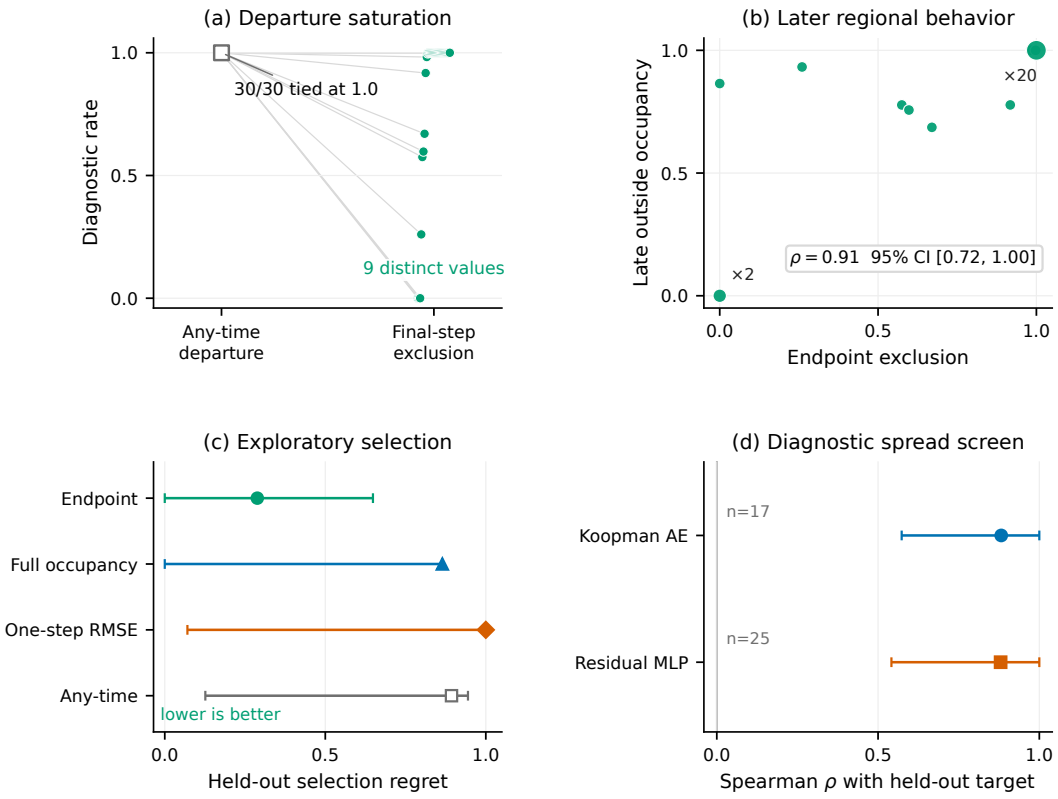



Figure 2: **Information recovered after departure saturation.** (a) All 30 primary models tie on any-time departure; endpoints give nine distinct values. (b) Endpoint exclusion predicts the later target on disjoint starts; bubble labels disclose coincident outcomes. (c) Exploratory expected regret under uniform random tie breaking, with crossed-factor intervals. (d) Association after screening diagnostic endpoint dispersion; this screen does not rule out later concentration. All intervals condition on the evaluation panels.

and later regional behavior in this cohort. They do not establish a general advantage of endpoint exclusion over other trajectory summaries.

### 5.2 The predictive signal survives diagnostic screening

Removing the 13 models with diagnostic endpoint spread at most $10^{-5}$ leaves a strong association in the remaining 17 models: $\rho = 0.882$ ($[0.573, 1.000]$). The same screen retains all 25 eligible residual models, with $\rho = 0.880$ ($[0.541, 1.000]$). This excludes near-identical endpoints at the diagnostic horizon as the sole explanation of the association. It does not certify noncollapse at the target horizon. The full primary cohort contains 20 joint endpoint/target ceiling outcomes; removing them leaves $\rho = 0.593$ on ten models. Appendix F.1 reports the complete dispersion, overflow, and ceiling sensitivities, including the small five-model spread subgroup.

### 5.3 Architecture and terminal-window comparisons

The separate architecture study supports endpoint association for both latent-linear and direct residual predictors (Table 2). Thus the regional signal is not confined to a Koopman representation. Averaging 5–50 terminal updates gives correlations from $0.941$ to $0.957$ for Koopman and $0.839$ to $0.882$ for residual models, but every paired contrast with the single endpoint includes zero (Figure 3). Longer windows therefore offer no statistically resolved association gain in this study; this does not prove equivalence of the summaries or optimality of the anchor horizon.

Table 2: **Additional architecture study on the mushroom system.** $\Delta R = R_{\text{any}} - R_{\text{end}}$; positive values favor endpoint selection. Intervals condition on the fixed panels. The residual selection contrast is unresolved.

| Model family | $n$ | $\rho$ [95% CI] | $R_{\text{any}}$ | $R_{\text{end}}$ | $\Delta R$ [95% CI] |
|---|---|---|---|---|---|
| Koopman AE | 30 | $0.941\ [0.698, 1.000]$ | 0.832 | 0.000 | $0.832\ [0.230, 0.935]$ |
| Residual AR MLP | 25 | $0.880\ [0.564, 1.000]$ | 0.156 | 0.002 | $0.155\ [-0.001, 0.207]$ |

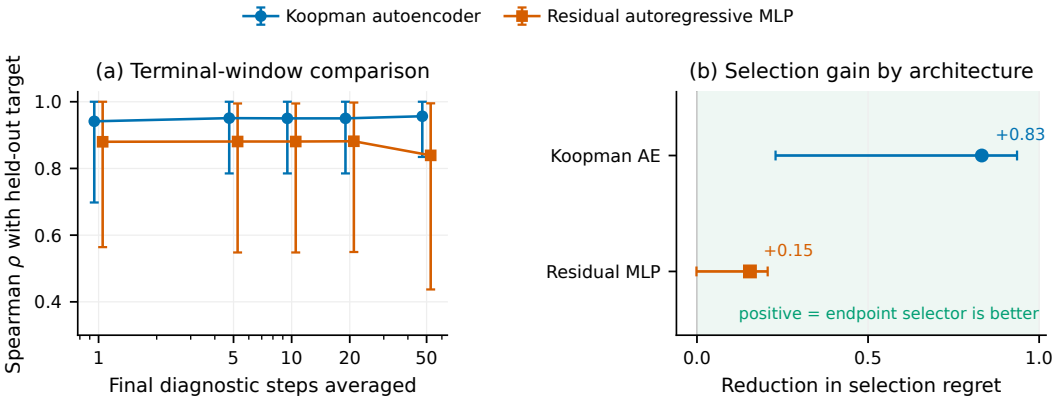



Figure 3: **What transfers across architectures.** (a) Endpoint and short terminal-window scores have similar association with the same later target. (b) Endpoint selection improves on the any-time comparator in the additional Koopman cohort; the residual confidence interval includes zero. The architectures use similar parameter counts and shared panels and rollout horizons (Section 4).

### 5.4 Exploratory selection: useful signal, explicit tie risk

In the primary cohort, uniform choice among endpoint minimizers gives expected regret $0.288$, compared with $0.893$ for the saturated any-time rule. Their exploratory paired difference is $[-0.934, -0.010]$, with lower endpoint regret in $97.9\%$ of bootstrap draws. Point regrets for whole-trajectory occupancy and one-step RMSE are $0.865$ and $1.000$; paired selection uncertainty does not resolve superiority over both. Association and selection are distinct estimands: a stronger rank correlation need not improve the selected minimum.

The selected tie set is consequential. Its three later occupancies are $0$, $0$, and $0.865$, so the expected value $0.288$ is neither the performance of every minimizer nor a guarantee for an unspecified tie break. Two models have concentrated held-out endpoints; the third later overflows (Table 6). The diagnostic dispersion screen leaves a unique endpoint minimizer with target zero, also selected by whole-trajectory occupancy, but that model concentrates later. Across all four recorded accuracy screens, endpoint selection has lower point regret than the any-time rule; all thresholds and comparator outcomes are reported in Table 3. These analyses support complementary regional information after a diagnostic screen, while leaving physical fidelity as a separate requirement.

The full-cohort minimum endpoint score is zero and the next distinct score is $0.260$. Descriptive near-tie tolerances corresponding to one, five, or ten additional endpoint exclusions out of 400 starts ($\eta = 0.0025, 0.0125, 0.025$) retain the same three-model tie set. This verifies separation in the observed score vector, while stability across newly sampled panels remains untested.

In the additional Koopman study, endpoint, whole-trajectory occupancy, and RMSE all attain zero point regret. In the residual study, endpoint and whole-trajectory occupancy select the same model. Consequently, the clearest selection result is the improvement over a saturated or nearly saturated alarm. The comparisons do not establish a unique advantage of endpoint selection over other informative alternatives.

Table 3: **Regional selection after diagnostic screening.** Exploratory point regrets use uniform random choice among diagnostic minimizers and the best held-out target within the same eligible pool. "Worst tie" is the worst endpoint-selected regret. Every recorded accuracy fraction and the diagnostic spread screen are shown; no threshold is selected using the target.

| Eligibility rule | $n$ | Any-time | RMSE | Occupancy | Endpoint | Worst tie |
|---|---|---|---|---|---|---|
| Lowest-RMSE 25% | 8 | 0.185 | 0.243 | 0.021 | 0.021 | 0.021 |
| Lowest-RMSE 50% | 15 | 0.853 | 1.000 | 0.865 | 0.432 | 0.865 |
| Lowest-RMSE 75% | 23 | 0.861 | 1.000 | 0.865 | 0.288 | 0.865 |
| Lowest-RMSE 100% | 30 | 0.893 | 1.000 | 0.865 | 0.288 | 0.865 |
| Diagnostic spread $> 10^{-5}$ | 17 | 0.878 | 0.757 | 0.000 | 0.000 | 0.000 |

**Terminal outcomes beyond the selection system.** The fixed mushroom checkpoint and smooth-flow control predominantly end outside, whereas standard-map excursions return under its finite-time oracle. Fresh mushroom training realizes both outcomes across seeds. Figure 4 and the per-seed tables in the appendix show every run. These tests broaden the evidence for the diagnostic distinction without extending the selection claim to untested systems.

## 6 Discussion and Conclusion

Any-time departure, endpoint status, and predictive fidelity answer different questions. The key result is that a saturated departure alarm can coexist with endpoint variation that predicts later regional behavior on disjoint starts. The controlled departure–return calculation explains why this can occur, and the selection bound states how sampling uncertainty, eligibility, ties, and temporal disagreement enter the decision. The practical recommendation is to retain the endpoint label whenever regional departure is measured, and to declare the regional target before using it to compare models.

The evidence remains finite-horizon and system-specific. It uses fixed evaluation panels, a small number of crossed seed levels, and one full selection system. Concentration and overflow show why a regional loss alone cannot certify a faithful simulator; a model that violates a hard safety constraint still fails an all-path criterion. A stronger selection study should pre-specify the fidelity screen and tie-breaking rule before evaluating new panels and horizons. Within the present scope, retaining endpoint status is inexpensive and provides reproducible information that the saturated any-time score discards. Its association with later regional behavior is observed in both tested architectures.

## Reproducibility Statement

The release includes five experiment records, model-level crossed-grid results, two fixed checkpoints, pinned dependencies, and source code. `code/reproduce_revision.py` replays all saved analyses in a new directory, verifies the numerical claims and proof examples, and regenerates the empirical figures and tables without retraining. `MANIFEST.sha256` records the current paper, code, and evidence hashes. Appendix K distinguishes saved-record verification from full experiment generation and documents the separate analysis stages.

## Data Availability

All data are simulated by the released generators; the records used for the reported results are included.

## AI Use Statement

Generative AI tools assisted with polishing the paper only. The authors take responsibility for the paper's text, claims, and code.

## REFERENCES


Odd O. Aalen, Ørnulf Borgan, and Håkon K. Gjessing. *Survival and Event History Analysis: A Process Point of View*. Statistics for Biology and Health. Springer, New York, 2008. doi: 10.1007/978-0-387-68560-1.

Franco Blanchini. Set invariance in control. *Automatica*, 35(11):1747–1767, 1999. doi: 10.1016/S0005-1098(99)00113-2.

Marko Budišić and Igor Mezić. Geometry of the ergodic quotient reveals coherent structures in flows. *Physica D: Nonlinear Phenomena*, 241(15):1255–1269, 2012. doi: 10.1016/j.physd.2012.04.006.

Leonid A. Bunimovich. Mushrooms and other billiards with divided phase space. *Chaos*, 11(4):802–808, 2001. doi: 10.1063/1.1418763.

Matthew J. Colbrook and Alex Townsend. Rigorous data-driven computation of spectral properties of Koopman operators for dynamical systems. *Communications on Pure and Applied Mathematics*, 77(1):221–283, 2024. doi: 10.1002/cpa.22125.

Gustav Conradie, Nicolas Boullé, Jean-Christophe Loiseau, Steven L. Brunton, and Matthew J. Colbrook. Trustworthy Koopman operator learning: Invariance diagnostics and error bounds. arXiv preprint arXiv:2603.15091, 2026.

Michael Dellnitz and Oliver Junge. On the approximation of complicated dynamical behavior. *SIAM Journal on Numerical Analysis*, 36(2):491–515, 1999. doi: 10.1137/S0036142996313002.

Carl P. Dettmann and Orestis Georgiou. Open mushrooms: Stickiness revisited. *Journal of Physics A: Mathematical and Theoretical*, 44(19):195102, 2011. doi: 10.1088/1751-8113/44/19/195102.

Peter Deuflhard and Marcus Weber. Robust Perron cluster analysis in conformation dynamics. *Linear Algebra and its Applications*, 398:161–184, 2005. doi: 10.1016/j.laa.2004.10.026.

Alexandre Donzé and Oded Maler. Robust satisfaction of temporal logic over real-valued signals. In *Formal Modeling and Analysis of Timed Systems*, volume 6246 of *Lecture Notes in Computer Science*, pp. 92–106. Springer, 2010. doi: 10.1007/978-3-642-15297-9_9.

Nathan Duignan and James D. Meiss. Distinguishing between regular and chaotic orbits of flows by the weighted Birkhoff average. *Physica D: Nonlinear Phenomena*, 449:133749, 2023. doi: 10.1016/j.physd.2023.133749.

Mahan Fathi, Clement Gehring, Jonathan Pilault, David Kanaa, Pierre-Luc Bacon, and Ross Goroshin. Course correcting Koopman representations. In *International Conference on Learning Representations*, 2024.

Gary Froyland and Michael Dellnitz. Detecting and locating near-optimal almost-invariant sets and cycles. *SIAM Journal on Scientific Computing*, 24(6):1839–1863, 2003. doi: 10.1137/S106482750238911X.

Samuel Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Irina Higgins, Peter Wirnsberger, Andrew Jaegle, and Aleksandar Botev. SyMetric: Measuring the quality of learnt Hamiltonian dynamics inferred from vision. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Ruoxi Jiang, Peter Y. Lu, Elena Orlova, and Rebecca Willett. Training neural operators to preserve invariant measures of chaotic attractors. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1202.

Pengzhan Jin, Zhen Zhang, Aiqing Zhu, Yifa Tang, and George Em Karniadakis. SympNets: Intrinsic structure-preserving symplectic networks for identifying Hamiltonian systems. *Neural Networks*, 132:166–179, 2020. doi: 10.1016/j.neunet.2020.08.017.

Bethany Lusch, J. Nathan Kutz, and Steven L. Brunton. Deep learning for universal linear embeddings of nonlinear dynamics. *Nature Communications*, 9:4950, 2018. doi: 10.1038/s41467-018-07210-0.

Andreas Mardt, Luca Pasquali, Hao Wu, and Frank Noé. VAMPnets for deep learning of molecular kinetics. *Nature Communications*, 9:5, 2018. doi: 10.1038/s41467-017-02388-1.

Nick McGreivy and Ammar Hakim. Invariant preservation in machine learned PDE solvers via error correction. arXiv preprint arXiv:2303.16110, 2023.

James D. Meiss. Symplectic maps, variational principles, and transport. *Reviews of Modern Physics*, 64(3):795–848, 1992. doi: 10.1103/RevModPhys.64.795.

Igor Mezić and Stephen Wiggins. A method for visualization of invariant sets of dynamical systems based on the ergodic partition. *Chaos*, 9(1):213–218, 1999. doi: 10.1063/1.166399.

Arnab Mondal, Siba Smarak Panigrahi, Sai Rajeswar, Kaleem Siddiqi, and Siamak Ravanbakhsh. Efficient dynamics modeling in interactive environments with Koopman theory. In *International Conference on Learning Representations*, 2024.

Yair Schiff, Zhong Yi Wan, Jeffrey B. Parker, Stephan Hoyer, Volodymyr Kuleshov, Fei Sha, and Leonardo Zepeda-Núñez. DySLIM: Dynamics stable learning by invariant measure for chaotic systems. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 43649–43684, 2024.

Peter J. Schmid. Dynamic mode decomposition of numerical and experimental data. *Journal of Fluid Mechanics*, 656:5–28, 2010. doi: 10.1017/S0022112010001217.

Naoya Takeishi and Yoshinobu Kawahara. Learning dynamics models with stable invariant sets. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(11):9782–9790, 2021. doi: 10.1609/aaai.v35i11.17176.

Albert Wilcox, Ashwin Balakrishna, Brijen Thananjeyan, Joseph E. Gonzalez, and Ken Goldberg. LS3: Latent space safe sets for long-horizon visuomotor control of sparse reward iterative tasks. In *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pp. 959–969, 2022. URL `https://proceedings.mlr.press/v164/wilcox22a.html`.

Matthew O. Williams, Ioannis G. Kevrekidis, and Clarence W. Rowley. A data-driven approximation of the Koopman operator: Extending dynamic mode decomposition. *Journal of Nonlinear Science*, 25(6):1307–1346, 2015. doi: 10.1007/s00332-015-9258-5.

# A PROOFS

**Additional label-disagreement identities.** For binary vectors, let $\delta(\mathbf{a}, \mathbf{b}) = n^{-1} \sum_i \mathbb{I}[a_i \neq b_i]$. For any $\mathbf{g} \in \{0,1\}^n$, write $\widehat{R}_G = n^{-1} \sum_i g_i$. The terminal decomposition also yields

$$\delta(\mathbf{d}, \mathbf{g}) \leq \widehat{R}_{\mathrm{tr}} + \delta(\mathbf{e}, \mathbf{g}), \qquad |\widehat{R}_F - \widehat{R}_G| \leq \delta(\mathbf{d}, \mathbf{g}). \tag{10}$$

*Proof of Proposition 1.* We first observe that $\mathbf{y}_{i,H}^F$ occurs among the states over which the maximum in Equation (2) is taken. The endpoint definition therefore gives $e_i \leq d_i$. Because both indicators are binary, $\tau_i = d_i - e_i \in \{0,1\}$ and $d_i = e_i + \tau_i$. Averaging this identity over the cell yields Equation (3). The coordinatewise inequality $e_i \leq d_i$ implies $\delta(\mathbf{d}, \mathbf{e}) = n^{-1} \sum_i (d_i - e_i) = \widehat{R}_{\mathrm{tr}}$. Applying the triangle inequality for normalized Hamming distance yields the first inequality in Equation (10). For the second inequality,

$$\left| \widehat{R}_F - \widehat{R}_G \right| = \left| \frac{1}{n} \sum_{i=1}^{n} (d_i - g_i) \right| \leq \frac{1}{n} \sum_{i=1}^{n} |d_i - g_i| = \delta(\mathbf{d}, \mathbf{g}), \tag{11}$$

which completes the proof. □

*Proof of Proposition 2.* The event of no departure consists of $H$ consecutive transitions from state zero to state zero, so its probability is $(1-a)^H$. Let $r_t = \Pr(Z_t = 1)$. The Markov recursion is $r_{t+1} = a(1-r_t) + (1-b)r_t = a + (1-a-b)r_t$, with $r_0 = 0$. Therefore

$$r_t = a\sum_{k=0}^{t-1}(1-a-b)^k = \frac{a}{a+b}\big[1-(1-a-b)^t\big]. \tag{12}$$

Since $a+b>0$, the denominator is nonzero. On $b \in [0, 1-a]$, $\lambda = 1-a-b$ lies in $[0, 1-a]$. For every $t \geq 2$, the polynomial $a\sum_{k=0}^{t-1}\lambda^k$ is strictly increasing in $\lambda$, because its linear coefficient is $a > 0$ and the remaining coefficients are nonnegative. It is consequently strictly decreasing in $b$. Averaging these strictly decreasing functions over any fixed nonempty later window preserves strict decrease. At $H = 1$, $r_1 = a$ does not distinguish return probabilities, which is why the strict-ranking conclusion requires $H \geq 2$. The restriction $a + b \leq 1$ prevents oscillatory label dynamics; no monotonicity claim is made outside it. □

*Proof of Proposition 3.* Let $V = \sum_s w_s^2/n_s$. For a fixed candidate, each independent summand $w_s e_{s,i}^{(j)}/n_s$ lies in an interval of width $w_s/n_s$. Weighted Hoeffding concentration therefore gives

$$\Pr\Big(|\widehat{R}_E^{(j)} - R_E^{(j)}| > \varepsilon\Big) \leq 2\exp(-2\varepsilon^2/V). \tag{13}$$

A union bound with $\varepsilon = \varepsilon_\delta$ gives simultaneous accuracy for all $M$ candidates. On this event, fix any nonempty subset $A$, tolerance $\eta \geq 0$, and $j \in I_\eta(A)$. For every $k \in A$,

$$R_E^{(j)} \leq \widehat{R}_E^{(j)} + \varepsilon_\delta \leq \widehat{R}_E^{(k)} + \eta + \varepsilon_\delta \leq R_E^{(k)} + 2\varepsilon_\delta + \eta. \tag{14}$$

Average over $j$ with weights $\pi_j$, and take $k$ to minimize $R_E^{(k)}$ over $A$. This proves the first bound, for every permitted subset and probability vector on the same uniform event; no further union bound over subsets is required. The triangle inequality inside the expectation in Equation (5) yields $|Q^{(j)} - R_E^{(j)}| \leq B^{(j)}$. Taking $k = j_Q^\star \in \arg\min_{k\in A} Q^{(k)}$ in Equation (14) and averaging gives

$$\begin{aligned}\sum_j \pi_j Q^{(j)} &\leq \sum_j \pi_j R_E^{(j)} + \sum_j \pi_j B^{(j)} \\ &\leq R_E^{(j_Q^\star)} + 2\varepsilon_\delta + \eta + \sum_j \pi_j B^{(j)} \\ &\leq Q^{(j_Q^\star)} + 2\varepsilon_\delta + \eta + \sum_j \pi_j B^{(j)} + B^{(j_Q^\star)}.\end{aligned} \tag{15}$$

Shared evaluation starts across candidates are allowed, since the union bound requires no independence across their scores. A screen can depend on these diagnostic observations because the event already holds for every candidate. Candidate training or adaptive construction using the evaluation observations would violate the fixed-candidate assumption. The guarantee is for the mixture's expected loss conditional on the diagnostic sample, not for every realized random choice at that same mixture bound. Each individual near-minimizer satisfies its own point-mass bound. □

**What a saturated score cannot identify.** The following is an information limit over compatible path arrays; it makes no claim that every array is generated by a specified autonomous architecture.

**Proposition 4** (Sharp departure-only regret)**.** *Let $M \geq 2$, $H \geq 2$, let the evaluation sample be nonempty, and let the oracle admit an inside and an outside state. Among selectors whose only candidate information is the departure-rate vector $(1, \ldots, 1)$, the minimax expected empirical endpoint regret over compatible path arrays is $1 - 1/M$. Uniform random selection attains it. Every deterministic such selector has worst-case regret one.*

*Proof.* On the constant observed vector, a randomized selector has fixed probabilities $p_1, \ldots, p_M$. Choose $j_\star$ with $p_{j_\star} \leq 1/M$. Construct every path to depart at update one; candidate $j_\star$ returns inside

from update two through $H$, and all other candidates remain outside. Endpoint risks are zero for $j_\star$ and one otherwise, so expected regret is $1-p_{j_\star} \geq 1-1/M$. Conversely, for arbitrary endpoint risks $r_j \in [0,1]$, uniform selection has regret $M^{-1}\sum_j r_j - \min_j r_j \leq (M-1)/M$. A deterministic selector chooses one index with probability one; assign endpoint loss one to that index and zero to another. □

**Membership stability away from the boundary.**

**Proposition 5** (Boundary-margin stability). *Let membership use one fixed convention $m = \mathbb{I}[\gamma \geq 0]$ or $m = \mathbb{I}[\gamma > 0]$, where $\gamma$ is $L_\gamma$-Lipschitz on the compared-state domain. Let paired finite paths from the same inside starts satisfy*

$$\max_{0\leq t\leq H} \|\mathbf{y}_{i,t} - \widetilde{\mathbf{y}}_{i,t}\|_2 \leq \varepsilon, \qquad i = 1,\dots,n. \tag{16}$$

*Define*

$$\mathcal{A}_\varepsilon = \{i : \min_{0\leq t\leq H} |\gamma(\mathbf{y}_{i,t})| \leq L_\gamma\varepsilon\}, \quad \mathcal{A}_{\varepsilon,H} = \{i : |\gamma(\mathbf{y}_{i,H})| \leq L_\gamma\varepsilon\}. \tag{17}$$

*Then, for the perturbed rates denoted by primes,*

$$|\widehat{R}_F - \widehat{R}'_F| \leq |\mathcal{A}_\varepsilon|/n, \quad |\widehat{R}_E - \widehat{R}'_E| \leq |\mathcal{A}_{\varepsilon,H}|/n, \quad |\widehat{R}_{\mathrm{tr}} - \widehat{R}'_{\mathrm{tr}}| \leq |\mathcal{A}_\varepsilon|/n. \tag{18}$$

*Proof of Proposition 5.* Suppose the membership labels of $\mathbf{y}_{i,t}$ and $\widetilde{\mathbf{y}}_{i,t}$ differ. Their signed margins have opposite signs or one is zero, so Lipschitz continuity and Equation (16) give $|\gamma(\mathbf{y}_{i,t})| \leq L_\gamma\varepsilon$. If $d_i \neq d'_i$, such a label change occurs at some time and $i \in \mathcal{A}_\varepsilon$. If $e_i \neq e'_i$, it occurs at the horizon and $i \in \mathcal{A}_{\varepsilon,H}$. Finally, $\tau_i \neq \tau'_i$ implies $d_i \neq d'_i$ or $e_i \neq e'_i$. Since $\mathcal{A}_{\varepsilon,H} \subseteq \mathcal{A}_\varepsilon$, this places $i$ in $\mathcal{A}_\varepsilon$. Each absolute rate difference is bounded by the normalized Hamming disagreement of its indicator vectors, which proves Equation (18). □

## B CELLWISE ENDPOINT PROTOTYPES

The analysis above retains the endpoint label for each trajectory. To examine spatial aggregation separately, define the circular-mean representative and its departure label by

$$\mathbf{c} = \boldsymbol{\psi}_\sigma\left(n^{-1}\sum_i \boldsymbol{\phi}_\sigma(\mathbf{y}^F_{i,H})\right), \qquad g = \ell(\mathbf{c}). \tag{19}$$

The implemented angular convention is specified in Appendix F. The directional mismatch rates are

$$a_{\mathrm{in}\to\mathrm{out}} = g(1-\widehat{R}_E), \qquad a_{\mathrm{out}\to\mathrm{in}} = (1-g)\widehat{R}_E. \tag{20}$$

The first rate counts inside endpoints represented by an outside prototype; the second counts outside endpoints represented by an inside prototype.

**Proposition 6** (Cellwise prototype disagreement). *Let $e_1,\dots,e_n,g \in \{0,1\}$, $\mathbf{g} = g\mathbf{1}_n$, and $\widehat{R}_E = n^{-1}\sum_i e_i$. Then*

$$\delta(\mathbf{e},\mathbf{g}) = a_{\mathrm{in}\to\mathrm{out}} + a_{\mathrm{out}\to\mathrm{in}} = |\widehat{R}_E - g|. \tag{21}$$

*If this disagreement is positive, exactly one directional rate is positive.*

*Proof.* For binary labels and a constant $g$,

$$\delta(\mathbf{e},\mathbf{g}) = n^{-1}\sum_i |e_i - g| = g(1-\widehat{R}_E) + (1-g)\widehat{R}_E = |\widehat{R}_E - g|. \tag{22}$$

Since $g(1-g) = 0$, at most one directional term is positive, and a positive sum makes exactly one term positive. □

## C TEMPORAL TRANSFER AND SIGNED MARGINS

**Why endpoint risk and later occupancy differ.** Let the oracle admit an inside state and an outside state, and fix $H \geq 2$ and a finite nonempty window $\mathcal{W} \subset \{H+1, H+2, \ldots\}$. A first path can depart at update one, be inside at $H$, and be outside throughout $\mathcal{W}$. A second can depart at update one, be outside at $H$, and be inside throughout $\mathcal{W}$. Both any-time rates equal one. Their endpoint losses are zero and one, respectively, so terminal minimization selects the first path. Its later occupancy is one, whereas that of the second path is zero. The resulting later-window regret is one. A later-occupancy guarantee therefore requires a temporal-transfer condition in addition to endpoint risk.

**A sufficient transfer condition.** A margin condition connects the transfer term to state displacement. Distances below use the same norm or periodic metric as the Lipschitz assumption. All states are finite in this statement.

**Proposition 7** (Margin control of temporal transfer)**.** *Fix a candidate $j$, a finite nonempty window $\mathcal{W} \subset \{H+1, H+2, \ldots\}$, and $m = \mathbb{I}[\gamma \geq 0]$ or $m = \mathbb{I}[\gamma > 0]$, with a fixed convention. Suppose all path states are finite, $\gamma$ is $L_\gamma$-Lipschitz on a domain containing the paths, and, for $\mu$-almost every start $\mathbf{x}$, measurable bounds $a_{j,t}(\mathbf{x}) \geq 0$ satisfy $\|\mathbf{y}_t^{(j)}(\mathbf{x}) - \mathbf{y}_H^{(j)}(\mathbf{x})\| \leq a_{j,t}(\mathbf{x})$. Then*

$$B^{(j)} \leq \mathbb{E}_\mu \left[ |\mathcal{W}|^{-1} \sum_{t \in \mathcal{W}} \mathbb{I}\Big\{ |\gamma(\mathbf{y}_H^{(j)})| \leq L_\gamma a_{j,t}(\mathbf{x}) \Big\} \right]. \tag{23}$$

*Proof.* If the two membership labels differ, the margins have opposite signs or one is zero. Lipschitz continuity implies $|\gamma(\mathbf{y}_H^{(j)})| \leq |\gamma(\mathbf{y}_t^{(j)}) - \gamma(\mathbf{y}_H^{(j)})| \leq L_\gamma a_{j,t}(\mathbf{x})$. The indicator on the right therefore bounds the label disagreement for each time and start. Averaging over the window and integrating over $\mu$ proves the claim. □

For a linear latent rollout, write $G = \boldsymbol{\psi}_\sigma \circ D$ and $\mathbf{z}_H(\mathbf{x}) = \mathbf{K}^H E(\boldsymbol{\phi}_\sigma(\mathbf{x}))$. If $G$ is $L_G$-Lipschitz on the relevant latent domain, one admissible bound is

$$a_{j,t}(\mathbf{x}) = L_G \|(\mathbf{K}^{t-H} - \mathbf{I})\mathbf{z}_H(\mathbf{x})\|. \tag{24}$$

Applying this conditional bound to periodic unembedding requires the periodic metric and a decoded domain separated from angular singularities. The experiments evaluate regional outcomes directly; Equation (24) supplies an analytic sufficient condition for future occupancy.

**The finite-time standard-map oracle.** The implementation tests $|p_t| < 2$ for $t = 0, \ldots, K$. Consequently,

$$\gamma_K(x,p) = \min_{0 \leq t \leq K} (2 - |p_t(x,p)|), \qquad m_{\mathrm{std}}(x,p) = \mathbb{I}[\gamma_K(x,p) > 0]. \tag{25}$$

A smooth lift of the standard map has a bounded one-step Jacobian. Each finite iterate therefore has a finite Lipschitz constant; absolute value and a finite minimum preserve Lipschitz continuity. The resulting bound may grow rapidly with $K$. The strict inequality matches the implementation at boundary points. The oracle describes finite-time boundedness, whose margin is distinct from an infinite-time invariant-set boundary.

**Decoded endpoint concentration.** The tanh decoder has final affine layer $\mathbf{W}_{\mathrm{out}}\mathbf{h} + \mathbf{b}_{\mathrm{out}}$, with $|h_k| \leq 1$. In real arithmetic,

$$|D_r(\mathbf{z})| \leq |b_{\mathrm{out},r}| + \sum_k |W_{\mathrm{out},rk}|. \tag{26}$$

Latent growth and decoded divergence are therefore different properties. At a finite horizon, endpoint dispersion describes variation across starts, while arithmetic overflow records numerical range. Asymptotic classification requires additional dynamical information.

# D KOOPMAN AUTOENCODER DETAILS

The training set consists of adjacent embedded state pairs $\mathcal{D}_{\text{train}} = \{(\mathbf{u}_j, \mathbf{v}_j)\}_{j=1}^{N}$, with $\mathbf{u}_j, \mathbf{v}_j \in \mathbb{R}^3$. The predictor contains an encoder $E_{\boldsymbol{\theta}_E} : \mathbb{R}^3 \to \mathbb{R}^q$, a decoder $D_{\boldsymbol{\theta}_D} : \mathbb{R}^q \to \mathbb{R}^3$, and a bias-free latent matrix $\mathbf{K} \in \mathbb{R}^{q \times q}$. For each training pair,

$$\mathbf{z}_j = E_{\boldsymbol{\theta}_E}(\mathbf{u}_j), \qquad \widehat{\mathbf{u}}_j = D_{\boldsymbol{\theta}_D}(\mathbf{z}_j), \qquad \widehat{\mathbf{z}}_j^{+} = \mathbf{K}\mathbf{z}_j, \qquad \widehat{\mathbf{v}}_j = D_{\boldsymbol{\theta}_D}(\widehat{\mathbf{z}}_j^{+}). \tag{27}$$

The implementation averages squared error over examples and coordinates:

$$\begin{aligned}
\mathcal{L}_{\text{rec}} &= \frac{1}{3N} \sum_{j=1}^{N} \left\| D_{\boldsymbol{\theta}_D}(E_{\boldsymbol{\theta}_E}(\mathbf{u}_j)) - \mathbf{u}_j \right\|_2^2, \\
\mathcal{L}_{\text{pred}} &= \frac{1}{3N} \sum_{j=1}^{N} \left\| D_{\boldsymbol{\theta}_D}(\mathbf{K} E_{\boldsymbol{\theta}_E}(\mathbf{u}_j)) - \mathbf{v}_j \right\|_2^2, \\
\mathcal{L}_{\text{lin}} &= \frac{1}{qN} \sum_{j=1}^{N} \left\| \mathbf{K} E_{\boldsymbol{\theta}_E}(\mathbf{u}_j) - E_{\boldsymbol{\theta}_E}(\mathbf{v}_j) \right\|_2^2.
\end{aligned} \tag{28}$$

The training objective is

$$\underset{\boldsymbol{\theta}_E, \boldsymbol{\theta}_D, \mathbf{K}}{\text{minimize}} \quad \mathcal{L} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{pred}} + \mathcal{L}_{\text{lin}}. \tag{29}$$

The hidden layers use $\tanh$ activations. Minibatch Adam produces the fitted encoder, decoder, and latent matrix used in Equation (8). We set $\operatorname{atan2}(0, 0) = 0$. For the mushroom billiard, the decoded angle is converted to arc length and momentum is clipped to $[-1, 1]$.

# E ORACLE AND SIMULATOR DETAILS

## E.1 MUSHROOM BILLIARD

The billiard boundary consists of a semicircular cap, horizontal shoulders, vertical stem walls, and a stem floor. Its nondimensional geometric constants are

$$R = 1, \qquad h_{\text{stem}} = 1, \qquad r = \cos\left(\frac{5 + \sqrt{2}}{23}\pi\right). \tag{30}$$

A boundary state is represented by arc length $s$ and Birkhoff momentum $p$, the tangential component of the unit outgoing velocity. The simulator computes the next positive-time intersection with every admissible boundary piece, selects the earliest intersection, and applies specular reflection.

For point $\mathbf{q}$ and outgoing velocity $\mathbf{v}$, angular momentum about the cap center is

$$L(\mathbf{q}, \mathbf{v}) = q_1 v_2 - q_2 v_1. \tag{31}$$

A regular state lies on the cap or a shoulder and satisfies $|L| \geq r$. Stem-wall and stem-floor states are assigned to the sea. This piece-aware definition applies to the initial and decoded states, thereby restricting the cap criterion to its geometric domain.

The relative-depth parameter maps to target momentum according to

$$\delta \in [0, 1], \qquad |L| = r + \delta(1 - r). \tag{32}$$

The evaluation bands are

$$\mathcal{B}_{\text{edge}} = [0.05, 0.25], \qquad \mathcal{B}_{\text{deep}} = [0.60, 0.90]. \tag{33}$$

Initial cap collisions are constructed at the sampled angular momentum and converted to $(s, p)$ before evaluation.

### E.2 STANDARD MAP

The standard map is implemented as

$$p_{t+1} = p_t + \kappa \sin x_t, \qquad x_{t+1} = (x_t + p_{t+1}) \bmod 2\pi, \qquad \kappa = 0.97. \tag{34}$$

Candidate evaluation states and the operational oracle are specified by

$$\rho_{\mathrm{eval}} = 1.4(1-\delta), \qquad \mathbf{x}_\star = (\pi, 0), \qquad m(\mathbf{x}) = \mathbb{I}[|p_t| < 2 \text{ for } t = 0, \ldots, 5000]\,. \tag{35}$$

The learned rollout, both endpoint controls, persistence control, and true orbit are evaluated through this oracle at every supplied state in their evaluation sequences. Here $\delta$ defines radial sampling. Reusing the intervals in Equation (33) yields an outer radial stratum and an inner radial stratum. The radial coordinate defines sampling around the elliptic fixed point, and the finite-time oracle supplies membership.

The training distribution combines local resolution around the elliptic fixed point with global phase-space coverage:

$$\begin{aligned} \mu_{\mathrm{train}} &= 0.4\,\mathcal{U}_{\mathrm{local}} + 0.6\,\mathcal{U}_{\mathrm{global}}, \\ \vartheta_{\mathrm{local}} &\sim \mathcal{U}[0, 2\pi], \qquad \rho_{\mathrm{local}} \sim \mathcal{U}[0, 1.4], \\ \mathcal{U}_{\mathrm{global}} &= \mathcal{U}([0, 2\pi] \times [-2.5, 2.5])\,. \end{aligned} \tag{36}$$

Evaluation starts and membership labels are excluded from the training objective.

### E.3 SMOOTH HAMILTONIAN BAND

The band boundary, transition width, and amplitude scale are

$$p_0 = 0.5, \qquad w = 0.35, \qquad a = 0.5. \tag{37}$$

Define

$$f(t) = \begin{cases} 0, & t \le 0, \\ \exp(-1/t), & t > 0, \end{cases} \qquad S(t) = \begin{cases} 0, & t \le 0, \\ \dfrac{f(t)}{f(t) + f(1-t)}, & 0 < t < 1, \\ 1, & t \ge 1. \end{cases} \tag{38}$$

The perturbation amplitude is

$$A(p) = aS\left(\frac{|p| - p_0}{w}\right), \qquad \mathcal{H}(x, p) = \frac{p^2}{2} + A(p)\cos x. \tag{39}$$

The implemented vector field is

$$\dot{x} = p + A'(p)\cos x, \qquad \dot{p} = A(p)\sin x. \tag{40}$$

Inside $|p| \le p_0$, both $A$ and its derivative vanish, so

$$\dot{x} = p, \qquad \dot{p} = 0. \tag{41}$$

Hence the analytical flow leaves $|p| \le p_0$ invariant. The evaluated starts lie strictly inside this band; the released reference-control record confirms that the stated midpoint implementation remains inside over the reported horizon.

Let $\mathbf{q} = (x, p)^\top$ and let $\mathbf{f}$ denote the vector field in Equation (40). The time-one map uses the implicit-midpoint update

$$\mathbf{q}_{k+1} = \mathbf{q}_k + h\mathbf{f}\left(\frac{\mathbf{q}_k + \mathbf{q}_{k+1}}{2}\right), \qquad k = 0, \ldots, N_{\mathrm{sub}} - 1. \tag{42}$$

The numerical constants are

$$N_{\mathrm{sub}} = 12, \qquad h = N_{\mathrm{sub}}^{-1}, \qquad N_{\mathrm{fp}} = 60, \qquad \varepsilon_{\mathrm{fp}} = 10^{-10}, \qquad \Delta p = 10^{-7}. \tag{43}$$

Fixed-point iteration uses $N_{\mathrm{fp}}$ and $\varepsilon_{\mathrm{fp}}$; the centered difference for $A'(p)$ uses $\Delta p$.

# F Model, Training, and Evaluation Protocol

Every primary model uses the Koopman autoencoder in Equation (27) and the rollout in Equation (8) (Lusch et al., 2018). The shared architecture is

$$\mathrm{widths}(E) = (3, 64, 64, 16), \qquad \mathrm{widths}(D) = (16, 64, 64, 3), \qquad \mathbf{K} \in \mathbb{R}^{16\times 16}, \tag{44}$$

with $\tanh$ on the first two hidden layers of each network and zero bias in $\mathbf{K}$. The resulting model has 11,155 trainable parameters. The embedding uses

$$(\theta, \sigma) = \begin{cases} (2\pi s/P, 1), & \text{mushroom billiard,} \\ (x, 2.5), & \text{standard map,} \\ (x, 2), & \text{smooth Hamiltonian flow.} \end{cases} \tag{45}$$

The decoder restores $\theta$ with $\mathrm{atan2}$, and mushroom momentum is clipped to its Birkhoff range. Training minimizes Equation (29) for 300 epochs in float32 on CPU with Adam: learning rate $10^{-3}$, $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$, zero weight decay, and batch size 512.

Each fresh primary data set contains 2,000 trajectories with 60 retained states, yielding 118,000 adjacent pairs. Initialization seeds are $s = 0, \ldots, 4$; the paired mushroom and standard-map data seeds are $42 + s$ and $7 + s$, respectively. The smooth-flow data set contains 2,000 trajectories with 60 transitions, yielding 120,000 pairs. Its generator, model, and evaluation seeds are 2026, 0, and 1,000. Smooth-flow starts satisfy $x \sim \mathcal{U}[0, 2\pi]$ and $|p| \sim \mathcal{U}[0, 1.6]$, with equiprobable momentum sign. Every primary evaluation uses seed 1,000, 40 starts per stratum, and 200 decoded updates. The endpoint controls follow Equations (2) and (19).

## F.1 Held-out matched-departure persistence panel

The held-out mushroom panel uses a specified crossed design that is separate from the primary fresh panel. For each training-data realization, all initialization realizations train on the same simulated transition set. Diagnostic starts and held-out starts are generated separately from the exact mushroom island using the depth bands in Equation (33). The scalar reference simulator remains confined on every selected start before learned-model evaluation.

The diagnostic statistic is the terminal-exclusion rate at its fixed horizon. The held-out target is late-window outside occupancy on the disjoint start cell. The matching predicate uses diagnostic any-time departure exclusively. The primary statistic is the model-level Spearman rank correlation between terminal exclusion and the held-out target. Its interval independently resamples the data and initialization factors and reconstructs the corresponding crossed grid. It is conditional on the diagnostic and held-out start panels generated by the fixed evaluation seeds. The protocol attempts 2,000 resamples; percentile intervals use the finite correlation coefficients, including 1,985 for the primary statistic. Each diagnostic draws its own resample stream, seeded by $20260826$ plus its index in the reported list. Because some resamples have constant ranks, each interval reports its finite-draw count.

The saved protocol crosses data seeds $42, \ldots, 47$ with initialization seeds $0, \ldots, 4$. Each training cell uses 2,000 generic billiard trajectories of 60 states, 300 epochs, and batch size 512. The diagnostic cell uses seed 17,291, 200 starts per depth band, and horizon 200. The held-out cell uses seed 31,415, 200 starts per depth band, and horizon 2,000; its target averages outside occupancy after update 1,000. Matching requires diagnostic any-time departure of at least $0.975$. The declared support rule requires at least 12 matched models and a positive lower percentile bound for the primary association. Exact mushroom trajectories remain in the regular island on both cells.

**Paired diagnostic comparison.** A secondary reanalysis uses common crossed-factor resamples for terminal exclusion and whole-diagnostic-trajectory occupancy. The statistic is the difference of their Spearman correlations with the same held-out target. Both coefficients must be finite for a draw to enter the percentile interval. The output records the seed, attempted and retained draws, and a hash of the unchanged source cohort. This comparison is conditional on the same fixed start panels.

**Permutation reference distributions.** We enumerate data-label permutations shared across all initialization columns, and separately initialization-label permutations shared across all data rows.

These are one-sided conditional references under exchangeability of the respective target-grid labels relative to the diagnostic grid. Each preserves the grouping of the other factor. The crossed bootstrap provides the primary uncertainty summary, and the permutation references condition on the stated exchangeability model. Table 5 gives the exact counts and results.

Table 4: **Model-level association summaries.** The intervals use separate crossed-factor bootstrap streams and fixed evaluation-start panels. The paired comparison is reported separately.

| Diagnostic statistic | association with held-out late occupancy | crossed-factor interval |
|---|---|---|
| terminal exclusion | $\rho = 0.905$ | $[0.719,\ 1.000]$ |
| diagnostic outside occupancy | $\rho = 0.758$ | $[0.440,\ 0.901]$ |
| mean first-departure index (censored) | $\rho = -0.226$ | $[-0.703,\ 0.283]$ |
| one-step RMSE | $\rho = 0.259$ | $[-0.205,\ 0.663]$ |
| endpoint spread | $\rho = -0.385$ | $[-0.811,\ 0.397]$ |
| any-time departure | constant | – |

Terminal re-entry is $1 - \widehat{R}_E$ in this cohort, so its rank association is exactly the negative of the first row. Within-data model pairs: 28 informative; 32 tied.

Table 5: **Paired comparison and factor-preserving permutation references.** The correlation difference uses common crossed-factor resamples. Exact permutation references relabel one factor across the entire grid.

| Quantity | result |
|---|---|
| Paired correlation difference | 0.148 |
| Crossed-factor percentile interval | $[0.011,\ 0.407]$ |
| Finite / attempted paired draws | 9,920 / 10,000 |
| Shared data-label permutations | $p = 0.0014$ (720 permutations) |
| Shared initialization-label permutations | $p = 0.0083$ (120 permutations) |

Table 6: **Models tied for the minimum diagnostic terminal rate.** Endpoint spread is measured across held-out starts at the final horizon. A dash marks arithmetic overflow.

| Data | initialization | late occupancy | endpoint spread | behavior |
|---|---|---|---|---|
| 43 | 3 | 0.865 | – | overflow |
| 45 | 2 | 0.000 | $3.40 \times 10^{-6}$ | concentrated |
| 46 | 0 | 0.000 | $2.98 \times 10^{-6}$ | concentrated |

Table 7: **Sensitivity analysis for the crossed-seed cohort.** Concentrated denotes held-out endpoint dispersion below $10^{-5}$; overflow denotes undefined dispersion; spread covers the remaining models. Table 4 gives the primary crossed-factor interval.

| Check | retained-record result |
|---|---|
| Crossed design | 6 data $\times$ 5 initialization realizations |
| Leave-one-data-realization-out | $\rho \in [0.866,\ 0.932]$ |
| Data-realization means | $\rho = 0.829$ |
| Joint-ceiling excluded | $\rho = 0.593$ (10 models) |
| Joint ceiling | 20 models |
| Endpoint behavior | 24 concentrated, 5 spread, 1 overflow |
| Association by endpoint group | $\rho = 0.806$ (25 concentrated or overflow); $\rho = -0.400$ (5 spread) |
| Overflow-target rank sweep | $\rho \in [0.811,\ 0.918]$ |
| Within-data pairs | 26/28 directionally concordant; 32 tied |

**Overflow and rank sensitivity.** One grid model, at data realization 43 and initialization 3, reaches arithmetic overflow between the diagnostic and held-out horizons, so its endpoint dispersion is

recorded as null. The scoring convention in Section 3 maps these states to departure. Because Spearman association depends on ranks, the sensitivity analysis places this target at every admissible rank among the remaining target values. The association remains positive throughout the sweep (Table 7).

The released cohort retains every model-level diagnostic and held-out record. Its deterministic inference workflow regenerates the paired comparison and the shared-factor references in a new output location. The main cohort retains aggregate diagnostics for every model; the two distributed checkpoints belong to the separate fixed-model panel.

**Selection policy and diagnostic eligibility screens.** The reported tie mean is the expected outcome of uniform random selection from all exact diagnostic minimizers, using absolute numerical tolerance $10^{-12}$. A realized selected model need not achieve that mean. Table 3 exposes the worst endpoint tie alongside the mean and all existing diagnostic screens. The accuracy pools retain the lowest-RMSE fractions $0.25, 0.50, 0.75, 1.00$, rounded up, with seed order resolving ties at the pool boundary. These fractions and the spread threshold already occur in the saved secondary analysis; none was chosen using the revised paper's held-out comparison. Every regret is relative to the best held-out target in the same retained pool, so regrets from different pools do not share a common benchmark minimum.

No confidence interval or prospective-validation claim is attached to the additional point comparisons in Table 3. The revision also reports a descriptive near-tie check at tolerances of one, five, and ten endpoint labels out of the 400 diagnostic starts. These tolerances all retain the original three minimizers; this does not measure new-panel sampling variation.

The spread screen selects model $(45, 2)$, which later has endpoint spread $3.40 \times 10^{-6}$. Thus this diagnostic-only screen removes concentration at $H = 200$ but does not certify dispersion at $H = 2000$. In the full cohort, 24 models have concentrated held-out endpoints, one overflows, and five remain spread under the stated threshold. The five-model association $\rho = -0.400$ is descriptive and too small a subgroup to resolve a general relationship. These records constrain the interpretation to the regional target rather than to overall simulator quality.

**Provenance of the inferential claims.** The saved primary plan concerns endpoint–target rank association. The paired correlation contrast, selection bootstrap, diagnostic concentration filter, and accuracy screens are secondary analyses of those same records. The released selection code labels them exploratory. The additional architecture study has its own saved configuration and plan hash, but shares the diagnostic and held-out panel seeds with the original study. The present revision does not convert any of these records into a new prospective confirmation. Original 2,000-draw association intervals and the 10,000-draw secondary intervals use different recorded random streams and need not agree exactly.

**Additional architecture and window protocol.** The added study contains all 30 planned data/initialization combinations for each of the Koopman and residual MLP families. The residual update is a direct autoregressive network with two width-100 hidden layers; the Koopman encoder has widths 3–64–64–16. The parameter counts are 10,803 and 11,155, respectively. Both use the same data labels and start-panel definitions. Eligibility is diagnostic any-time departure at least $0.975$. The five ineligible residual models remain in the released record. Each terminal window ends at update 200. In the additional Koopman cohort, endpoint, whole-trajectory occupancy, and RMSE each achieve zero point regret. In the residual cohort, endpoint and whole-trajectory occupancy both have regret $0.0016625$ and choose model $(46, 0)$. Therefore the architecture extension supports the association across model families without establishing unique selection superiority of endpoint scoring. Complete model records and all window comparisons are provided by the saved-record reproduction workflow.

Fresh models couple their data and initialization seeds and share an evaluation generator. The fixed-checkpoint panel uses the distributed checkpoints with that same evaluation generator.

The checkpoint identifiers begin with `443c5c5c080fc3b9` for the mushroom model and `0c4721c9f46f1671` for the standard-map model. The package manifest stores their complete hashes.

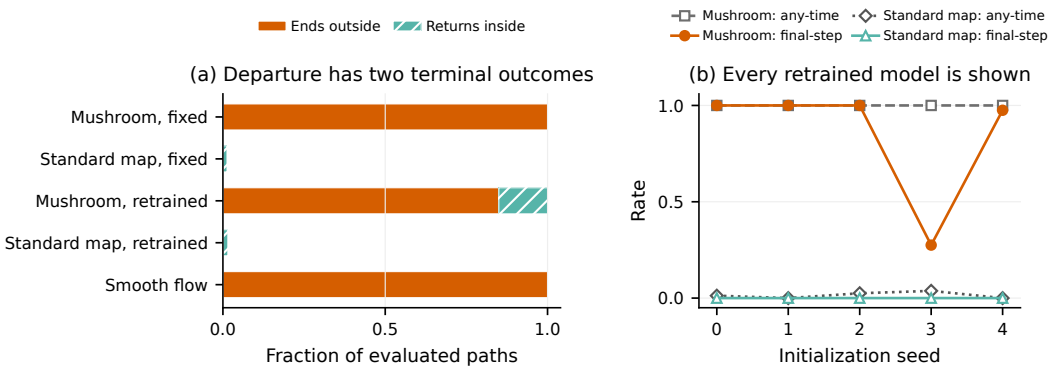



Figure 4: **Both terminal outcomes occur across systems and retraining realizations.** (a) The evaluated paths are partitioned into terminal exclusion and departure followed by return; the unfilled remainder never departs. (b) Every retrained model is shown. These stress tests use the declared system-specific oracles and do not measure model-selection gains outside the mushroom cohorts.

The smooth-flow generator samples angle and signed momentum independently. Evaluation constructs each depth band by relative distance from the invariant-band boundary.

One-step RMSE is computed in the three-coordinate embedded state space. Endpoint dispersion is the largest coordinatewise standard deviation of the learned endpoints in that representation. First departure is indexed from the shared input state. All selected primary starts are confined under their reference dynamics.

## G Supporting Decompositions

The binary labels induce four empirical event rates. For a prototype label $g$ shared within a model, system, and start stratum,

$$\widehat{R}_E = \frac{1}{n}\sum_i \mathbb{I}[d_i = e_i = 1], \qquad \widehat{R}_{\mathrm{tr}} = \frac{1}{n}\sum_i \mathbb{I}[d_i = 1, e_i = 0],$$
$$a_{\mathrm{in}\to\mathrm{out}} = \frac{1}{n}\sum_i \mathbb{I}[g = 1, e_i = 0], \quad a_{\mathrm{out}\to\mathrm{in}} = \frac{1}{n}\sum_i \mathbb{I}[g = 0, e_i = 1]. \tag{46}$$

Table 8: **Fixed-checkpoint decomposition.** Mushroom entries are near-boundary / deep; standard-map entries are outer / inner radial. Terminal re-entry is the rollout count minus the own-endpoint count. Median $h^F$ is computed among departing paths.

| System | RMSE | rollout | own endpoint | mean endpoint | re-entry | median $h^F$ |
|---|---|---|---|---|---|---|
| Mushroom | .112 | 40/40 / 40/40 | 40/40 / 40/40 | 40/40 / 40/40 | 0/40 / 0/40 | 3 / 2 |
| Standard map | .006 | 1/40 / 0/40 | 0/40 / 0/40 | 0/40 / 0/40 | 1/40 / 0/40 | 6 / – |

Table 9: **Fresh-training decomposition.** Counts aggregate the fresh models. Mushroom entries are near-boundary / deep; standard-map entries are outer / inner radial.

| System | models | rollout | own endpoint | mean endpoint | re-entry |
|---|---|---|---|---|---|
| Mushroom | 5 | 200/200 / 200/200 | 168/200 / 172/200 | 200/200 / 200/200 | 32/200 / 28/200 |
| Standard map | 5 | 6/200 / 0/200 | 0/200 / 0/200 | 0/200 / 0/200 | 6/200 / 0/200 |

Table 10: **Per-seed terminal decomposition.** Each row pools two system-specific strata; spread reports the strata separately.

| System | seed | rollout | terminal | re-entry | mean | spread (stratum 1 / 2) |
|---|---|---|---|---|---|---|
| Mushroom | 0 | 80 | 80 | 0 | 80 | 0.000 / 0.000 |
| Mushroom | 1 | 80 | 80 | 0 | 80 | 0.046 / 0.037 |
| Mushroom | 2 | 80 | 80 | 0 | 80 | 0.000 / 0.000 |
| Mushroom | 3 | 80 | 22 | 58 | 80 | 0.955 / 0.968 |
| Mushroom | 4 | 80 | 78 | 2 | 80 | 0.836 / 0.554 |
| Standard map | 0 | 1 | 0 | 1 | 0 | 0.002 / 0.002 |
| Standard map | 1 | 0 | 0 | 0 | 0 | 0.002 / 0.001 |
| Standard map | 2 | 2 | 0 | 2 | 0 | 0.017 / 0.005 |
| Standard map | 3 | 3 | 0 | 3 | 0 | 0.008 / 0.001 |
| Standard map | 4 | 0 | 0 | 0 | 0 | 0.007 / 0.002 |

Table 11: **Smooth-flow decomposition.** The columns separate terminal status, terminal re-entry, prototype agreement, endpoint spread, and event timing.

| Band | rollout | own endpoint | mean endpoint | re-entry | own/mean | spread | median $h^F$ |
|---|---|---|---|---|---|---|---|
| Near-edge | 40/40 | 40/40 | 40/40 | 0/40 | 40/40 | 1.5895 | 11 |
| Deep | 40/40 | 40/40 | 0/40 | 0/40 | 0/40 | 1.6326 | 14.5 |

## H Per-Seed Fresh-Training Results

Table 12 contains every fresh primary run and is generated directly from the frozen experiment record.

Table 12: **Fresh models by seed.** Paired entries denote near-boundary / deep for the mushroom and outer / inner radial for the standard map. Rollout/own and own/mean are agreement counts; spread is endpoint dispersion in the periodic embedding.

| System | seed | RMSE | rollout | own endpoint | mean endpoint | rollout/own | own/mean | endpoint spread |
|---|---|---|---|---|---|---|---|---|
| Mushroom | 0 | 0.1121 | 40/40 / 40/40 | 40/40 / 40/40 | 40/40 / 40/40 | 40/40 / 40/40 | 40/40 / 40/40 | 0.0000 / 0.0000 |
| Mushroom | 1 | 0.1231 | 40/40 / 40/40 | 40/40 / 40/40 | 40/40 / 40/40 | 40/40 / 40/40 | 40/40 / 40/40 | 0.0457 / 0.0366 |
| Mushroom | 2 | 0.1135 | 40/40 / 40/40 | 40/40 / 40/40 | 40/40 / 40/40 | 40/40 / 40/40 | 40/40 / 40/40 | 0.0003 / 0.0003 |
| Mushroom | 3 | 0.1083 | 40/40 / 40/40 | 10/40 / 12/40 | 40/40 / 40/40 | 10/40 / 12/40 | 10/40 / 12/40 | 0.9545 / 0.9676 |
| Mushroom | 4 | 0.1037 | 40/40 / 40/40 | 38/40 / 40/40 | 40/40 / 40/40 | 38/40 / 40/40 | 38/40 / 40/40 | 0.8358 / 0.5539 |
| Standard map | 0 | 0.0036 | 1/40 / 0/40 | 0/40 / 0/40 | 0/40 / 0/40 | 39/40 / 40/40 | 40/40 / 40/40 | 0.0025 / 0.0017 |
| Standard map | 1 | 0.0109 | 0/40 / 0/40 | 0/40 / 0/40 | 0/40 / 0/40 | 40/40 / 40/40 | 40/40 / 40/40 | 0.0021 / 0.0009 |
| Standard map | 2 | 0.0090 | 2/40 / 0/40 | 0/40 / 0/40 | 0/40 / 0/40 | 38/40 / 40/40 | 40/40 / 40/40 | 0.0165 / 0.0048 |
| Standard map | 3 | 0.0142 | 3/40 / 0/40 | 0/40 / 0/40 | 0/40 / 0/40 | 37/40 / 40/40 | 40/40 / 40/40 | 0.0083 / 0.0009 |
| Standard map | 4 | 0.0038 | 0/40 / 0/40 | 0/40 / 0/40 | 0/40 / 0/40 | 40/40 / 40/40 | 40/40 / 40/40 | 0.0072 / 0.0024 |

Across these mushroom runs, endpoint spread ranges from a nearly point-valued endpoint cloud to a broad cloud in the periodic representation. The broadest clouds occur in rows with terminal re-entry; Table 12 gives the stratum-level counts. Every mushroom prototype departs. Standard-map departures occur in the outer radial stratum and end inside; their individual and mean endpoints remain confined in every cell.

## I Fixed-Checkpoint and Smooth-Flow Records

The fixed record stores one-step RMSE, endpoint location, endpoint dispersion, departure count, and median first-departure time for the learned rollout, both endpoint controls, persistence, and true dynamics. The mushroom decision vectors are uniformly one for all three learned-output arms. The standard-map endpoint vectors remain confined, while its rollout vector contains an outer-stratum departure followed by terminal re-entry.

The smooth record stores embedded-state one-step RMSE, both endpoint summaries, true and persistence controls, agreements, and endpoint dispersion. The exact reference dynamics preserve momentum inside the band. The supporting tables read this record directly and keep the local-error, departure, prototype, and spread summaries in separate columns.

# J CONTROLS AND DECISION SEMANTICS

**True dynamics.** The true-dynamics arm begins from each held-out state and applies the reference simulator. Its role is to verify that the selected starts and the stated oracle define a confined reference sequence.

**Persistence.** The persistence arm repeats the held-out initial state at every evaluation time. It verifies the membership of time zero and fixes the convention that all arms share their initial state.

**Own endpoint.** This arm repeats each learned terminal state after the shared input. Because that state appears in the learned rollout, its departure label is a subset of the rollout departure label. Their difference is the exact terminal-re-entry count.

**Mean endpoint.** This arm repeats the cellwise circular mean of the learned endpoints after the shared input. It has one representative endpoint per model, system, and start stratum. Its agreement with the own-endpoint arm measures preservation of the individual terminal labels.

The two replay paths are therefore

$$\mathbf{y}^{E}_{i,t} = \begin{cases} \mathbf{x}_i, & t = 0, \\ \mathbf{y}^{F}_{i,H}, & t = 1, \dots, H, \end{cases} \qquad \mathbf{y}^{C}_{i,t} = \begin{cases} \mathbf{x}_i, & t = 0, \\ \mathbf{c}, & t = 1, \dots, H. \end{cases} \tag{47}$$

**Path information.** First-departure time, intermediate membership sequence, path length, residuals, and reversibility remain trajectory-level quantities. Proposition 1 isolates terminal-exclusion and terminal-re-entry labels while leaving these path quantities distinct.

For any arm $A \in \{F, E, C, T, P\}$, the first-departure index is

$$h^{A}_{i} = \min\{t \in \{0, \dots, H\} : \ell(\mathbf{y}^{A}_{i,t}) = 1\}, \qquad \min \varnothing = \infty. \tag{48}$$

Local predictive accuracy is measured on $M$ held-out embedded transition pairs by

$$\mathrm{RMSE}_1 = \left[ \frac{1}{3M} \sum_{j=1}^{M} \left\| D_{\widehat{\boldsymbol{\theta}}_D}\left( \widehat{\mathbf{K}} E_{\widehat{\boldsymbol{\theta}}_E}(\mathbf{u}_j) \right) - \mathbf{v}_j \right\|_2^2 \right]^{1/2}. \tag{49}$$

Let $\mathbf{r}_i = \boldsymbol{\phi}_\sigma(\mathbf{y}^{F}_{i,H})$ and $\overline{\mathbf{r}} = n^{-1} \sum_i \mathbf{r}_i$. Endpoint dispersion is

$$s_{\mathrm{end}} = \max_{k \in \{1,2,3\}} \left[ \frac{1}{n} \sum_{i=1}^{n} (r_{i,k} - \overline{r}_k)^2 \right]^{1/2}. \tag{50}$$

## J.1 EVALUATION ALGORITHM

Algorithm 1 states the implemented sequence for one model, system, and start stratum.

**Algorithm 1** Oracle-anchored evaluation of confinement for one held-out cell

**Require:** Predictor $F$, reference map $T$, oracle $m$, horizon $H$, candidate starts
1: Construct $\mathcal{S} = \{\mathbf{x}_i\}_{i=1}^{n}$ satisfying Equation (1)
2: **for** $i = 1, \dots, n$ **do**
3: Generate $\mathbf{y}^{F}_{i,0:H}$ using Equation (8)
4: Compute $d_i$, $e_i$, and $\tau_i = d_i - e_i$ using Equation (2)
5: **end for**
6: Construct $\mathbf{c}$ using Equation (19)
7: Compute $\mathbf{g}$, $\delta(\mathbf{d}, \mathbf{e})$, and $\delta(\mathbf{e}, \mathbf{g})$
8: Evaluate the reference and persistence arms in Equation (9)
9: Record departure counts, matched agreements, $h^{A}_{i}$, $\mathrm{RMSE}_1$, the endpoint coordinates, and $s_{\mathrm{end}}$
10: **return** The cell-level evaluation results

## K Reproduction Procedure

The supplementary package contains the saved experiment records, two fixed-panel checkpoints, dependency specification, notebooks, and generators. From the package root, `python3code/reproduce_revision.py--output-rootNEW_DIRECTORY` requires every saved input and writes to a new directory. It replays the original four records, recomputes the paired analysis, reruns the main-story and additional architecture analyses, and regenerates the empirical figures and LaTeX tables, including the diagnostic-screen table. It also independently checks rank correlations, selected model identities, tie ranges, screen outcomes, and the new proof examples. The 30 primary-cohort checkpoints are not retained; saved-record replay does not rerun their training.

The legacy `code/run_all.py` covers the original four experiment records. Its explicit `--recompute` mode retrains those experiments. The additional 60-model architecture grid has a separate generator, `paper_code.window_architecture_experiment`; the README gives the full commands. The revision workflow never silently invokes training when a saved input is missing. All new outputs record relative source paths and SHA-256 hashes. Saved-record consistency is distinct from independent scientific replication or confirmation of an exploratory analysis.

`MANIFEST.sha256` binds the current manuscript, code, and evidence files. It can be checked with `python3code/package_integrity.py--verify`. The TikZ schematic and manuscript are compiled from their LaTeX sources; the statistical figures also have color and grayscale exports. The revision records document the source snapshot, numerical and proof checks, final page boundaries, and inspection of the rendered PDF.

**Ethics and broader impact.** The experiments use simulated dynamical systems. The contribution is an evaluation protocol for learned simulators.