Active FractalRabbit: A Synthetic Benchmark for Belief Filtering Under Sparse Waypoint Observations

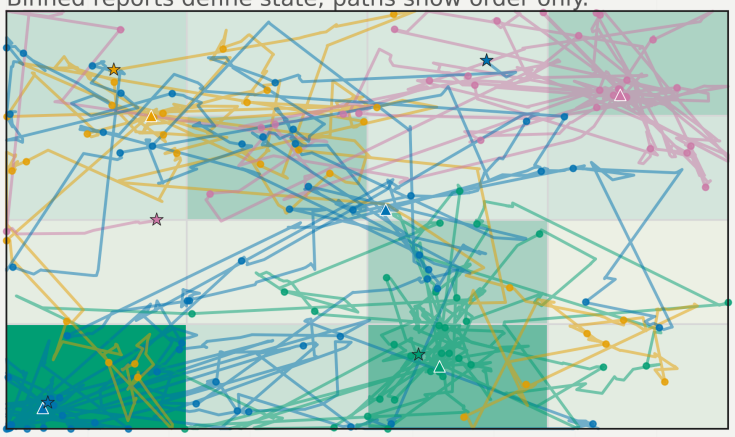
Daniel Ari Friedman
ORCID: 0000-0001-6232-9096
DOI: 10.5281/zenodo.21330636
July 12, 2026

Active FractalRabbit

2400 reports
5 travellers

Synthetic benchmark for belief filtering under sparse waypoint observations
Primary mechanism: soft marginalization under noisy emissions; AIF diagnostics remain regime-qualified.

Dominant synthetic waypoint lattice
Binned reports define state; paths show order only.



Where each model family earns its keep
The observation regime decides; there is no global winner.

CLEAN / DIRECT
Point estimates suffice
Persistence + Markov — belief adds no general gain

NOISY / HIDDEN
Preserve uncertainty
Soft · HMM · particle · AIF — paired tie vs strongest non-AIF

SWITCHING / GAPPED
Adapt or stay simple
Online base rate + controls — fixed AIF does not promote

EVIDENCE BOUNDARY
Synthetic gates complete · promotion remains blocked

Artifact-bound surface

reports	travellers	latent cells	observed loss	withheld mass	pymdp delta
2400	**5**	**4x4 / 16**	**5.719**	**0.069**	**4.0e-07**

Canonical synthetic artifacts only • no global SOTA, empirical mobility, or operational claim

Contents

Abstract

Sparse waypoint analysis is privacy-sensitive: it must separate movement from irregular reporting, missingness, spatial coarsening, and corruption while preserving uncertainty about hidden location. Active FractalRabbit provides a controlled, artifact-bound benchmark whose headline lane uses a deterministic project-local synthetic FractalRabbit-format fixture; a separately retained lane exercises pinned open-source software from the National Security Agency as an independent simulator surface. The benchmark converts sporadic reports into categorical evidence, fits explicit hidden-state generative models, and compares transparent temporal, Markov, sequence, state-space, neural, latent-state, and active inference predictors under matched information sets. Under noisy partial-observability, Active Inference is the lowest-loss implemented predictor: it clearly leads point-estimate and raw-observation families and sits in a statistical tie with the strongest non-AIF belief-preserving comparator. The shared mechanism is soft Bayesian marginalization, which preserves probability across plausible cells instead of committing early to one state. Point estimates suffice for clean observations, an online base-rate predictor leads under regime switching, transparent temporal and disclosed kinematic controls anchor sparse reporting gaps, and withholding location sharply limits specific-cell recovery from metadata. The partially observable Markov decision process (POMDP) formulation also exposes variational and expected-free-energy diagnostics for belief, minimization, and integrity. These results establish a regime-specific synthetic model map and a reproducible evidence chain. The present contract covers synthetic software behavior; separate evidence protocols govern privacy and empirical evaluation. Code, fixtures, manuscript source, and the release manifest are public at github.com/ActiveInferenceInstitute/active_fractal_rabbit.

1 Introduction: Sparse Waypoints Need Belief Filters and Claim Guards

1.1 Sparse Waypoint Evidence Is Fragmentary, Not a Track

Sparse waypoint evidence is a privacy-sensitive analytic problem because the raw material is incomplete, irregular, and easy to overinterpret. Mobility traces can remain highly identifying, can be matched across datasets, and cannot be treated as safe merely because they are sparse or synthetically represented [de Montjoye et al., 2013, Kondor et al., 2020, Buchholz et al., 2024]. A device may report only when an application wakes, a vehicle may emit a location only at sparse event boundaries, a consented field study may retain only coarse checkpoints, or an analyst may intentionally minimize direct location while keeping timing and metadata. In each case the observed record is not a track. It is a sequence of fragments shaped by movement, reporting cadence, missingness, spatial binning, and possible corruption. A few reports can therefore look like a route, a reporting artifact, a routine break, or an integrity failure depending on which uncertainty model and which null comparison the analyst brings to the data.

The practical challenge is to make those alternatives separable. A sparse waypoint system should be able to say when a location observation is genuinely informative, when timing or burstiness is only a reporting-process cue, when a method is quietly using oracle-like information, when a minimization choice removes too much utility, and when an apparent anomaly is just a corruption or null-model failure. The modeling problem is therefore not merely to draw lines between reports or maximize a next-cell score. It is to expose the evidence chain from raw emissions to categorical observations, hidden-state beliefs, predictive losses, minimization tradeoffs, and claim boundaries.

This manuscript uses synthetic data because that evidence chain needs controlled ground truth before it should be tried on real traces. Transparent nulls and base-rate models establish what occupancy alone predicts. Persistence and Markov baselines measure the value of previous-cell information and short memory (sec. 14.2). Hidden Markov, particle, and state-space models infer an unobserved cell sequence through a noisy observation channel. Sequence and neural comparators test whether richer predictors improve prequential next-cell loss under the same anti-leakage contract [Dawid, 1984, Gneiting and Raftery, 2007]. Partially observable Markov decision process (POMDP) and active-inference models make beliefs, observations, preferences, policy evaluation, and uncertainty diagnostics explicit. The comparison therefore places active inference alongside HMM, particle, and state-space filters as a first-order uncertainty-aware model family [Kaelbling et al., 1998, Rabiner, 1989, Arulampalam et al., 2002, Doucet et al., 2001, Patterson et al., 2008, 2017]. This shared arena reveals which information regime rewards each family and measures minimization choices before empirical use.

A longer historical vocabulary helps only if it is kept in its lane. Early work on chance, conjecture, inverse probability, utility under risk, perception, induction, experimental caution, population records, aggregate statistics, legal intrusion categories, inspection, and enumeration provides context for why evidence, uncertainty, observation, and governance have to be named carefully [Huygens, 1657, Bernoulli, 1713, de Moivre, 1718, Bernoulli, 1738, Bayes and Price, 1763, Laplace, 1774, Berkeley, 1709, Locke, 1690, Hume, 1748, Kant, 1781, Boyle, 1661, Newton, 1729, Graunt, 1662, Halley, 1693, Sinclair, 1791, Blackstone, 1765, Bentham, 1791, United States Census Bureau, 1790]. These sources are historical anchors, not technical ancestors of this implementation: they do not anticipate POMDP tensors, active inference, modern information privacy, real-mobility validation, or surveillance readiness.

> **Scope.** The evidence covers synthetic software behavior under declared generator, emission, split, and scoring contracts. Minimization and linkage quantities are synthetic risk proxies; differential privacy, legal compliance, empirical human-mobility validity, and operational use each require their own evaluation protocol [Dwork, 2006, European Parliament and Council, 2016, Buchholz et al., 2024]. The many model gates remain exploratory, while the unanimous partial-observability construction carries its separate mechanism-level replication contract. Real-trace evaluation follows sec. 6.5.

That boundary matters because sparse waypoint analysis now sits near real commercial location-data and connected-vehicle governance debates. Commercially available information policy materials and location-data enforcement actions make the policy relevance visible, while surveillance-theory work explains why recombined traces can acquire meaning beyond any one dataset [Haggerty and Ericson, 2000, Office of the Director of National Intelligence, 2022, 2024, Federal Trade Commission, 2024, 2026]. This manuscript uses those sources only to motivate minimization, uncertainty, and responsible interpretation; they do not validate the technical benchmark or support operational readiness.

Synthetic mobility generation is useful for that arena because it gives the analyst controlled ground truth without importing the consent, provenance, and re-identification risks that real traces carry [Barbosa et al., 2018, Kulkarni et al., 2018, Dirmeier et al., 2024, Wei et al., 2021]. But synthetic data are not automatically useful or private just because their summary statistics resemble a target domain. Synthetic urban-mobility reviews emphasize practical applicability limits, while recent synthetic-trajectory work emphasizes task-based evaluation and separate privacy analysis: the generated trace must support the specific inference question being tested, downstream utility can fail in real-life tasks even when summaries look plausible, privacy-preserving daily-trajectory synthesis remains a separate mechanism and data-source claim, and the benchmark must say which claims remain outside the evidence [Kapp et al., 2023, Deng et al., 2025, Kapp and Mihaljevic, 2023, Ozaki et al., 2025, Buchholz et al., 2024].

1.2 Research Questions and Evidence Boundary

The benchmark is organized around three questions. First, when does preserving a distribution over hidden cells improve prediction relative to committing to a point estimate? Second, where does the active-inference implementation sit after the locked split, transition source, observation channel, and information set are matched to the strongest fair baselines? Third, how much recoverability remains when location is withheld or coarsened for minimization?

The answers form a regime map. Soft marginalization improves prediction when a noisy emission hides the current cell. Within that regime, active inference ranks first among the implemented predictors, clearly ahead of models that commit to a point estimate or raw observation and statistically tied with the strongest non-AIF belief filter (sec. 3.1.1.1). Clean observations favor simpler point predictors; latent switching favors online adaptation; and location minimization sharply reduces specific-cell recovery. The external simulator remains a separate reproducibility surface, and every result resolves to a generated artifact or formal contract.

The current external landscape is broader than any one generator. GPT-style mobility generation, trajectory-LLM survey taxonomies, reversible trajectory-to-CNN transforms, large synthetic mobility datasets, and LLM-guided diffusion generators across cities all define future comparator families or evaluation directions [Haydari et al., 2024, Xu et al., 2025, Merhi et al., 2024, Yuan et al., 2026, Liu et al., 2026a]. They do not validate FractalRabbit, and they are not scored here; their role is to keep the future baseline map honest about sequence, grid, diffusion, and language-model families the current synthetic fixture does not cover.

FractalRabbit is useful for this purpose because the public, open-source National Security Agency software produces synthetic sparse waypoint reports from a stochastic point process, a retro-preferential trajectory process, and a sporadic reporting process [Darling, 2018, National Security Agency, 2026a]. The headline

experiments use a deterministic project-local FractalRabbit-format fixture with controlled ground truth; the separately retained upstream lane exercises the pinned NSA software and preserves its provenance as an independent simulator surface. Together these lanes keep the analysis centered on missing reports, return structure, observation uncertainty, and reproducible software behavior.

That same design prevents a common overreach: simulator output is not empirical mobility validation against known regularities, scaling behavior, or predictability limits in real human trajectories [Gonzalez et al., 2008, Song et al., 2010a,b]. Those predictability limits are themselves task- and binning-dependent — estimates shift with the prediction target and temporal aggregation [Ikanovic and Mollgaard, 2016] — which is a further reason to treat simulator output as a controlled testbed, not a statement about real mobility. FractalRabbit lets this manuscript ask questions about uncertainty, minimization, and integrity without making surveillance claims; the supplement documents the pinned simulator boundary, while the main text uses it only as the setting for a bounded methodological comparison.

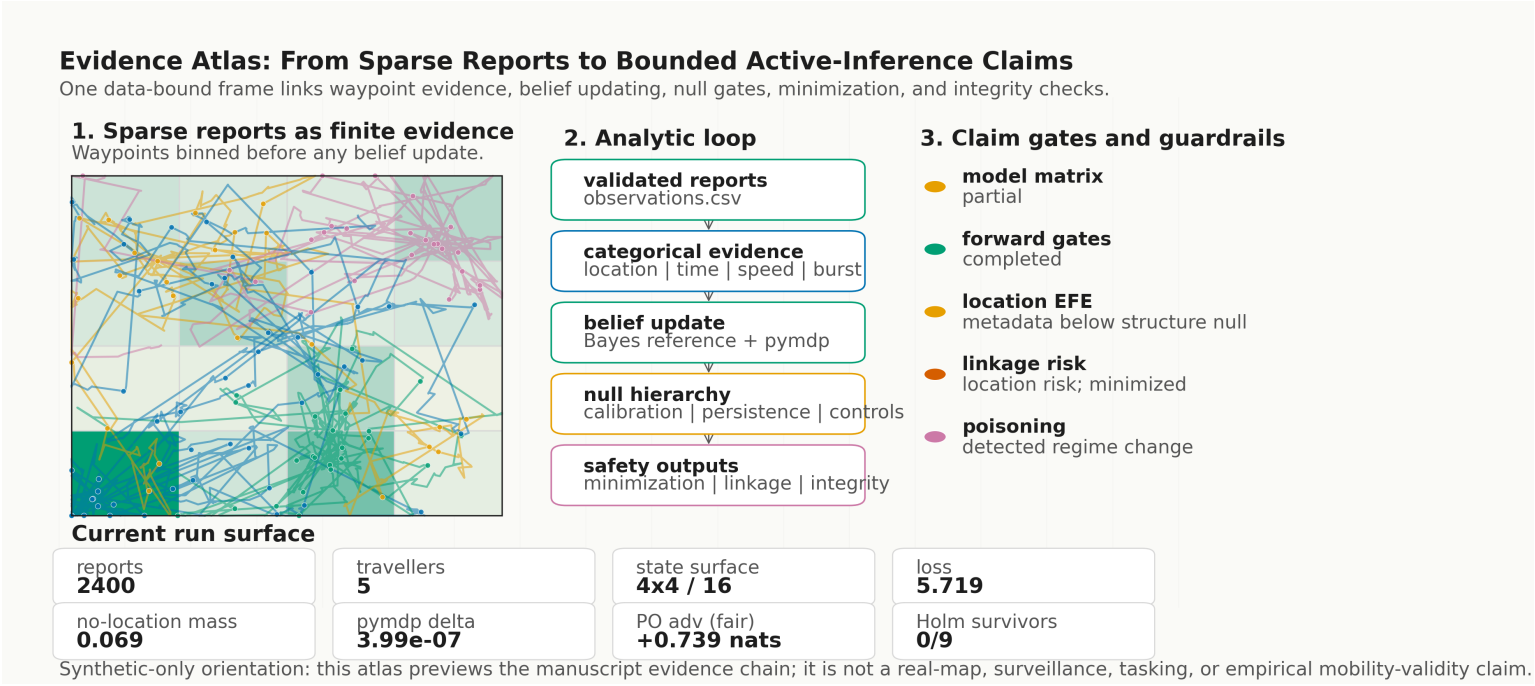


Figure 1: Sources observations.csv, discrete_features.csv, metrics.json, spatial_resolution_frontier.json, model_gate_matrix.json, forward_gates_summary.json, efe_location_frontier.json, linkage_risk_frontier.json, and poisoning_robustness.json feed one orientation atlas. It traces sparse synthetic reports through categorical evidence, belief updating, model gates, minimization, and integrity checks. The chip strip places the partial-observability positive beside the Holm multiplicity result, so the reader sees the main finding and its claim discipline together. Evidence: front-loaded orientation to the artifact-bound synthetic evidence chain and safety guardrails. Boundary: surveillance readiness, real-map inference, operational tasking, or empirical mobility validation.

fig. 1 previews the whole evidence path before the technical sections unpack it: synthetic waypoint reports become categorical evidence, categorical evidence updates hidden-state beliefs, belief traces face null gates, and the remaining outputs are bounded by minimization and integrity checks.

1.3 Variational Diagnostics Expose Belief, Preference, and Minimization

Active inference supplies a useful diagnostic vocabulary for this kind of benchmark because it separates what the analyst believes, what the analyst observes, what the analyst prefers, and which policies would reduce uncertainty or satisfy preferences. In discrete-state POMDP form, the generative model distinguishes hidden states, observations, transitions, priors, preferences, and policy evaluation within the broader free-energy and active-inference literature [Kaelbling et al., 1998, Friston, 2010, Friston et al., 2017, 2023a, Da Costa et al., 2020, Sajid et al., 2021a, Champion et al., 2024, Parr et al., 2022, Smith et al., 2022]. Navigation-specific active-inference work has used maze planning, synthetic spatial foraging, and hierarchical robot-navigation settings to study planning as inference [Kaplan and Friston, 2018, Neacsu et al., 2022, Catal et al., 2021]. Those studies are related-work context only: they do not validate this generator, this waypoint task, or any performance claim. Partial-observability work outside mobility likewise treats hidden-state prediction as an explicit modeling problem, including attention-guided state-prediction approaches that are useful comparator context but are not implemented in this benchmark [Limanjaya and Kang, 2026]. For sparse waypoints, that separation matters: a high-confidence cell posterior, a high-surprise report, a policy preference, a minimization decision, and an information-seeking action are different objects and should not be collapsed into a single "tracking quality" score. The benchmark therefore treats active inference as a structured model class to explore alongside simpler belief filters and baselines, not as a foregone performance winner.

The same scholarship also sharpens the negative space around the implementation. Expected-free-energy work contains canonical decompositions, reformulations, and links to Bayesian design and decision-theoretic objectives; those sources support careful notation, not rhetorical promotion [Millidge et al., 2021, Sajid et al., 2021b, Champion et al., 2024, Sweeney et al., 2026]. This manuscript therefore keeps the real `pymdp` expected-free-energy trace, the transparent preference cross-entropy proxy, and the fair predictive-loss gates as distinct surfaces. An EFE diagnostic can make belief, preference, and epistemic value inspectable, but it cannot substitute for a locked-split predictive win, a formal privacy guarantee, or empirical mobility validation.

Variational inference makes uncertainty and minimization inspectable within this framework. The benchmark uses active inference as an instrument panel rather than as a license for stronger claims. Variational free energy records report-level surprise under the current synthetic generative model. Expected free energy decomposes policy value into epistemic and pragmatic components, making it possible to ask whether location evidence is informative, whether metadata secretly recovers location, and whether an acquisition policy beats a blind schedule while remaining a non-operational benchmark question, not a collection recommendation. Belief trajectories show how much uncertainty remains when direct location is withheld. These diagnostics are especially useful in privacy-sensitive analysis because they surface uncertainty and minimization instead of hiding them behind a single accuracy number.

The `pymdp` implementation gives this manuscript a real active-inference lane for discrete state spaces while the transparent reference filter keeps the arithmetic auditable [Heins et al., 2022, infer-actively, 2026a,b]. The `pymdp` lane is run inside the same experiment surface as the reference lane, and its state inference is validated against the exact Bayes posterior at every step (sec. 11.1). When the two lanes diverge under policy selection, that divergence is treated as evidence about action-conditioned priors and expected-free-energy control, not as a hidden implementation win.

1.4 The Contribution Is a Bounded Evidence Chain

Active FractalRabbit: A Synthetic Benchmark for Belief Filtering Under Sparse Waypoint Observations connects these pieces into an artifact-bound methods study. It discretizes synthetic sporadic waypoints into analysis cells, constructs categorical evidence for location, time, speed, and reporting bursts, and compares

transparent nulls, Markov and persistence baselines, sequence/state-space families, and active-inference lanes. The transparent reference lane makes belief updating auditable. The `pymdp` lane adds variational free energy, policy posterior, expected-free-energy surfaces, and action-conditioned prior traces (sec. 11.1). The predictive ladder then identifies where latent-state inference improves loss, which component supplies the gain, and where simpler models remain sufficient.

The current contribution is the artifact-bound claim discipline around that comparison. Spatial-resolution analysis separates observed-location fit from no-location recovery. Calibration moves from weak uniform floors to stricter base-rate and persistence nulls in the synthetic benchmark. Gap filling is scored against a persistence oracle rather than celebrated as path reconstruction. Reporting-process gates test whether timing structure is movement information or reporting behavior. Public next-location datasets can contain substantial trajectory overlap, with performance varying sharply by overlap status [Luca et al., 2023]. Here, traveller-disjoint splits prevent one synthetic traveller's history from contributing to both parameter estimation and holdout scoring, while temporal holdouts separately test forward prediction; prequential scoring preserves prediction-before-observation order [Dawid, 1984]. These controls reduce specific leakage paths without proving generalization. Factorial `pymdp` diagnostics expose movement and synthetic reporting-mode factors without converting them into real cognitive, intent, or behavior claims. Safety-boundary artifacts then ask what location contributes to epistemic value, what minimization does to aggregate linkage risk, and whether variational free energy can flag crude input corruption. A family-wise multiplicity audit (Holm and Benjamini–Hochberg) over every directional comparison in the clustered-bootstrap family closes that discipline: none survives correction, so the one fair-by-construction positive — established outside that family by unanimous sign-stability across independent generator draws — stands alone, and the location-minimization frontier is reported together with its predictive-utility dual so the governance tradeoff is visible rather than asserted.

The contribution is a synthetic evidence chain for responsible analytic development: what each model infers on a controlled trace, which predictors lead under each information regime, which apparent gains dissolve against stronger nulls, and how minimization changes recoverability. Adjacent hidden Markov and state-space movement models provide the immediate comparison context [Patterson et al., 2008, 2017], while consented empirical validation defines the next evidence layer.

The rest of the manuscript follows that sequence. The simulator boundary is defined before the observation model; the observation model is inspected before active-inference diagnostics; the diagnostics are tested against nulls before performance language; and roadmap items become executable gates before they are promoted as claims. The central noisy-emission belief-filtering result is specified in sec. 2.6.2 and reported first in sec. 3.1.1. In the partial-observability regime, soft Bayesian marginalization beats a hard point estimate as the emission degrades: from 0 nats at a clean channel to 0.739 nats at emission flip probability 0.400 on the primary fixture draw, replicating at a mean of 0.495 nats and sign-stable across 10 independent generator draws — a synthetic software measurement that never beats the oracle-fed baseline. Its fair-comparison ladder, explicit model-selection map, and bounded Bayesian-mechanics analogy are synthesized in sec. 4.1 and sec. 4.5. The remaining handoffs are the fractal-dimension control in sec. 10.2, the measured `pymdp` distinction in sec. 11.1, the sequence and reporting-process gates in sec. 14.2, and the forward gates in sec. 6.3.

2 Methods: Artifact-First POMDP Benchmarking on Synthetic Waypoints

2.1 The Research Loop Writes Evidence Before Claims

The pipeline is organized as a closed research loop rather than a one-shot parser, following reproducible-computational-research practice that keeps inputs, scripts, intermediates, and software identifiers inspectable [Sandve et al., 2013, Wilson et al., 2017, Smith et al., 2016]. After the opening evidence atlas in fig. 1, this section narrows the same story to its artifact contract: a configured run resolves a FractalRabbit-compatible CSV, validates and sorts waypoint rows, derives categorical observations, constructs normalized POMDP tensors, filters latent spatial-state beliefs, scores inference quality, and ranks the next simulator candidate.

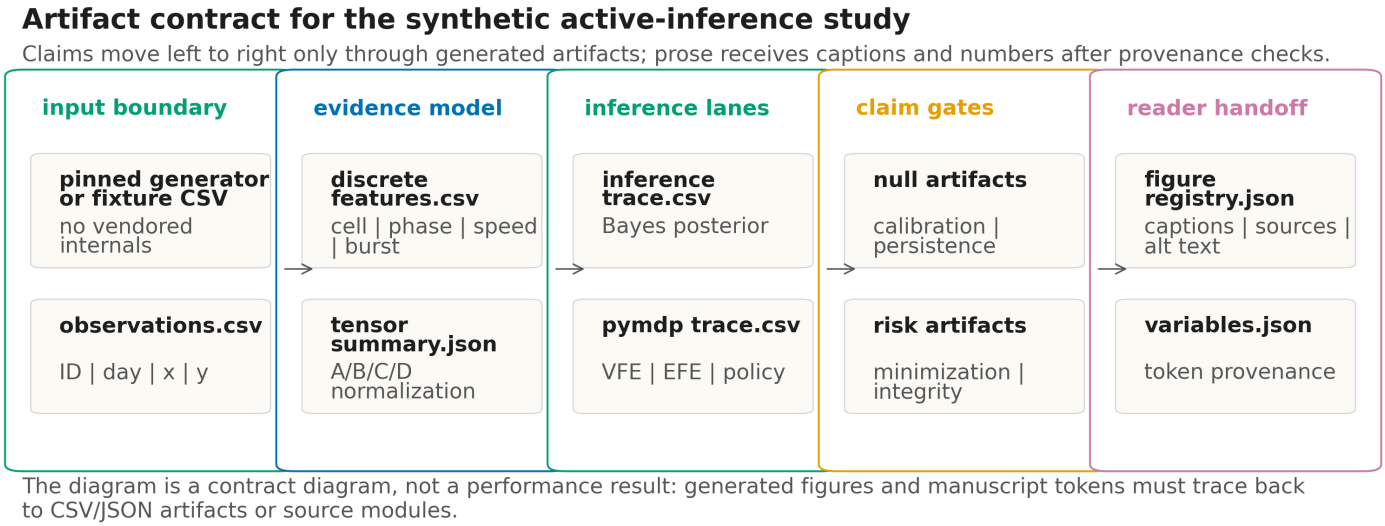


Figure 2: From the figure generator and from source pipeline modules, this panel summarizes the rendered artifact dataflow: pinned simulator or checked fixture, evidence-model artifacts, dual inference lanes, claim-gate artifacts, figure registry, and manuscript variables. Every drawn box resolves to an auditable local artifact, so the reader can trace claims without trusting prose alone. Evidence: traceability from simulator boundary through generated manuscript variables. Boundary: trust in prose without artifacts.

fig. 2 is generated from the source pipeline modules and summarizes the artifact contract as five columns: an input boundary (the pinned generator or fixture CSV, yielding only waypoint observations), an evidence model (discretized categorical features and tensor summaries), the inference lanes (the reference Bayes-filter trace and the pymdp trace), the claim gates (null and risk artifacts), and the reader handoff (the figure registry and manuscript variables). No step in this pipeline imports private simulator internals or real-world mobility labels.

2.2 The Default Configuration Defines the Tensor Surface

The default configuration uses the `grid` spatial discretizer with 4 bins per axis, 4 time-of-day bins, speed thresholds 10, 50, 150 km/day, burst rule `median_positive`, and controls `empirical_continuation`, `high_entropy_exploration`. The emitted POMDP likelihood tensors have shapes `16 x 16`, `4 x 16`, `4 x 16`, `2 x 16`, while the transition tensor has shape `16 x 16 x 2`.

2.3 Estimands, Information Sets, and Fairness

The primary estimand is the mean paired held-out next-cell log-loss difference between two predictors under the same synthetic split. Traveller is the resampling unit: reports from one traveller remain clustered, and the holdout is traveller-disjoint from training. Transition parameters are learned from the training portion only. A known observation channel is a declared synthetic condition for the partial-observability mechanism test, not information silently learned from the holdout. The oracle-fed persistence comparator is retained as an information-asymmetry ceiling and is never presented as a fair competitor.

Every fair comparison receives the same observations, candidate-cell support, transition source, scoring rule, and reporting protocol; only the inferential operation named by the comparison is changed. Point estimates commit to the most likely cell, whereas soft filters retain the full emission posterior. Active-inference diagnostics are therefore interpreted at two levels: predictive loss for the locked comparison, and variational or expected-free-energy quantities for mechanism inspection. Bootstrap intervals, independent-generator replication, and multiplicity records qualify which differences can be promoted. The optional external lane is kept separate from the fixture estimand because it changes simulator provenance and is not evidence about real mobility.

2.4 Statistical Procedures Are Pinned Before Any Verdict

Every active-inference-versus-fair directional comparison that enters the multiplicity family is defined as a paired predictive-score difference: the same held-out synthetic reports are scored by two predictors, then the loss differential is summarized rather than comparing unmatched aggregate losses [Diebold and Mariano, 1995]. The interval is traveller-clustered because reports within one traveller are not exchangeable independent rows: whole travellers are resampled with replacement, per-step losses are pooled only after the traveller draw, and the 95% interval is the percentile interval of the bootstrapped mean paired difference [Efron and Tibshirani, 1994, Field and Welsh, 2007, Cameron and Miller, 2015]. The one exception is deliberate: the partial-observability soft-versus-hard positive sits outside this bootstrap family and is instead established by unanimous sign-stability across independent generator draws (sec. 3.2.1.1). The one-sided p-value is the Davison-Hinkley achieved significance level $(1+\#\{\bar{d}_b^* \le 0\})/(B+1)$, read from the same bootstrap distribution as the interval so the two can never disagree.

The regime-search family uses B = 2000 resamples per comparison over a 40-traveller synthetic population and reports each regime at its worst (largest) p over its 2 realization seeds - a pinned, conservative choice; the cross-generator lanes use B = 2000. Multiplicity is controlled at significance level 0.050 by Holm-Bonferroni (primary; valid under arbitrary dependence) with Benjamini-Hochberg at q = 0.100 as a secondary false-discovery report [Holm, 1979, Benjamini and Hochberg, 1995]. The smallest reportable p at these replicate counts is 5.00e-04, which sits below the family-wide Bonferroni per-comparison threshold of 5.56e-03 - correction survival was achievable in principle, so the corrections are not vacuous by construction. These intervals and p-values are descriptive software measurements on the pinned synthetic generator. They are not population inference, do not estimate field deployment performance, and do not add a Diebold-Mariano test to the pipeline.

2.5 Roadmap Analyses Run as Executable Gates

The current study treats roadmap ideas as executable stages rather than prose promises. The fractal-dimension sweep in sec. 10.2, the circadian-phase prior gate in sec. 14.7, the sequence and reporting-process gates in sec. 14.2, and the benchmark matrix in sec. 11.1 all write JSON artifacts before figures and manuscript variables are generated.

This ordering is intentional: an idea enters the manuscript as a gate, then a generated artifact, and only then a bounded claim if the artifact clears its own nulls. That keeps speculative directions separate from promoted claims while still advancing the research program, and it makes claim review possible without relying on prose memory.

2.6 Formal Model: Belief Filtering Under Bounded Policy Proxies

2.6.1 The POMDP Surface Separates Hidden Cells from Observations

The active-inference analysis uses a discrete POMDP-style generative model, following the convention that hidden states generate observations and policies induce transitions [Kaelbling et al., 1998, Friston, 2010, Friston et al., 2017, 2023a, Da Costa et al., 2020, Parr et al., 2022, Smith et al., 2022, Heins et al., 2022]. In this scaffold the hidden state is the discretized spatial cell from eq. 21.

For each observation modality m, the likelihood tensor is:

$$A^{(m)}_{o,s} = P(o^{(m)} = o \mid s), \qquad \sum_o A^{(m)}_{o,s} = 1. \tag{1}$$

The configured transition controls define:

$$B^{(u)}_{s',s} = P(s_{i+1} = s' \mid s_i = s, u_i = u), \qquad \sum_{s'} B^{(u)}_{s',s} = 1. \tag{2}$$

Prior preferences and initial-state priors are represented as:

$$C^{(m)}_o = \log P_{\text{pref}}(o^{(m)} = o), \qquad D_s = P(s_0 = s). \tag{3}$$

The default modality set (location, time-of-day, speed, reporting-burst), the control set, and the emitted tensor shapes are pinned in sec. 2.2; in the default configuration the preferences are uniform, a fact whose consequences for the policy score are made explicit below. The temporal prior used at observation i is propagated explicitly:

$$D_0(s) = D_s, \qquad D_i(s) = \sum_{s'} B^{(u_{i-1})}_{s,s'} q_{i-1}(s') \quad i > 0. \tag{4}$$

The active-inference diagnostic chooses u_{i-1} as the argmax of the previous policy posterior eq. 14 (deterministic selection, not sampling), while the fixed-transition baseline filter uses `empirical_continuation` for the full sequence.

Given observed categorical outcomes o_i, the filtering update is:

$$q_i(s) = \sigma\left(\log D_i(s) + \sum_m \log A^{(m)}_{o_i^{(m)},s}\right), \tag{5}$$

where σ normalizes a vector onto the simplex. The update is evaluated only after smoothing: model construction applies positive pseudocounts and probability floors to $D_i(s)$ and to every likelihood entry that can be logged, so eq. 5 never relies on $\log 0$.

This recursion is the common Bayesian-filtering operation: the transition predicts a distribution over hidden states and the likelihood updates that distribution after the next observation. Discrete active inference expresses it within a POMDP generative model, HMM filtering supplies the corresponding finite-state recursion, and particle filtering approximates the same operation with weighted samples when exact propagation is impractical [Da Costa et al., 2020, Kaelbling et al., 1998, Rabiner, 1989, Arulampalam et al., 2002].

The one-step predictive loss uses the configured conditional-independence approximation over observation modalities:

$$P(o_i \mid s) = \prod_m A^{(m)}_{o_i^{(m)},s}. \tag{6}$$

$$\ell_i = -\log\left(\sum_s P(o_i \mid s) D_i(s)\right). \tag{7}$$

Posterior uncertainty is reported as Shannon entropy [Shannon, 1948, Cover and Thomas, 2006]:

$$H(q_i) = -\sum_s q_i(s) \log q_i(s). \tag{8}$$

2.6.2 The Partial-Observability Protocol Tests Marginalize Versus Commit

The default benchmark sets the observation equal to the discretized latent cell, so the emission is the identity and a point estimate is sufficient — there is no hidden state left to marginalize. This is why no active-inference advantage survives on the default surface, and it is the central reason the formal model is introduced *before* the headline result rather than after it.

Identity-emission collapse. If the location emission is the identity, then observing o_i sets the posterior over the latent cell to the one-hot vector $\mathbf{e}_{o_i} = \mathbf{e}_{s_i}$ after normalization. The soft predictor, hard predictor, observed-Markov predictor, and oracle-fed predictor therefore all propagate the same previous-cell vector through the same transition. Any measured advantage from recursive marginalization must disappear on that surface; a nonzero advantage requires a genuinely noisy or missing emission.

To isolate the regime in which a recursive belief over a hidden state earns its keep, let K denote the number of latent cells in this protocol and let P denote the learned transition matrix shared by every fair predictor. We observe a latent cell sequence through an explicit symmetric emission with flip probability ε,

$$P(o_i = c \mid s_i) = (1-\varepsilon)\,\mathbb{1}\{c = s_i\} + \frac{\varepsilon}{K-1}\,\mathbb{1}\{c \neq s_i\}, \tag{9}$$

and crucially the generator draws each flipped observation uniformly over the $K-1$ cells *other than* the true cell — a K-ary symmetric channel in the information-theoretic sense [Cover and Thomas, 2006] — so the data-generating channel is exactly eq. 9 and the filter is correctly specified rather than mismatched. A standard Bayes filter maintains the belief $b_i \propto P(o_i \mid \cdot) \odot (P\, b_{i-1})$ from past observations only, and three predictors form the next-cell distribution from the *same* learned transition P, the *same* emission, the *same* observations, and the *same* priors, differing only in how they use the belief:

$$\hat{p}_i^{\text{soft}} = P\, b_{i-1}, \qquad \hat{p}_i^{\text{hard}} = P\, \mathbf{e}_{\arg\max b_{i-1}}, \qquad \hat{p}_i^{\text{obs}} = P\, \mathbf{e}_{o_{i-1}}, \tag{10}$$

where $\mathbf{e}_k$ is the k-th unit vector. Soft marginalizes the full belief; hard collapses it to its maximum-a-posteriori cell before propagating; observed-Markov commits to the raw noisy observation (eq. 10). Because soft and hard are identical except for marginalize-versus-commit, no hyperparameter can be asymmetric between them — the comparison is fair by construction, which is precisely the property the retracted observation-noise frontier of sec. 6.4 lacked (an earlier version used an asymmetric noise knob that flattened one predictor's prior while leaving the other's untouched). Each predictor is scored by the prequential next-latent-cell negative log-loss on held-out travellers [Dawid, 1984, Gneiting and Raftery, 2007]; the transition is estimated from training travellers only, and predictions at step i use observations through $i-1$, so the comparison is leakage-free against the true latent target. The direction of this comparison is anchored by standard theory: the log score is strictly proper [Gneiting and Raftery, 2007], so among predictors evaluated under the model, the model's own posterior predictive is the unique minimizer of expected loss, and any competitor q pays a model-expected excess of exactly $D_{\mathrm{KL}}(\hat{p}_i^{\text{soft}} \,\|\, q) \geq 0$ — committing to the MAP cell is a lossy compression of the belief. At $\varepsilon = 0$ the belief is a one-hot vector, so marginalize and commit coincide identically; the zero-at-clean flag (yes) records an algebraic identity, not a fortunate sample. What theory does not guarantee — because the shared transition is estimated from training travellers rather than known — is the sign under the true generator at finite sample; that residual is precisely what the all-positive resampling check and the 10-draw sign-stability check establish empirically. Concretely, at each traveller's first report the belief is initialized by normalizing the emission likelihood under a uniform prior over the K cells, and that step is never scored; the shared transition is a column-stochastic matrix estimated from training-traveller transitions only, with a symmetric pseudocount of 0.050 on every entry; and the MAP collapse of the hard predictor breaks exact ties toward the lowest cell index.

Uncertainty is reported as a robustness band rather than a generalization interval. Each noise level aggregates |noise seeds| × |split seeds| paired soft-versus-hard resamplings on the same latent draw; we report the mean, the full min–max envelope, the fraction of resamplings with a positive advantage (a paired sign-test proportion), and an ordinary-least-squares slope of the mean advantage against ε with its R^2. Because this run scores a single latent draw, the spread characterizes sensitivity to the emission noise draw and the holdout split, not between-dataset variance. The summary scalars — ε-resolved advantage, clean-noise advantage, the noise slope and R^2, the worst-case and standard-deviation band at the strongest noise, and the monotonicity and all-positive flags — are bound to manuscript variables through the provenance contract of sec. 11.

2.6.3 The Policy Score Is a Proxy, Not Full Expected Free Energy

The policy-proxy diagnostic evaluates each control by propagating the current posterior through eq. 2 and projecting expected observations through eq. 1. For a candidate control u, write $p_u^{(m)} = A^{(m)} B^{(u)} q_i$ for the predicted observation distribution and $\pi^{(m)} = \sigma(C^{(m)})$ for the configured preference distribution. The implemented score is the projected preference cross-entropy proxy:

$$S(u) = \sum_m \left[H\left(p_u^{(m)}\right) + D_{\mathrm{KL}}\left(p_u^{(m)} \,\middle\|\, \pi^{(m)}\right)\right] = \sum_m H\left(p_u^{(m)}, \pi^{(m)}\right) \tag{11}$$

The entropy-plus-divergence display is algebraic bookkeeping: $H(p) + D_{\mathrm{KL}}(p\|\pi) = -\sum_o p(o) \log \pi(o)$. It is not two independently estimated active-inference terms, and it is not canonical expected free energy. For one modality, a standard one-step expected-free-energy decomposition can be written as a risk-plus-ambiguity score [Millidge et al., 2021, Parr et al., 2022]:

$$G(u) = D_{\mathrm{KL}}(p_u \,\|\, \pi) + \mathbb{E}_{q(s'|u)}[H(A(\cdot \mid s'))]\,. \tag{12}$$

The corresponding marginal-entropy identity is:

$$H(p_u) = \mathbb{E}_{q(s'|u)}[H(A(\cdot \mid s'))] + I(o; s' \mid u), \qquad S(u) = G(u) + I(o; s' \mid u). \tag{13}$$

The implemented multimodal score in eq. 11 sums this modality-wise bookkeeping across m. The proxy therefore differs from expected free energy by exactly the sign-flipped epistemic-value term in this formal diagnostic: minimizing S penalizes controls whose predicted observations are informative about the hidden state, whereas canonical expected free energy rewards them (eq. 12; eq. 13; sec. 11.1). This is one concrete reason the pymdp lane and the proxy lane can legitimately select different controls from identical tensors (sec. 11.1). We use the term proxy deliberately: this score is an implementation diagnostic over the configured tensors, not a complete derivation of expected free energy with epistemic-value terms or a normative optimal-design rule [Millidge et al., 2021, Sajid et al., 2021b, Champion et al., 2024, Sweeney et al., 2026, van Oostrum et al., 2024].

Policy posterior probabilities are then:

$$Q(u \mid o_{\leq i}) = \sigma(-S(u)). \tag{14}$$

The softmax is taken implicitly under a flat policy prior and unit precision (no temperature parameter), over a depth-one control space: controls here are single-step transition choices, not the multi-step policy sequences of the standard treatment [Friston et al., 2017, Parr et al., 2022]. One degenerate case deserves explicit statement because the default run sits in it: with uniform preferences (the default and only main-lane configuration), the preference divergence reduces to the log-cardinality of each modality minus the predictive entropy, so the two displayed components cancel and the proxy score is exactly constant across controls on any data. The policy posterior is then exactly uniform (mean policy entropy 0.693 nats over the two controls), the reported policy tie is an algebraic identity of uniform preferences rather than an empirical property of the fixture, and the per-step control choice of the diagnostic reduces to floating-point tie-breaking.

This completes the generative-model and partial-observability scaffold. How these tensors behave empirically across regimes is reported in the Results (sec. 3), and the implementation-level performance and interpretability boundary of the engine is detailed in the supplement (sec. 11.1).

3 Results: Belief Helps Only When Emissions Are Noisy

This benchmark tests active inference across many regimes and finds exactly one token-bound positive regime — partial observability under a noisy emission — while a hierarchy of honest nulls and information-asymmetry bounds marks the far larger set of regimes where it does *not* clear its fair directional gates, and where, on the identity-emission surface specifically, it provably cannot help. That contrast is the result, and the sections are ordered to state it as such, in three tiers: the single positive regime first, then the null hierarchy and bounds that surround it, then the diagnostic, gate-verification, and safety surface on which both rest. Leading with the positive regime is not promoting it over the nulls — it is naming the one regime that survived a discipline built to dissolve apparent advantages into confounds, which is exactly what that discipline exists to establish. The proportion is the point: one regime helps, many do not.

This spine keeps the stable results anchor and states the reading order for the split results modules. The detailed evidence lives in the adjacent noisy-emission and null-hierarchy result sections, with fixture diagnostics, program extensions, privacy, external-lane boundaries, and confound audits kept in the supplements. That structure keeps the main Results focused on the one positive mechanism and the null discipline surrounding it, while preserving the generated artifacts that support deeper review.

The reading rule is simple: primary evidence comes before diagnostic extension. The partial-observability comparison identifies the mechanism; the fair ladder and oracle control identify its information-set limits; and the remaining figures test robustness, minimization, reporting-process structure, switching dynamics, and future-model readiness. Read together, the results do not rank one universal winner: point estimates suffice on clean observations, belief-preserving families lead as a cluster under noisy hidden states, adaptive baselines regain the lead under regime switching, and metadata does not replace withheld location. Those extensions sharpen what the benchmark can say, but none upgrades a regime-qualified synthetic finding into blanket active-inference superiority.

3.1 Noisy Emissions Reward Belief Marginalization, Not Point Estimates

3.1.1 Soft Marginalization Beats Point Estimation as Emissions Degrade

Under a known noisy emission, uncertainty-preserving belief marginalization lowers predictive loss by 0.739 nats on the primary fixture draw. This is the benchmark's one replicated positive mechanism-level regime.

Active inference realizes that mechanism through recursive hidden-state belief updating in a discrete POMDP generative model [Da Costa et al., 2020]. HMM, particle, state-space, and soft-filter methods can implement the same belief-preserving mechanism when they propagate a calibrated posterior over the hidden state [Kaelbling et al., 1998, Rabiner, 1989, Arulampalam et al., 2002, Doucet et al., 2001, Patterson et al., 2008, 2017]. Their performance can still diverge through model misspecification, approximate inference, finite-sample estimation, adaptation, and policy structure [Da Costa et al., 2020, Arulampalam et al., 2002, Patterson et al., 2017]. When the observation directly names the discretized latent cell, the posterior is already concentrated and a point estimate supplies the same information.

The protocol of sec. 2.6.2 isolates the condition that rewards recursive belief: a latent state, a noisy emission, and prediction of the true latent next cell. Soft (marginalize the full belief), hard (commit to the MAP cell), and observed-Markov (commit to the raw noisy observation) share the same learned transition, emission, observations, and priors. Marginalize-versus-commit is the sole treatment difference, eliminating the asymmetry that invalidated the earlier noise frontier (sec. 6.4). The log score rewards the full predictive distribution [Gneiting and Raftery, 2007], so this lane tests hidden-state filtering and probabilistic prediction. It does not isolate an expected-free-energy policy-selection advantage [Millidge et al., 2021].

Across 16 latent cells, the soft belief advantage over the point estimate is 0 nats at zero noise (zero-at-clean: yes) and rises monotonically with the observation-noise level (monotone: yes), reaching 0.739 ± 0.127 nats at $\varepsilon = 0.400$ (fig. 3). The oracle-fed reference remains the predictive ceiling; the measured gain belongs to the declared synthetic emission contract.

The growth is close to linear in the flip probability: an ordinary-least-squares fit of the mean advantage against ε gives a slope of 1.860 nats per unit ε with $R^2 = 0.995$. The near-linearity has a mechanistic reading. Flips are independent across steps, and each flip costs a committing predictor a roughly fixed penalty: it propagates the wrong row of the transition. The soft belief retains $(1-\varepsilon)$ of its mass on the true cell, so to first order the excess loss scales with the flip rate. On this reading the slope estimates the average per-flip price of committing.

The spread is a robustness band, not a generalization interval: each level aggregates 15 paired soft-versus-hard resamplings (noise draws × locked holdout splits), and *every* resampling at *every* nonzero noise level is positive. The worst single resampling at the strongest noise still favors the soft belief filter by 0.574 nats; the clean channel carries no advantage by construction, as the zero-at-clean row above records.

The between-draw check repeats the construction across 10 independent synthetic generator draws with 16 travellers each. The noisy-emission sign remains stable (yes), and at $\varepsilon = 0.400$ the replicated advantage is smaller than the headline: the cross-draw mean is 0.495 nats (range 0.429–0.578 nats across draws, largest between-draw SD 0.049 nats). The 0.739-nat headline is therefore the primary 5-traveller fixture-draw value, not the replicated magnitude; the replication establishes the sign and a more conservative effect size.

The advantage varies in magnitude and remains unanimous in sign, while the broader benchmark gate stays `not_demonstrated`. The fair-comparison ladder and traveller×noise power map locate the active-inference row within this mechanism in the Discussion (sec. 4).

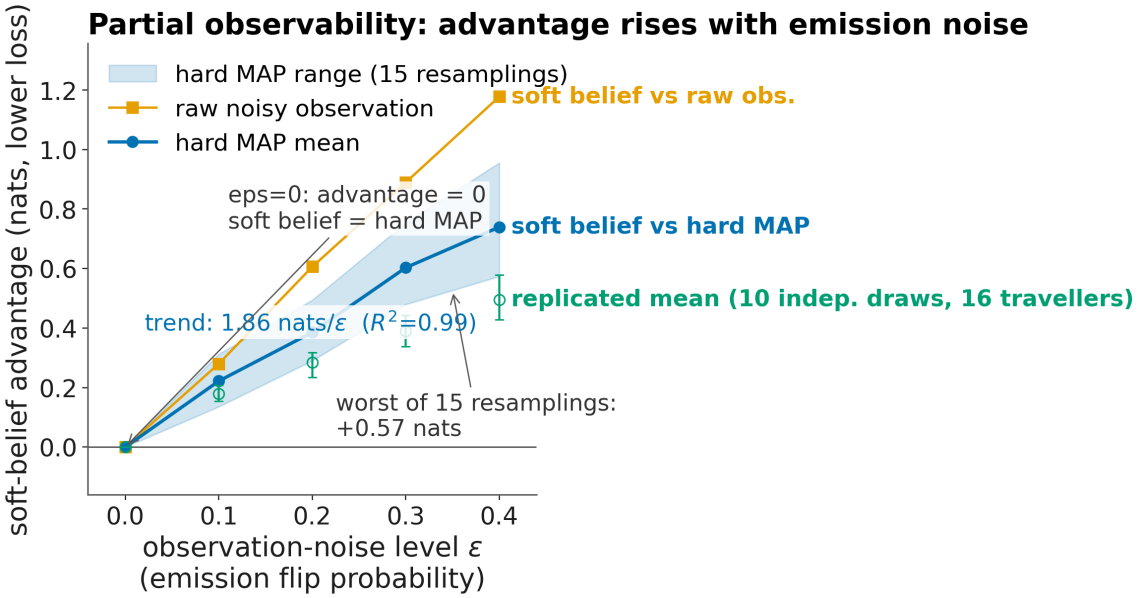


Figure 3: Source partial_observability_analysis.json compares soft belief, hard MAP, and observed Markov under the same known symmetric emission with flip probability epsilon, transition, observations, and priors. The soft-belief advantage is zero for a clean channel and grows monotonically with noise. The green overlay separates the replicated cross-draw range and mean from the primary fixture draw, identifying belief marginalization as the mechanism behind the gain. Evidence: a fair-by-construction isolation of the genuine partial-observability regime where belief marginalization beats committing to a point estimate under a known noisy emission, zero at zero noise and growing monotonically with it. Boundary: active-inference-specific superiority, a state-of-the-art claim, an advantage under full observability, or real-mobility validity.

3.1.1.1 Where Belief-Marginalizing Methods Lead: The Within-Regime Model Comparison

fig. 4 places that mechanism into the full implemented roster. Active inference achieves the lowest fair next-cell loss, rank 1 of 14, on the locked noisy-emission lane. The 4 predictors that maintain a belief over the latent cell and marginalize the noisy emission — active inference, its soft-filter reference, the switching-noise HMM, and the particle filter — occupy the leading cluster. That cluster leads the best point-estimate or raw-observation method by 0.192 nats. The uniform-chance loss is 3.584 nats, so the ranking distinguishes among genuine predictive models rather than chance-level variants.

Active inference's advantage over the strongest non-AIF method (`switching_noise_hmm`) is 0.005 nats, with a clustered interval that crosses zero. It is therefore the point leader and a statistical member of the top belief-preserving cluster. The separated result is belief preservation versus early commitment to a point estimate or raw observation; active inference does not clear a resolved margin over the strongest non-AIF belief filter, and that comparison stays a statistical tie. This ordering is specific to the noisy hidden-state lane, remains below the oracle-fed ceiling, and leaves the cross-generator gate at `sota_program_not_cleared`.

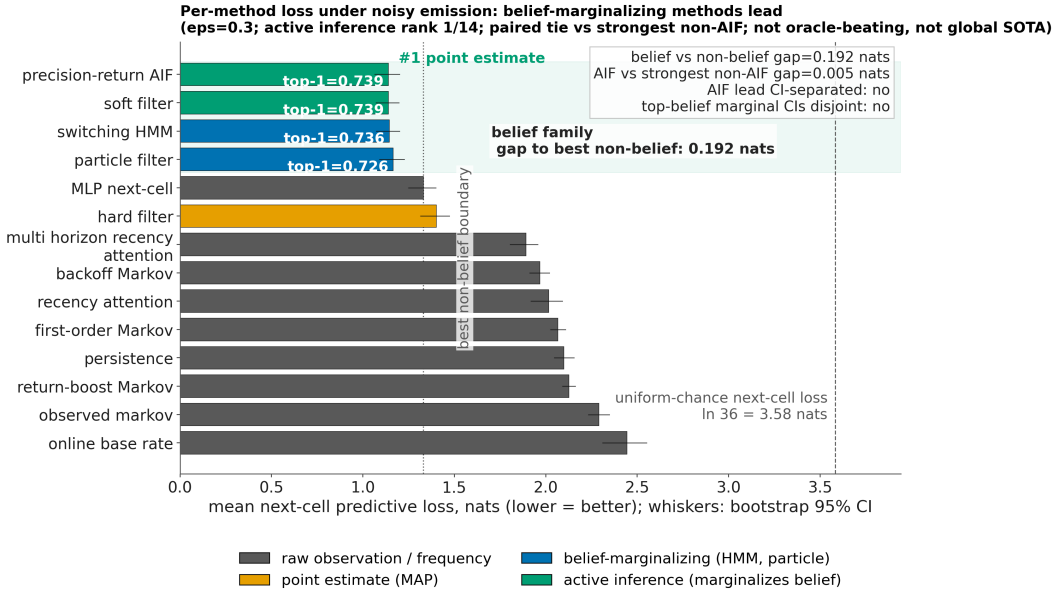


Figure 4: Source model_comparison_leaderboard.json ranks every implemented next-cell predictor by fair loss on the locked synthetic noisy-emission FractalRabbit lane. Active inference, soft filter, switching-noise HMM, and particle filter form the leading belief-preserving cluster. The plot marks uniform chance, the belief-family band, the best non-belief boundary, top-row accuracy, the belief-versus-nonbelief gap, and active inference's paired interval against the strongest non-AIF method. Marginal whiskers are descriptive; the paired interval determines separation. Evidence: a within-regime ranking showing belief-marginalizing methods, active inference included, lead the implemented predictors under noisy emission. Boundary: a global state-of-the-art claim, an oracle-beating result, or a real-mobility claim.

3.2 Nulls, Calibration, and Multiplicity Keep Gains Bounded

3.2.1 Recovery Calibration Faces Uniform, Base-Rate, and Persistence Nulls

This section frames state recovery and model performance against explicit baselines: ground-truth recovery on observed and latent configurations, skill scoring against uniform-chance and persistence nulls, and proper-score calibration diagnostics that decompose predictive loss into per-modality contributions and honest uncertainty pricing.

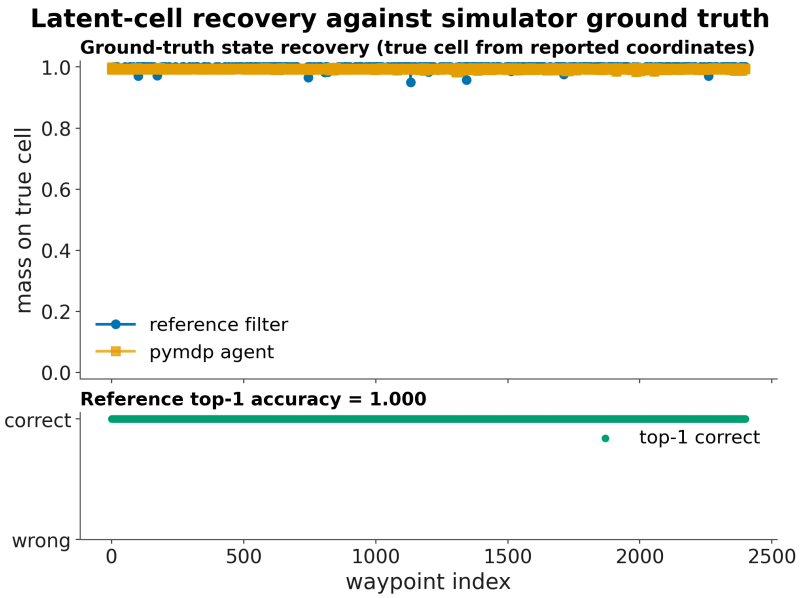


Figure 5: Sources ground_truth.csv and inference traces score latent-state recovery against the simulator's own truth. Posterior mass is plotted for both lanes when available, while top-1 correctness is plotted for the reference lane; with location observed, high recovery is a sanity bound (reference mean mass 0.995, top-1 1.000), not an achievement. Calibration and location-withheld recovery carry the substantive burden elsewhere. Synthetic software evaluation, not empirical validity. Evidence: synthetic-truth recovery sanity checks and calibration handoff. Boundary: empirical validity.

fig. 5 reports ground-truth state recovery. Because the true latent cell of every report is computable exactly from the reported coordinates, any posterior can be scored against it. On the default configuration — where location is itself an observation modality — recovery is near-perfect by construction (reference mean mass on the true cell 0.995, top-1 accuracy 1; pymdp 0.994 and 1), and we state that plainly rather than presenting it as an achievement. The non-trivial result is the deep-latent configuration from the sensitivity lane: with the location modality removed, the filter recovers mean mass 0.135 on the true cell with top-1 accuracy 0.132 from time-of-day, speed, and reporting-burst evidence alone, at mean posterior entropy 2.241 nats. Latent location structure is partially recoverable without any direct location report on this synthetic configuration — though, as the calibration analysis below shows, most of this recovered mass reflects occupancy structure and mobility autocorrelation (the base-rate and persistence baselines below) rather than the time-of-day, speed, and reporting-burst evidence itself. Within the modeled situation this is simultaneously the analytic point and a privacy-relevant observation; whether the same recoverability holds for real mobility traces is exactly the empirical question that sec. 8 leaves open, not something these synthetic numbers settle.

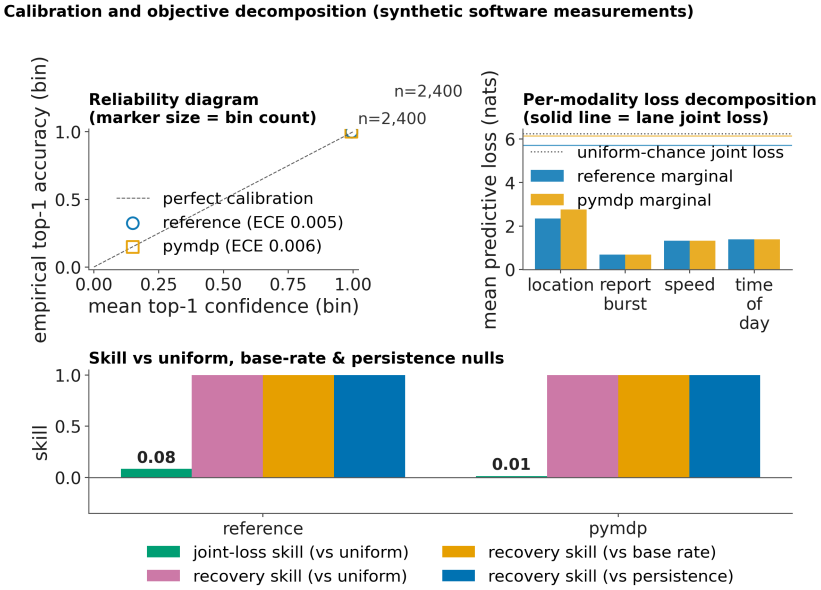


Figure 6: Source calibration.json decomposes predictive fit, confidence, and null-relative recovery for both inference lanes on shared axes. The reliability panel compares top-1 confidence with realized top-1 accuracy; the loss panel splits joint predictive loss into modality margins and their interaction gap; the skill panel moves from a uniform floor to base-rate and temporal-persistence nulls, with bin counts and tiny joint-loss skills printed directly. Persistence is the stricter null for autocorrelated synthetic mobility, so apparent by-construction wins are not overpromoted. Evidence: calibration floors, null hierarchy, and persistence-relative skill checks. Boundary: promotion of by-construction location wins as substantive recovery.

fig. 6 and tbl. 1 decompose and calibrate the headline objectives rather than leaving them as single numbers. First, the joint one-step predictive loss is split into per-modality marginal losses under the running prior. For any observed modality m, the joint predictive probability is bounded above by its marginal because

$$\sum_s D_i(s) \prod_j A^{(j)}_{o_i^{(j)},s} \leq \sum_s D_i(s) A^{(m)}_{o_i^{(m)},s}, \tag{15}$$

so $-\log P(o_i^{(m)}) \leq -\log P(o_i)$ for the same row. A checked invariant enforces this bound, and the summed-marginal-minus-joint gap — 0.023 nats on the reference lane — measures how much the modalities corroborate one another through the latent cell.

Second, each lane is scored against an explicit uniform-chance floor (6.238 nats joint loss): the reference filter attains a joint-loss skill of 0.083 and the pymdp closed loop 0.015, where a skill of one is perfect and zero is chance. This synthetic null floor makes the reference lane's clear predictive signal and the pymdp loop's smaller positive signal comparable against the same floor.

Third, recovery is calibrated, not just averaged, using proper-score and calibration diagnostics as scoring context [Brier, 1950, Guo et al., 2017, Gneiting and Raftery, 2007, Gneiting and Katzfuss, 2014]. The reference lane's Brier score is 0.000 with expected calibration error 0.005 — sharply calibrated largely by construction, since location is observed — while the pymdp closed loop scores Brier 0.000 with ECE 0.006 and runs under-confident (mean top-1 confidence 0.994 against top-1 accuracy 1). That is the direct signature of policy-induced prior divergence inflating its posterior spread without destroying its ranking. The substantive calibration number is again the deep-latent one: at full report density the deep-latent filter scores Brier 0.933 (Brier skill 0.004 against the uniform floor), quantifying how honestly the model prices its own uncertainty when location is never reported. Expected calibration error is read here as a binning-sensitive, descriptive diagnostic rather than a definitive metric, so it is interpreted alongside the strictly-proper Brier score and the calibration-versus-sharpness tradeoff rather than on its own [Nixon et al., 2019, Gneiting and Katzfuss, 2014].

Fourth, because a uniform floor is a weak null for state recovery — beating it is trivial when location is itself observed — we also score recovery against a tighter base-rate null: the constant marginal-frequency predictor, whose Brier score is the Gini impurity of the true-cell distribution and which is therefore never a looser null than the uniform floor, and a strictly tighter one whenever cell occupancy is uneven. For a one-hot true cell Y drawn from marginal distribution π and the constant predictor π,

$$\mathbb{E}\left[\sum_k (\pi_k - \mathbb{1}\{Y = k\})^2\right] = 1 - \sum_k \pi_k^2, \tag{16}$$

which is exactly the Gini impurity. Against this base-rate null the location-observing reference lane scores Brier skill 1.000 and the deep-latent lane -0.162 (positive = beats merely knowing how often each cell is occupied). The reference lane's location evidence trivially clears this null, while the deep-latent lane's margin measures how much of its above-uniform recovery is occupancy structure rather than observation-driven signal — on the current expanded fixture that margin sits at the occupancy floor, which is exactly the overreach this stricter baseline exists to catch.

Fifth, even the base-rate null is loose for mobility, because reports are highly autocorrelated — travellers stay in or return to the same cell — so we add the strongest honest null, temporal persistence (predict the previous report's cell, falling back to the base rate at each traveller's first report). Against persistence the deep-latent lane's recovery skill falls to -19.745 (reference lane 1.000), the conservative measure of how much the time-of-day, speed, and reporting-burst evidence adds beyond simply assuming no movement. The base-rate-versus-persistence ordering is itself data-dependent — persistence is the tighter null only where stay-put autocorrelation exceeds marginal predictability — so we report skill against both. All calibration estimates are computed over 2400 fixture reports with per-bin counts displayed in fig. 6, so thinly occupied confidence bins are visible rather than averaged away.

Table 1: Per-modality marginal predictive losses against each lane's joint loss, from `calibration.json`.

modality	reference marginal loss (nats)	pymdp marginal loss (nats)
`location`	2.342	2.770
`time_of_day`	1.385	1.386
`speed`	1.321	1.322
`report_burst`	0.693	0.693
joint (all modalities)	5.719	6.147

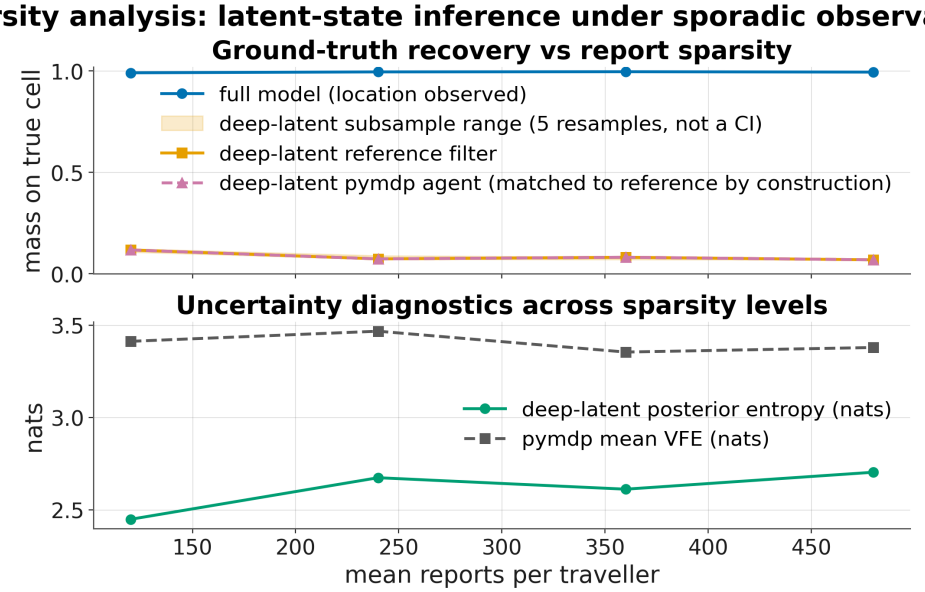


Figure 7: Source sparsity_analysis.json subsamples each traveller's reports and re-estimates both the full model and the deep-latent no-location model at each sparsity level. Mean true-cell posterior mass prices direct location evidence, while entropy and pymdp variational free energy show how uncertainty and closed-loop cost change as sequences lengthen. These fixture-lane measurements separate filter recovery from policy-induced prior divergence and are not empirical claims. Evidence: how recovery and uncertainty change under fixture-lane report thinning. Boundary: population generalization across real sampling regimes.

fig. 7 and tbl. 2 give the comprehensive form of the latent-state question: recovery as a function of report density across 4 subsampling levels. With the full report set the deep-latent reference filter holds mean mass 0.069 on the true cell, against 0.118 at the sparsest level; the full model dominates every deep-latent curve at every level, and the gap between them prices the value of direct location evidence at each report density. The pymdp column of the table is a variational-free-energy trace from the validated state updater, not an independent closed-loop recovery curve: its deep-latent recovery reuses the matched reference posterior by construction once the sentinel agreement check passes, so the table carries one recovery curve per configuration. The accompanying uncertainty diagnostics — deep-latent posterior entropy and pymdp variational free energy — are reported per level rather than asserted to follow a fixed direction, because their movement depends on how subsampling reshapes the interval and speed evidence. Recovery-versus-sparsity curves over deterministic subsamples of the checked fixture. They quantify how this configured model degrades as reports thin; they are software measurements on synthetic data, not empirical mobility findings. The full-model curve observes the true cell directly through the location modality, so its near-ceiling recovery is a sanity bound by construction; only the deep-latent curves test genuine inference. The per-traveller subsample keeps at least two reports, so actual_fraction is the honest sparsity level when it diverges from the requested fraction. The pymdp lane validates the released state updater on deterministic sentinel rows at each sparsity level; full-row deep-latent recovery uses the matched reference posterior after that agreement check. The deep-latent recovery band is the min-max spread of the mean mass on the true cell across 5 independent

subsamples at the same fraction — a subsample-resampling robustness spread over this one fixture, not an inferential or population confidence interval; at the densest (full-report) level every subsample keeps all reports, so the band collapses to a point by construction.

Table 2: Recovery versus report sparsity emitted by `scripts/08_run_sparsity_analysis.py`.

reports/traveller	full mass	latent mass	latent top-1	latent entropy (nats)	pymdp VFE (nats)
480	0.995	0.069	0.082	2.704	3.380
360	0.996	0.082	0.113	2.612	3.355
240	0.996	0.074	0.081	2.675	3.468
120	0.991	0.118	0.212	2.449	3.414

3.2.1.1 Multiplicity-Corrected Claims: No Directional Win Survives the Comparison Family A benchmark that runs many directional comparisons must price in the size of the family before reading any single clearance as a result. We do this without introducing a new researcher degree of freedom. The comparison family is enumerated by a fixed rule from the live artifacts: every active-inference-versus-fair comparison that reports a one-sided traveller-clustered bootstrap p-value, namely each synthetic regime of the regime search and each lane of the cross-generator program. Each row compares paired predictive-score differences on the same held-out reports rather than unmatched summary losses [Diebold and Mariano, 1995]. Each regime enters at its worst (largest) one-sided p over its realization seeds, so a regime clears only when even its least favourable realization does, the conservative direction. The family is fixed by this rule, since selecting which comparisons "count" would itself be the kind of confound this benchmark exists to expose.

Bootstrap intervals are used as descriptive software-measurement intervals, with the traveller as the resampled dependence unit rather than individual steps [Efron and Tibshirani, 1994, Field and Welsh, 2007, Cameron and Miller, 2015]. This holds at every interval feeding the comparison family: regime-search candidate and paired intervals, cross-generator lane intervals, and the traveller×noise regime map. The only percentile constructions in the study that are not traveller-clustered are permutation-null thresholds, which resample labels under the null rather than sampling units, and per-injection-seed robustness spreads; neither is read as a traveller-level confidence interval.

Each cross-generator lane resamples 26 held-out travellers, and each regime comparison draws from a 40-traveller synthetic population. With cluster counts of this order the percentile interval is a serviceable but not exact-coverage summary [Field and Welsh, 2007, Cameron and Miller, 2015], one more reason the corrections below are read as bounding over-interpretation rather than certifying population inference. Applying formal corrections to descriptive software-measurement p-values is deliberate: the corrections do not upgrade the measurements into population inference, they bound how many apparent clearances resampling noise plus a family of 9 could manufacture on their own.

Across the resulting 9 comparisons (fig. 8), computed with B = 2000 traveller-resampled replicates per comparison (smallest reportable p 5.00e-04), 0 clears the uncorrected significance threshold (0.050). Every comparison's own confidence interval, not merely its raw p-value, is required to separate from zero before it counts as a clearance at all, so a comparison whose interval still straddles zero is read as unresolved rather than as a marginal win.

Under Holm–Bonferroni correction, which bounds the family-wise error rate under arbitrary dependence, 0 survive [Holm, 1979]. That is the honest assumption given that the regimes and lanes share generators, candidate code, and the active-inference fit. Under a Benjamini–Hochberg false-discovery control, valid under independence or positive regression dependence, 0 survive [Benjamini and Hochberg, 1995]. We state that dependence assumption rather than claiming it.

Across the family the mean advantages range from -0.259 to 0.020 nats, so most comparisons favor the fair baseline outright, and none clears its own uncorrected threshold, let alone survives correction. The regime-qualified gate of sec. 6.3 is therefore read here as not clearing even before multiplicity is priced in. That is the honest reading across the whole family: no directional win, corrected or uncorrected, survives.

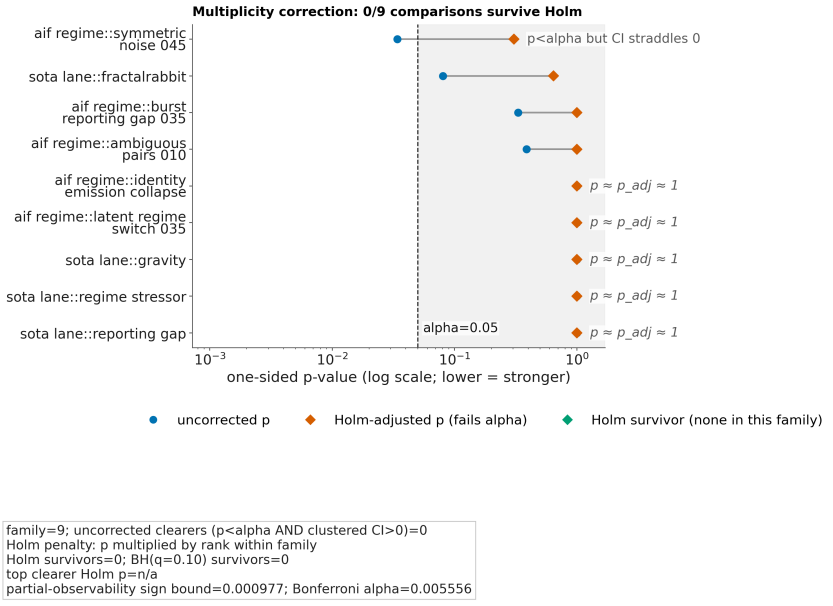


Figure 8: Source multiplicity_ledger.json enumerates every synthetic AIF-versus-fair directional comparison with a one-sided clustered-bootstrap p-value. Each dumbbell joins the raw and Holm-adjusted p-values on a log axis against alpha; markers identify rows whose raw p-value and clustered interval both clear. The annotation reports family size, uncorrected clearers, Holm and Benjamini-Hochberg survivors, the best adjusted p-value, and the separate partial-observability Bonferroni sign-test bound. Evidence: a multiplicity audit reporting how many directional comparisons survive Holm and Benjamini-Hochberg correction across the deterministically enumerated family. Boundary: a state-of-the-art claim, an oracle-beating result, or a real-mobility claim.

The single fair-by-construction positive — soft marginalization under a noisy emission (sec. 3.1.1) — sits outside this bootstrap family because it is established by a stricter criterion: unanimous sign-stability across 10 independent generator draws, a replication requirement that is multiplicity-robust by construction.

Its validity is moreover mechanism-level rather than frequentist. Soft Bayesian marginalization is provably at least as good as a hard point estimate under a known emission, in the model-expected sense made precise in sec. 2.6.2, with a strictly positive advantage as the channel degrades. A multiple-testing correction on a sign statistic is therefore the wrong lens for the mechanism. The direction of that advantage is fixed *a priori* rather than chosen after seeing the draws, so the one-sided sign-test bound is the appropriate null.

Reported conservatively all the same, the sign-test bound for its 10 unanimous draws is 9.77e-04, below the family-wide Bonferroni per-comparison threshold of 5.56e-03 (the minimum sufficient is 8 unanimous draws). The one positive therefore survives even the most conservative family-wide correction the paper applies, over and above the mechanism-level argument, which does not rest on the sign test at all. The synthesis is the honest one: of every directional comparison the

benchmark makes, none survives family-wise correction as a measured win, and the one advantage that does hold is the mechanism-level result that never claimed to beat the oracle or to transfer to real mobility.

4 Discussion: Belief Filtering Earns a Regime-Bound Role

The benchmark identifies one positive mechanism: uncertainty-preserving belief-state filtering improves prediction when a noisy emission hides the current cell. The effect reaches 0.739 nats on the primary fixture draw and replicates at a mean of 0.495 nats across 10 independent higher-power draws. Active inference realizes this mechanism through recursive belief updating and achieves the lowest implemented loss in the matched noisy-emission comparison [Da Costa et al., 2020]. HMM, particle, state-space, and soft-filter approaches can target the same posterior predictive distribution when models and information sets are matched; approximation, learning, adaptation, and policy structure can still separate them [Kaelbling et al., 1998, Rabiner, 1989, Arulampalam et al., 2002, Doucet et al., 2001, Patterson et al., 2008, 2017]. The active-inference row occupies a precise position: lowest implemented point estimate under noisy partial observability, below the oracle-fed ceiling, within a statistically tied leading belief-filtering cluster.

The partial-observability construction of sec. 3.1.1 isolates the mechanism: soft Bayesian marginalization ties the point estimate when the observation reveals the latent cell, then gains monotonically as the emission degrades. The wider evidence assigns that gain. The fair ladder separates single-step marginalization from recursive belief; the traveller×noise map locates the regime; the within-regime comparison (sec. 3.1.1.1, fig. 4) places active inference first among the implemented noisy-emission predictors; and the oracle, clean-emission, switching, and cross-generator surfaces identify its limits.

The active-inference row leads by point estimate and shares the top statistical tier with the strongest non-AIF belief filter. The historical confound taxonomy remains supplementary audit evidence in sec. 16. The central answer is direct: preserve and recursively update a belief when observations leave the state uncertain; use a simpler predictor when the observation already resolves it.

4.1 A Reader's Model Map: Match the Predictor to the Information Regime

The results are easiest to interpret as a model-selection map rather than as one global ranking: there is no global winner. The decisive variable is what information the predictor receives and whether uncertainty about the current cell is genuine, transient, or already resolved by a simpler statistic.

Information regime	What works	Observed limit	Why
Clean or directly observed cell	Persistence, Markov, and point-estimate predictors are sufficient reference models.	Recursive belief offers no general gain.	The posterior is already concentrated; recursion mostly reintroduces transition-estimation and smoothing error.
Noisy emission over a hidden cell	Soft Bayesian marginalization, HMM, particle/state-space, soft-filter, and active-inference families form the leading implemented cluster.	The strongest non-AIF belief method is statistically tied with active inference; the oracle remains lower.	Preserving probability mass over plausible cells avoids propagating a prematurely committed state; most of the gain comes from marginalization, with recursion helping as the emission degrades.
Latent regime switching	An adaptive online base-rate predictor is the strongest implemented fair comparator on the tested switching lane.	Online base-rate adaptation leads fixed-generative-model AIF.	Fast changes in occupancy favor online adaptation; a fixed transition model carries stale structure across regime boundaries.
Sparse reporting gaps	Persistence and disclosed kinematic baselines remain indispensable controls; gap-aware gains stay synthetic and cue-qualified.	Timing alone stays below promotion; displacement-derived speed narrows the task.	Timing can describe the reporting process rather than motion, while displacement-derived speed directly reveals whether movement occurred.
Location withheld, metadata retained	Occupancy and trained-dynamics baselines explain most recoverable broad structure.	Specific-cell recovery remains at the strongest dynamics/persistence null surface.	Time, speed, and burst metadata carry limited incremental information once occupancy and autocorrelation are priced in; withholding location therefore remains strongly protective.

The model-selection rule is practical: use the simplest predictor whose information set resolves the state; preserve a distribution over states when the emission leaves genuine uncertainty; prefer adaptation when the generator changes; and attribute apparent gains to an oracle, cue, occupancy pattern, or unmatched information set whenever those explanations account for the result.

4.1.1 Online Adaptation Repairs Fixed Dynamics Without Winning the Switching Lane

The AIF-deepening audit in fig. 9 separates modular surfaces by predictive admission: factorial tensor provenance and multiscale soft-belief backoff remain audit-only, while online transition adaptation enters `Candidate.predict(PredictContext)` and is scored through a genuinely nested train/validation contract. Protocol candidates currently admitted: 1; predictive gate clears: no. Across 5 distinct synthetic generator/noise realizations sharing the locked split rule, the adaptive candidate improves the separately tuned fixed-transition AIF by a mean 0.146 nats (range 0.068 to 0.280), with traveller-clustered separation on 4 realizations. It beats the predeclared online-base-rate comparator with clustered separation on 0 realizations, so the repeated predictive gate remains no. This advances the generative model and localizes the remaining gap without clearing promotion.

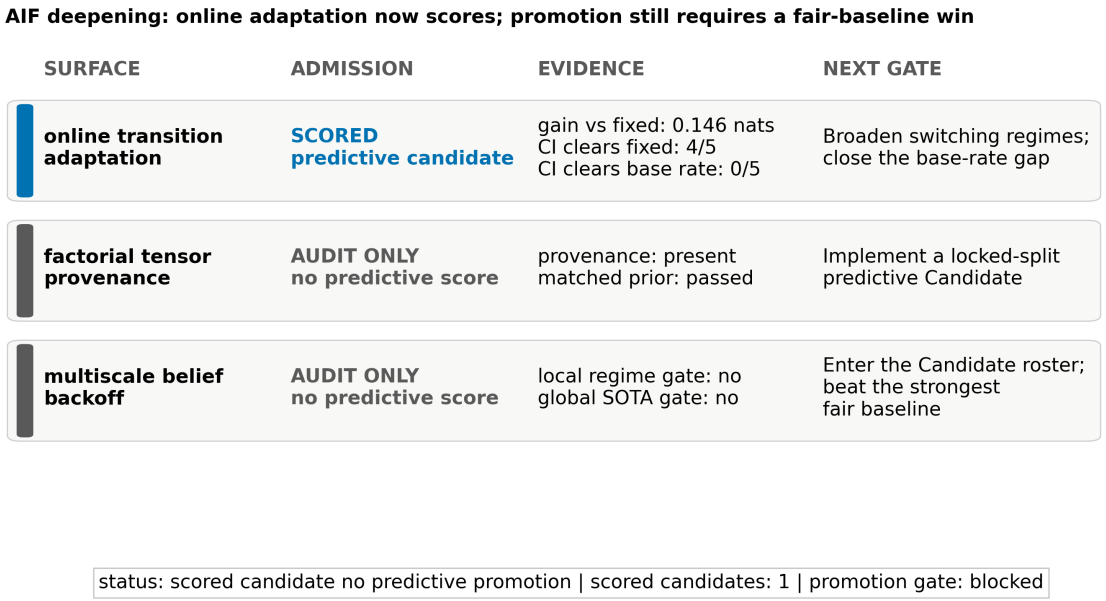


Figure 9: Source aif_deepening_candidates.json tracks active-inference deepening surfaces by predictive admission: factorial tensor provenance, multiscale soft-belief backoff, and online adaptation. Rows show whether each surface has entered the locked Candidate.predict(PredictContext) protocol and whether any predictive gate clears. Online adaptation is now a scored candidate that improves the fixed AIF ablation but not the predeclared online-base-rate comparator; audit-only rows remain visible so tensor, EFE, and VFE ideas cannot substitute for prediction. Evidence: a predictive-admission map where online adaptation is scored while factorial and multiscale surfaces remain audit-only. Boundary: predictive improvement, SOTA promotion, or operational claims from EFE, VFE, tensor, or adaptation surfaces alone.

4.2 The Fair Ladder Assigns Most Gain to Soft Marginalization

The active-inference lane uses the same locked, traveller-disjoint split as the transparent baselines and is scored by prequential next-cell loss [Dawid, 1984, Luca et al., 2023, Gneiting and Raftery, 2007]. Across the structural sweep, no active-inference configuration beats the strongest baseline: the best inference loss is 0.909 nats, above the strongest baseline's lower confidence bound, and the claim gate stays `sota_not_cleared`. That comparison is information-asymmetric because the baselines receive the true previous cell. The fair ladder instead gives every predictor the same locked split and transition source, varying only how the previous cell is inferred from the emission (fig. 10). A hard MAP estimate scores 0.936 nats; full soft marginalization scores 0.788 nats; recursive active inference scores 0.835 nats; and the oracle-fed reference scores 0.606 nats.

Soft Bayesian marginalization supplies the dominant inference gain, about 0.147 nats over the hard point estimate. Recursive temporal belief contributes conditionally: single-step marginalization leads at the clean setting, while recursion turns positive as the channel degrades and reaches about 0.356 nats at the noisiest setting.

The direction of that crossover is what estimation error predicts. At a clean emission the single-step posterior is already nearly one-hot, so the recursive prior contributes no information the observation lacks but does inject the smoothing mass and finite-sample error of the learned transition into every step. Once the emission degrades, the prior carries information the observation no longer supplies, and recursion's contribution turns positive. Recursive belief therefore earns its keep specifically under observation noise, not in general.

At the clean setting, single-step soft marginalization leads the recursive active-inference filter (0.788 versus 0.835 nats). The per-traveller visited-cell return prior worsens both predictors on the regenerated fixture (soft 1.129 nats; filter 1.085 nats), although the recursive filter absorbs less of that penalty.

Positive controls verify the comparison: an identity emission collapses every inferring predictor onto the oracle, and the oracle-fed baseline remains the ceiling. The fair ladder therefore assigns most of the gain to single-step soft marginalization and the remaining noise-dependent gain to recursion. The broader benchmark gate remains `sota_not_cleared`.

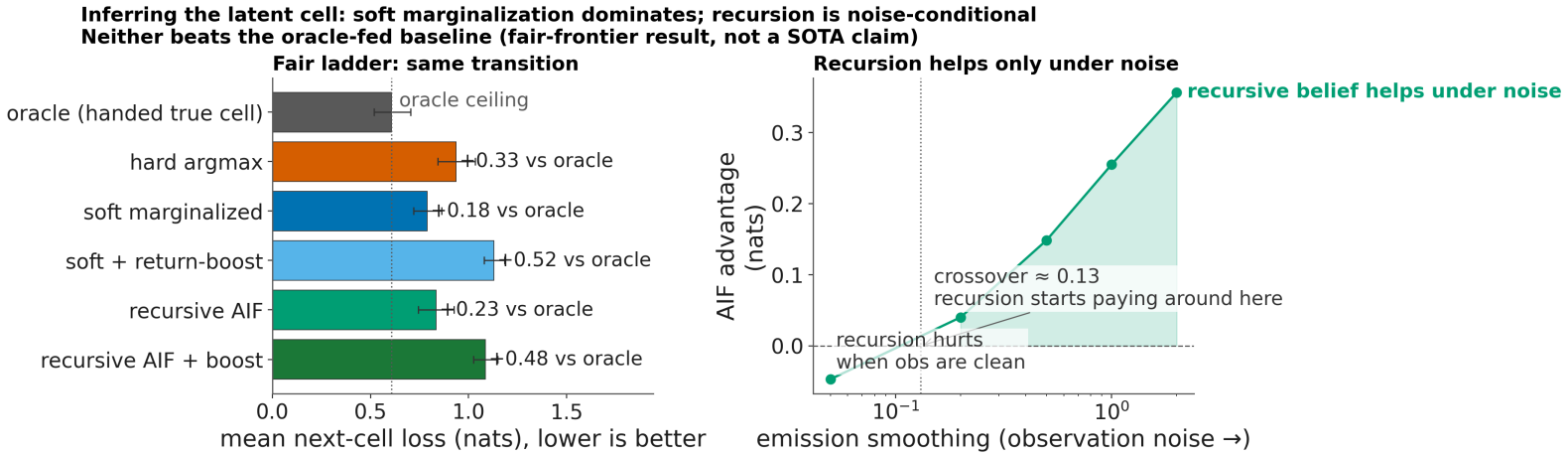


Figure 10: Source fair_inference.json scores an oracle-stripped ladder in which every predictor learns the same transition from the same observed cells. The left panel ranks oracle, hard argmax, soft marginalization, return-boosted soft prediction, and recursive active inference with bootstrap whiskers. The right panel traces recursion's incremental advantage as emission noise rises. Most of the gain comes from soft marginalization; recursion adds value conditionally as the observation channel degrades. Evidence: a fair, transition-source-matched comparison of how much latent-cell inference helps, decomposed into soft marginalization and recursion. Boundary: a state-of-the-art claim, an oracle-access result, or real-mobility validity.

4.3 The Regime Map Keeps the Advantage Sparse and Noisy

The fair question has a measured answer, but it must be read from the tokens, and with its sample size in the same breath. Scored under the pre-specified locked next-cell loss, chosen before this comparison and not to favor any predictor, and aggregated across multiple seeds at a holdout whose power the next paragraph qualifies, the artifact reports the number of seeds in which an active-inference predictor is the lowest-loss fair predictor as 0/8. Its mean advantage is -0.348 nats over the best non-active-inference fair predictor, where a negative advantage means the active-inference predictor trails that baseline.

Those seeds reshuffle the same handful of travellers, with as few as 2 distinct held-out travellers per split. Only a unanimous fixture-local rate with a positive advantage would support even bounded wording about this fixture's tested predictor set. The current hydrated values do not support that wording: they report a non-unanimous rate and a negative mean advantage, consistent with the fair-ladder finding above that the active-inference filter is not the best fair predictor. Either way the comparison is directional and provisional, does not beat the oracle, and is not an operational claim.

We deliberately report no confidence interval. Per-step bootstrap intervals treat within-traveller-correlated steps as independent and overstate precision, while a traveller-clustered interval is degenerate at this distinct-traveller count. No valid interval exists, so CI-separation is withheld rather than merely wide. A positive control confirms the oracle-fed baseline is still lower in every seed, so the comparison is non-degenerate; the prerequisite for a precision claim is more held-out travellers and larger, more strongly regime-structured synthetic data, not a cleverer filter.

Independently of where that between-predictor comparison lands, the predictor's own calibration can still be improved. Precision-weighting the active-inference belief ($\gamma = 1.500$, fixture-selected rather than held out) lowers its mean fair loss from 0.958 to 0.759 nats in 8/8 seeds with lower expected-calibration error observed on this fixture (mean ECE 0.367 to 0.148), because the over-smoothed base belief is under-confident. This is an in-sample inference-internal improvement, not a generalization claim, and it is deliberately excluded from the between-predictor comparison above so it cannot inflate the active-inference margin. Cross-vendor audit is a constraint in this project record: the confound taxonomy marks 2 apparent advantages caught only by that audit, so the responsible reading is constraint, not confirmation.

The session's central limitation — too few held-out travellers to interval the advantage — is itself testable. We generate deterministic synthetic streams at a fixed spatial resolution with up to 32 travellers and interval the fair advantage with a *traveller-clustered* bootstrap, resampling whole travellers as the correlated unit rather than steps. Because that bootstrap still only sees one generator draw, we repeat the whole sweep across 5 independent draws and call a cell a *seed-robust win* only when its clustered interval separates from zero in every draw. The across-draw spread is the between-dataset variance the bootstrap alone cannot capture.

The result is a regime map over traveller count and observation noise. With few travellers the active-inference filter's advantage over a return-boosted soft Markov is large and seed-robust (0.408 nats at a clean emission). In this smallest column the holdout contains only 2 travellers per draw — too few for a valid traveller-clustered interval, which is why we withhold CI-separation language for this column (see the power analysis above). Its seed-robust designation should therefore be read as across-draw sign unanimity with a large margin (worst draw 0.378 nats across 5 independent draws), not as within-draw interval precision; the interval-backed cells are those with more travellers and at least four holdout clusters.

As travellers grow the clean-emission advantage shrinks to the noise floor and slightly reverses (-0.005 nats at the largest count, not a seed-robust win). At a fixed cell grid, traveller count and per-cell density are deliberately coupled, and the vanishing advantage is the shared population transition model becoming sufficient as data-per-cell grows — a density effect, not a count-intrinsic law. Its onset is resolution-dependent: a finer grid delays it, but it is not a fixed-box artifact, since it persists when anchor density is held constant.

The advantage re-emerges under observation noise as a monotone gradient across the noise axis even at that scale, reaching a seed-robust 0.054 ± 0.003 nats at the noisiest emission (noise restores the advantage: yes). Because these cells are selected from a 2D sweep, the monotone gradient, not any single cell clearing zero, is the claim. The magnitudes are kept distinct: the few-traveller effect (0.408 nats) is roughly an order of magnitude larger than the at-scale-noisy one (0.054 nats).

For the target sparse-waypoint regime, recursive temporal belief earns its keep when the emission is uninformative, while a simpler boosted Markov suffices when locations are dense and clean. The oracle-fed baseline remains the ceiling, and the advantage follows the noise gradient rather than extending across regimes (fig. 11).

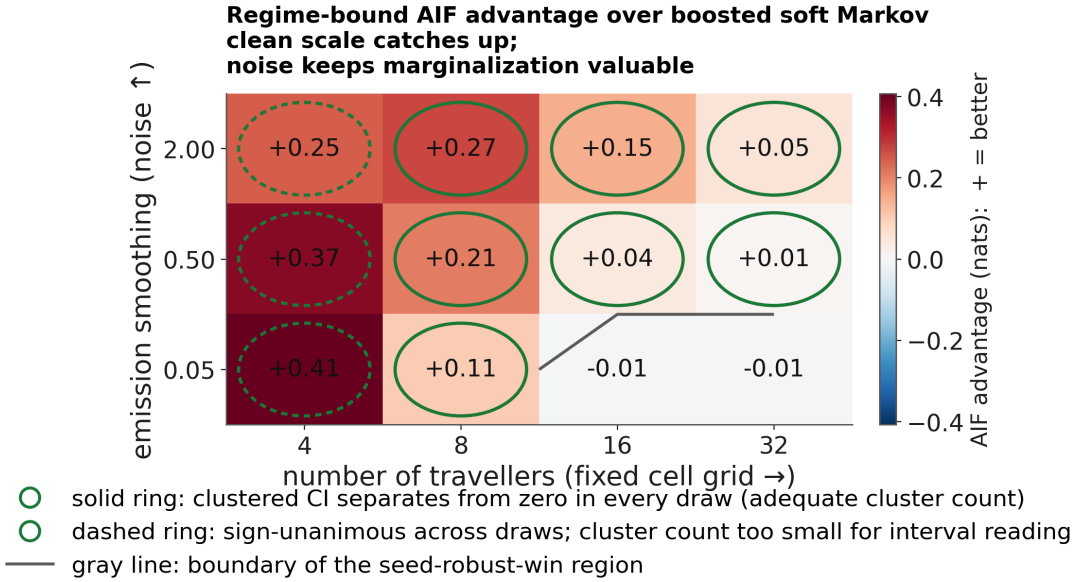


Figure 11: Source power_analysis.json maps active inference minus return-boosted soft Markov across traveller count and observation noise on deterministic synthetic streams. Color reports signed advantage in nats. Solid rings mark traveller-clustered interval separation when enough holdout clusters exist; dashed rings mark cross-seed sign stability where within-draw intervals are underpowered. The map shows large small-sample margins, vanishing clean-emission margins at scale, and renewed advantage as observation noise rises. Evidence: a clustered-bootstrap regime map showing the fair AIF advantage is large with few travellers, vanishes at scale under a clean emission, and re-emerges under observation noise. Boundary: a blanket state-of-the-art claim, an oracle-access result, or real-mobility validity.

4.4 Oracle Asymmetry and Cross-Generator Nulls Block Promotion

The same discipline marks the boundaries of the claim as sharply as its centre. The persistence null of sec. 3.2.1 is an oracle baseline: it is handed the true previous cell. The current location-blind lane therefore loses under an explicitly weaker information set; this is a benchmark-specific information-asymmetry bound, reported plainly as a synthetic-regime fact rather than reframed as an artifact.

The higher-power, 4-lane benchmark of sec. 6.3 tests cross-regime generality. Against strong fair baselines on locked splits with traveller-clustered intervals across `fractalrabbit`, `gravity`, `reporting_gap`, `regime_stressor`, its gate stays `sota_program_not_cleared`.

The synthesis is precise. Under the declared noisy emission, active inference achieves the lowest implemented point estimate and participates in a replicated hidden-state filtering advantage: 0.739 nats on the primary fixture draw and 0.495 nats on average across 10 higher-power draws. Its margin over the strongest non-AIF belief filter remains statistically unresolved, and the oracle-fed baseline remains lower. On the default observed-state surface, `pymdp` contributes interpretability through its variational and policy diagnostics (sec. 7.3). These are synthetic software results; a separate empirical protocol covers real-data evaluation (sec. 6.5).

4.5 Bayesian Mechanics Frames the Waypoint Particle as a Bounded Analogy

Bayesian mechanics gives this result a useful conceptual grammar, but not a stronger evidential warrant. In that literature, systems with a particular partition can be described in terms of internal states whose trajectories encode beliefs about external states, mediated by blanket states and interpreted through variational free energy [Ramstead et al., 2023, Da Costa et al., 2021, Friston, 2019]. Related Markov-blanket work connects this partition vocabulary to autonomy, information geometry, and stochastic thermodynamics [Kirchhoff et al., 2018, Parr et al., 2020], while the path-integral formulation treats particle trajectories as paths subject to a variational principle [Friston et al., 2023b]. The analogous object here is deliberately smaller: a synthetic waypoint belief over a hidden cell path, updated through sparse reports and a declared observation model. Calling that object a waypoint particle or cognitive-particle analogue is therefore a metaphor for an abstract belief-carrying trajectory in a controlled state space, not a claim that the synthetic trace is cognitive, living, sentient, autonomous, or empirically tracked.

The formal analogy is a typed bookkeeping device, not an ontology. A Bayesian-mechanics partition is often written as external, sensory, active, and internal states; in this benchmark the only defensible mapping is: hidden cell (x_t) as the external variable of interest, sparse report (o_t) as the sensory channel, no embodied active state, and the software posterior (q_t(x_t)) as the internal coordinate. The operative update is the ordinary discrete filter:

$$q_t(x_t) = p_m(x_t \mid o_{1:t}) \propto p_m(o_t \mid x_t) \sum_{x_{t-1}} p_m(x_t \mid x_{t-1}) q_{t-1}(x_{t-1}). \tag{17}$$

where (m) is the synthetic model contract. This is enough to say that the waypoint particle is a belief path over latent cells. It is not enough to assert a realized Markov blanket, a conditional synchronization map from internal to external states, a nonequilibrium steady state, a path integral over an organism, or a strange-particle/sentience claim. Those stronger readings require assumptions the current artifact neither states nor tests, and scope critiques of the FEP make that caution substantive rather than merely rhetorical [Aguilera et al., 2022].

The analogy is still useful because it sharpens the role of sporadic waypoints. The report stream is not a dense path; it is a sparse coupling between hidden movement and observed categorical evidence. The filter's internal state is the posterior over the latent cell, and the successful noisy-emission regime is precisely the case where preserving that distribution is better than committing to a point estimate. Read this way, the benchmark is a toy path-tracking problem: the internal belief path follows a hidden stochastic process only through the information the synthetic report channel supplies. That is compatible with the broader free-energy and active-inference vocabulary already used for the formal model [Friston, 2010, Friston et al., 2017, 2023a, Parr et al., 2022], while staying below a full Bayesian-mechanics derivation. The manuscript does not define a biological Markov blanket, prove a nonequilibrium steady state, or infer any actual mental state; it defines a software generative model and measures how belief updates behave under its synthetic evidence.

That boundary prevents the attractive but wrong conclusion. Bayesian Mechanics can motivate asking whether sparse observations let an internal state track a hidden process; it cannot turn a synthetic waypoint fixture into real cognition, real mobility validation, or a promotion claim. The right claim is therefore conceptual and local: this benchmark offers a source-bound example of belief-state path tracking under sparse synthetic reports, and its result remains the same bounded one stated above — uncertainty-preserving inference helps only in the noisy partial-observability regime and stays below the oracle-fed ceiling.

4.6 Governance Sources Bound Interpretation, Not Deployment

The governance lesson is a boundary discipline, not a new empirical result. Sparse waypoint traces are policy-relevant because location data can be unique, matched across datasets, aggregated, sold, repurposed, or recombined, and because commercially available information and sensitive-location enforcement sources treat trace provenance and minimization as live governance questions [de Montjoye et al., 2013, Kondor et al., 2020, Haggerty and Ericson, 2000, Office of the Director of National Intelligence, 2022, 2024, Federal Trade Commission, 2024, 2026]. Those sources support the relevance of the problem setting only. They do not validate FractalRabbit, do not establish legal compliance, and do not turn the synthetic benchmark into an operational or surveillance-ready system. Formal privacy mechanisms and trajectory-specific privacy evaluations remain separate from this benchmark's synthetic risk proxies [Dwork, 2006, Buchholz et al., 2024, National Institute of Standards and Technology, 2020].

The responsible interpretation is therefore analyst-facing and uncertainty-preserving, not deployment-facing. The model should be read as a structured aid to inference: it exposes posterior uncertainty, null failures, calibration gaps, and minimization costs, while keeping human judgment responsible for problem framing and for any decision to seek empirical data [Heuer, 1999]. Automation-bias work makes the same caution sharper: a clean trace, a calibrated probability, or a low-loss row can become misleading when readers over-trust an automated aid or mistake a bounded benchmark for deployment evidence [Horowitz and Kahn, 2024]. Human oversight is necessary, but it is not a magic solvent; the manuscript's stronger safeguard is that the artifacts keep uncertainty and non-claims visible.

Official AI, cyber, and responsible-AI frameworks supply vocabulary for that discipline: governability, traceability, reliability, risk mapping, and measurement are useful labels for the code gates, ledgers, and artifact provenance [National Institute of Standards and Technology, 2023, 2024, U.S. Department of Defense, 2021]. They are conceptual anchors, not certification. In this manuscript, governance means refusing to let a regime-qualified AIF advantage outrun its evidence: no global SOTA claim, no real-trace claim, no tasking claim, and no claim that policy sources certify the model.

5 Limitations: Synthetic Evidence Does Not Promote Real-Trace Claims

5.1 The Generator and Lanes Bound Reproducibility

The following limitations concern the synthetic-data substrate and the gap between latent-state recovery and ground truth. Active FractalRabbit: A Synthetic Benchmark for Belief Filtering Under Sparse Waypoint Observations has deliberate limitations. Although the external lane now builds and runs the full stochastic simulator at the pinned commit, the upstream CSV parameter surface exposes no random seed, so individual external runs are not bit-reproducible; reproducibility at that boundary is statistical (replicate dispersion is reported) rather than exact. The checked fixture remains the deterministic regression substrate.

The latent state is a discretized spatial cell. It is useful for a discrete POMDP analysis, but it is not a recovered FractalRabbit internal state, a real location semantic, or a cognitive state. In the default configuration location remains an observation modality, so the flagship lane stays close to an observed-state filter; the deep-latent configuration (`model.modalities` without `location`, exercised by the sensitivity lane and a checked config fixture) removes direct location evidence and demonstrably raises posterior uncertainty, but it is a configuration option rather than the default reporting lane.

The spatial-resolution frontier narrows this limitation rather than removing it. Its current verdict is `default_resolution_not_best_feasible_loss`: the 4x4 default remains fully occupied and auditable, but the canonical fixture no longer makes it the best feasible observed-location loss among the measured candidates. The deterministic sample-size ladder (2, 4, 8 times the canonical reports per traveller, largest n=19200) is useful because it retests the same frontier under larger synthetic samples; it reports `larger_sample_changes_frontier_decision` at the largest run and a guardrail-stable status of yes across the ladder. It refines the observed-location fit verdict, so it must be read as sensitivity evidence rather than population replication. It is still generated from the same fixture family and is not a population confidence interval or empirical validation. Because the ladder has strict feasible candidate counts 2, 0, 0, its selected-bin series 4, 8, 4 is a fallback best-measured-loss series wherever that count is zero, not proof that the larger lanes satisfy the canonical occupancy floor. The no-location guardrail still

moves against fine-cell recovery (canonical default-vs-coarse mass change -0.045; largest-sample change -0.036), so the model gains nominal cell granularity only where location is observed; it does not gain an unsupported ability to infer fine-grained cells from metadata alone.

The active-inference, fixed-transition, and pymdp traces are synthetic-data software diagnostics. The external sensitivity lane now resimulates stochastic waypoints per variant, which broadens the evidence beyond a single fixture, but nothing here shows that FractalRabbit output reproduces empirical human-mobility regularities such as recurrent returns, scaling laws, population-level trajectory structure, or task- and aggregation-dependent predictability limits [Barbosa et al., 2018, Gonzalez et al., 2008, Song et al., 2010a,b, Ikanovic and Mollgaard, 2016].

5.2 Performance Language Remains Blocked by SOTA Gates

The `pymdp` lane is not a state-of-the-art performance result. The benchmark matrix reports `not_demonstrated`, and the measured predictive-loss delta versus the reference filter is 0.428 nats. The present evidence supports implementation validity and interpretability, not predictive superiority. A performance claim would need the baselines and empirical gates listed in sec. 6.3.

The closed-loop selector is a transparent heuristic for comparing declared synthetic simulator configurations. It does not optimize a real collection policy, and it should not be used as operational guidance without an empirical design that defines sampling, harms, privacy, and validation endpoints independently of this scaffold [de Montjoye et al., 2013, Kondor et al., 2020]. The same boundary applies to commercially available or connected-vehicle location-data debates: official policy and enforcement sources make the governance context visible, but they do not transform this scaffold into a CAI workflow, legal-compliance result, or operational system [Office of the Director of National Intelligence, 2022, 2024, Federal Trade Commission, 2024, 2026].

5.3 Scoring and Holdout Limits Keep Nulls Local

The remaining limitations concern measurement uncertainty, the distinction between internal validation and real-world generalization, and honest null findings that do not extend beyond their measurement regime. The fixture sensitivity lane is a deterministic reprocessing lane and its normal-approximation intervals summarize configured variants only. The external sensitivity lane does resimulate stochastic waypoints per variant, but its replicate family is small and the upstream seed is uncontrollable, so its dispersion numbers are descriptive spread, not calibrated sampling uncertainty.

The holdout and negative-control lanes are software-falsification evidence, not external validity. Holdout scoring re-estimates tensors on training rows but shares discretization bin edges with the full lane, so it bounds tensor-estimation generalization only. The negative controls confirm that the pipeline responds to destroyed structure in the configured direction; they do not show that intact structure corresponds to any real mobility process.

Several program-extension findings (sec. 14.7, sec. 14.6) are honest nulls rather than capability wins, and the limitation is precisely that they do not generalize beyond the regime in which they were measured. Structure recovery does not beat its dynamics-only baseline on the expanded fixture lane (the only lane the current structure-recovery artifact contains): the deep-latent occupancy total variation (0.171) does not improve on propagating the trained transition model alone (0.042), so the sparse non-location metadata adds little to the preferred-place structure a fitted dynamics model already encodes. Mission-conditioned active sensing is scored against a pure mission-pragmatic selection rule on the expanded benchmark (0.590 for expected-free-energy selection versus 0.554 for pragmatic selection), and the budgeted information-gain schedule beats only a fraction 0 of blind random schedules. We report these regenerated values without reframing nulls as wins.

The claim-level evaluation contract (sec. 8) has a deliberately narrow reach, and that narrowness is itself a limitation rather than a guarantee. The static contract catches only the statically-encodable share of this project's honesty failures, through its five shipped guards: wrong-regime persistence skill (G1), undisclosed speed cues (G2), unsupported leaderboard rhetoric (G3), unbounded active-inference-help or uniqueness wording (G4), and deployment, surveillance-readiness, or real-mobility-transfer wording that is not explicitly negated (G5). A learned model underperforming its own baseline, headline numbers that did not reproduce on real data, a fabricated library call, and a scoring artifact visible only on full re-scoring were all caught only by executing the analysis. The contract is therefore the smaller, claim-level tier of a two-tier honesty stack; a clean contract pass does not certify the analysis is honest, only that the statically-checkable claim-level errors are absent. As with every result here, all of these remain synthetic-data software measurements on the pinned generator within the oversight-positive, no-operational-tasking boundary, never empirical mobility validation.

5.4 Roadmap Artifact Boundaries Block Deployment Evidence

The new robustness and roadmap artifacts are still software evidence, not deployment evidence. Cross-realization variance bands summarize repeated synthetic draws, not population uncertainty; upstream sensitivity can block when no comparable prior commit builds; the fractal-dimension sweep varies a generator control rather than estimating a real trajectory dimension; and the circadian-phase artifact tests a candidate prior rather than real circadian behavior. The sequence-baseline ladder is a stronger prequential null, not a complete movement-model benchmark. The reporting-process gate tests synthetic inter-report timing; without empirical evidence about the observation mechanism, it cannot establish missing-at-random assumptions or rule out informative reporting [Rubin, 1976, Lin et al., 2004]. The privacy frontier reports unicity and co-travel coincidence only as synthetic minimization-risk proxies, not k-anonymity, differential privacy, or any other formal privacy guarantee. The hardened anomaly probes improve the baseline, but they remain perturbation tests over generated data rather than an alerting or investigation method. The privacy-utility frontier now reports a synthetic monotone minimization frontier and a source-cleared minimum-collection operating point on the larger canonical fixture, but that still does not imply an operational minimization recommendation, real-mobility evidence, or privacy guarantee; synthetic mobility utility itself still requires task-based checks before it can be generalized [Sweeney, 2002, Dwork, 2006, Kapp and Mihaljevic, 2023]. Privacy-preserving synthetic daily-trajectory pipelines are useful future comparators, but their aggregation assumptions, mechanism design, and deployment claims do not transfer to this sparse waypoint fixture [Ozaki et al., 2025]. Cross-domain synthetic-data reviews make the same caution sharper: privacy and utility metrics are task-dependent, not interchangeable certification labels, and many synthetic-data evaluations understate privacy risk when privacy is not tested directly [Kaabachi et al., 2025]. The timing-channel integrity probe is read from its regenerated artifact rather than from a fixed null narrative: on the current hardened-anomaly artifact the timing-manipulation probe shows a modest above-chance variational-free-energy separation while the online Markov baseline stays near chance, and that is a synthetic perturbation-response diagnostic on this fixture, not a detector to deploy and not a timing-integrity capability claim.

6 Roadmap: Promotion Stays Blocked by Evidence Gates

This roadmap is a work queue, not a changelog. Current results, audits, and release artifacts set the boundary conditions for future work; historical sequencing belongs in `tasks.yaml`, the claim ledger, generated manifests, and git history. The open gates below are therefore framed as falsifiable next questions: what evidence would be needed to promote a claim, what would kill it, and which synthetic-only boundary remains in force.

6.1 Current Evidence Stops at Synthetic Artifacts

The current evidence surface is already reported with provenance-checked tokens in Results and Discussion: fixture diagnostics and spatial-resolution choice in sec. 13.3, sequence/reporting/model gates in sec. 14.2 and sec. 14.3, forward gates in sec. 14.4, safety and minimization boundaries in sec. 15.2, `pymdp` diagnostics in sec. 13.2, and synthesis in sec. 4. This section deliberately does not repeat verdicts. Its role is to keep the future bar visible without turning open work into a delivered-work recap.

6.2 Near-Term Model Gates Need Fresh External Evidence

- **External-realization model gates.** Promote any fixture-positive sequence, reporting-process, return-plus-phase, forward-gate, or safety-boundary result beyond fixture language only after fig. 34 and fig. 35 show current-run external evidence instead of an unavailable lane.

- **Multiscale spatial state.** Replace the measured single-grid default with a hierarchy of coarse-to-fine states, so uncertainty can be interpretable at multiple resolutions rather than forced into one cell size. This should extend `spatial_resolution_frontier.json`, not bypass it.
- **Continuous or state-space dynamics.** Add continuous, multiscale, or particle-filter dynamics only as regenerated artifacts with the same null hierarchy, runtime surface, and interpretability audit used by the current gates.

6.3 State-of-the-Art Gates Remain Uncleared

The current status is not a state-of-the-art performance claim. The first stronger-baseline rung, the lane-resolved matrix, the factorial pymdp comparison in sec. 15.1, and the promoted forward gates mostly narrow the claim space. A latent-regime hidden Markov model overfits out of sample; the dependency-light neural sequence model does not beat the particle/state-space mixture; and learned Dirichlet or two-level hierarchical priors are refuted, with only a marginal persistence-rate survivor whose paired bootstrap interval (-0.016 to -0.001) excludes zero but does not clear the pre-registered promotion threshold, so its verdict remains `honest_regime_null`. Those negative gates remain claim blockers in the artifact surface, so a future claim cannot quietly resurrect a refuted direction.

First principles make the future bar simple: a method would need to assign a better predictive distribution to the next sparse waypoint cell on a locked task. Lower variational free energy, richer expected-free-energy diagnostics, or a more legible policy posterior can make the model easier to audit, but those surfaces do not by themselves establish predictive leadership. The primary future contract should therefore be prequential next-cell loss under sparse reports, with calibration, Brier/ECE, persistence-relative gap filling, runtime, and minimization or integrity regressions carried as secondary gates. The artifacts that should eventually encode this contract are `sota_task_contract.json`, `sota_benchmark_matrix.json`, `sota_leaderboard.csv`, and `sota_claim_gate.json`.

The canonical fixture locked-split contract partitions travellers disjointly into train and holdout so no traveller identity crosses the split, scores the transparent baseline ladder under prequential next-cell loss with calibration, runtime, location-minimization, and integrity columns, and emits all four artifacts above. The claim gate currently reports `sota_not_cleared`: no implemented candidate beats the strongest honest baseline, whose mean predictive loss is 0.606 nats on the locked split, and the required foundation-model and language-model baseline families are declared but unavailable, so performance-leading language stays blocked. The false-claim oracle makes this guard legible rather than asserted: a leakage probe that peeks at the true next cell reaches mean loss 0.000 nats — far below every honest baseline — yet is excluded by the leakage guard, so a leakage-driven win cannot clear the gate. The same gate rejects a fixture-only win, a single-metric win that loses on calibration, and any win that depends on direct location or fails the poisoning-integrity diagnostics. A positive control confirms the gate is not vacuous: a clean, calibrated, externally confirmed win with no minimization or integrity regression would clear it, which is exactly the condition no current method meets. This matrix remains a leakage and null guard; it is not the direct active-inference comparison because its rows do not score the active-inference candidate on that same expanded program surface.

A higher-power, robustness-expanded instance of this contract sharpens the verdict rather than softening it. The locked benchmark was re-run in the partial-observability regime with 64 travellers (26 held out) across 4 canonical lanes (`fractalrabbit`, `gravity`, `reporting_gap`, `regime_stressor`), so that every candidate faces the same traveller-disjoint split, known noisy-emission channel, train-only transition source, and traveller-clustered interval. The stronger sequence and state-space candidates — a switching-noise HMM and a particle filter — are scored directly on the locked split rather than marked ineligible, and the roster now includes dependency-light recency and multi-horizon attention sequence baselines alongside the improved active-inference candidate (a precision-weighted belief with a calibrated return prior selected on a train-internal split only). fig. 12 is the direct FractalRabbit-lane comparison: it places `precision_return_aif` and the `soft_filter` active-inference reference on the same locked-split surface as switching HMM, particle filtering, neural and attention comparators, Markov variants, persistence, and base rate. The gate stays `sota_program_not_cleared` (clears: no). On the FractalRabbit lane, the active-inference row is 1.139 nats against the strongest fair baseline `switching_noise_hmm` at 1.144 nats, with paired clustered interval [-0.002, 0.012] nats; on the gravity lane it is 3.030 nats against `online_base_rate` at 2.771 nats. The reporting-gap lane is scored: yes, and the larger regime-stressor lane is scored: yes. A separate diagnostic robustness grid now scores high-motion adversarial missingness and an independent repeated regime-stressor realization through the same candidate interface, but it is an audit surface, not an expanded claim gate. The bounded reading is set by the artifact itself: Reporting-gap and regime-stressor lanes are scored synthetic robustness evidence, not empirical transfer, operational readiness, or SOTA clearance. The oracle-fed reference is excluded as a ceiling rather than a competitor, no baseline was weakened to manufacture the comparison, and the candidate was calibrated on training travellers only — so the blocked gate is the evidence speaking, not a missing implementation.

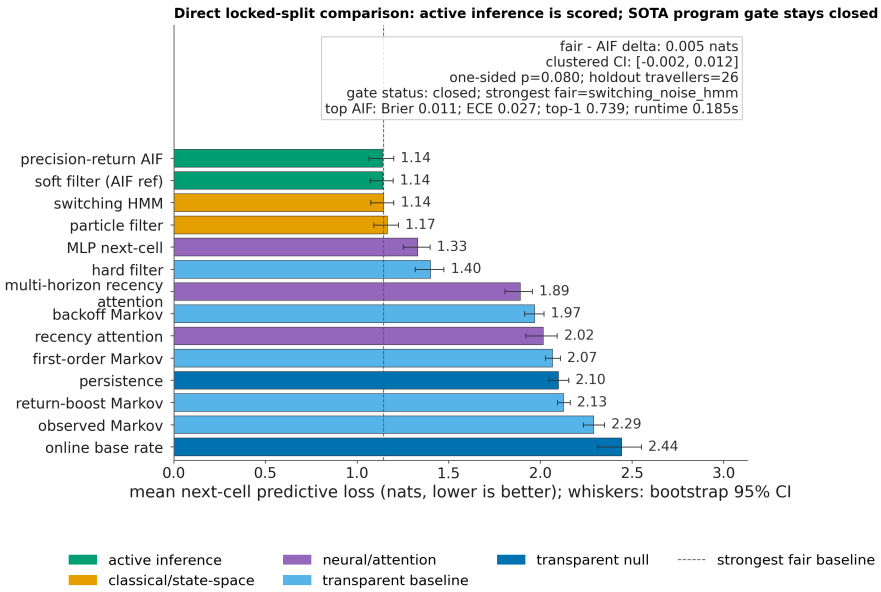


Figure 12: Source sota_program_leaderboard.csv ranks every locked-split synthetic FractalRabbit row by mean next-cell predictive loss. Bootstrap whiskers share one loss axis; the annotation adds Brier score, ECE, top-1 accuracy, runtime, paired delta, one-sided p-value, and holdout-traveller count. Precision-return AIF and soft-filter AIF are scored beside switching HMM, particle, neural, attention, Markov, persistence, and base-rate comparators. The comparison is direct: active inference is scored directly, and the strongest fair baseline and blocked program gate remain visible. Evidence: a direct locked-split SOTA-program comparison that includes active inference and the strongest fair baselines on the same FractalRabbit lane. Boundary: a global state-of-the-art performance claim, operational tasking, or real-mobility validity.

A separate regime-qualified active-inference search asks the narrower question the global gate is not allowed to answer: are there specific synthetic latent-emission regimes where the best active-inference family candidate beats the strongest implemented non-AIF fair baseline under the same locked split, train-only transition source, known observation channel, and traveller-clustered CI rule? fig. 13 maps that local gate by regime and realization seed. Its status is `regime_qualified_sota_not_cleared` (clears: no; cleared regimes: 0). The artifact's cleared-regime fields hydrate as: best cleared regime `none`; best AIF candidate(s) unavailable; strongest fair baseline unavailable; minimum advantage (nats) unavailable; conservative CI edge (nats) unavailable. When no regime clears, those fields render their

explicit `none`/unavailable sentinels — the bounded null — rather than numbers. Either way the boundary is fixed by the artifact: This gate is separate from the canonical global SOTA program. A positive regime result is limited to the exact synthetic regime IDs, realization seeds, candidate roster, and observation-channel contract in this artifact; it does not imply external, empirical, operational, surveillance, foundation-model, or global SOTA performance.

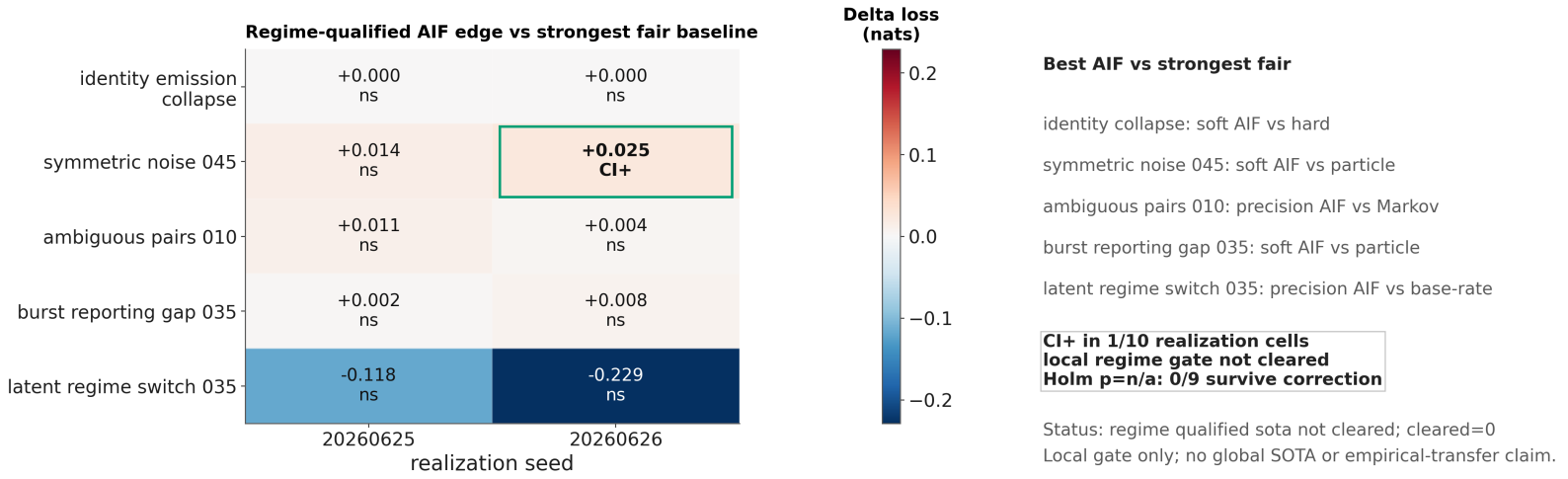


Figure 13: Source aif_regime_sota_gate.json maps the separate regime-qualified synthetic search for active-inference advantage. Each cell is loss(strongest fair baseline) minus loss(best AIF-family candidate) for one synthetic regime and realization seed on a common signed scale; positive values favor AIF. Green rings mark traveller-clustered CI-separated local wins that also clear the oracle/leakage floor, while the side panel reports p/max-CI clearance and the Holm-adjusted familywise failure from multiplicity_ledger.json. Evidence: a regime-local synthetic search for where AIF-family belief marginalization beats the strongest implemented fair baseline with clustered-CI separation. Boundary: global SOTA, external simulator SOTA, empirical transfer, operational readiness, or real-mobility validation.

The baseline map also needs to become wider than the current fixture gate. Future comparison work should re-check the mobility-prediction literature at implementation time rather than freezing this manuscript's date-stamped view of the field. Useful seed families include deep-learning mobility surveys, memory-and-competition mobility models, pretrained and continual-learning mobility transformers, universal mobility predictors, self-supervised disentangled next-POI models, sparse transformer-style trajectory predictors, GPT-style mobility generation, trajectory-LLM taxonomies, spatially aware mixture-of-experts foundation models, instruction-tuned and open-source mobility language models, intention-guided and coordinate-regression next-location systems, training-free LLM-agent evidence-gathering predictors, LLM-agent synthetic-trajectory generators, reversible trajectory-to-CNN generators, systematic synthetic-urban-mobility generator taxonomies, large-scale diffusion-style synthetic mobility datasets, aggregated-input privacy-preserving daily-trajectory synthesis, cross-city LLM-guided mobility generators, train/test-overlap-aware benchmark protocols, HuMob/GISCUP-style sparse-challenge systems, higher-order path-validity checks, semantic mobility question-answering benchmarks, plausibility-versus-realism audits for LLM-based urban simulators, and task-based synthetic-mobility privacy/utility evaluations [Luca et al., 2021, Yan et al., 2017, Wu et al., 2024, Haydari et al., 2024, Gao et al., 2022, Xu et al., 2025, Shi et al., 2026, Long et al., 2024, Solatorio, 2023, Yuan et al., 2025, Han et al., 2025, Tang et al., 2024, Qin et al., 2026, Liu et al., 2026b, 2024, Chen et al., 2026, Li et al., 2026, Merhi et al., 2024, Kapp et al., 2023, Yuan et al., 2026, Ozaki et al., 2025, Liu et al., 2026a, Luca et al., 2023, Fang et al., 2025, LaRock et al., 2026, Asano et al., 2025, Santos et al., 2026, Deng et al., 2025, Wu et al., 2025, Kapp and Mihaljevic, 2023, Mishra et al., 2025]. Inside this project, those references are not evidence that the current scaffold is competitive, privacy-preserving, realistic, or useful in real-life downstream tasks; they define external baseline categories, leakage/overlap checks, synthetic-validity checks, semantic-evaluation gaps, and task-utility/privacy checks that any future state-of-the-art or release-utility claim would need to face.

The remaining forward positives stay bounded. Cross-generator transfer shows persistence strength is a self-transition-rate regime property across synthetic generators, while the expanded held-out reporting-process gate demotes the earlier burst-cadence survivor to an honest null. What the study still lacks is consented empirical traces under the separate validation protocol, plus enough repeated external and stressor realizations to stress the negative gates fairly. The fair-ladder external-realization manifest records this block explicitly: fixture language cannot be promoted until every required current external realization is present and agrees on the oracle-boundary controls. Each future candidate must report predictive loss, calibrated recovery, persistence-null skill, runtime, interpretability surface, failure modes, locked-split performance, and current-lane external evidence. Only a method that improves performance while preserving the claim-level null discipline should be described as performance-leading.

The claim guard is stricter for national-security-adjacent language. Synthetic performance, even if improved, cannot unlock real waypoint, deployment, surveillance-readiness, or operational tasking claims. Those remain blocked by the empirical-validation protocol, consent and provenance records, minimization endpoints, privacy review, and a non-operational publication plan.

6.4 AIF Deepening Must Admit Predictive Candidates

The active-inference lane remains under the prototype-before-claim discipline. Two current methodology surfaces bound future work. First, the generative model is an auditable artifact rather than a side effect of construction code: a machine-checkable specification (the generative model card) states the factor and observation structure, the dependency graph, the column-stochastic normalization invariants, and the conditional-independence approximation, and a conformance validator confirms that the implemented tensors are the model the manuscript describes. Second, the location-minimization question is answered in the free-energy machinery's own currency: `metadata_recovery_not_above_structure_null`, with location share 1.000, metadata null p-value 0.417, and residual localization 0. The oversight-positive reading, scoped to this generative model and offered as neither a universal claim nor a collection-priority recommendation, is narrower after the 4x4 expanded fixture: the artifact verdict decides whether metadata clears the structure null, while withholding location still sharply reduces localizability. Real mobility may couple metadata and place far more strongly, so the result does not transfer without the empirical protocol.

Current active-inference scholarship sharpens that discipline rather than relaxing it. Scalable POMDP planning variants such as Active Inference Tree Search, sparse-reward/world-model R-AIF agents, and formal-equivalence reviews of active inference all motivate richer future candidates; none is a scored row here, none makes the present proxy canonical, and none upgrades a local synthetic regime result into global SOTA language [Maisto et al., 2025, Nguyen et al., 2024, Sweeney et al., 2026].

Two further active-inference levers were prototyped and returned honest nulls. Multi-step expected-free-energy planning does not improve recovery: when the cell is observed the chosen control is moot, and in the location-withheld regime any apparent gain is not robust and does not survive a smoothing sweep — a synthetic control where multi-step planning provably should help confirms the machinery works, so the result is a genuine null rather than a power failure. Policy precision is performance-inert under deterministic selection, and an apparent gain under stochastic selection is explained away by a random-action confound rather than by precision itself; precision is retained only as an interpretive readout, never to sharpen any location inference.

The lane was then made eligible on the same locked split the transparent baselines use, by scoring the POMDP filter's prequential next-cell loss under a traveller-disjoint train/holdout split. The measured answers that follow from that eligibility — the fair-comparison ladder (how much principled latent-cell inference helps once every predictor shares the transition source, and how that help splits between single-step soft marginalization and recursion) and the traveller×noise regime

map that bounds where the advantage lives — are reported in the Discussion (sec. 4.2 and sec. 4.3, with fig. 10 and fig. 11). What remains genuinely forward here is everything those results do *not* settle: a cross-vendor audit of the single-vendor result, more held-out travellers and more strongly regime-structured synthetic data to interval the advantage, and the multiscale and continuous-dynamics extensions below.

The noise-conditional reading of those results is where an earlier session overclaimed, via the retracted observation-noise frontier described above. The corrected settlement uses soft versus hard predictors that differ only in marginalize-versus-commit under an emission matched to the generator by construction, so no knob *can* be asymmetric; the protocol is specified in sec. 2.6.2 and the result, with its robustness band, is reported as a measured result in sec. 3.1.1 (fig. 3). It remains a synthetic software measurement that does not beat the oracle-fed baseline and is not a real-mobility claim, and the benchmark gate is unchanged.

The completed online-adaptation ablation and its remaining fair-baseline deficit are reported in sec. 4.1.1. What remains forward is narrower: admit factorial or multiscale candidates only after they implement `Candidate.predict(PredictContext)`, use the same nested estimation contract, and plausibly address a measured residual rather than adding audit-only complexity. EFE, VFE, and tensor surfaces cannot substitute for locked-split predictive improvement.

6.5 Real-Trace Claims Require a Separate Empirical Gate

The empirical-validation protocol remains separate (`docs/empirical_validation_protocol.md`). It is the prerequisite for any real waypoint-data claim, requiring consent, provenance, sampling, minimization endpoints, and privacy review defined independently of this scaffold [Patterson et al., 2017, Deng et al., 2025, de Montjoye et al., 2013]. The synthetic numbers here settle nothing about real mobility and are not a step toward an operational system.

Throughout, the boundary is unchanged: these are synthetic-data directions on the pinned FractalRabbit generator, oversight-positive and location-minimization-framed, with no operational surveillance, tasking, or evasion surface. Each item is a falsifiable question with a kill-gate, not an intended capability claim.

7 Conclusion: A Synthetic Benchmark Can Clarify Without Promoting

7.1 Recovery and Minimization Stay Null-Bounded

The signature contribution of Active FractalRabbit: A Synthetic Benchmark for Belief Filtering Under Sparse Waypoint Observations is a claim-bounded benchmark discipline for the public NSA-hosted FractalRabbit synthetic generator. The discipline rests on a hierarchy of synthetic nulls: uniform chance, base-rate prediction, and temporal persistence handed the true previous cell (sec. 3.2.1). That last null is an oracle baseline, and the central methodological lesson is an information-asymmetry bound: the location-blind filter is scored with less state information than the baseline receives for free. This bound is reported in Results as a synthetic-regime fact, not reframed as an artifact or limitation to bypass.

The same discipline applies across three front-line tests. Routine-break detection clears its synthetic positive controls, with VFE surprise AUC 0.649 for rare-cell breaks and 0.667 for random-cell breaks. Structure recovery from non-location metadata remains regime-contingent, depending on learned transition structure rather than on metadata alone. Mission-conditioned collection advances from prose to executable comparison on the expanded fixture, not to an operational workflow.

The minimization frontier gives the main synthesis: with all modalities the filter holds mean mass 0.995 on the true cell, but withholding location collapses it to 0.069, an oversight-positive statement of what location minimization protects on this generator. Its utility-side dual (sec. 15.4, fig. 48) now reports a monotone synthetic frontier on the larger canonical fixture: the endpoint pair moves from 3.933 bits and 0.389 risk to 1.747 bits and 0 risk, with a minimum-collection operating point at grid_2x2.

That operating point is not a privacy guarantee; it is a generated-task frontier that would need separate k-anonymity, differential-privacy, legal, and empirical utility analyses before any real-data use [Sweeney, 2002, Dwork, 2006, Kapp and Mihaljevic, 2023].

The spatial-resolution result follows the same rule. The default 4x4 grid remains auditable because all 16 default cells are occupied, but the frontier now reports `default_resolution_not_best_feasible_loss`: the default is not the best feasible observed-location loss in the canonical run, and its full-location loss changes by 0.867 nats versus the coarse comparator. A deterministic synthetic sample-size ladder (2, 4, 8 times the canonical reports per traveller; largest n=19200) reports `larger_sample_changes_frontier_decision` at the largest run, with selected bins 4, strict feasible candidate count 0, and default-vs-coarse loss change -1.069; that ladder refines the observed-location fit verdict rather than turning it into population replication. The no-location guardrail moves by -0.045 in the canonical run and -0.036 in the largest check, with guardrail stability yes across the ladder, so the paper does not claim improved fine-cell recovery from metadata-only evidence.

7.2 Roadmap Advances Still Leave Performance Gates Closed

The new roadmap artifacts advance the paper from "future work" to executable research gates. The spatial-resolution frontier reports default bins 4, best feasible bins 3, and pymdp sentinel agreement 3.20e-07. The return-structure gate reports 0.186 nats of move-step entropy reduction and verdict `promote_return_conditioned_prior`. Cross-realization bands place gap-filling skill between 0.219 and 0.268 across synthetic draws. The fractal-dimension sweep reports status `fixture_control_unvaried`; the circadian-phase analysis reports verdict `honest_null`; the sequence-baseline gate reports best model `first_order_markov` and gate `stronger_sequence_baseline_beats_pymdp_on_this_run`; and the lane-resolved model-gate matrix reports status `partial`.

The promoted forward gates narrow the next-performance story. Latent-regime HMM is `latent_regime_hmm_does_not_beat_markov_null_em_overfits`; neural sequence is `neural_sequence_beats_weak_nulls_not_particle_mixture`; learned stay-rate is `honest_regime_null`; held-out reporting is `honest_null`; and cross-generator persistence is `regime_dependent_on_self_transition_rate`. The safety-boundary gates report `metadata_recovery_not_above_structure_null`, `location_carries_relink_risk_that_minimization_removes`, and `poisoning_detection_regime_changed`. The future roadmap for moving this toward state-of-the-art performance is now narrower and more concrete: current-run external evidence for the fixture positives, multiscale or continuous dynamics, and consented empirical validation under the separate protocol.

The benchmark's clearest answer concerns noisy partial observability. Holding model, emission, observations, and priors fixed while varying marginalize-versus-commit, soft Bayesian marginalization ties the point estimate when the latent cell is directly observed and gains strictly as the emission degrades (1.860 nats per unit ε, $R^2 = 0.995$). Every one of the 15 resamplings is positive at every nonzero noise level.

Active inference realizes this mechanism through recursive belief updating. In the full noisy-emission comparison it ranks 1 of 14, leads the best point-estimate/raw-observation family by the belief-cluster gap of 0.192 nats, and leads the strongest non-AIF belief filter by 0.005 nats with a clustered interval crossing zero. Active inference is therefore the implemented point leader and a statistical member of the top belief-preserving cluster. HMM, particle, state-space, and soft-filter methods share the core marginalization mechanism.

The wider model map completes the answer. Under latent regime switching, online transition adaptation improves fixed AIF by 0.146 nats on average, while the adaptive online base-rate comparator remains strongest. Under reporting gaps, persistence and disclosed kinematic controls anchor the comparison (sec. 14.2). With location withheld, occupancy and trained dynamics explain broad structure while specific-cell recovery collapses. The practical rule is conditional: commit when the state is effectively observed, marginalize when it is genuinely hidden, and adapt when the generating regime changes.

The oracle-fed baseline remains the predictive ceiling. The higher-power, 4-lane program scores 64 travellers across `fractalrabbit`, `gravity`, `reporting_gap`, `regime_stressor` and reports `sota_program_not_cleared`. The present contribution is therefore a regime-specific synthetic model result: Active Inference leads the implemented noisy hidden-state lane, shares that lead statistically with the strongest non-AIF belief filter, and supplies an auditable variational account of the uncertainty that produces the gain.

7.3 pymdp Adds Interpretability Rather Than Performance Leadership

The `pymdp` lane provides a validated variational active-inference implementation. It agrees with the exact Bayes posterior under matched priors to maximum absolute difference 3.99e-07 and exposes variational free energy, policy posterior, expected free energy per policy, runtime, and action-conditioned prior divergence. Relative to the transparent reference filter, the benchmark matrix reports status `not_demonstrated`, predictive-loss delta 0.428 nats, and posterior-entropy delta 0.009 nats. Its established contribution on the default observed-state surface is therefore interpretability; its predictive advantage appears in the separate noisy hidden-state comparison above.

7.4 The Final Boundary Is an Honesty Stack

Holding the whole arc together is a two-tier honesty stack. The execution tier supplies checked fixtures, negative controls, holdout scoring, a checksummed release manifest, tokenized quantities, and source-artifact provenance. The claim tier uses five guards (G1-G5) to bind every performance statement to its regime, information set, cue disclosure, and evidence class [Sandve et al., 2013, Wilson et al., 2017, Smith et al., 2016].

Active FractalRabbit: A Synthetic Benchmark for Belief Filtering Under Sparse Waypoint Observations therefore delivers an auditable synthetic model map: Active Inference records its clearest predictive result under noisy hidden-state observations; simpler predictors lead when the state is clean or the generator changes faster than the fixed model; and location minimization materially constrains recovery. That map is the basis for the next empirical contract.

8 Supplementary S1 - Validation Gates Preserve Synthetic-Only Claims

8.1 Tests Protect Artifacts Before Prose Claims

The validation strategy is artifact-first: every manuscript-facing claim is meant to remain tied to retained inputs, scripts, generated artifacts, and software identifiers rather than to an unrecorded analyst session [Sandve et al., 2013, Wilson et al., 2017, Smith et al., 2016]. The synthetic validation suite uses unit tests to validate config parsing, FractalRabbit parameter ordering, CSV parsing, first-observation sentinel handling, observation determinism, stochastic-matrix normalization, JAX-compatible tensor shapes, policy-proxy component sums, fixed-transition baseline trace emission, holdout split and scoring behavior, seeded negative-control mutations, candidate-score ordering, figure registry completeness, PNG integrity, release-manifest checksums, manuscript token coverage, fractal-dimension sweep summaries, circadian-phase signal formulas, sequence-baseline loss formulas, reporting-process log-loss formulas, and the `pymdp` benchmark matrix. Fixture tests run without Java or Maven; the external FractalRabbit integration test is explicitly marked and skips with an explicit reason naming the missing tool or Java version.

8.2 Guardrails Block Claim Drift and Evidence Laundering

The primary failure modes are concrete:

Table 3: Validation threat model for the synthetic waypoint analysis.

risk	guardrail
FractalRabbit source drift	pinned commit `9933449c4f4fe1b26b6ac7bfdeeac76583085df5` and config validation
CSV contract drift	parser tests for `ID`, `Days`, `x(km)`, `y(km)`
observation category drift	deterministic feature-table tests
invalid probability tensors	column-sum tests for eq. 1 and eq. 2
ambiguous temporal prior	explicit eq. 4 and trace-level baseline comparison
selector opacity	emitted tbl. 5 and deterministic tie-breaks
manuscript stale numbers	token-generation gate and rendered-output checks
manuscript stale source artifacts	token provenance tuples from eq. 30 with checksum and value matching
pipeline insensitive to destroyed structure	seeded negative-control lane with directional probes (fig. 14, tbl. 7)
in-sample-only scoring	additive per-traveller chronological holdout metrics in `metrics.json`
roadmap artifact overpromotion	fractal-dimension, circadian-phase, sequence-baseline, reporting-process, forward-gate, and `pymdp` benchmark artifacts carry explicit status and caveats
claim-ledger stale artifacts	`scripts/12_check_claim_ledger_artifacts.py` parses generated `output/` paths cited in `docs/claim_ledger.md` and fails on missing artifacts
release bundle drift	checksummed `release_manifest.json` over configs, fixtures, run artifacts, variables, figures, and rendered PDF

8.3 Real Traces Require Empirical and Privacy Review

The red-team boundary is also explicit. A persuasive plot is not sufficient evidence for real-world mobility inference. The project therefore treats fig. 17 through fig. 26 as software diagnostics over generated or fixture data. Any future empirical study must introduce separate consent, provenance, sampling, privacy, and external-validity checks before reusing the manuscript language for real traces, especially because human mobility data can remain re-identifiable after coarse spatiotemporal reduction and can be vulnerable to cross-dataset trace matching [de Montjoye et al., 2013, Kondor et al., 2020, Mishra et al., 2025]. Formal privacy guarantees are separate mechanism claims, not something produced by this benchmark automatically [Dwork, 2006]. Synthetic trajectory generation also should not be treated as a privacy guarantee by itself without explicit privacy evaluation: recent trajectory-generator work separates utility from adversarial privacy testing, studies memorization mitigation in diffusion generators, and shows that differential privacy can impose measurable utility costs under constrained mechanisms [Wang et al., 2023, Buchholz et al., 2024, Bouras et al., 2026, Cherigui et al., 2026, Guepin et al., 2026, Buchholz et al., 2025, Yadav et al., 2024]. Cross-domain synthetic-data metric reviews reinforce that privacy and utility must be evaluated separately and that untested privacy dimensions should remain out of claim scope rather than inferred from utility or resemblance [Kaabachi et al., 2025]. The manuscript's "minimization" language follows the data-minimization principle and privacy-by-design posture codified in data-protection law and risk-management guidance [European Parliament and Council, 2016, National Institute of Standards and Technology, 2020], but its minimization and linkage numbers are synthetic risk *proxies* on generated traces, not compliance determinations, formal guarantees, or differentially private releases. AI risk, cyber-governance, responsible-AI, commercially available information, and location-data enforcement sources extend that policy framing only: they support provenance, minimization, oversight, and responsible interpretation language, not technical validity, legal compliance, operational readiness, or a real-trace workflow [National Institute of Standards and Technology, 2023, 2024, U.S. Department of Defense, 2021, Office of the Director of National Intelligence, 2022, 2024, Federal Trade Commission, 2024, 2026].

8.4 Risk and Integrity Gates Stay Synthetic Proxies

Three defensive, oversight-positive analyses sharpen the privacy and integrity boundary, each gated with the project's structure-null plus positive-control discipline so that a reassuring number means "the synthetic proxy passed its structure-null and positive-control check", not "the test was too weak". The EFE location frontier in fig. 43 reports verdict `metadata_recovery_not_above_structure_null`: location carries share 1.000 of the joint epistemic value, the metadata null

p-value is 0.417, and the residual fraction localized without location is 0. That artifact is a minimization boundary for this synthetic generator: the current verdict determines whether the metadata residual clears the structure null, while withholding location still collapses localization and is not a collection-priority recommendation.

The linkage-risk frontier in fig. 44 reports verdict `location_carries_relink_risk_that_minimization_removes`. Full-location linkage is 1, while the no-location regime exceeds its null band: no. The statistic is an aggregate per-regime risk number over synthetic identities, computed to show what data-minimization buys back; it is not, and does not yield, a re-identification, surveillance, or tasking capability over any real person. The boundary is gated in code: the data-loading entry point derives and validates a synthetic-generator provenance marker, so running on real data requires a deliberate, auditable misrepresentation rather than an accident.

The poisoning gate in fig. 45 asks whether the analyst's own active-inference generative model notices tampered input. Its verdict is `poisoning_detection_regime_changed`: gross teleport poisoning detected no, subtle-drift null versus Markov yes, and the teleport mid-rate VFE AUC is 0.636. The defensive reading is verdict-bound: the 4x4 artifact reports which crude stressors separate from the Markov baseline in this configuration, not a general label-flip or subtle-drift assurance. The stressors are integrity probes, never an attack technique, and no operational capability is produced.

8.5 The Claim Contract Catches Manuscript Overreach

FractalRabbit exists so that sparse-mobility-analytics algorithms can be developed and compared on a controlled, reproducible, ground-truth-bearing synthetic testbed without real traces [Darling, 2018, National Security Agency, 2026a]. The central methodological hazard for any such benchmark is the one this project was built to detect: a number that looks like recovered structure but is an artifact of the wrong null, the wrong regime, a self-selected coverage subset, or an undisclosed cue.

We therefore promote the project's honesty discipline from prose to a machine-checkable *evaluation contract* (`src/active_fractal_rabbit/eval_contract.py`, enforced by `tests/test_eval_contract.py`). It is a static linter over the manuscript and its token bindings, encoding the discipline as falsifiable claim-level guards: a null-regime guard (G1) that flags any sentence asserting positive skill against the persistence null while citing a location-blind token — the precise error this manuscript itself avoids, since persistence receives the true previous cell and the location-blind filter does not; a cue-disclosure guard (G2) that requires any recovery claim leaning on the speed cue to disclose that speed is derived from inter-report displacement; a performance-claim guard (G3) that requires state-of-the-art or performance-leading wording to be explicitly negated, future-scoped, or bound to a cleared gate-status token; an active-inference-help guard (G4) that requires active-inference help or uniqueness wording to be negated or bounded to its measured regime and comparator family; and a deployment/surveillance-claim guard (G5) that requires deployment, surveillance-readiness, or real-mobility-transfer wording to be explicitly negated, so synthetic progress cannot unlock that language.

The contract is the same negative-control philosophy that gates the pipeline, lifted to the level of the *claim*: it complements, and does not replace, the execution-level harness (negative controls, full-trajectory scoring, the binding gate, the no-hardcoded-numbers guard) that catches honesty failures visible only by running the analysis.

We apply the project's own prototype-before-claim rule to the contract itself, and report its reach honestly rather than overstating it: of this project's documented honesty failures, only the wrong-regime class was statically catchable from prose and token bindings, while the others — a learned model that underperformed its own baseline, headline numbers that did not reproduce on real data, a fabricated library call, a scoring artifact visible only on full re-scoring, and external simulator and render-readiness failures that only appear when the full artifact pipeline runs — were caught only by executing the analysis. The static contract is therefore the smaller, claim-level tier of a two-tier honesty stack, and it says so. It ingests nothing but this project's own manuscript and artifacts — a reusable, oversight-positive evaluation layer rather than an instrument over any real data.

8.6 Token Provenance Runs Before Rendering

The execution tier now includes fail-loud manuscript-variable provenance. `scripts/11_check_manuscript_variable_provenance.py` rebuilds the selected token map, verifies that every result-critical source artifact and JSON path exists, checks the source checksum, and byte-compares the formatted artifact value to the rendered token. Missing optional external quantities are allowed only as explicit unavailable text; silent zeros are not an acceptable fallback. This is narrower than full semantic review, but it closes the stale-token class before the PDF render is trusted.

9 Supplementary S2 - Reproducibility Runs Through Artifacts and Render Checks

9.1 Fixture-First Commands Rebuild the Study Surface

The reproducibility design is artifact-first and executable, following computational-research and software-citation guidance that favors versioned inputs, retained intermediates, scriptable workflows, plain-text project structure, and specific software identifiers [Sandve et al., 2013, Wilson et al., 2017, Smith et al., 2016]. The default reproducibility lane is fixture-first:

```
uv run python scripts/00_preflight.py
uv run python scripts/02_run_pipeline.py --config configs/default.yaml
uv run python scripts/03_run_sweep.py --config configs/sweep.yaml
uv run python scripts/05_run_negative_controls.py --config configs/default.yaml
uv run python scripts/08_run_sparsity_analysis.py --config configs/default.yaml
uv run python scripts/04_generate_figures.py --run output/runs/default
uv run python scripts/14_check_visual_quality.py
uv run python scripts/20_run_aif_regime_search.py
uv run python scripts/z_generate_manuscript_variables.py
uv run python scripts/11_check_manuscript_variable_provenance.py
uv run python scripts/15_check_claim_audit.py
uv run python scripts/16_check_visual_claim_audit.py
uv run python scripts/12_check_claim_ledger_artifacts.py
uv run pytest tests/ --cov=src --cov-fail-under=90

# or run both lanes end to end (external lane runs when JDK 21 + Maven resolve):
uv run python scripts/07_run_full_study.py --lane both
```

The render lane runs from the sibling template checkout:

```
uv run python -m infrastructure.orchestration link-projects
uv run python scripts/03_render_pdf.py --project working/active_fractal_rabbit

# back in the project checkout, bind the rendered outputs to their exact inputs:
uv run python scripts/29_finalize_publication_package.py
uv run python scripts/29_finalize_publication_package.py --verify-only
```

9.2 Code and Data Availability

The full source tree, checked fixtures, manuscript source, and generated release manifest are public at github.com/ActiveInferenceInstitute/active_fractal_rabbit under the licenses recorded in `LICENSE` (code, MIT) and `LICENSE-DOCS.md` (documentation and scholarly outputs, `CC-BY-4.0`). Third-party notices for the pinned external simulator and installed dependencies are recorded in `THIRD_PARTY_NOTICES.md`. A tagged GitHub release and an archival Zenodo deposit provide a versioned, citable snapshot; citation metadata is recorded in `CITATION.cff` and, once minted, the archival digital object identifier appears on the manuscript cover and in the Zenodo record linked from the repository.

9.3 Inventories and Checksums Bind the Release Surface

The primary generated artifacts are grouped by role:

- **Run traces and tensors:** `observations.csv`, `discrete_features.csv`, `tensor_summary.json`, `posterior_states.csv`, `inference_trace.csv`, `policy_trace.csv`, `baseline_trace.csv`, `baseline_metrics.json`, `candidate_scores.json`, `pymdp_trace.csv`, `ground_truth.csv`, and `provenance_manifest.json`.
- **Fixture, robustness, and diagnostic analyses:** `sensitivity_summary.json`, `sensitivity_summary_external.json`, `negative_controls_summary.json`, `sparsity_analysis.json`, `spatial_resolution_frontier.json`, `anticipatory_analysis.json`, `active_sensing_analysis.json`, `anomaly_analysis.json`, `structure_recovery_analysis.json`, `mission_collection_analysis.json`, `frontier_analysis.json`, `partial_observability_analysis.json`, `return_structure_analysis.json`, `variance_bands.json`, `upstream_sensitivity.json`, `fractal_dimension_sweep.json`, `circadian_phase_analysis.json`, canonical `sequence_baseline_analysis.json`, lane-qualified `sequence_baseline_analysis_*.json`, canonical `missingness_model_analysis.json`, and lane-qualified `missingness_model_analysis_*.json`, `return_phase_prior_analysis*.json`, `model_gate_matrix.json`, `model_conformance.json`, `forward_gates_summary.json`, `latent_regime_hmm.json`, `neural_baseline.json`, `learned_stay_rate.json`, `heldout_reporting_process.json`, and `cross_generator_transfer.json`.
- **Safety, minimization, benchmark, and challenge surfaces:** `efe_location_frontier.json`, `linkage_risk_frontier.json`, `poisoning_robustness.json`, `pymdp_benchmark_matrix.json`, `sota_task_contract.json`, `sota_benchmark_matrix.json`, `sota_leaderboard.csv`, `sota_claim_gate.json`, `sota_program_gate.json`, `sota_program_leaderboard.csv`, `sota_program_robustness_grid.json`, `sota_program_robustness_grid.csv`, `aif_regime_search.csv`, `aif_regime_search.json`, `aif_regime_sota_gate.json`, `aif_sweep.json`, `aif_deepening_candidates.json`, `fair_inference.json`, `fair_inference_external_realizations.json`, `power_analysis.json`, `open_challenge_pack.json`, `multiplicity_ledger.json`, `confound_taxonomy.json`, `model_comparison_leaderboard.json`, `privacy_frontier_analysis.json`, `privacy_utility_frontier.json`, `privacy_utility_frontier_repeats.json`, `synthetic_task_utility_privacy.json`, and `hardened_anomaly_analysis.json`.
- **Manuscript, visual, and release contracts:** `figure_registry.json`, `active_fractal_rabbit_cover.png`, `visual_quality_audit.json`, `claim_audit.json`, `visual_claim_audit.json`, `manuscript_variables.json`, `manuscript_variable_provenance.json`, and `release_manifest.json`.

The manuscript values in sec. 3 are hydrated from these artifacts. If a metric changes, the rendered manuscript changes with it, and the provenance check fails if a token source is missing, stale, incomplete, or byte-mismatched. The release manifest additionally records a SHA-256 checksum and byte size for source modules, scripts, tests, documentation, configs, fixtures, default-run artifacts, cover art, figures, variable files, required release artifacts, and rendered outputs, so a shipped bundle can be verified byte-for-byte and fails when a required or discovered release entry is absent. The resolved manuscript embeds a content-addressed input fingerprint into both PDF and HTML; final verification recomputes that fingerprint, so changing a manuscript or visual input and merely updating output timestamps cannot re-certify an unchanged render. For reviewers who want to rebuild the headline scalars without rerunning the whole simulator, `scripts/z_generate_manuscript_variables.py` rehydrates every manuscript token from the retained run artifacts and `scripts/11_check_manuscript_variable_provenance.py` verifies the source path, JSON key, checksum, and rendered value behind each token.

The visual-quality audit reports `pass` for `53` registered figures. Its minimum rendered extent is `1339 x 724` pixels, its minimum normalized pixel standard deviation is `0.129`, and `53` captions carry source or boundary language. The audit also records aspect-ratio extremes and names the shortest, narrowest, lowest-variance, and widest/tallest-aspect figures so the contact sheet can be manually inspected where static layout risk is highest. This is not an aesthetic proof, but it prevents blank, undersized, caption-orphaned, source-orphaned, or silently hard-to-read figures from entering the manuscript unnoticed.

The figure layer treats visual design as part of the reproducibility contract rather than decoration. Quantitative comparisons use aligned axes, common-position encodings, and generated source tables where possible; comparison patterns are graphed only when the plot clarifies a relation that would be harder to read in a dense table; and uncertainty marks are paired with source artifacts plus explicit non-claim boundaries [Cleveland and McGill, 1984, Gelman et al., 2002, Munzner, 2014, Hullman, 2020, Spiegelhalter et al., 2011]. These sources justify the visual audit's reader-task, data-shape, encoding, and uncertainty checks. They do not make a figure self-validating: the manuscript claim still has to survive the artifact, variable-provenance, claim-ledger, and visual-claim audits.

The claim and visual-claim audits add the RedTeam layer above those mechanical checks. `claim_audit.json` inventories manuscript claims by section and verifies high-risk families against the claim ledger. `visual_claim_audit.json` records a mini-brief for every registered figure plus the cover, checking story job, reader task, source artifact, data shape, encoding, caption support, and explicit non-claim boundary before render.

Table 4: Figure registry sources for the default generated figure set.

figure key	source artifact
`evidence_atlas`	`output/runs/default/metrics.json`
`waypoint_paths`	`output/runs/default/observations.csv`
`spatial_bins`	`output/runs/default/discrete_features.csv`
`spatial_resolution_frontier`	`output/runs/spatial_resolution_frontier.json`
`reporting_bursts`	`output/runs/default/discrete_features.csv`
`speed_categories`	`output/runs/default/discrete_features.csv`
`transition_heatmap`	`output/runs/default/discrete_features.csv`
`likelihood_panels`	`output/runs/default/discrete_features.csv`
`posterior_entropy`	`output/runs/default/inference_trace.csv`
`policy_proxy`	`output/runs/default/policy_trace.csv`
`candidate_scores`	`output/runs/default/candidate_scores.json`
`negative_controls`	`output/runs/negative_controls_summary.json`
`pymdp_agreement`	`output/runs/default/pymdp_trace.csv`
`pymdp_diagnostics`	`output/runs/default/pymdp_trace.csv`
`state_recovery`	`output/runs/default/ground_truth.csv`
`sparsity_recovery`	`output/runs/sparsity_analysis.json`
`calibration`	`output/runs/default/calibration.json`
`anticipatory`	`output/runs/anticipatory_analysis.json`
`efe_decomposition`	`output/runs/active_sensing_analysis.json`
`anomaly_detection`	`output/runs/anomaly_analysis.json`
`recoverability_frontier`	`output/runs/frontier_analysis.json`
`return_structure`	`output/runs/return_structure_analysis.json`
`variance_bands`	`output/runs/variance_bands.json`
`upstream_sensitivity`	`output/runs/upstream_sensitivity.json`

figure key	source artifact
fractal_dimension_sweep	output/runs/fractal_dimension_sweep.json
circadian_phase_signal	output/runs/circadian_phase_analysis.json
partial_observability	output/runs/partial_observability_analysis.json
sequence_baselines	output/runs/sequence_baseline_analysis.json
missingness_model	output/runs/missingness_model_analysis.json
model_gate_matrix	output/runs/model_gate_matrix.json
m10_forward_gates	output/runs/forward_gates_summary.json
factorial_pymdp_modes	output/runs/factorial_pymdp_analysis_fixture.json
latent_mode_recovery	output/runs/factorial_pymdp_analysis_fixture.json
factorial_model_gate_matrix	output/runs/factorial_pymdp_analysis_fixture.json
efe_location_frontier	output/runs/efe_location_frontier.json
linkage_risk_frontier	output/runs/linkage_risk_frontier.json
poisoning_robustness	output/runs/poisoning_robustness.json
privacy_frontier	output/runs/privacy_frontier_analysis.json
hardened_anomaly_detection	output/runs/hardened_anomaly_analysis.json
sota_leaderboard	output/runs/sota_program_leaderboard.csv
aif_regime_map	output/runs/aif_regime_sota_gate.json
fair_inference	output/runs/fair_inference.json
fair_power_regime	output/runs/power_analysis.json
multiplicity_ledger	output/runs/multiplicity_ledger.json
confound_taxonomy	output/runs/confound_taxonomy.json
regime_mechanism_synthesis	output/runs/regime_mechanism_synthesis.json
regime_switch_phase_diagnostic	output/runs/regime_switch_phase_diagnostic.json
privacy_utility_frontier	output/runs/privacy_utility_frontier.json
privacy_utility_repeats	output/runs/privacy_utility_frontier_repeats.json
synthetic_task_utility_privacy	output/runs/synthetic_task_utility_privacy.json
model_comparison_under_noise	output/runs/model_comparison_leaderboard.json
aif_deepening_candidates	output/runs/aif_deepening_candidates.json
dataflow	source pipeline modules

9.4 The External Toolchain Boundary Stays Stochastic and Pinned

External FractalRabbit runs require the toolchain boundary documented by preflight: Maven plus a JDK 21 runtime, which the boundary layer resolves automatically from the environment, PATH, Homebrew keg locations, or `java_home`. The external integration tests skip with explicit named reasons when that resolution fails, and run for real when it succeeds. Because the upstream simulator exposes no seed, external waypoint outputs are stochastic; the release manifest checksums pin the exact artifacts behind any given rendered manuscript.

Render input fingerprint: `f9a806b2a7e62501c1c237224f4b5052`.

10 Supplementary S3 - FractalRabbit Becomes Categorical Evidence, Not Real Mobility

This supplement documents the exogenous FractalRabbit generator contract and the observation-discretization pipeline that converts raw waypoints into the categorical evidence consumed by the main-line inference layer (sec. 2).

10.1 The Pinned Upstream Contract Defines the Fixture Lane

FractalRabbit is treated as an upstream stochastic data generator, not as project-local source. The pinned external lane writes the published parameter rows with labels matching `parameters.csv`, invokes the built jar as `java -cp fractalrabbit-1.0-jar-with-dependencies.jar fractalRabbitGenerator.MainClassFR parameters.csv <output_prefix>` (the simulator appends `XY.csv` and `PLACES.csv` to the prefix rather than taking an output filename), and parses only the documented output record shape [National Security Agency, 2026a,c,b]. Because the upstream repository currently has no release archive, the manuscript cites commit-specific URLs and stores the exact commit in the generated provenance manifest, following software-citation principles of specificity and persistence [Smith et al., 2016]. The fixture lane uses a checked CSV with the same columns, allowing Python tests and manuscript generation to run when Java and Maven are unavailable. The boundary layer resolves a JDK 21 runtime explicitly (environment override, PATH, Homebrew keg locations, then `java_home`) and exports `JAVA_HOME` to the Maven and Java subprocesses, because Homebrew installs `openjdk@21` keg-only. On the current host the external lane status is: completed on this host with the pinned upstream build — an artifact-presence status derived from the retained pinned-build metrics file, not a statement that the lane completed within the current full-study invocation (the current-run gate artifacts report their own external-lane status independently).

10.2 Fractal Dimension Is a Generator Knob, Not a Real-Trajectory Estimate

The simulator-facing parameter vector is:

$$\theta_{\mathrm{FR}} = (d, h, N_{\mathrm{pts}}, N_{\mathrm{trav}}, c, T_{\mathrm{days}}, N_{\mathrm{rep}}), \tag{18}$$

where d is the ambient space dimension, h the fractal dimension, N_{pts} the generated point count, N_{trav} the traveller count, c the co-traveller count, T_{days} the duration in days, and N_{rep} the mean report count. eq. 18 is a wrapper around the upstream CSV surface; it is not a fitted statistical parameter vector in this project.

The dimension parameter obeys the generator-side range:

$$0 < h \leq d, \tag{19}$$

with this scaffold restricted to the upstream two-dimensional waypoint output. We use h as a controlled synthetic roughness knob: lower values concentrate generated support into a thinner set of places, while higher values spread the support through more of the ambient plane. Movement-ecology and human-mobility models make clear why this distinction matters: generator controls, individual movement mechanisms, and observation processes are different objects even when their outputs can be plotted as trajectories [Nathan et al., 2008, Barbosa et al., 2018]. The project therefore asks how downstream inference diagnostics respond when h changes; it does not estimate a box-counting, Hausdorff, or correlation dimension from observed traces, and it does not claim that any generated h matches real movement.

10.3 CSV Parameters Expose Only Part of the Upstream Simulator

The pinned upstream source implements three explicit tiers, each with its own parameters. The retro-preferential tier is useful benchmark context for exploration-and-return mobility ideas, but this project treats it as a simulator boundary rather than a fitted empirical law [Song et al., 2010a, Darling, 2018]. The CSV interface exposes seven of these parameters; the remaining internals are fixed constants in the upstream main class, documented here so the configurability boundary is exact rather than implied:

tier (upstream class)	role	CSV-exposed parameters	upstream-fixed internals
`AgoraphobicPoints`	heavy-tailed fractal place process	space dimension d, fractal dimension h, point count N_{pts}	clump restart rate $\theta = 0.75$
`Retropreferential`	retro-preferential trajectory process over places	traveller count N_{trav}, co-traveller count c	distance exponent -2.0, exploration mean $\phi = 10$, trajectory-length mean 100
`SporadicReporter`	sporadic report-time process along trajectories	duration T_{days}, mean report count N_{rep}	speed bound 50 units/day, 200 km/unit, Pareto cutoff 0.001, tail -1.5

Every CSV-exposed parameter is configurable through `fractalrabbit.params`; the upstream-fixed internals cannot be changed without forking the pinned commit, and this project deliberately does not fork it. The simulator emits two row-aligned output streams per run: the waypoint stream `*XY.csv` consumed as observations, and the ground-truth stream `*PLACES.csv` (`Traveler ID,Days,Place ID`) naming the true `AgoraphobicPoints` place behind every report. The pipeline ingests both: places feed the revisit-structure statistics and the ground-truth state recovery evaluation in sec. 3.

10.4 The Parser Consumes a Narrow Waypoint Record Contract

The parser accepts rows of the form:

$$w_i = (\text{id}_i, t_i, x_i, y_i), \tag{20}$$

with t_i measured in days and positions measured in kilometers. The parser sorts by traveller and time and writes `output/runs/*/observations.csv` as the raw sorted post-parser artifact; the downstream observation-discretization step (sec. 10.5) subsequently validates nonnegative time before any categorical features are derived.

10.5 The Observation Model Converts Waypoints into Finite Evidence

With the simulator boundary fixed, the observation layer turns the raw waypoint stream above into the finite categorical evidence the generative model of sec. 2.6 consumes.

10.5.1 Categorical Features Bound the Evidence Surface

The observation layer maps continuous, irregular waypoint records into finite categorical modalities. The default emitted feature table has columns `ID`, `Days`, `x(km)`, `y(km)`, `is_first_observation`, `x_bin`, `y_bin`, `spatial_state`, `time_bin`, `speed_km_per_day`, `interval_days`, `speed_category`, `burst_category` and is the source for the spatial, temporal, speed, and burst figures in sec. 3.

10.5.2 Spatial States Are Analysis Cells, Not Simulator Internals

For each sorted waypoint, spatial discretization maps coordinates into a finite cell:

$$s_i = b_y(y_i) \cdot G + b_x(x_i), \tag{21}$$

where G is the configured spatial bin count per axis and b_x, b_y are the configured spatial binning functions. The current implementation supports equal-width grids, fixed-bound grids, and empirical quantile bins. The latent state in eq. 21 is a categorical analysis state, not the hidden point identity inside FractalRabbit; because location is also emitted as an observation modality, the default fixture is close to a directly observed-state filter rather than a deep latent-uncertainty benchmark.

10.5.3 Temporal, Kinematic, and Burst Features Remain Heuristic Cues

Time-of-day is represented as:

$$b_t : [0,1) \to \{0, \dots, G_t - 1\}, \qquad \tau_i = b_t(t_i \bmod 1), \tag{22}$$

where G_t is the configured number of time-of-day bins and b_t is the explicit temporal binning map. Within each traveller, inter-report interval and speed are:

$$\Delta t_i = t_i - t_{i-1}, \qquad v_i = \begin{cases} \dfrac{\sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2}}{\Delta t_i} & \Delta t_i > 0, \\ 0 & \Delta t_i \le 0. \end{cases} \tag{23}$$

Non-positive intervals are not floored to a positive denominator; the implementation instead leaves v_i at its zero-initialized sentinel value, distinct from the observation-noise probability in sec. 2.6.2. The first waypoint for each traveller has no previous within-traveller report, so the emitted feature table marks 5 rows with `is_first_observation=1` and assigns the zero interval, zero speed, and non-burst sentinel. This sentinel convention keeps all waypoint rows in the categorical summaries while avoiding an artificial positive interval.

Speed categories are thresholded from v_i, while burst categories are assigned by the configured interval rule. In the default run, the burst rule is `median_positive`, so short intervals are defined relative to the positive-interval median rather than a universal biological or operational threshold. We use "burst" in the temporal-dynamics sense of clustered reporting separated by longer waiting times; the current median-positive rule is a heuristic discretizer, not a full burstiness statistic, and finite event sequences require particular caution before treating a burst score as stable [Barabasi, 2005, Goh and Barabasi, 2008, Kim and Jo, 2016].

10.5.4 Fractal-Dimension Sweeps Shift Occupancy Through the Same Discretizer

The sweep introduced in sec. 10.2 changes the upstream support before this observation layer sees any rows. The downstream state count, occupancy profile, revisit rate, and predictive losses are therefore consequences of the same discretizer applied to generator outputs with different support roughness. This is why the sweep is reported as generator sensitivity rather than a direct fractal-dimension estimator.

11 Supplementary S4 - Formal Diagnostics Stay Proxies and Provenance Checks

This supplement collects the implementation-level formalisms — robustness, privacy-proxy, and provenance contracts, plus the pymdp performance/interpretability boundary — that underpin but are not part of the main-line generative-model argument (sec. 2.6).

The robustness extensions keep the same artifact discipline as the generative model of sec. 2.6. A return-conditioned prior introduces a visited-destination feature

$$r_i = \mathbb{1}\{s_i \in V_{i-1}\}, \qquad V_i = V_{i-1} \cup \{s_i\}, \tag{24}$$

and promotes it only when the move-step conditional entropy reduction

$$\Delta H_{\text{return}} = H(s_i \mid s_{i-1}) - H(s_i \mid s_{i-1}, r_i) \tag{25}$$

clears the predeclared artifact gate. Otherwise the result remains an explicit null rather than a hidden negative result. Cross-realization robustness uses empirical bands over repeated synthetic-generator realizations:

$$\mathcal{B}_\alpha(z) = \left[Q_\alpha(\{z_j\}), Q_{1-\alpha}(\{z_j\})\right], \tag{26}$$

where z_j is a scalar verdict metric from realization j. These bands are descriptive synthetic-realization intervals, not population confidence intervals. They should be read like bootstrap-style uncertainty summaries over generated software runs, not like a guarantee that the same interval would cover a real-world population quantity [Efron and Tibshirani, 1994].

The privacy frontier reports only oversight and minimization quantities. For traveller a, a coarsened signature is

$$\Gamma_a(M) = \{(\tau_i, c_i^M) : \text{traveller}(i) = a\}, \tag{27}$$

where M is a modality-minimization regime and c_i^M is the retained categorical code. Synthetic unicity and co-travel coincidence are then

$$U(M) = \frac{|\{\Gamma_a(M)\}|}{|\{a\}|}, \qquad C(M) = \frac{1}{|\mathcal{P}|} \sum_{(a,b)\in\mathcal{P}} \mathbb{1}\{\Gamma_a(M) \cap \Gamma_b(M) \neq \emptyset\}. \tag{28}$$

The set-valued signature is intentional: it discards multiplicity and ordering to make a coarse oversight proxy, not a full reconstruction attack, k-anonymity mechanism, differential-privacy mechanism, or formal privacy proof [Sweeney, 2002, Dwork, 2006, Buchholz et al., 2024]. If a later analysis prices repeated visits, ordered routines, or temporal sequence alignment, $\Gamma_a(M)$ must become a multiset or sequence and the corresponding frontier must be regenerated.

The hardened anomaly probe compares pymdp-style surprise with a fair online Markov-surprise baseline:

$$\mathcal{S}_i^{\text{Markov}} = -\log \widehat{P}_{i-1}(s_i \mid s_{i-1}), \tag{29}$$

and reports both AUC and operating-point metrics so separability is not summarized by area alone. Finally, result-critical manuscript tokens are bound to source artifacts by a provenance tuple

$$\pi_k = (k, p_k, \eta_k, h_k, v_k), \tag{30}$$

where k is the token, p_k the source artifact path, η_k the JSON key path, h_k the source checksum, and v_k the rendered value. Rendering is trusted only if the artifact exists, the key resolves, the checksum is current, and the formatted artifact value byte-matches the token.

11.1 pymdp Adds Interpretability Without Predictive Leadership

The same `A`, `B`, `C`, and `D` arrays are also consumed by the actual pinned `pymdp` agent. The benchmark matrix reports `not_demonstrated` for state-of-the-art performance and `diagnostic_gain_demonstrated` for diagnostic value — a distinction that is load-bearing. `pymdp` currently helps by exposing variational free energy, policy posteriors, and expected-free-energy terms in a common active-inference grammar; reaching performance leadership would require clearing the sequence-baseline gate in sec. 14.2 and the broader roadmap gates in sec. 6.3, not just replacing the transparent reference filter with the library agent.

When `model.engine` is `both`, the pipeline constructs a batched JAX `Agent` from the tensors above and runs a sequential perception-action loop: variational state inference under the current empirical prior, policy inference returning the policy posterior and expected free energy, deterministic action selection, and an empirical-prior update through the selected control's transition slice [Heins et al., 2022, infer-actively, 2026a,b]. Two comparisons are emitted. First, a matched-prior validation: at every step the pymdp variational posterior is compared against the exact Bayes posterior computed from the same prior and tensors; on the default run the maximum absolute difference is 3.99e-07, which validates the inference implementations against each other. Second, a closed-loop comparison: the full trajectories diverge (maximum absolute posterior difference 0.044 on this run) through policy-induced prior divergence — differing action-selection criteria (expected free energy versus the preference-cross-entropy proxy score of eq. 11) select different controls, which propagate different empirical priors, so subsequent posteriors legitimately differ. Because the posteriors agree under matched priors at every step, the closed-loop divergence is fully explained by differing action sequences rather than any latent inference discrepancy. Its magnitude is trajectory-specific, not a stable property of the implementations. Both traces are emitted side by side in `output/runs/default/pymdp_trace.csv` and `output/runs/default/policy_trace.csv`. The pymdp trace additionally records, per timestep, the agent's own variational free energy (the final fixed-point iterate returned by `infer_states`), the wall-clock step runtime, the policy posterior, and the expected free energy of every policy, so the full diagnostic surface of the pinned library is a run artifact rather than an internal quantity.

12 Supplementary S5 - Closed-Loop Ranking Remains a Non-Controller

This supplement describes the closed-loop candidate-ranking heuristic, the recovery-versus-sparsity lane, and the additive holdout-scoring protocol — secondary experimental lanes extending the main-line evaluation.

12.1 Candidate Ranking Is a Heuristic, Not a Controller

The closed loop is an experiment-level feedback loop: simulate or load waypoints, infer latent-state beliefs, score the run, then rank declared candidate simulator configurations. It is not an online controller acting on a real mobility system.

For a configured objective J and candidate k, the selector records:

$$\text{score}_k = J(\text{run}) \cdot \frac{r_{\text{current}}}{r_k}, \tag{31}$$

when the objective direction is minimization. Here r is the configured mean report count. eq. 31 is a transparent ranking heuristic for comparing declared simulator candidates under the current objective: denser reporting is favored when the measured inference objective is high. Every candidate score is written

to `candidate_scores.json` and `candidate_scores.csv`. It is not presented as an optimal experimental-design rule; formal active-inference treatments connect expected free energy to Bayesian optimal design and expected utility under assumptions that this scalar fixture diagnostic does not satisfy [Sajid et al., 2021b]. The current default ranking is:

Table 5: Candidate scores emitted by the default closed-loop run.

rank	candidate	metric value	reporting ratio	score
1	`higher_reports`	5.718819	0.500000	2.859410
2	`lower_reports`	5.718819	1.500000	8.578229

The selected next candidate is `higher_reports`, with score 2.859410 and reporting ratio 0.500000. Because tbl. 5 is generated from run artifacts, it changes when the objective, candidate set, or simulator parameters change.

12.2 Sporadic-Observation Fit Stays Synthetic

The modeled situation deserves explicit statement. FractalRabbit was published by NSA Research as a sparse-mobility synthetic generator; this project uses it only as a pinned software benchmark. Its traces are observed through sporadic, irregular reports, where most of a trajectory is never seen [Darling, 2018, National Security Agency, 2026a]. The analytic question that situation poses — how much latent movement structure is recoverable from sparse reports, with what calibrated uncertainty, and which next data-collection configuration would be most informative — is exactly the shape of problem a discrete active-inference agent addresses: state estimation under partial observability, uncertainty quantified per step as posterior entropy and variational free energy, and candidate evaluation through expected free energy [Da Costa et al., 2020, Sajid et al., 2021b]. The ground-truth recovery evaluation in sec. 3 makes this concrete: it measures, against the simulator's own latent structure, how much of that structure each configured model actually recovers. Two boundaries hold throughout. The evaluation is synthetic-only software evidence, and the same recoverability that makes sparse analytics tractable is what makes coarse spatiotemporal data re-identifiable — which is why the privacy threat model of sec. 8 and the separate empirical protocol are load-bearing parts of this project rather than disclaimers [de Montjoye et al., 2013, Kondor et al., 2020].

12.3 Recovery Is Measured Across the Sparsity Lane

A dedicated sparsity lane turns the question into a quantitative check. Each traveller's reports are deterministically subsampled (seeded by the root `seed`) to a configured fraction of the full set; at every level the full and deep-latent models are re-estimated from scratch and scored against the exact ground-truth cell, and the pymdp lane validates the released state updater on deterministic sentinel rows at each sparsity level rather than running a full deep-latent agent pass. The result is a recovery-versus-sparsity curve rather than a single operating point — and the curve's direction is read from the artifact, not assumed: under this deterministic subsampling the deep-latent recovery mass does not fall monotonically as reports thin (densest-level mean mass on the true cell 0.069 against 0.118 at the sparsest level, which currently scores highest), so the lane measures how sparsity reprices the available evidence rather than demonstrating degradation.

12.4 Holdout Scoring Remains Additive Under Fixed Discretization

The experiment layer also supports holdout scoring as an additional pass. When `experiment.holdout_fraction` is positive, the last configured fraction of each traveller's chronologically ordered waypoints is held out (each traveller always keeps at least two training rows), the `A/B/D` tensors are re-estimated from the training rows only, and the filter is re-run over the full ordered sequence while aggregating predictive loss and posterior entropy only at holdout indices. The primary full-data metrics are never replaced by this pass; the holdout keys are additive. Discretization bin edges remain shared with the full lane, so the holdout numbers measure out-of-sample tensor estimation under a fixed discretization rather than a fully leakage-free protocol; this boundary is restated in sec. 5.

13 Supplementary S6 - Fixture Diagnostics Bound Sensitivity and pymdp Agreement

This supplement audits the experimental fixture — sensitivity sweeps, negative controls, and pymdp implementation agreement — verifying the setup without advancing the headline claim (sec. 3).

13.1 Sensitivity and Negative Controls Falsify Easy Wins

The fixture sensitivity lane emits 10 runs across deterministic repetitions, report-count variants, spatial-bin settings, burst-rule settings, smoothing, and transition-control choices. This lane runs the sweep configuration (`configs/sweep.yaml`), which deliberately differs from the canonical default configuration: among other declared differences it configures a third `identity` transition control alongside the two canonical controls and uses a different spatial discretizer, seed, and holdout setting. Its policy-entropy and loss summaries therefore audit the sweep configuration and are not directly comparable to the canonical run's values reported later in this supplement. This deterministic sensitivity lane reprocesses the checked fixture. Simulator parameter variants affect declared run metadata and candidate scoring, but they do not resimulate waypoint rows unless the external FractalRabbit lane is available. Fixture-lane replicates are bit-identical reruns: their spread is exactly zero by construction, so std and normal_approx bounds in this lane are determinism checks, never sampling uncertainty. Sensitivity variants use the reference engine; the canonical lane carries the full pymdp trace. The interval column in tbl. 6 is a normal-approximation interval over deterministic variants, not an inferential population confidence interval. Policy entropy is summarized across 10 run(s) with 3 control(s), max entropy 1.099 nats.

Table 6: Sensitivity summary emitted by `scripts/03_run_sweep.py`.

metric	n	mean	std	min-max	normal-approx interval
`mean_predictive_loss`	10	5.112	1.003	[3.392, 6.360]	[4.490, 5.734]
`mean_posterior_entropy`	10	0.321	0.842	[0.005, 2.710]	[-0.201, 0.842]
`mean_policy_entropy`	10	1.099	0	[1.099, 1.099]	[1.099, 1.099]
`baseline_delta_mean_predictive_loss`	10	1.357	0.893	[-0.006, 2.335]	[0.803, 1.910]

The default run also reports per-traveller chronological holdout scoring with configured fraction 0.250: 600 of 2400 waypoint rows are scored by tensors estimated only from the remaining 1800 training rows, yielding holdout mean predictive loss 5.774 nats and holdout mean posterior entropy 0.068 nats. Because the holdout pass is additive, the headline full-data metrics above are identical whether or not the knob is enabled.

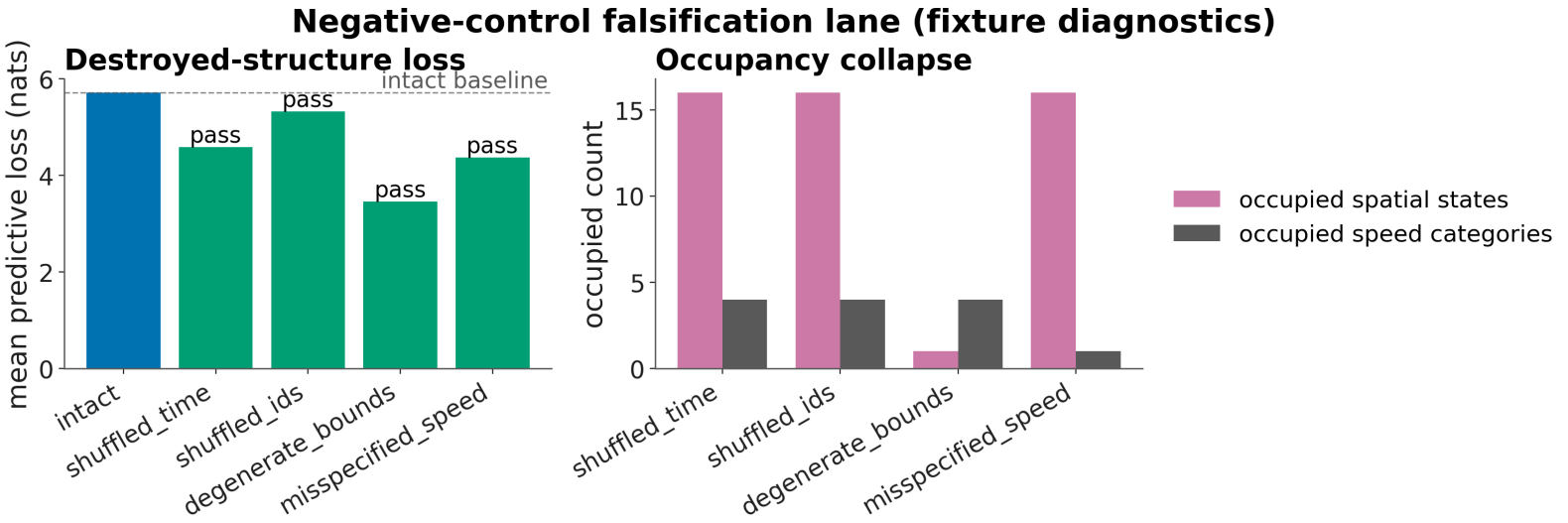


Figure 14: Source negative_controls_summary.json reports 4/4 directional expectations held. Loss changes under shuffled report times and shuffled identities test destroyed structure, while occupancy-collapse probes test degenerate bounds and misspecified speed thresholds. Because collapsed modalities can be easier to predict, the oracle is control-specific rather than a universal loss direction. Fixture-lane software diagnostics, not empirical mobility evidence. Evidence: artifact-level falsification checks for destroyed or degenerate structure. Boundary: empirical mobility evidence.

The negative-control lane intentionally destroys one structural property at a time and checks a directional expectation against the resulting artifacts. fig. 14 and tbl. 7 report the current outcome: 4 of 4 controls behaved as their falsification hypotheses require. Negative controls falsify pipeline insensitivity to destroyed structure. They are software diagnostics on the fixture lane, not empirical mobility evidence.

Table 7: Negative-control outcomes emitted by `scripts/05_run_negative_controls.py`.

control	probe	delta mean predictive loss	passed
`shuffled_time`	speed_km_per_day column differs from the intact run	-1.135	yes
`shuffled_ids`	speed_km_per_day column differs from the intact run	-0.390	yes
`degenerate_bounds`	occupied spatial states == 1 (observed 1)	-2.256	yes
`misspecified_speed`	occupied speed categories == 1 (observed 1)	-1.343	yes

13.2 pymdp Agreement Is an Implementation Check, Not a Win

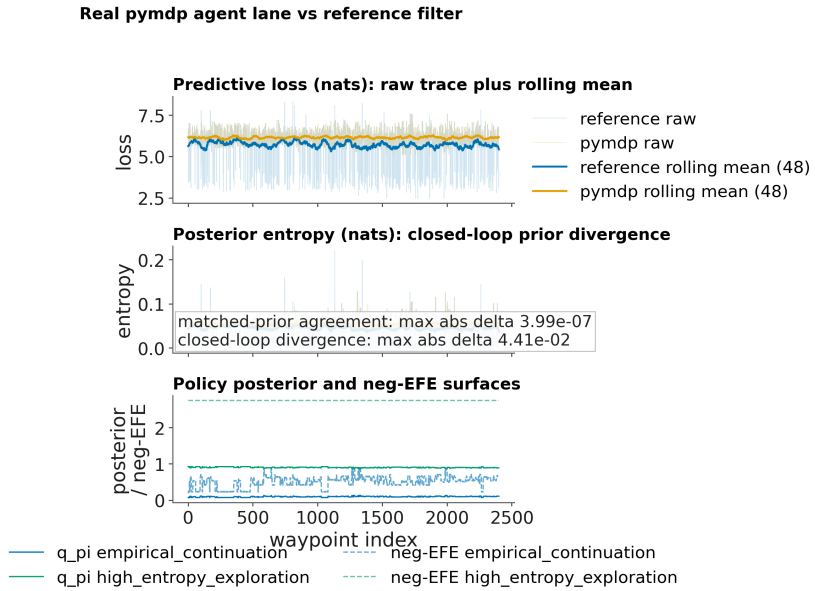


Figure 15: Sources pymdp_trace.csv and inference_trace.csv compare the pinned pymdp agent with the transparent reference on identical tensors. Under matched priors the state posteriors agree to max abs delta 3.99e-07, validating the inference implementation; closed-loop posteriors diverge to 4.41e-02 through policy-induced prior divergence because expected-free-energy action selection and the proxy-score rule propagate different priors. This is a synthetic action-selection diagnostic, not an inference bug. Evidence: implementation agreement and policy-induced prior divergence diagnostics. Boundary: a state-of-the-art performance claim.

The real `pymdp` agent lane (fig. 15) reports mean predictive loss 6.147 nats and mean posterior entropy 0.055 nats on the default run. The matched-prior validation gives a maximum absolute state-posterior difference of 3.99e-07 against the exact Bayes posterior, so the two independent inference implementations agree at numerical precision; the closed-loop traces nevertheless diverge (maximum absolute posterior difference 0.044 on this run) through the policy-induced prior divergence mechanism described in sec. 2.6. The matched-prior control localizes that divergence entirely to action selection.

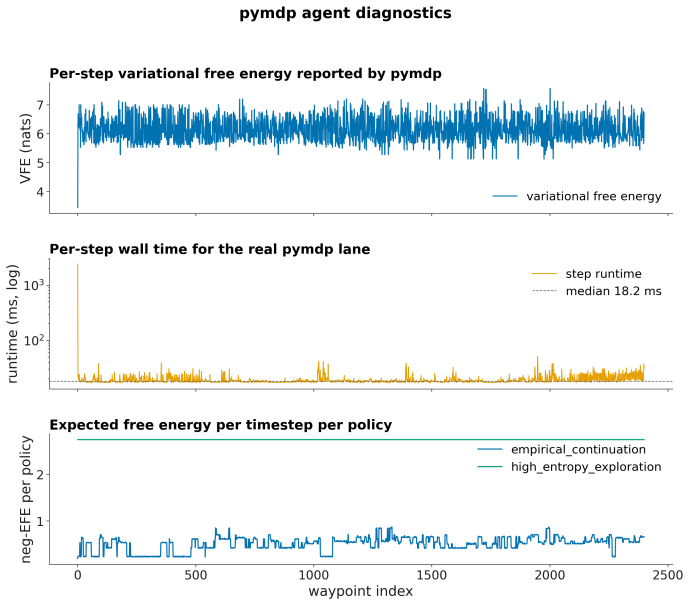


Figure 16: Source pymdp_trace.csv makes the pinned agent's internal surface visible. variational free energy per step (mean 6.147 nats) shows fit-plus-complexity surprise, wall time (median 18.159 ms) reports computational cost, and expected free energy per policy shows the quantity used for policy ranking. The panel supports interpretability claims, not state-of-the-art predictive performance on synthetic waypoints. Evidence: interpretability of the real pymdp lane through VFE, EFE, policy, and runtime surfaces. Boundary: predictive-performance leadership.

fig. 16 exposes the full per-step diagnostic surface of the pinned agent: variational free energy (mean 6.147 nats on the default run), wall-clock runtime (median 18.159 ms per step, maximum 2430.487 ms, 48.486 s total against 0.988 s for the reference filter), and the expected free energy of every policy at every timestep. These numbers make the computational cost and the policy-evaluation dynamics of the active-inference layer auditable artifacts.

The performance-versus-interpretability answer from sec. 11.1 is summarized in tbl. 8. The current state-of-the-art performance status is `not_demonstrated`: `pymdp` changes predictive loss by 0.428 nats and posterior entropy by 0.009 nats relative to the transparent reference filter, so the current contribution is not a predictive-performance win. The diagnostic status is `diagnostic_gain_demonstrated`, because the library lane exposes VFE, policy posterior, expected free energy, and action-conditioned prior dynamics as inspectable artifacts. The roadmap gates are: use the lane-resolved model gate matrix before performance-leading language; learn hierarchical or continuous-space dynamics instead of fixed cells only; calibrate missingness and reporting processes explicitly; validate on consented empirical traces before real-world claims.

Table 8: Current `pymdp` performance and interpretability matrix, emitted by `pymdp_benchmark_matrix.json`.

model	implemented	current role	interpretability surface
`reference_filter`	yes	transparent Bayes-style performance baseline	posterior trace, tensor provenance, calibration
`pymdp_agent`	yes	variational active-inference implementation and policy diagnostic	VFE, policy posterior, expected free energy, action-conditioned prior
`factorial_pymdp_agent`	yes	multi-factor latent-mode active-inference diagnostic	cell, movement-mode, and cognitive-proxy posteriors plus VFE/EFE per step
`persistence_and_base_rate_nulls`	yes	honesty floor for autocorrelated waypoint recovery	explicit null hierarchy
`markov_surprise_baseline`	yes	transition-aware anomaly comparator	online transition surprise
`prequential_sequence_baselines`	yes	stronger next-cell predictive-performance gate	named online nulls and per-baseline loss/skill rows
`lane_resolved_model_gate_matrix`	yes	fixture/external evidence gate for model claims	lane status, unavailable handling, and per-gate verdict rows
`return_plus_phase_prior`	yes	first promoted-or-null prior extension after return and phase gates	pooled, phase, return, combined, persistence, and permuted comparator rows
`future_sota_baselines`	planned	required before any state-of-the-art performance claim	to be compared under the same null hierarchy

13.3 Spatial-Resolution Diagnostics Choose the Reporting Grid

The remaining sections are the supporting battery — the diagnostics, null hierarchy, gate verifications, and safety boundaries against which the headline result was tested. They begin with the fixture and discretization diagnostics that every later lane depends on.

The default fixture run writes 18 run artifacts and 53 registered figures. From this baseline, the measured summary is: 2400 observations, 16 latent states, mean predictive loss 5.719, final posterior entropy 0.001, and mean policy entropy 0.693. The fixed-transition baseline filter uses `empirical_continuation`, yielding mean predictive loss 3.845 and an active-minus-baseline loss delta of 1.874. On this fixture, the fixed-transition baseline has lower mean predictive loss than the active policy-proxy trace, so the current active layer should be read as an auditable diagnostic rather than a performance improvement — and the delta has a mechanical explanation: with the proxy exactly tied under uniform preferences (sec. 2.6.3), tie-breaking selects the uniform `high_entropy_exploration` transition on a fraction 0.818 of steps, flattening the temporal prior that the fixed `empirical_continuation` baseline keeps sharp.

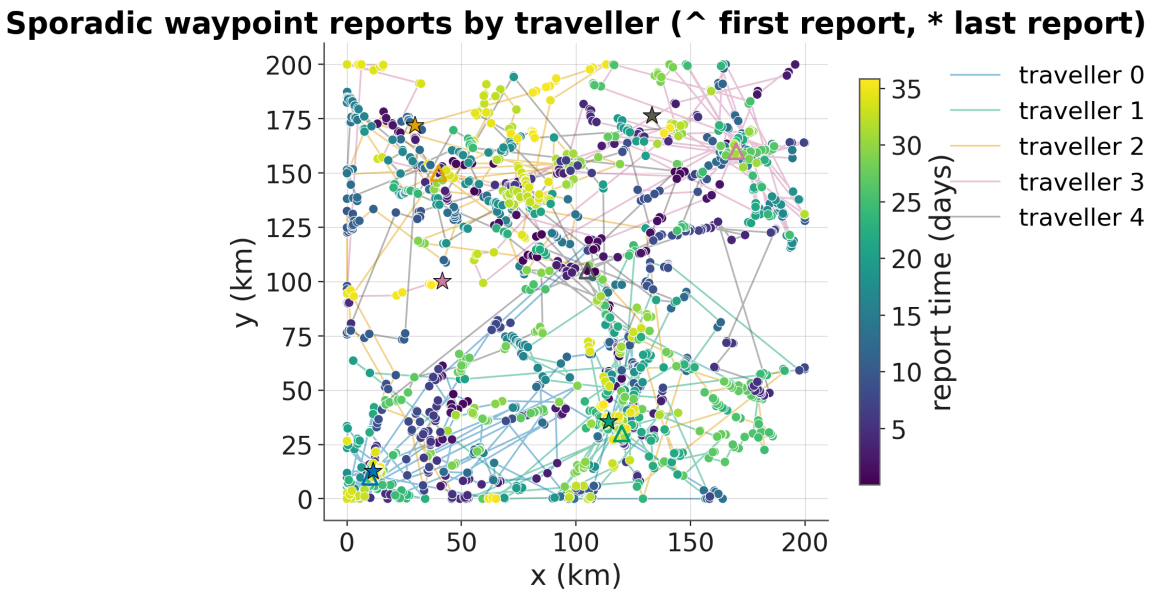


Figure 17: Source observations.csv is shown as 2400 synthetic sporadic waypoint reports for 5 travellers. Lines connect reports only in report order; they are a visual index, not reconstructed continuous tracks. Color encodes report time, the open triangle marks each first report, and the star marks the final report. The visible clustering and gaps define the sparse-observation regime that the benchmark scores, not empirical mobility. Evidence: inspection of sparse report ordering and coverage gaps. Boundary: continuous-track reconstruction, surveillance, or empirical mobility inference.

fig. 17 shows that the fixture preserves the FractalRabbit-style sparse waypoint contract while remaining small enough for deterministic tests.

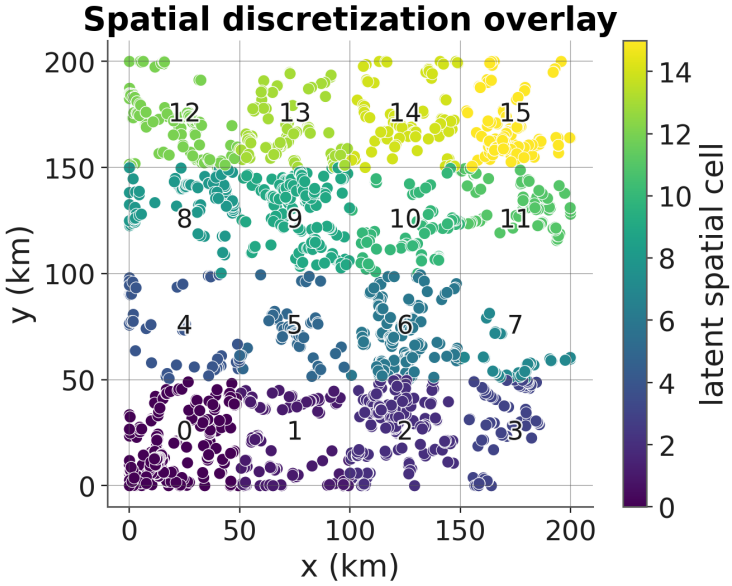


Figure 18: Source discrete_features.csv maps all n=2400 reports from continuous coordinates into the model's latent spatial cells. Color and printed cell labels show the configured grid that the location modality actually observes: reports sharing a cell become equivalent evidence even when their coordinates differ. This figure fixes the analysis resolution before inference and is a synthetic discretization audit, not a map of real places. Evidence: the fixed synthetic analysis-cell boundary used by the tensors. Boundary: geographic interpretation of places or real-world map validity.

fig. 18 audits the spatial discretization used in eq. 21. The plotted cells are analysis bins, not FractalRabbit internal point-process states.

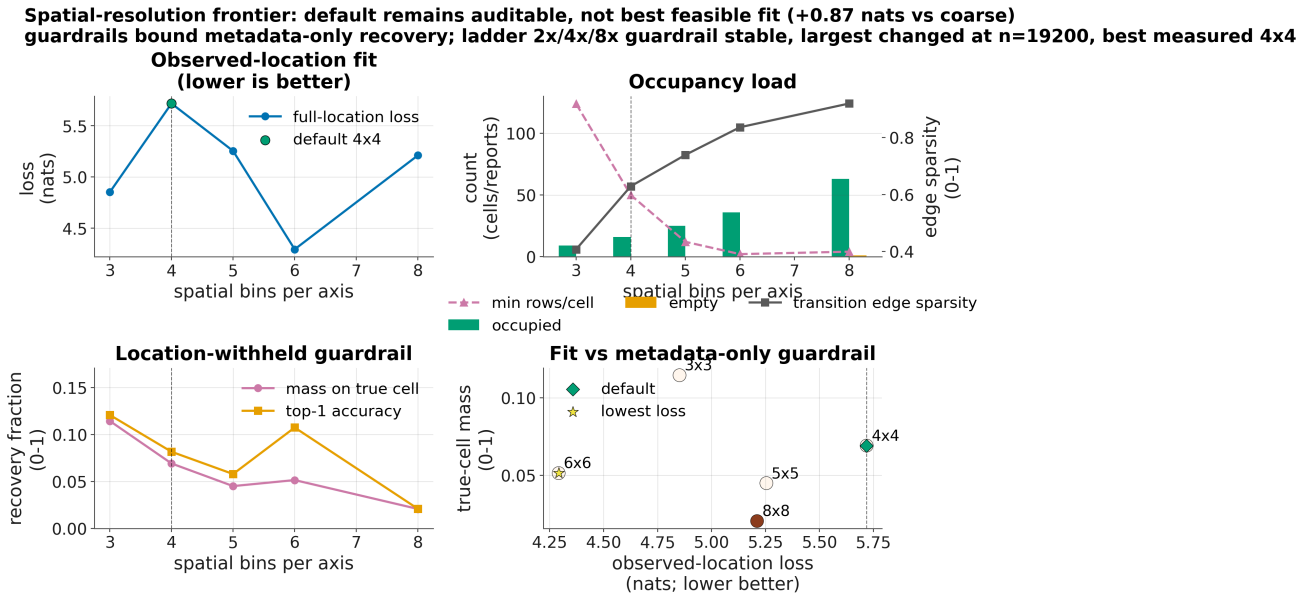


Figure 19: Sources spatial_resolution_frontier.json and spatial_resolution_larger_sample_confirmation.json compare candidate square grids on occupancy, transition sparsity, full-location predictive loss, no-location recovery, real-pymdp agreement, and deterministic sample-size sensitivity. The selected default is the canonical guardrail survivor. When the larger lane has no strict-feasible grid, the marked bin is explicitly a best-measured fallback. The two recovery rows keep observed-location fit separate from metadata-only recovery. Evidence: the grid-selection guardrail for observed-location fit and metadata-only recovery. Boundary: promotion of no-location recovery beyond the artifact frontier.

fig. 19 makes the discretization choice first-class. In fig. 19, the artifact evaluates candidate square grids against occupancy, transition sparsity, full-location predictive loss, location-withheld recovery, and bounded real-`pymdp` matched-prior agreement. The default grid is 4x4: 16 states with 0 empty cells, full-location loss 5.719, and default-vs-coarse loss change 0.867. On the canonical fixture the emitted verdict is `default_resolution_not_best_feasible_loss`, with best feasible bins 3. A deterministic sample-size ladder then regenerates 2, 4, 8 times the canonical reports per traveller from the same fixture generator family, with the largest 8x lane at n=19200 (3840 per traveller); the parallel largest-sample tokens resolve to n=19200 and 3840 reports per traveller for the same lane. The largest lane reports `larger_sample_changes_frontier_decision`: status `changed`, strict feasible candidate count 0, selected bins 4, default-vs-coarse loss change -1.069, and default-vs-coarse no-location mass change -0.036. Across the ladder, verdict stability is no, no-location guardrail stability is yes, selected bins are 4, 8, 4, strict feasible candidate counts are 2, 0, 0, and the generated interpretation is `guardrail_stable_but_fit_selection_changes`. When a strict feasible count is zero, the selected bins are the best measured loss among the candidate grids rather than a strict occupancy claim. The ladder refines the observed-location fit verdict at larger synthetic report counts, while the no-location guardrail keeps the same direction against fine-cell recovery (canonical default-vs-coarse mass change -0.045). It is not evidence that time, speed, and reporting-burst metadata recover fine spatial cells better, and it is no longer phrased as a universal best-fit grid.

That no-location guardrail is a *change* between the default and coarse discretizers. It is not the same scalar as the absolute deep-latent recovery mass in fig. 5 or the densest sparsity recovery mass in fig. 7, so those values should not be compared as if they measured one regime.

Table 9: Spatial-resolution frontier emitted by `spatial_resolution_frontier.json`; the pymdp column is the matched-prior agreement check on deterministic sentinel rows. The sample-size ladder is emitted separately as `spatial_resolution_larger_sample_confirmation.json` so it can test the qualitative decision without changing the canonical fixture rows in this table.

bins	states	occupied	empty	full loss	no-location mass	pymdp delta
3	9	9	0	4.852	0.114	2.08e-07
4	16	16	0	5.719	0.069	3.20e-07
5	25	25	0	5.254	0.045	2.68e-07
6	36	36	0	4.292	0.051	4.82e-07
8	64	63	1	5.211	0.021	3.62e-07

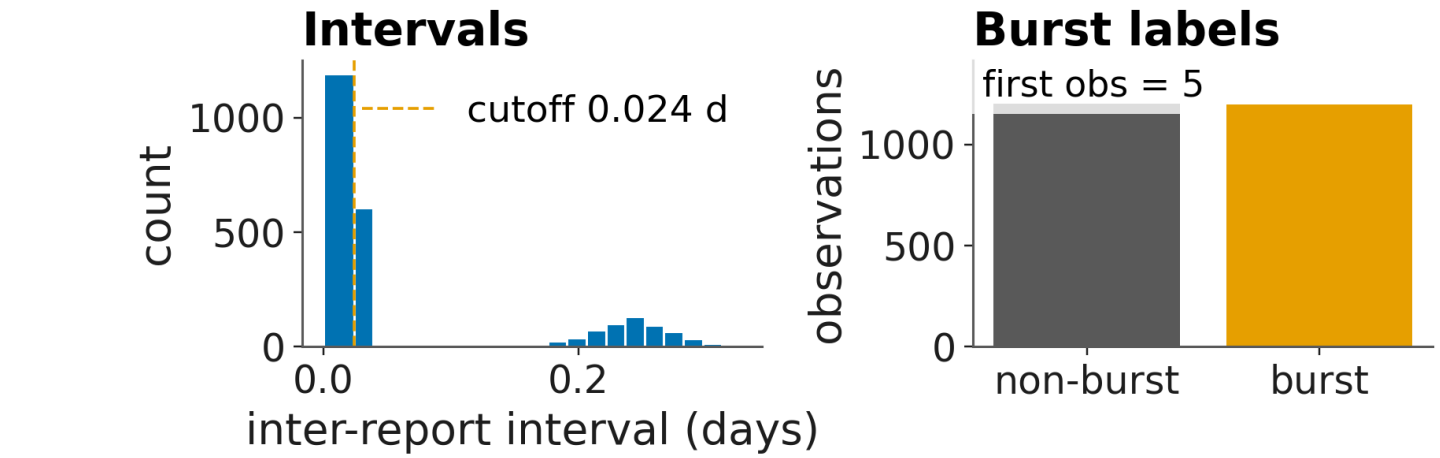


Figure 20: Source discrete_features.csv exposes the reporting process rather than treating missingness as background noise. The left panel plots positive inter-report intervals against the configured burst cutoff; the right panel counts burst labels after the 5 first observations are assigned zero-interval non-burst sentinels. The heavy tail and short-interval cluster are the synthetic SporadicReporter texture that later missingness gates must explain. Evidence: the synthetic reporting-process texture used by later missingness gates. Boundary: a behavioral account of real reporting cadence.

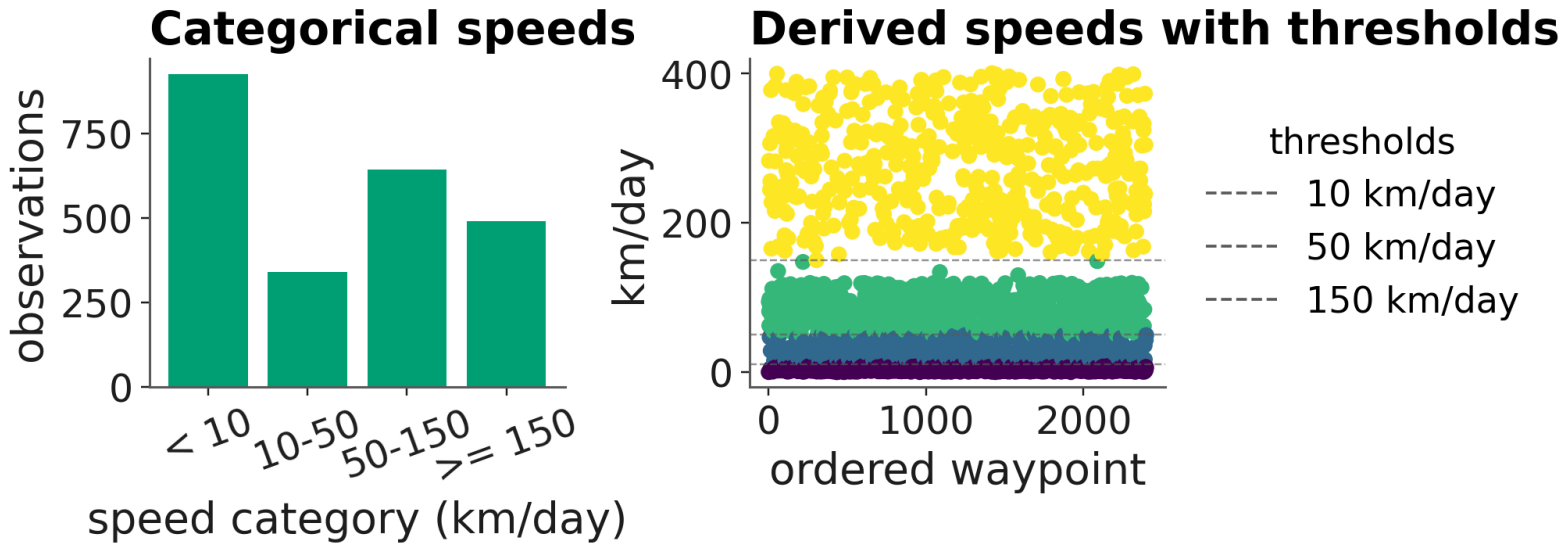


Figure 21: Source discrete_features.csv converts consecutive reports into speed evidence for n=2400 synthetic reports. Bars summarize category counts; the trace shows each derived speed against thresholds (10, 50, 150 km/day), with first observations fixed at zero by convention. This is one non-location modality used by the deep-latent configuration, and it remains derived evidence rather than direct ground truth about unobserved movement. Evidence: the derived kinematic evidence channel used by the deep-latent lane. Boundary: ground-truth claims about unobserved movement between reports.

fig. 20 and fig. 21 inspect the derived temporal and motion categories from eq. 23. They are category diagnostics, not empirical claims about real movement speed.

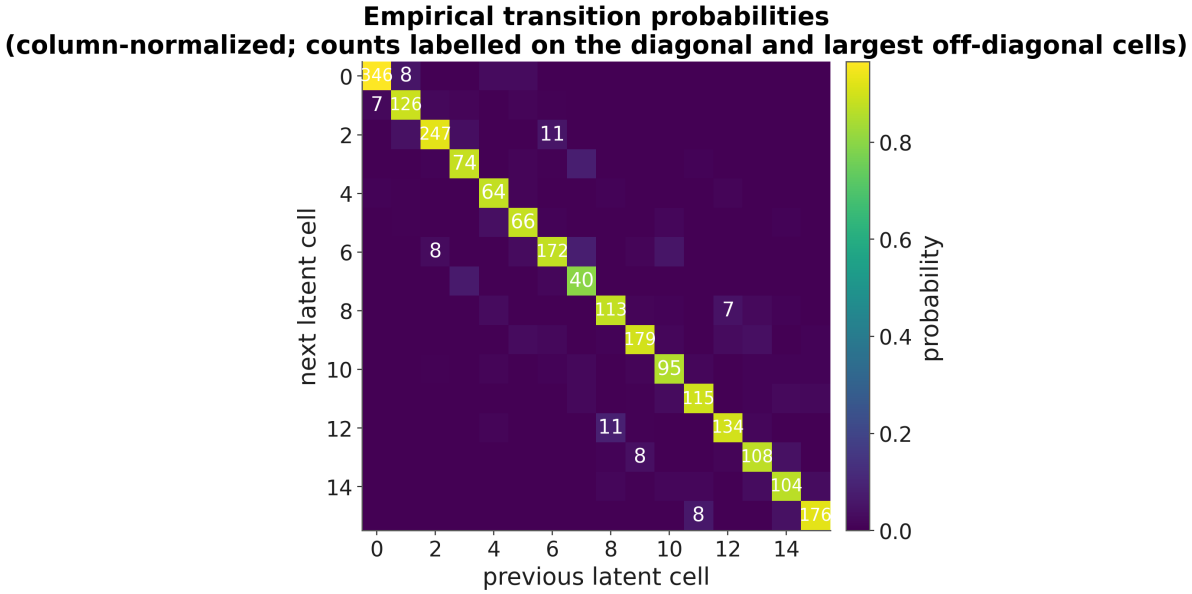


Figure 22: Source discrete_features.csv estimates the empirical latent-cell transition surface before tensor smoothing, excluding traveller-boundary (first-observation) pairs exactly as the model's B-tensor estimator does. Each column is P(next cell | previous cell), with raw counts printed over the diagonal and the largest off-diagonal cells; the diagonal makes dwell-and-return structure visible, while off-diagonal mass carries movement evidence. This panel is the normalization and sparsity audit for downstream synthetic B tensors before smoothing=0.05 is applied. Evidence: normalization and sparsity checks for the synthetic B-tensor surface. Boundary: external movement-model validation.

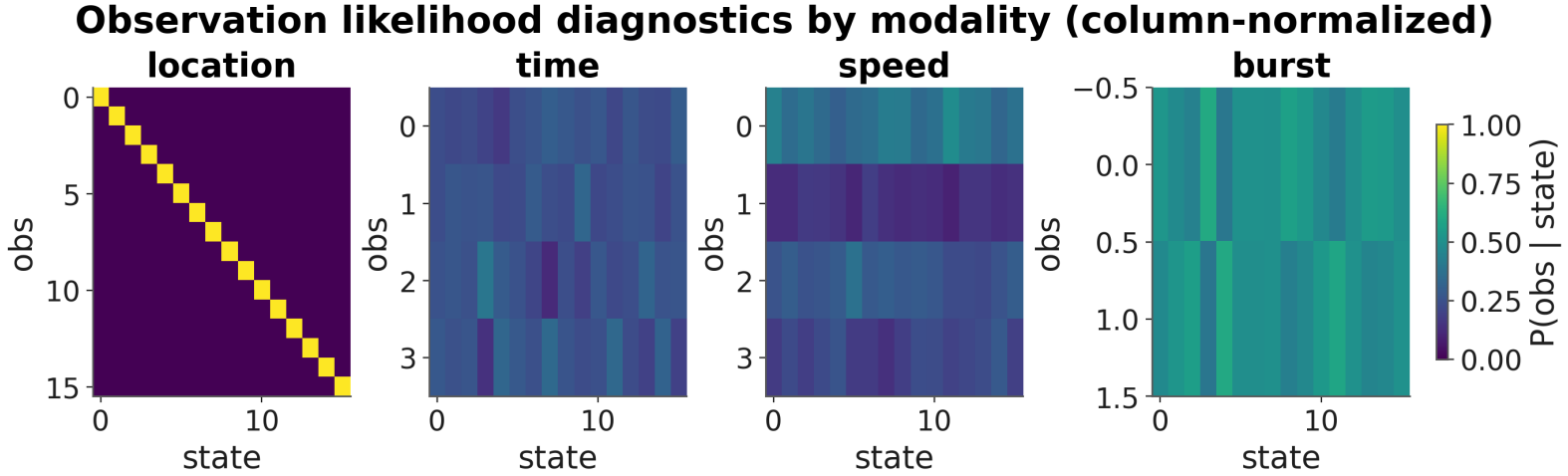


Figure 23: Source discrete_features.csv builds column-normalized observation likelihoods P(observation | latent cell) for location, time, speed, and reporting-burst modalities. The near-diagonal location panel is expected by construction; the time, speed, and burst panels are the evidence channels the deep-latent configuration uses when location is withheld. The panels are the audit point for the conditional-independence approximation before Model A smoothing=0.05 is applied, not empirical sensor validation. Evidence: the likelihood-audit boundary for location, time, speed, and burst modalities. Boundary: empirical sensor validation or independence claims outside the configured approximation.

fig. 22 and fig. 23 expose the empirical structure that feeds eq. 2 and eq. 1. These diagnostics are useful because normalization bugs in `A` or `B` can otherwise produce plausible-looking posterior traces.

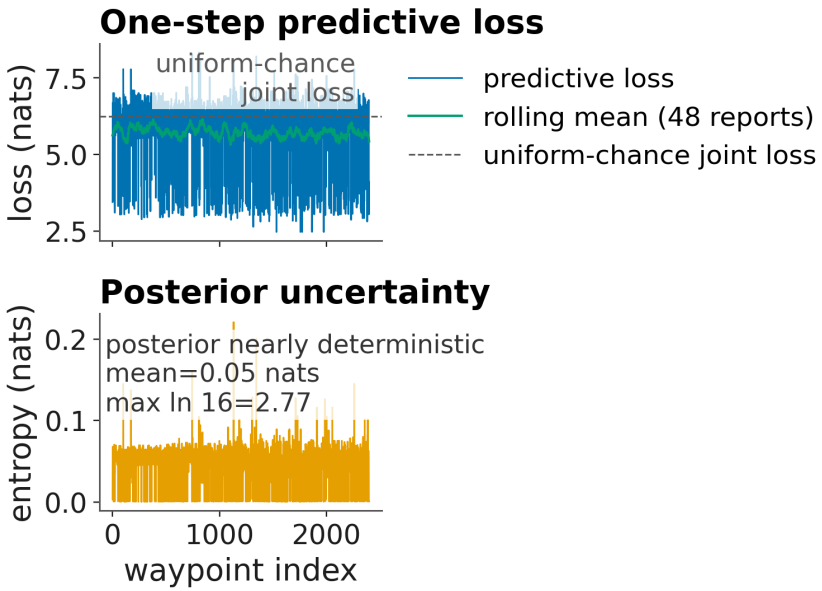


Figure 24: Source inference_trace.csv records the transparent reference filter over n=2400 reports. The top panel shows one-step predictive loss, the negative log prior-predictive probability assigned to the observed report, against the uniform-chance joint-loss floor; the bottom panel shows posterior entropy after updating. Spikes identify synthetic reports that surprise the model, while low sustained entropy marks confident cell tracking rather than proof of real trajectory reconstruction. Evidence: reference-filter surprise and uncertainty diagnostics over synthetic reports. Boundary: proof of real trajectory reconstruction.

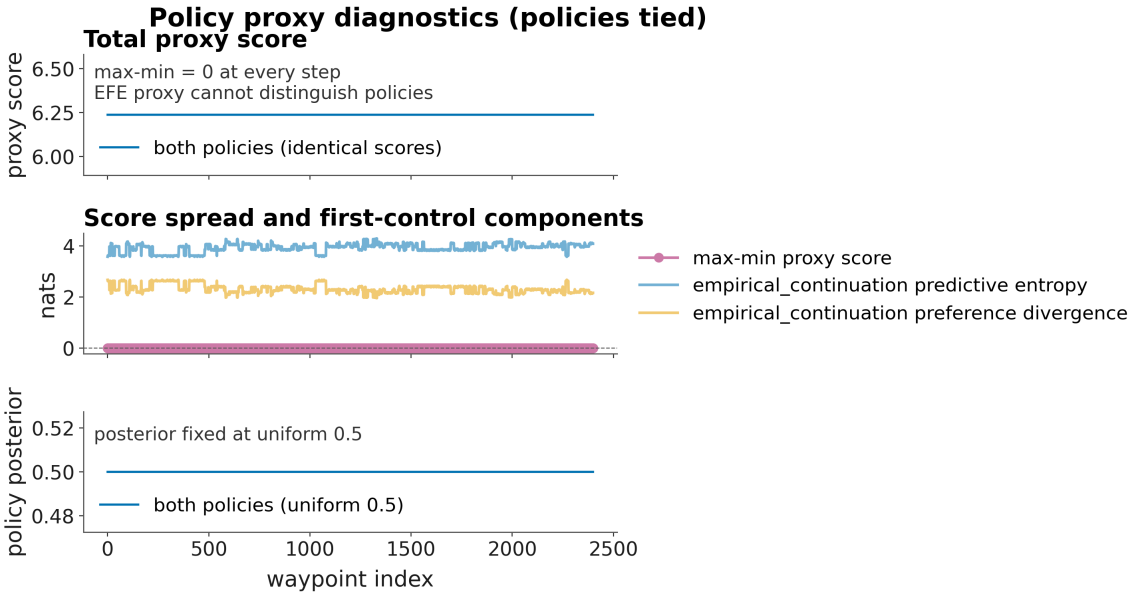


Figure 25: Source policy_trace.csv exposes the proxy policy diagnostic for the 2 configured transition control(s). Total score, score spread, entropy and preference components, and the softmax policy posterior show whether candidate controls separate under the current preferences; identical proxy scores are drawn as one line when the loaded artifact makes the tie exact. Near-uniform posteriors are reported as a synthetic observational-equivalence finding, not as a controller failure or operational recommendation. Evidence: candidate-control separability under the proxy scoring rule. Boundary: operational control advice or controller validation.

fig. 24 reports the measured loss and uncertainty sequences from eq. 7 and eq. 8. fig. 25 reports the configured-control diagnostics from eq. 11 and eq. 14. For the default fixture, near-maximal policy entropy; under the default uniform preferences the proxy score is constant across controls by algebra, so the configured policies are exactly tied rather than empirically close.

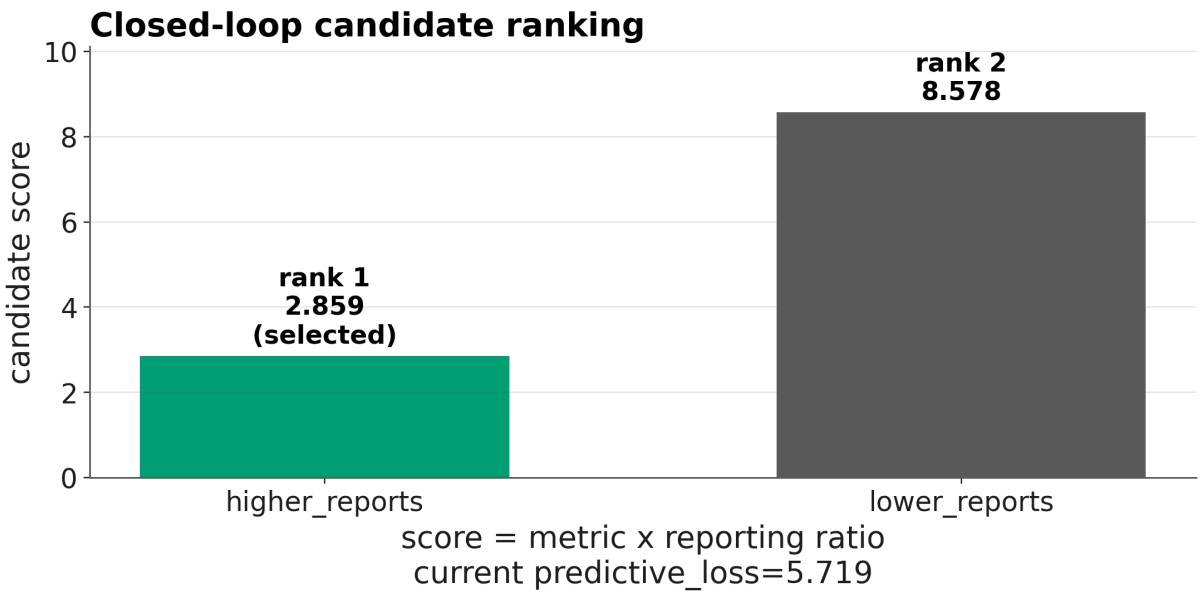


Figure 26: Source candidate_scores.json ranks the n=2 declared next-run simulator candidates by the measured objective scaled by relative reporting density. Rank labels and scores make the closed-loop choice recomputable from artifacts. This is an audit trail for candidate selection inside the synthetic fixture lane, not an external simulator-performance benchmark. Evidence: recomputable closed-loop candidate ranking inside the fixture lane. Boundary: external simulator performance leadership.

fig. 26 visualizes tbl. 5. The selected next candidate is `higher_reports`; the displayed formula is minimize `predictive_loss`: score = 5.718819 x 0.500000 = 2.859410. It is deliberately a score explanation figure: it supports auditability of the next-run choice, not an external performance benchmark.

14 Supplementary S7 - Program Extensions Add Gates, Not Promotion

This supplement holds the secondary result lanes — anticipatory gap-filling, expected-free-energy decomposition, program extensions, robustness probes, sequence baselines, and roadmap-gate verifications — that support but do not constitute the main-line findings (sec. 3).

14.1 Robustness Artifacts Record Synthetic Realization Spread

The robustness expansion asks whether those verdicts survive stronger nulls and broader software checks. The return-structure probe in fig. 27 applies eq. 24 and eq. 25 on move steps only. The measured move-step entropy reduction is 0.186 nats, persistence headroom is 0, and the current verdict is `promote_return_conditioned_prior`; this either promotes a return-conditioned prior as a concrete next model component or records a bounded null in the same artifact. The figure's third bar, `return_skill_vs_persistence`, is a secondary, prototype-only Brier-skill diagnostic of the return-conditioned predictive prior itself, not of the location-observed POMDP posterior; it is negative on the current run and plays no role in the promotion gate above, which is decided solely by the entropy reduction and the persistence-headroom sign, so a negative bar underneath a `promote_return_conditioned_prior` title is expected rather than a contradiction of that verdict. Cross-realization bands in fig. 28 repeat the expanded configuration across 10 synthetic realizations. The gap-filling skill band spans 0.219 to 0.268 (interquartile range 0.230–0.246, SD 0.014 — min–max over ten draws is extreme-value-dominated, so the rank- and moment-based spreads are reported alongside it), so single-run conclusions can be read against an empirical synthetic-realization spread rather than treated as if one draw were a stable law. The same repeated-realization pass also plots five further verdict metrics discussed at the single-run level elsewhere in this document: the location-blind-limit skill against persistence (`blind_limit_skill`, sec. 14.5), the no-location frontier recovery mass (`frontier_no_location_mass`, sec. 14.7), the deep-latent structure-recovery total variation (`structure_deep_latent_tv`, sec. 14.7), the random-cell anomaly ROC AUC (`anomaly_random_auc`, sec. 14.7), and the mission EFE-minus-pragmatic recovery margin (`mission_efe_minus_pragmatic`, sec. 14.7). Because `blind_limit_skill` is the Brier-type skill score in the harder location-blind limit, already reported as strongly negative in sec. 14.5. Unlike the other four metrics (which stay within a bounded unit-scale range), it is unbounded below — so its band sits roughly an order of magnitude beneath the rest of the figure due to its scale, not a plotting error or a stale realization. These bands are robustness summaries over generated software realizations, not population confidence intervals [Efron and Tibshirani, 1994]. Upstream sensitivity in fig. 29 reports `blocked` after 5 attempted prior commits; a blocked result is rendered as evidence about build comparability, not as a missing figure.

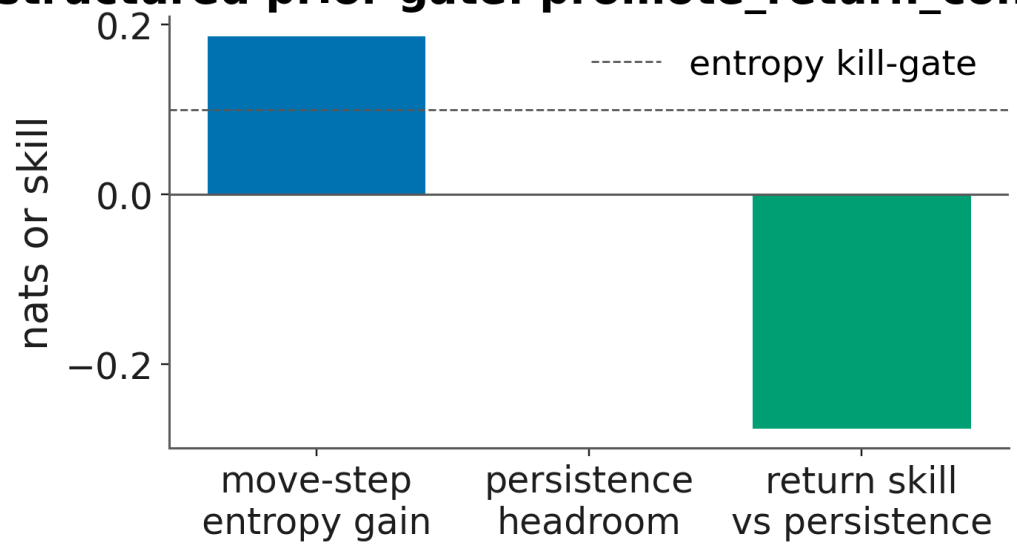


Figure 27: Source return_structure_analysis.json tests whether a visited-cell return feature lowers move-step next-cell entropy beyond persistence headroom. The bars show entropy reduction, available null headroom, and the prototype-before-claim score. A promoted result becomes a candidate prior; a null remains an explicit bound on this synthetic trace rather than a hidden failure. Evidence: whether return features merit promotion to a candidate prior. Boundary: a hidden behavioral law of movement.

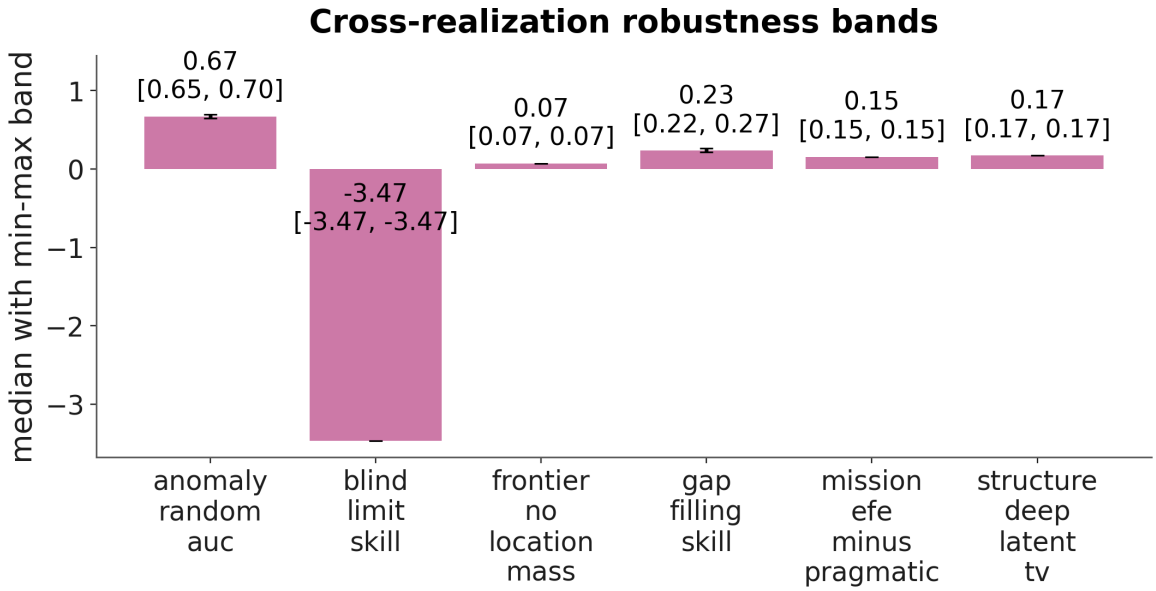


Figure 28: Source variance_bands.json repeats key analyses across synthetic-generator realizations and plots medians with min-max bands. Gap-filling skill, anomaly separability, and minimization-frontier recovery are shown as realization-sensitive program diagnostics. These bands describe synthetic reruns, not population confidence intervals or empirical uncertainty. Evidence: sensitivity of key diagnostics across synthetic-generator reruns. Boundary: population confidence intervals.

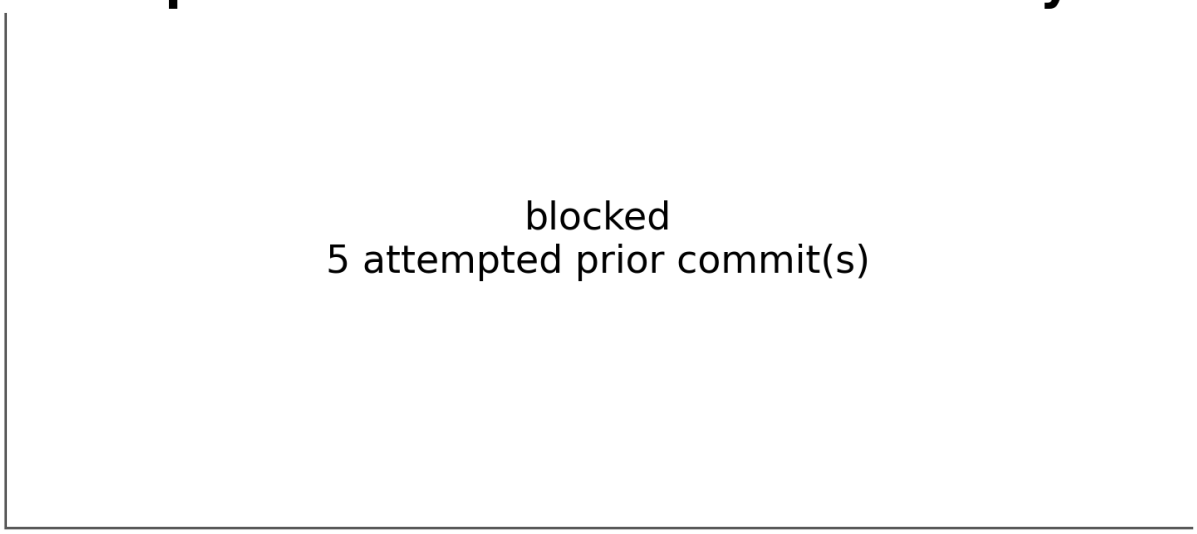


Figure 29: Source upstream_sensitivity.json records whether a second buildable FractalRabbit commit produced comparable tensor or recovery distances under the current JDK/Maven toolchain. If no comparable build completes, the blocked state is rendered explicitly. The captioned artifact prevents silent upstream drift from masquerading as validated simulator robustness. Evidence: explicit blocking or comparison of buildable upstream simulator commits. Boundary: silent robustness to upstream drift.

The implemented roadmap probes from sec. 6.1 extend that robustness layer. The fractal-dimension sweep in fig. 30 varies the generator control introduced in sec. 10.2; its status is `fixture_control_unvaried`, the best predictive-loss dimension is 1.100, the highest revisit-rate dimension is unavailable, and the predictive-loss sweep range is 5.719 to 5.719. The circadian-phase signal in fig. 31 tests whether a phase-conditioned occupancy prior is worth promoting: the current verdict is `honest_null`, with phase gain 0.011, permuted gain 0.007, and maximum Jensen-Shannon separation 0.004. Against a 200-permutation label null (floored at one over the draw count plus one), the phase gain carries permutation p-value 0.199 and the maximum Jensen-Shannon separation p-value 0.214, so neither clears at the conventional threshold. Both artifacts are roadmap gates, not behavioral claims.

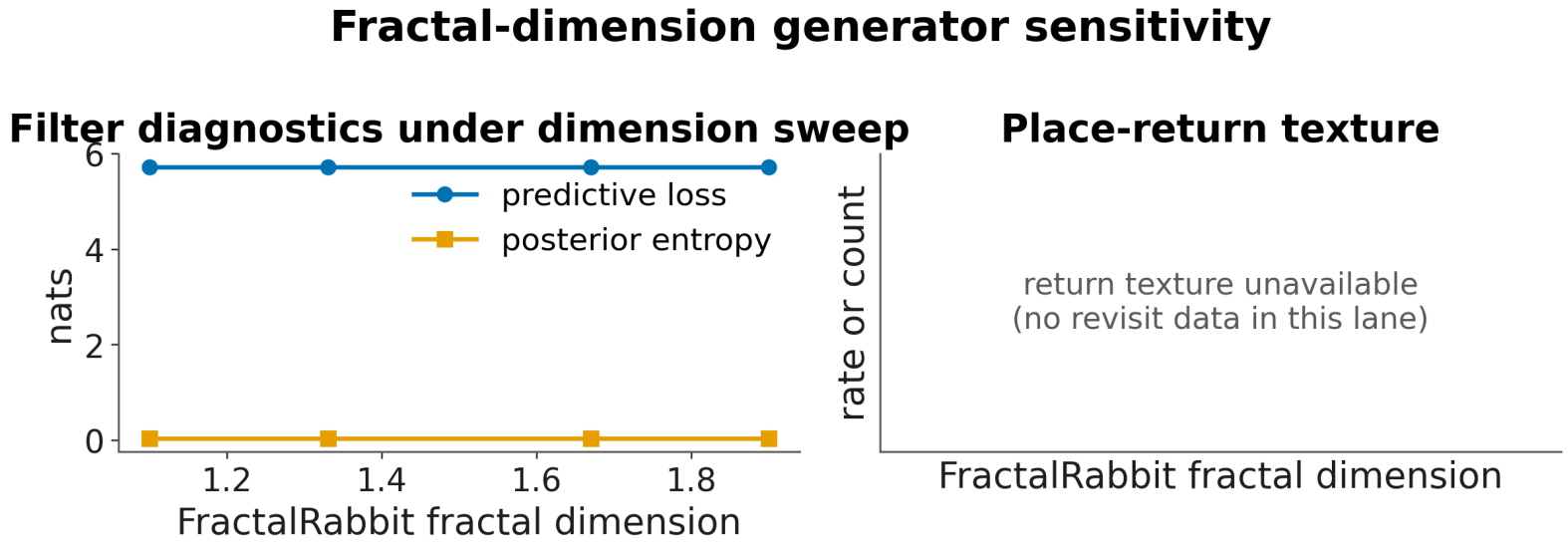


Figure 30: Source fractal_dimension_sweep.json varies the FractalRabbit dimension parameter while holding the expanded configuration fixed, then reports filter diagnostics (left panel) and, when the lane populates it, place-return texture (right panel; an explicit 'no revisit data in this lane' marker is drawn when the return texture is unavailable). The parameter is treated as a synthetic generator knob over support roughness and revisit structure. It is not estimated fractal dimension for real mobility. Evidence: a fixture-replay invariance check over the nominal dimension knob (the external generator was unavailable, so metrics are constant by construction). Boundary: estimation of real-mobility fractal dimension.

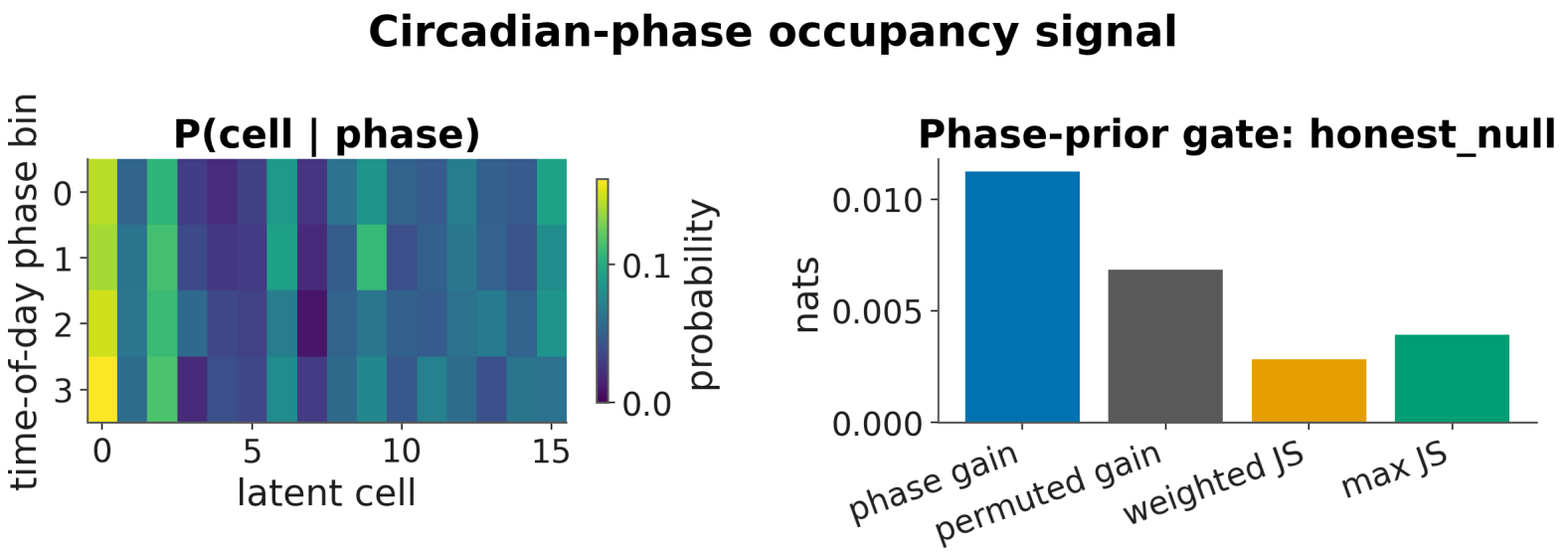


Figure 31: Source circadian_phase_analysis.json tests whether phase-conditioned occupancy carries predictive structure. The heatmap shows cell occupancy by phase, and the bars compare phase-conditioned gain with a permuted-phase null plus Jensen-Shannon separation. A positive result can become a candidate prior; a null remains bounded synthetic evidence rather than a behavioral claim. Evidence: phase-conditioned prior candidacy against a permuted-phase null. Boundary: a behavioral circadian claim.

14.2 Sequence and Reporting Gates Strengthen the Nulls

The sequence-baseline gate in fig. 32 implements a prequential ladder of transparent next-cell predictors: base-rate, persistence, first-order Markov, phase-conditioned Markov, metadata-conditioned Markov, and return-boosted variants [Dawid, 1984]. Its current best model is `first_order_markov`, with mean predictive loss 0.570 nats and skill 0.787 against the online base-rate null. Relative to the default reference filter, the best sequence baseline changes predictive loss by -5.149 nats; relative to the `pymdp` lane, it changes predictive loss by -5.577 nats. The current performance gate is `stronger_sequence_baseline_beats_pymdp_on_this_run`, which is deliberately harder than asking whether `pymdp` agrees with a matched-prior Bayes implementation.

Table 10: Prequential next-cell sequence baselines emitted by `sequence_baseline_analysis.json`.

model	mean predictive loss (nats)	skill vs base-rate
`base_rate`	2.677	0
`persistence`	0.839	0.687
`first_order_markov`	0.570	0.787
`phase_markov`	0.680	0.746
`metadata_markov`	1.111	0.585
`return_boost_markov`	0.592	0.779
`phase_return_markov`	0.782	0.708

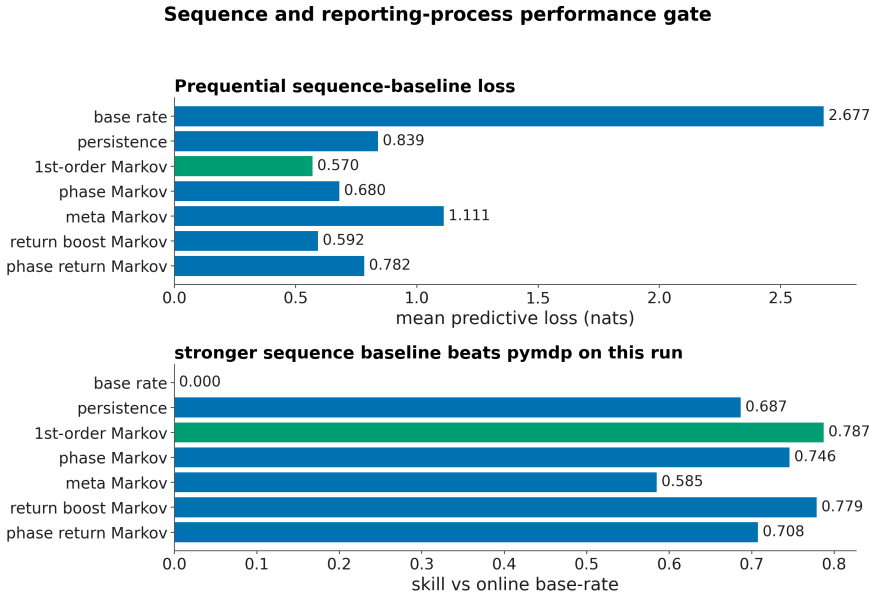


Figure 32: Source sequence_baseline_analysis.json implements the prequential baseline ladder before any stronger performance language is allowed. Online base-rate, persistence, Markov, phase-conditioned, metadata-conditioned, and return-boosted variants are scored on next-cell predictive loss before observing the current report. pymdp must clear these transparent synthetic sequence nulls before the manuscript can claim predictive leadership. Evidence: the prequential baseline ladder required before performance language. Boundary: pymdp predictive leadership until the gate is cleared.

The reporting-process gate in fig. 33 asks whether sparse reporting itself has phase or traveller structure. Treating observation timing as ignorable requires explicit assumptions: classical missing-data theory conditions ignorability on the missingness mechanism, and irregular outcome-dependent observation can bias ordinary

marginal analyses when visit intensity and the outcome process are associated [Rubin, 1976, Lin et al., 2004]. The synthetic gate scores 2395 positive inter-report intervals, defines short gaps from the generated interval distribution, and reports verdict `honest_null`. The short-gap rate is 0.500, interval coefficient of variation is 1.316, phase gain over a pooled reporter is -0.005, traveller gain is -0.007, combined phase-traveller gain is -0.029, and permuted-phase gain is -0.006. It tests report timing in generated waypoints; it neither establishes ignorability nor characterizes real nonresponse behavior.

Table 11: Reporting-process log-loss gains emitted by `missingness_model_analysis.json`.

reporting context	log-loss gain vs pooled
phase	-0.005
traveller	-0.007
phase + traveller	-0.029
permuted phase	-0.006

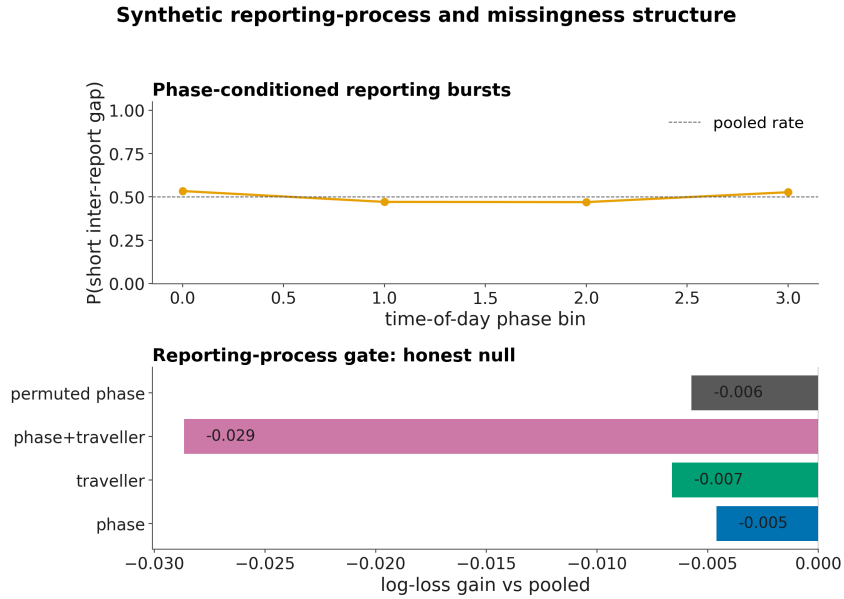


Figure 33: Source missingness_model_analysis.json models reporting cadence as its own synthetic process. Phase-conditioned short-gap probabilities and online log-loss gains for phase, traveller, and combined contexts are compared with a pooled Bernoulli reporter and a permuted-phase null. A promoted result is a candidate reporting-process model, not a claim about real nonresponse or missingness behavior. Evidence: reporting-process structure against pooled and permuted nulls. Boundary: claims about real nonresponse behavior.

14.3 Lane-Resolved Gates Keep Fixture and External Evidence Separate

The singleton sequence and reporting-process artifacts remain for backward compatibility, but the manuscript-facing comparison surface is now the lane-resolved model gate matrix. fig. 34 and tbl. 12 put fixture and external readiness side by side for the sequence-baseline ladder, the synthetic reporting-process gate, and the return-plus-phase prior gate. The matrix status is `partial`; on the external lane the sequence gate is `unavailable`, the missingness verdict is `unavailable`, and the return-plus-phase verdict is `unavailable`. If the external lane did not complete in the current invocation, the external cells render explicit `unavailable` statuses rather than stale numbers or numeric zeroes.

Table 12: Lane-resolved model gate matrix emitted by `model_gate_matrix.json`.

lane	gate	status	model	metric	verdict
`fixture`	`sequence_baseline_ladder`	completed	`first_order_markov`	0.570	`stronger_sequence_baseline_beats_pymdp_on_this_run`
`fixture`	`reporting_process_missingness`	completed	`phase_traveller_reporter`	-0.029	`honest_null`
`fixture`	`return_plus_phase_prior`	completed	`return_plus_phase_prior`	-1.797	`honest_null`
`fixture`	`factorial_pymdp_latent_modes`	completed	`particle_state_space`	5.296	`factorial_pymdp_cleared_current_lane`
`external`	`sequence_baseline_ladder`	unavailable	`unavailable`	unavailable	`unavailable`
`external`	`reporting_process_missingness`	unavailable	`phase_traveller_reporter`	unavailable	`unavailable`
`external`	`return_plus_phase_prior`	unavailable	`unavailable`	unavailable	`unavailable`
`external`	`factorial_pymdp_latent_modes`	unavailable	`unavailable`	unavailable	`unavailable`

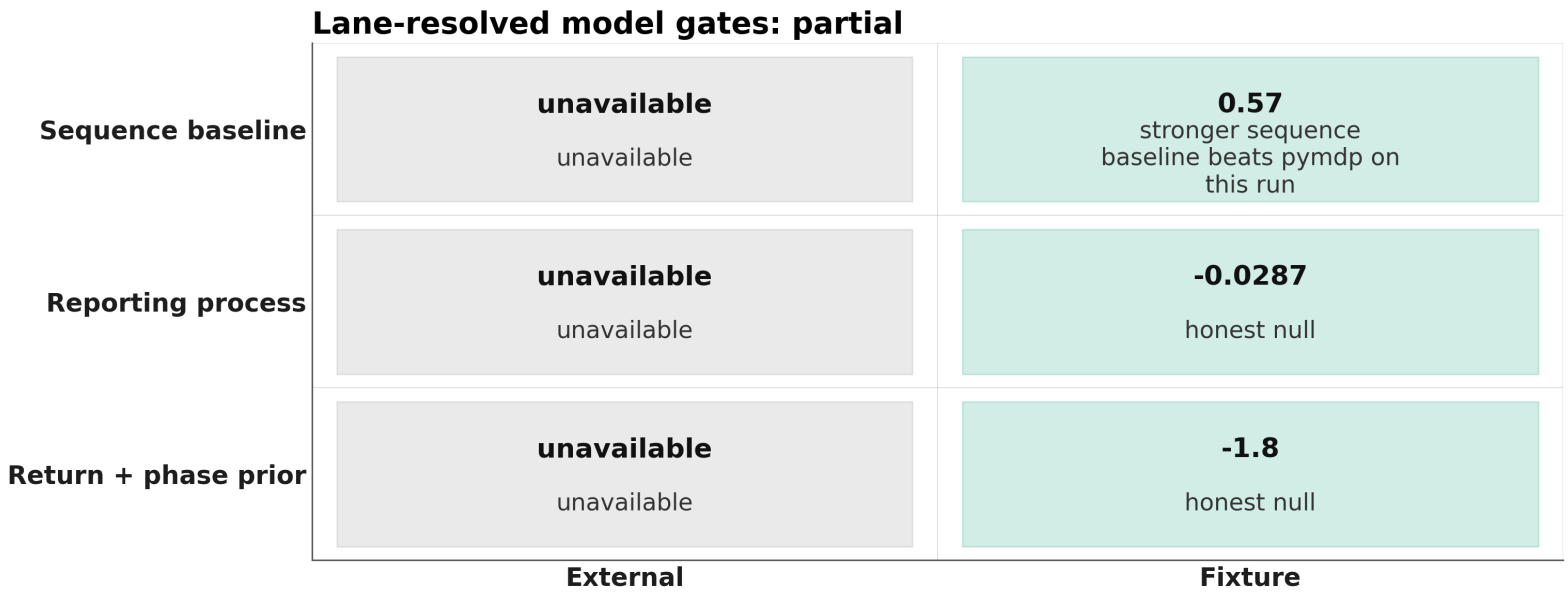


Figure 34: Source model_gate_matrix.json puts fixture and external lanes on one readiness contract for the sequence baseline ladder, reporting-process structure, and return-plus-phase prior. Completed cells show current-lane metrics and verdicts; unavailable lanes render unavailable rather than zero. The matrix preserves lane honesty while newer promoted synthetic families move to the forward-gates artifact. Evidence: lane readiness for promoted fixture and external evidence families. Boundary: treating unavailable external lanes as failures or successes.

The return-plus-phase prior is the first concrete M9 model idea. In tbl. 13, it compares pooled, phase-only, return-only, and return-plus-phase priors against persistence and permuted-phase nulls before promotion. The current canonical-lane verdict is `honest_null`; the best model is `return_plus_phase_prior` with mean predictive loss 2.636, gain -1.797 over persistence, and gain 0.004 over the permuted-phase comparator. A promoted verdict would justify adding the factor to the generative prior; an honest null keeps the manuscript from treating return and circadian texture as a model improvement merely because they are plausible.

Table 13: Return-plus-phase prior gate emitted by `return_phase_prior_analysis.json`.

prior	mean predictive loss (nats)	skill vs base-rate
`base_rate_null`	2.677	0
`persistence_null`	0.839	0.687
`pooled_prior`	2.677	0
`phase_prior`	2.738	-0.023
`return_prior`	2.658	0.007
`return_plus_phase_prior`	2.636	0.015
`permuted_return_plus_phase_prior`	2.640	0.014

14.4 Forward Gates Promote Only Bounded Synthetic Positives

The forward gates in fig. 35 promote the deeper roadmap analyses from prototype claims into regenerated artifacts. The summary status is `completed`, with 8 completed fixture gates; external rows are either current-lane evidence from the same full-study invocation or explicit `unavailable` records. The latent-regime HMM remains a bounded null: its gate is `latent_regime_hmm_does_not_beat_markov_null_em_overfits`, best honest held-out loss 2.615, and overfit gap 0.492 [Rabiner, 1989, Zucchini et al., 2016]. The neural sequence baseline reports `neural_sequence_beats_weak_nulls_not_particle_mixture`, with loss 0.618 against particle/state-space loss 0.550 [Patterson et al., 2008, Doucet et al., 2001, Arulampalam et al., 2002].

The learned-prior and reporting gates separate positives from nulls. The learned-stay-rate verdict is `honest_regime_null`; the best variant is `stay_move`, with best-minus-particle difference -0.008 and bootstrap interval -0.016 to -0.001. The held-out reporting-process verdict is `honest_null`, with runlen gain 0.001 on leave-one-traveller-out scoring and 0.002 on the temporal split. Cross-generator transfer reports persistence as `regime_dependent_on_self_transition_rate`, particle-vs-hard-persistence as `transfers`, and return structure as `transfers`; this keeps broad mechanism language bounded to the two synthetic generators actually tested.

The table deliberately keeps lane disagreements visible rather than merging lanes: an external row that did not complete in the current invocation renders an explicit `unavailable` record instead of inheriting a fixture value or a stale prior-run number, and the manuscript conclusion follows the fixture-scoped token unless a current-run external result is explicitly named.

Table 14: Promoted forward-gate summary emitted by `forward_gates_summary.json`.

lane	gate	status	metric	verdict	artifact
`fixture`	`latent_regime_hmm`	completed	2.615	`latent_regime_hmm_does_not_beat_markov_null_em_overfits`	`output/runs/latent_regime_hmm.json`
`fixture`	`neural_baseline`	completed	0.618	`neural_sequence_beats_weak_nulls_not_particle_mixture`	`output/runs/neural_baseline.json`
`fixture`	`learned_stay_rate`	completed	-0.008	`honest_regime_null`	`output/runs/learned_stay_rate.json`
`fixture`	`heldout_reporting_process`	completed	0.001	`honest_null`	`output/runs/heldout_reporting_process.json`

lane	gate	status	metric	verdict	artifact
fixture	cross_generator_transfer	completed	30	persistence regime_dependent_on_self_transition_rate; particle transfers; return transfers	output/runs/cross_generator_transfer.json
fixture	efe_location_frontier	completed	1.000	metadata_recovery_not_above_structure_null	output/runs/efe_location_frontier.json
fixture	linkage_risk_frontier	completed	1	location_carries_relink_risk_that_minimization_removes	output/runs/linkage_risk_frontier.json
fixture	poisoning_robustness	completed	0.636	poisoning_detection_regime_changed	output/runs/poisoning_robustness.json
external	latent_regime_hmm	unavailable	unavailable	external lane was not completed in this run	output/runs/latent_regime_hmm_external.json
external	neural_baseline	unavailable	unavailable	external lane was not completed in this run	output/runs/neural_baseline_external.json
external	learned_stay_rate	unavailable	unavailable	external lane was not completed in this run	output/runs/learned_stay_rate_external.json
external	heldout_reporting_process	unavailable	unavailable	external lane was not completed in this run	output/runs/heldout_reporting_process_external.json
external	cross_generator_transfer	unavailable	unavailable	external lane was not completed in this run	output/runs/cross_generator_transfer_external.json
external	efe_location_frontier	unavailable	unavailable	external lane was not completed in this run	output/runs/efe_location_frontier_external.json
external	linkage_risk_frontier	unavailable	unavailable	external lane was not completed in this run	output/runs/linkage_risk_frontier_external.json
external	poisoning_robustness	unavailable	unavailable	external lane was not completed in this run	output/runs/poisoning_robustness_external.json

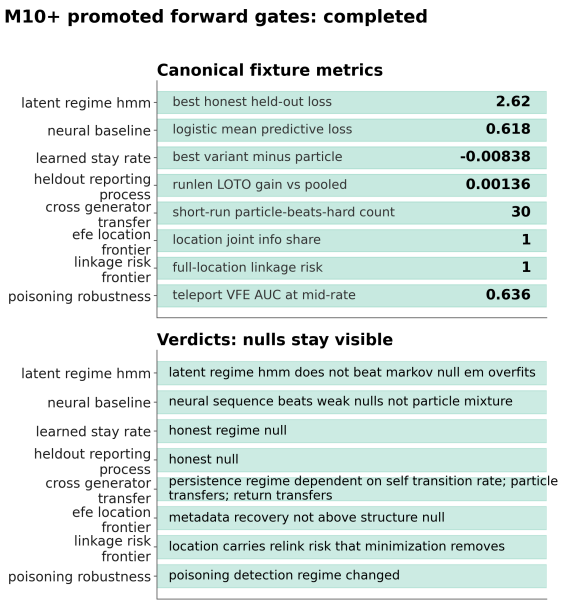


Figure 35: Source forward_gates_summary.json summarizes the promoted M10+ artifact families: latent-regime HMM, neural baseline, learned stay-rate, held-out reporting process, cross-generator transfer, EFE location frontier, linkage-risk frontier, and poisoning robustness. Metrics and verdicts are shown together, including honest nulls. External rows count only same-run synthetic evidence, not stale simulator results. Evidence: current M10+ synthetic gate outcomes including honest nulls. Boundary: stale or empirical simulator-validation conclusions.

14.5 Gap Filling Still Faces the Persistence Oracle

The persistence null deserves a closer look, because it is a deceptively strong baseline: it is *handed the true previous cell* and simply predicts it again.

Two regimes separate cleanly. In the **location-blind limit** — the location modality removed at every step — the filter's information set is not handed the true previous cell, while persistence receives that cell by construction. There the deep-latent skill against persistence is strongly negative under this weaker information set — the filter does not beat the persistence oracle, and the deficit is large rather than marginal (reference lane -19.745 on the fixture; the current pinned-fixture resimulation gives flat -3.422 and speed-conditioned -3.467): non-location evidence narrows the gap but does not close it. This is an honest benchmark-specific information asymmetry, not a universal theorem about location-blind methods; a different generator or target could make a location-blind predictor competitive.

The target synthetic gap-filling regime is **sporadic-reporting interpolation**: location is observed at sparse reports and missing in between, so both the filter and persistence know the last *reported* cell and the question is what happens across a gap. fig. 36 reports recovery scored on the 1199 held-out gap steps of the current pinned-fixture run (location hidden there). Persistence rigidly holds the last reported cell (Brier 0.304). Against that oracle the regenerated run scores skill 0.133 for the flat belief filter and 0.219 for the speed-conditioned variant (positive = clears persistence; speed-clamp edge 0.086) — on the current canonical regeneration both are positive, so both variants clear the persistence oracle here, with the speed-conditioned filter strongest. We do not promote these margins

as a win: they are single-run synthetic software measurements that lean on the disclosed displacement-derived speed cue, they sit outside the clustered-bootstrap multiplicity family of sec. 3.2.1.1, and the controlled harness's role remains proving the machinery can detect a planted speed-to-move coupling. Speed is derived from inter-report displacement, so it is a strong kinematic cue for *whether* the cell changed — we disclose this rather than hide it, report the speed-free flat filter alongside, and confine all of these to synthetic-data software measurements on the pinned simulator, never empirical mobility validation.

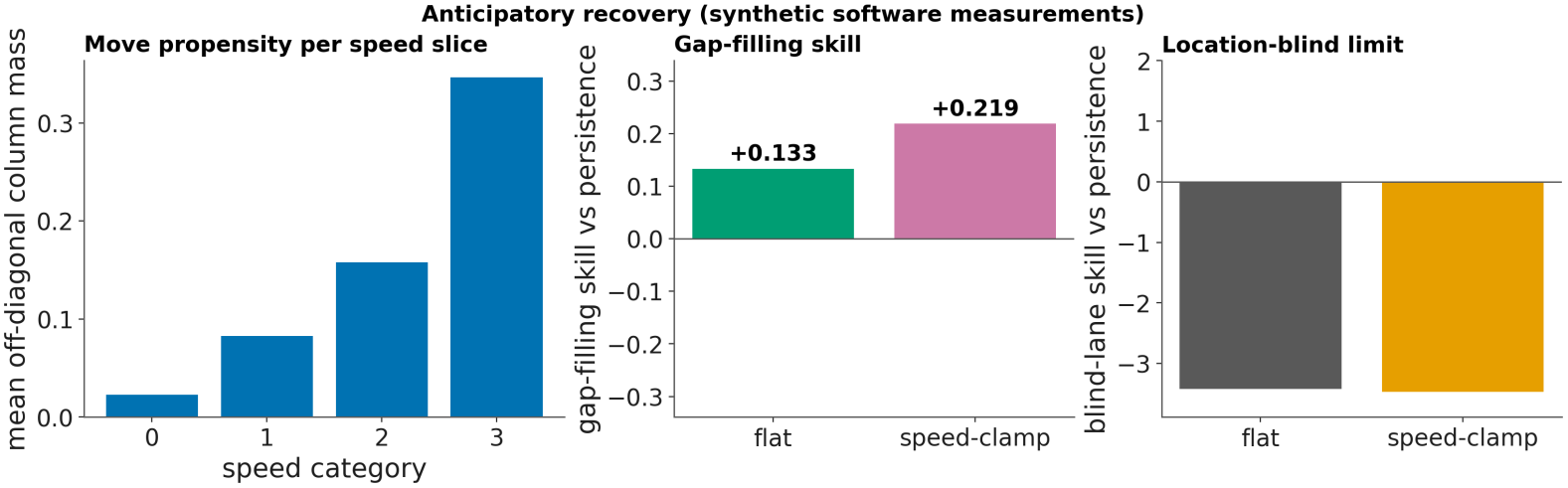


Figure 36: Source anticipatory_analysis.json tests whether belief propagation fills sparse gaps better than persistence. The left panel shows speed-conditioned off-diagonal transition mass; the middle panel reports gap-filling recovery skill against the persistence null, and the right panel keeps the location-blind limit on its own scale. Positive bars would support gap filling beyond the oracle previous-cell baseline, while negative bars are retained as bounded nulls. Speed is derived from reports, and the result is a synthetic software measurement. Evidence: gap-filling skill or bounded nulls against the persistence oracle. Boundary: claims that propagation predicts real hidden paths.

14.6 Expected Free Energy Separates Epistemic and Pragmatic Value

The active-sensing analysis ranks candidate location fixes by expected information gain / expected-free-energy components, which decompose into an epistemic term (the expected information gain about the latent cell from an observation) and a pragmatic term (the expected log-preference of the predicted observation under the preference vector C) [Millidge et al., 2021, Sajid et al., 2021b, Champion et al., 2024]. fig. 37 makes that decomposition explicit. This is separate from the transparent closed-loop simulator-candidate selector of sec. 12, which ranks declared simulator configurations from measured artifact metrics.

Because this project runs with uniform preferences (C all zero), the pragmatic term is identically zero and the agent's sensing value is *purely epistemic*: the mean epistemic share of a location fix is 1. Sharpening C toward a preferred observation introduces a pragmatic drive and the epistemic share falls — to 0.820 at preference strength 4 — which confirms the decomposition is a genuine measurement that separates the two value components rather than a constant. Under this project's synthetic, preference-free generative model, then, "active" inference here means purely information-seeking sensing, not operational collection.

We also tested whether that epistemic value is *useful* for scheduling: under a fixed observation budget, does ranking steps by expected information gain (spending location fixes where they most reduce cell uncertainty) recover the latent cell better than spending the same budget on a blind, budget-matched random schedule? Scored honestly on the full non-first trajectory — every schedule covers the identical evaluation set, so the metric cannot be flattered by a schedule that leaves only easy steps unobserved — the answer is an honest null: epistemic-led acquisition beats only a fraction 0 of random schedules on the current pinned-fixture run. The speed-conditioned dynamics already anticipate where the cell will change, so most of the uncertainty an information-greedy fix would target is uncertainty the transition model has already resolved; choosing *which* step to observe by information gain therefore adds little over choosing blindly. We report this null rather than search for a framing that inverts it.

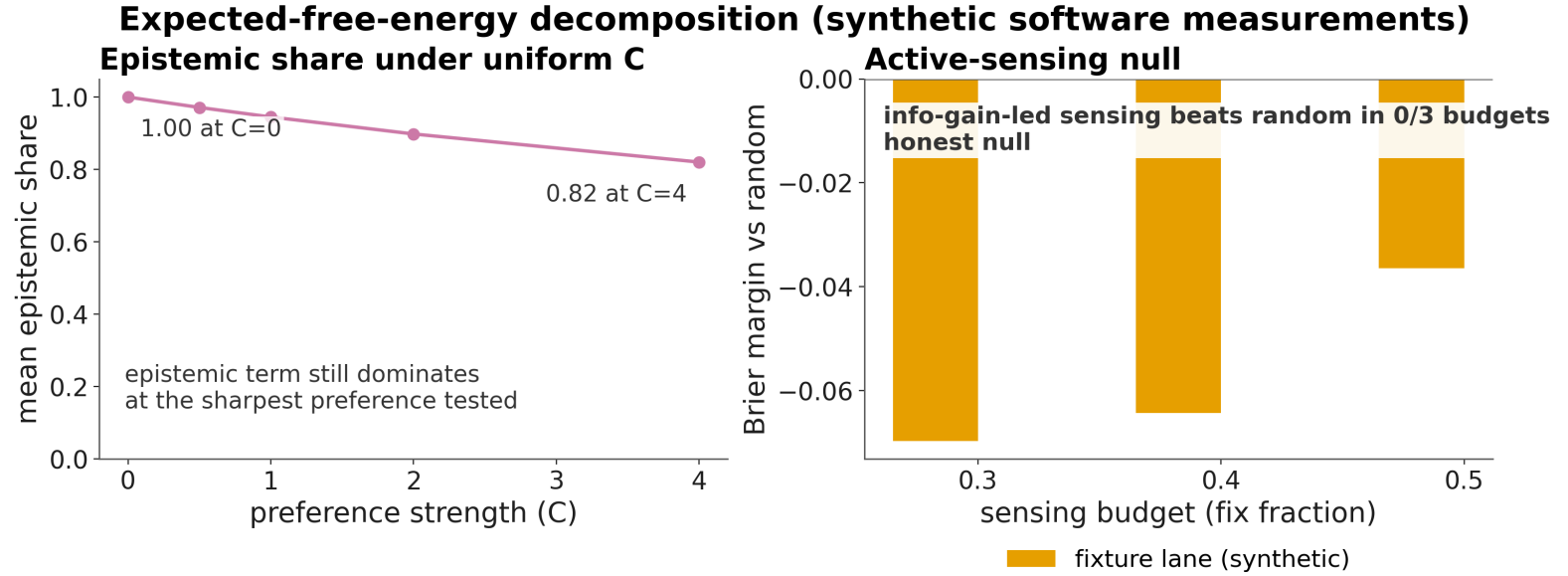


Figure 37: Source active_sensing_analysis.json decomposes expected free energy into epistemic and pragmatic value. As preferences sharpen, the location-fix epistemic share changes rather than remaining constant; the budgeted sensing panel compares information-led acquisition with a budget-matched random schedule. Near-zero or negative margins are reported as an active-sensing null, showing that information gain does not robustly beat blind acquisition on this synthetic trace. Evidence: the separation of epistemic and pragmatic value in synthetic active sensing. Boundary: a robust collection or sensing recommendation.

14.7 Detection, Structure Recovery, and Minimization Stay Program Extensions

Three further analyses map what the sparse-mobility benchmark can and cannot do, each reported against its honest baseline.

For routine-break detection, fitting the generative model on the clean routine and scoring per-step surprise (variational free energy) on a stream with injected breaks separates anomalies from normal steps with ROC AUC 0.649 for a rare-cell break and 0.667 for a random-cell break, against a first-occurrence novelty baseline of 0.500 (fig. 38). Surprise wins because it folds in the transition model, making an unreachable jump surprising in a way mere novelty cannot see; the comparison is a synthetic anomaly-as-surprise diagnostic, not a deployed detector [Pimentel et al., 2014].

For pattern-of-life structure recovery, the target is occupancy total-variation and preferred-place-set overlap rather than next-cell prediction. The deep-latent lane recovers the broad occupancy structure (preferred-set Jaccard 0.500) above a uniform null, but its honest baseline is dynamics-only propagation: on the current pinned-fixture run the non-location evidence (occupancy TV 0.171) does *not* beat simply propagating the trained dynamics (TV 0.042), so sparse metadata adds little to the preferred-place structure a fitted transition model already encodes. A between-draw check repeats this comparison on 10 independent synthetic generator draws (16 travellers each, mean occupancy TV 0.076 deep-latent vs. 0.075 dynamics-only): the deep-latent lane beats dynamics-only on only 0.500 of draws, corroborating rather than contradicting the single-fixture result above — this is a genuine coin-flip across independent draws, not a robust win or a robust loss, consistent with a trained transition model already carrying most of the preferred-place signal.

For mission-conditioned collection, allocating a fixed location-fix budget by expected free energy under mission-weighted preferences is scored against a mission-pragmatic comparator, not operational tasking, on the expanded benchmark: mission-weighted recovery is 0.590 for EFE versus 0.554 for pragmatic. The text should be read from the regenerated values rather than a fixed prior verdict.

Composing these, the recoverability frontier (fig. 39) maps recovery against the observed modality set: with all modalities the filter holds mean mass 0.995 on the true cell, but withholding location collapses it to 0.069 — an oversight-positive statement of how protective location-minimization is and how little the specific cell is recoverable from time, speed, and reporting-burst metadata alone. All of these are synthetic-data software measurements on the pinned generator, not empirical mobility findings.

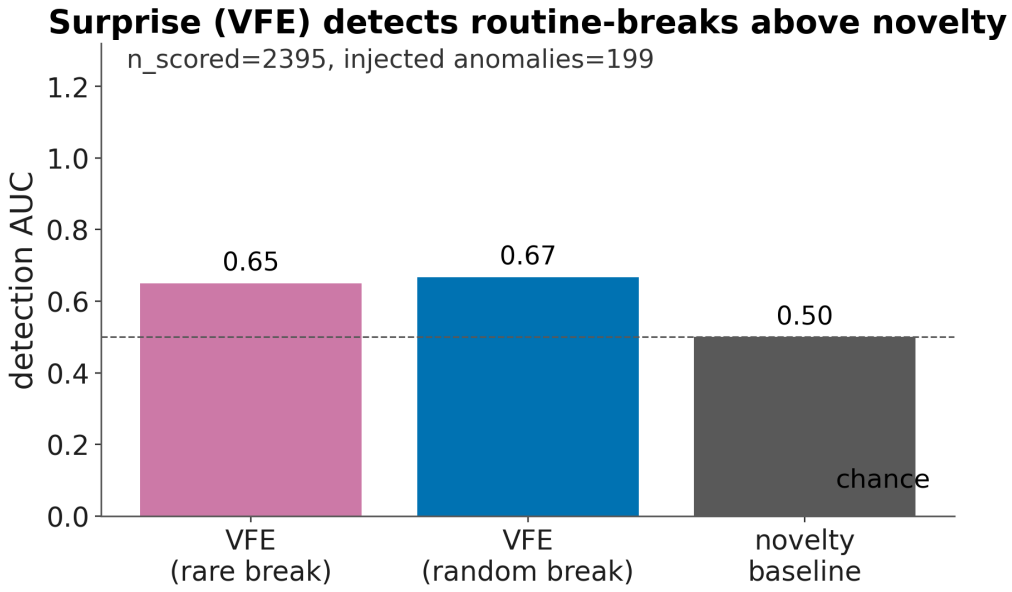


Figure 38: Source anomaly_analysis.json injects routine breaks into the synthetic observation stream after fitting the clean routine. Filter surprise, measured by variational free energy, is scored by ROC AUC against injected labels and compared with first-occurrence novelty for rare-cell and random-cell breaks. The result is a software diagnostic of synthetic perturbations, not empirical anomaly detection. Evidence: VFE surprise behavior under injected synthetic routine breaks. Boundary: deployment-ready anomaly detection.

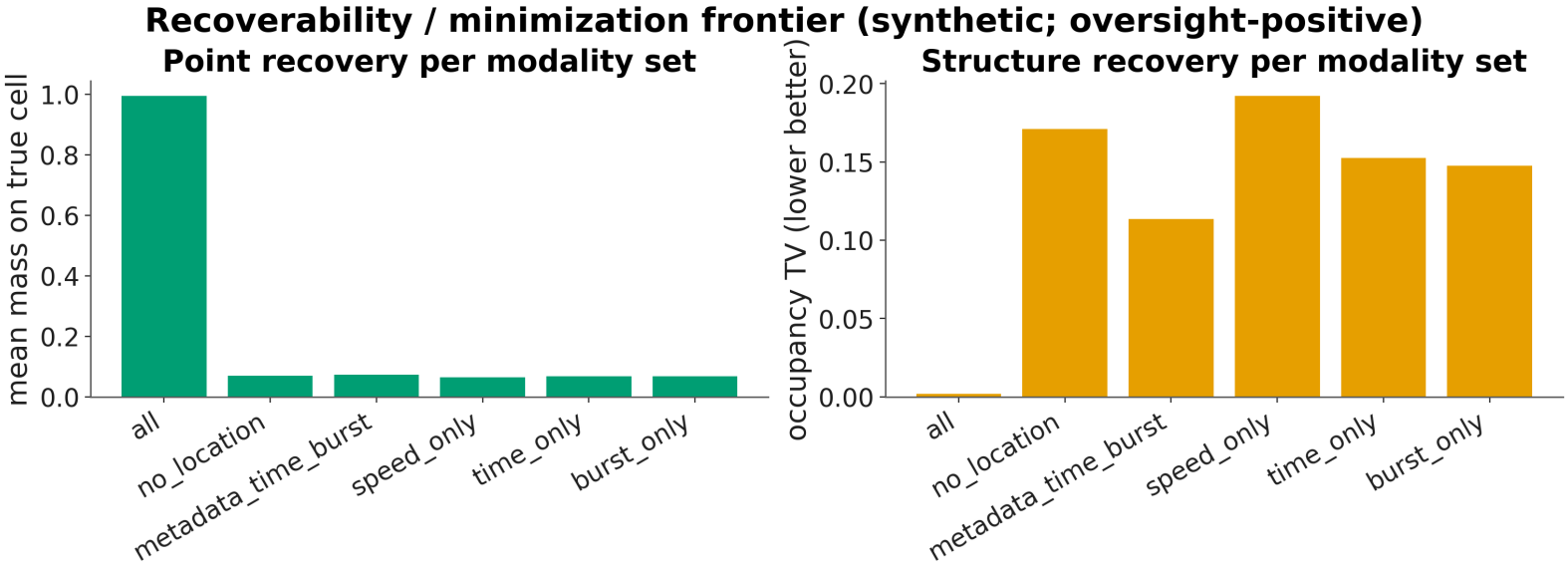


Figure 39: Source frontier_analysis.json compares observation-modality subsets on point recovery and occupancy-structure recovery. Mean posterior mass on the true cell shows what direct location buys; occupancy total-variation distance shows which aggregate structure survives minimization. The frontier is an oversight-positive map of synthetic recoverability, not evidence about real populations. Evidence: oversight-positive accounting of recovery versus minimization. Boundary: claims about real populations or individual recoverability.

15 Supplementary S8 - Safety and Privacy Analyses Stay Synthetic Risk Proxies

This supplement covers multi-factor pymdp interpretability, location-minimization and linkage-risk safety gates, privacy-utility analysis, and external-lane run evidence — diagnostic and data-governance surfaces underpinning the honest-negative thesis.

15.1 Factorial pymdp Diagnostics Expose Modes Without Cognitive Claims

The single-factor lanes treat the latent cell as the only hidden variable. The factorial pymdp lane adds interpretable structure: across its `3` hidden factors — the spatial cell, a movement mode, and a synthetic reporting mode — it factorizes the generative model so that movement texture and report-timing texture are inferred as separate, inspectable posteriors. The reporting-mode factor is a synthetic reporting-process construct derived from gap and burst structure; it represents `synthetic cognitive-proxy states, not real cognition or intent`, not a behavioral or psychological claim. fig. 40 shows the per-factor posterior entropy and the movement and reporting-mode confidence and label stability.

A multi-factor mean-field posterior is an approximation, and the lane reports the cost of that approximation rather than hiding it [Jordan et al., 1999, Blei et al., 2017]. The released pymdp variational engine is validated against an independent analytic mean-field solver: the pymdp-versus-mean-field agreement is 0.000, near machine precision, so the two implementations cross-validate. The mean-field-versus-exact marginal gap is 0.044 — the honest price of factorizing a posterior whose likelihoods couple factors — and it is reported as a diagnostic rather than folded into the agreement check. fig. 41 places these two quantities side by side with per-factor recovery sharpness.

Multi-factor interpretability is not a performance claim. fig. 42 and tbl. 15 compare the factorial lane against the transparent reference filter, single-factor pymdp, and stronger switching-HMM and particle/state-space baselines on mean next-observation predictive loss [Zucchini et al., 2016, Doucet et al., 2001, Arulampalam et al., 2002]. The factorial mean predictive loss is 5.296 nats, against 0.635 for the switching-HMM baseline and 0.550 for the particle/state-space baseline; the current best model is `particle_state_space` and the performance gate reports `factorial_pymdp_cleared_current_lane`. A non-leading gate is reported as such: the factorial lane earns its place through interpretable latent-mode diagnostics, not through a state-of-the-art predictive claim.

Table 15: Factorial pymdp model comparison emitted by `factorial_pymdp_analysis_fixture.json`.

model	implemented	mean predictive loss (nats)	interpretability surface
`reference_filter`	yes	5.719	posterior trace and tensor provenance
`single_factor_pymdp`	yes	6.147	VFE, EFE, policy posterior
`factorial_pymdp`	yes	5.296	cell, movement-mode, cognitive-proxy posteriors plus VFE/EFE
`switching_hmm`	yes	0.635	online transition regime baseline
`particle_state_space`	yes	0.550	transparent persistence/Markov mixture state-space baseline
`neural_sequence`	planned	blocked until explicit prequential training gate	future comparator only

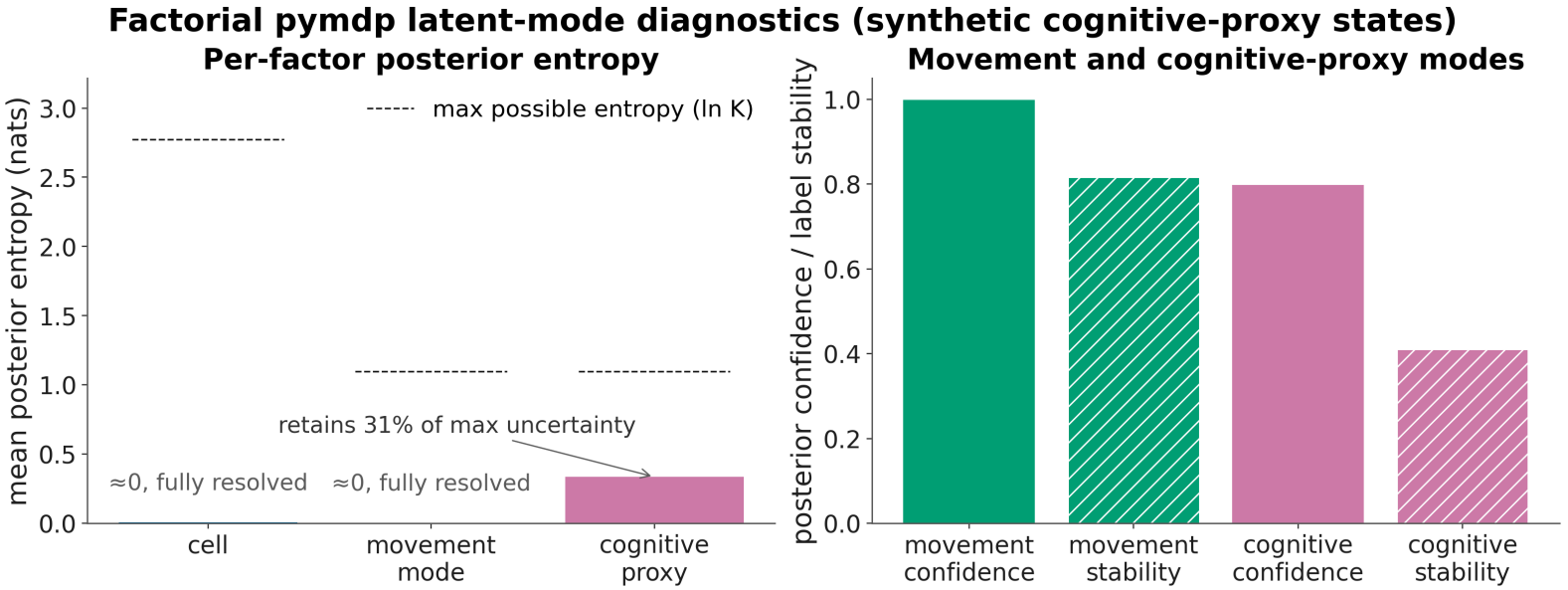


Figure 40: Source factorial_pymdp_analysis_fixture.json exposes the factorial pymdp hidden-factor diagnostics. Posterior entropy is shown against each factor's computed maximum entropy (all three factors), while mode confidence and label stability are shown for the movement-mode and synthetic cognitive-proxy reporting factors only. The cognitive-proxy factor is derived from gap and burst texture; it is not a claim about real cognition, intent, or mental state. Evidence: factor-specific posterior diagnostics in the factorial pymdp lane. Boundary: claims about real cognition, intent, or mental state.

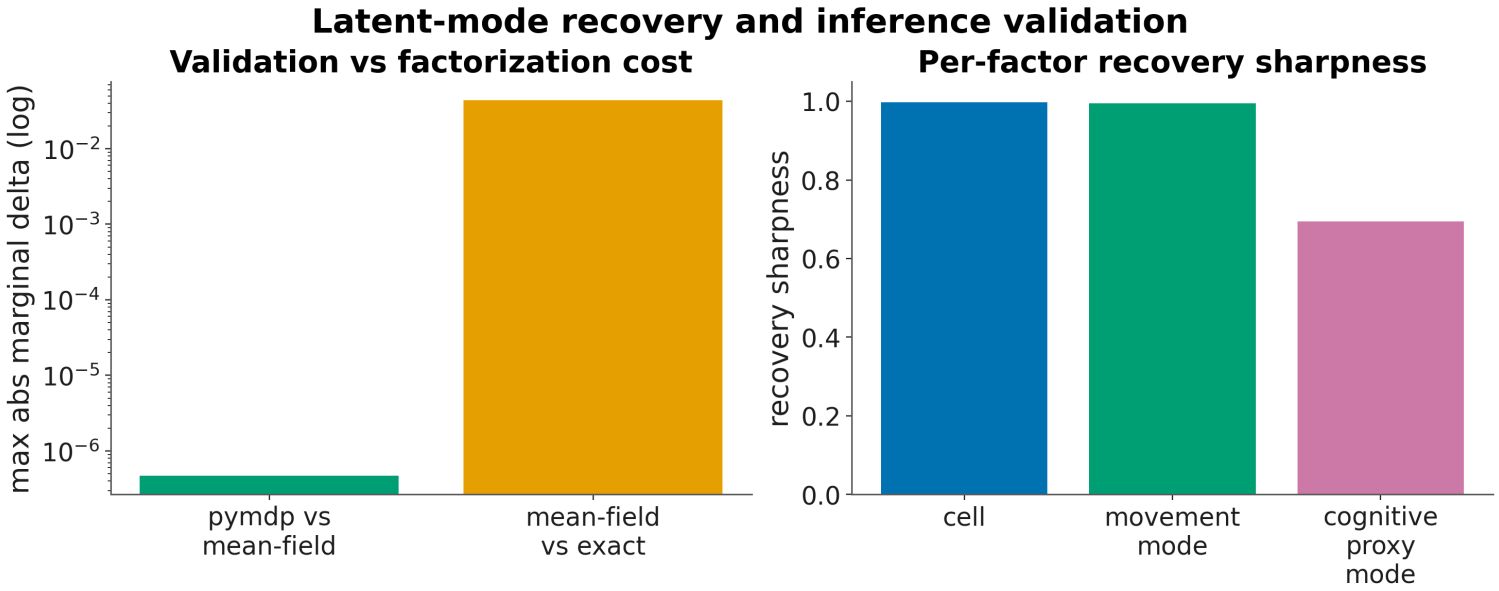


Figure 41: Source factorial_pymdp_analysis_fixture.json validates the factorial lane and reports the cost of factorizing coupled likelihoods. The pymdp variational posterior is checked against an independent analytic mean-field solver, the mean-field-versus-exact marginal gap is shown separately, and recovery sharpness is plotted per factor. Synthetic software measurements, not empirical claims. Evidence: mean-field agreement, exact-marginal gap, and factor recovery costs. Boundary: empirical latent-mode validation.

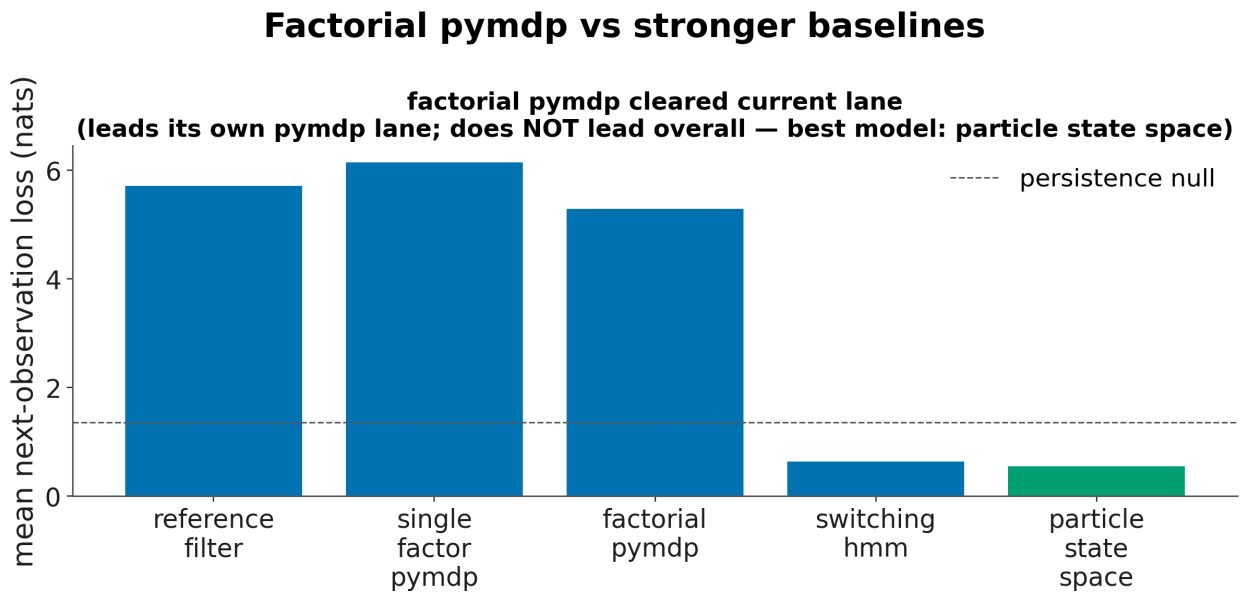


Figure 42: Source factorial_pymdp_analysis_fixture.json compares predictive loss for the transparent reference filter, single-factor pymdp, factorial pymdp, switching-HMM, and particle/state-space baselines, with persistence marked as a null. The performance gate reports whether the factorial lane leads; a non-leading verdict is part of the synthetic result. Multi-factor interpretability does not establish state-of-the-art predictive performance. Evidence: performance-gate comparison between factorial, single-factor, and sequence baselines. Boundary: state-of-the-art predictive performance when the gate is non-leading.

15.2 Location-Minimization and Integrity Gates Stay Oversight-Positive

The promoted safety-boundary artifacts in sec. 14.4 sharpen the old privacy and anomaly summaries. They are the point where model diagnostics become explicit minimization, linkage-risk, and data-integrity gates rather than general-purpose safety prose.

fig. 43 prices location in expected-free-energy terms: the location joint share is 1.000, the metadata-bundle structure-null p-value is 0.417, the residual fraction localized without location is 0, and the verdict is `metadata_recovery_not_above_structure_null`. The 4x4 expanded fixture therefore demotes any blanket metadata-recovery claim: the artifact verdict, not prior prose, decides whether the metadata residual clears the structure null, and withholding location still sharply reduces localizability on this generator.

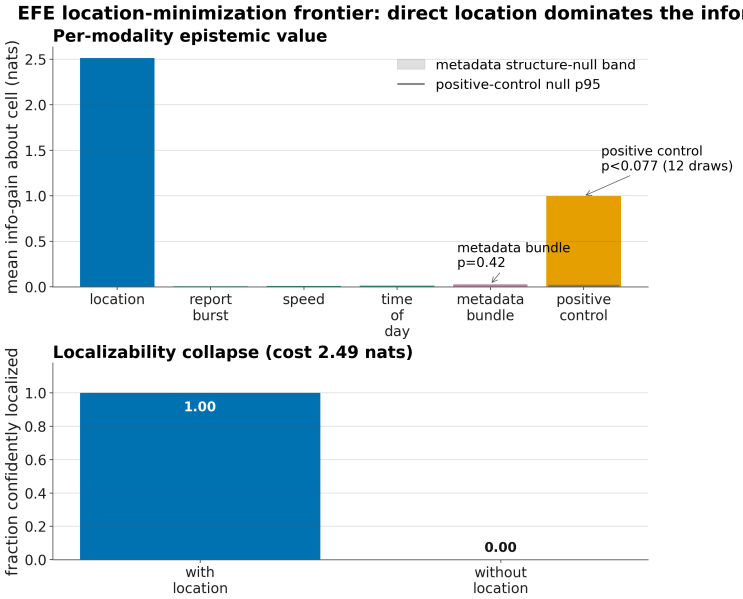


Figure 43: Source efe_location_frontier.json prices location in the model's epistemic currency. The figure compares the expected cell-information value of direct location with the non-location metadata bundle, checks metadata against a structure-destroyed null, and includes a positive control for test power. This is an oversight-positive synthetic boundary, not a collection recommendation. Evidence: the epistemic value and positive-control boundary for location minimization. Boundary: a collection recommendation.

The linkage-risk frontier in fig. 44 measures aggregate synthetic traveller re-linkability under minimization regimes. Full location has linkage 1 and verdict `location_carries_relink_risk_that_minimization_removes`; the no-location regime exceeds its structure-null band: no. This is a defensive data-minimization statistic over synthetic IDs, not a matcher, re-identification product, formal privacy mechanism, or real-population claim [de Montjoye et al., 2013, Dwork, 2006, Buchholz et al., 2024].

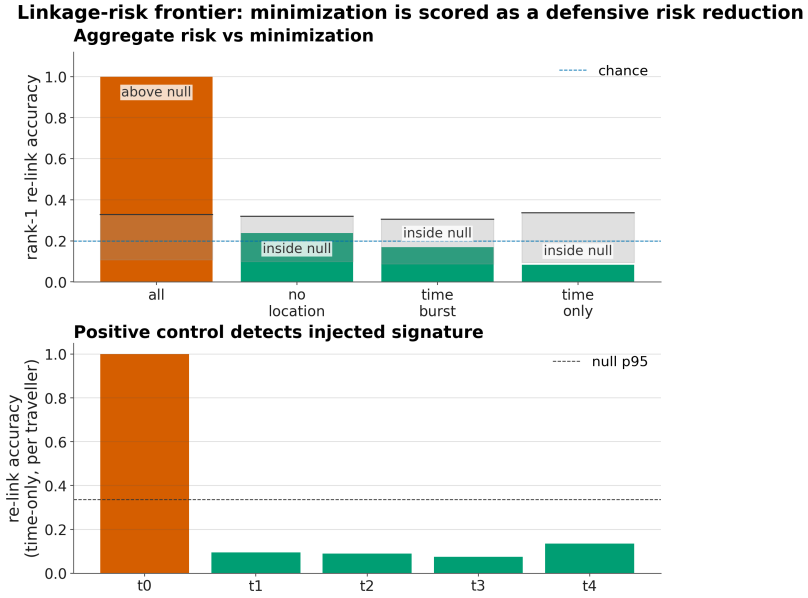


Figure 44: Source linkage_risk_frontier.json measures aggregate synthetic traveller re-link accuracy under full-location and minimized-metadata regimes, with structure-null bands and a positive control showing the weak linker can detect injected identity structure. The output is a defensive minimization-risk statistic, never a per-person linkage product or real-population claim. Evidence: defensive aggregate relink-risk accounting under minimization. Boundary: per-person linkage, tasking, or real-population inference.

The poisoning-robustness gate in fig. 45 checks whether the analyst's own VFE surprise flags corrupted inputs before downstream interpretation trusts them. Its verdict is `poisoning_detection_regime_changed`: gross teleport poisoning detected no, subtle drift remains a null against the Markov baseline yes, and the teleport mid-rate VFE AUC is 0.636. The 4x4 regeneration changes the prior fixture conclusion, so the verdict is deliberately reported as a regime-change diagnostic rather than a generalized assurance claim; which stressors separate from the Markov baseline is read from the artifact's per-mode gates, not carried over from an earlier fixture. The stressors are crude synthetic data-integrity probes and the output is a defensive self-monitoring diagnostic.

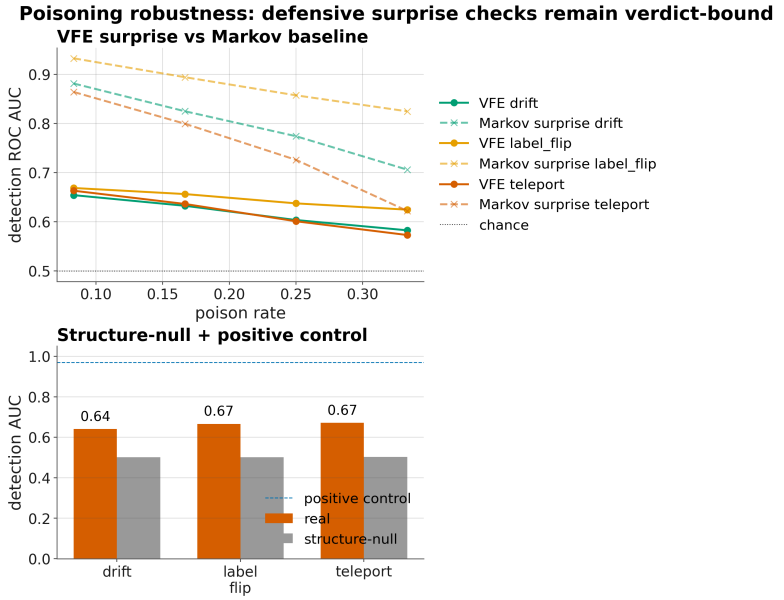


Figure 45: Source poisoning_robustness.json tests VFE surprise as a defensive data-integrity monitor under crude synthetic stressors. VFE is compared with a Markov-surprise baseline and structure-null plus positive-control checks; separated corruptions and remaining nulls are reported by the generated verdict. This is robustness triage, not a generalized assurance claim. Evidence: defensive VFE-integrity triage under crude synthetic corruptions. Boundary: generalized robustness assurance.

15.3 Privacy Frontier and External Evidence Do Not Certify Deployment

The older privacy and hardened-anomaly extensions complement the promoted artifact-bound safety checks in sec. 15.2. fig. 46 reports eq. 28: synthetic unicity is 1 with all modalities and 1 without location, while location-blind co-travel coincidence is 1. These are data-minimization risk indicators on synthetic traces, not association claims about people, privacy proofs, compliance determinations, or complete adversarial evaluations [Buchholz et al., 2024, Bouras et al., 2026, Cherigui et al., 2026]. fig. 47 compares filter surprise to the online Markov-surprise baseline from eq. 29 on subtler probes: the speed-regime shift scores VFE AUC 0.666 versus Markov AUC 0.923, and preferred-set drift scores VFE AUC 0.637. The point is not to produce an alerting system; it is to ensure the anomaly result still has signal when novelty is replaced by a transition-aware null.

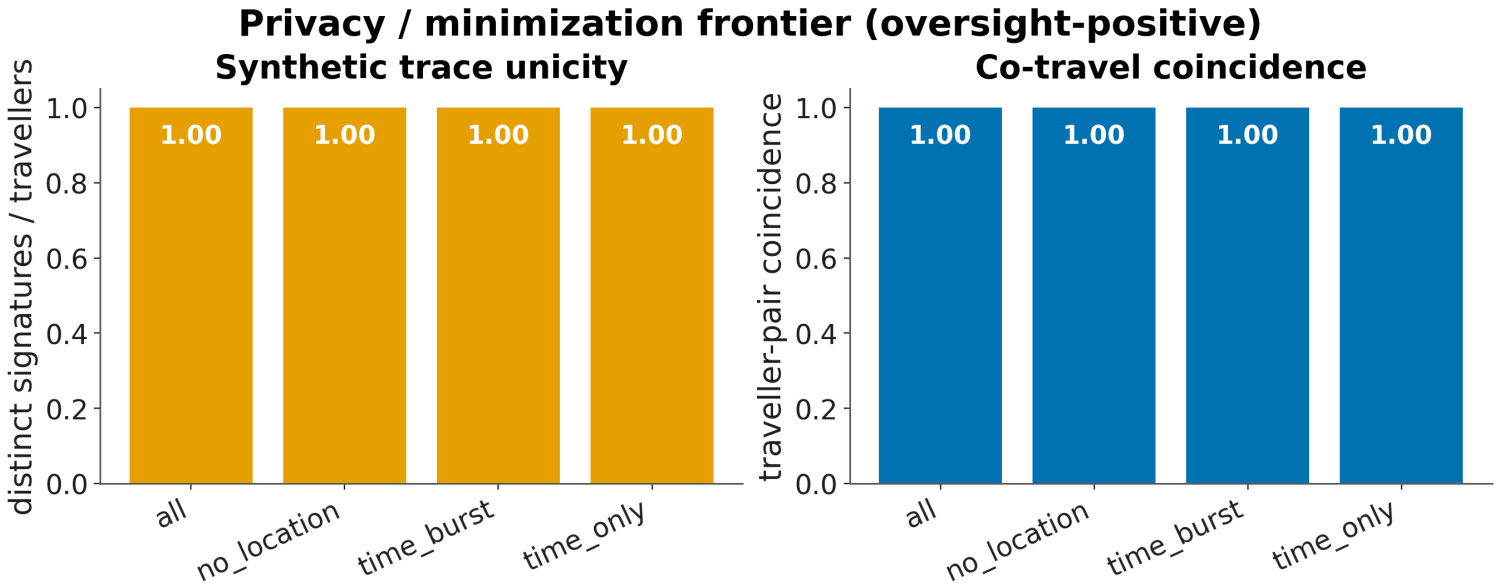


Figure 46: Source privacy_frontier_analysis.json quantifies what minimization buys back on the synthetic trace. Unicity under coarsened modality sets and co-travel coincidence rates are shown side by side for the same regimes. On this fixture unicity remains at the ceiling across every rendered coarsening, and co-travel coincidence does too except for the full 'all'-modality set, which dips slightly below ceiling (0.89) - an honest, near-no-relief finding rather than a uniform one. The figure supports oversight-positive risk accounting, not association tasking or claims about real people. Evidence: synthetic minimization accounting for unicity and co-travel coincidence risk. Boundary: association tasking or claims about real people.

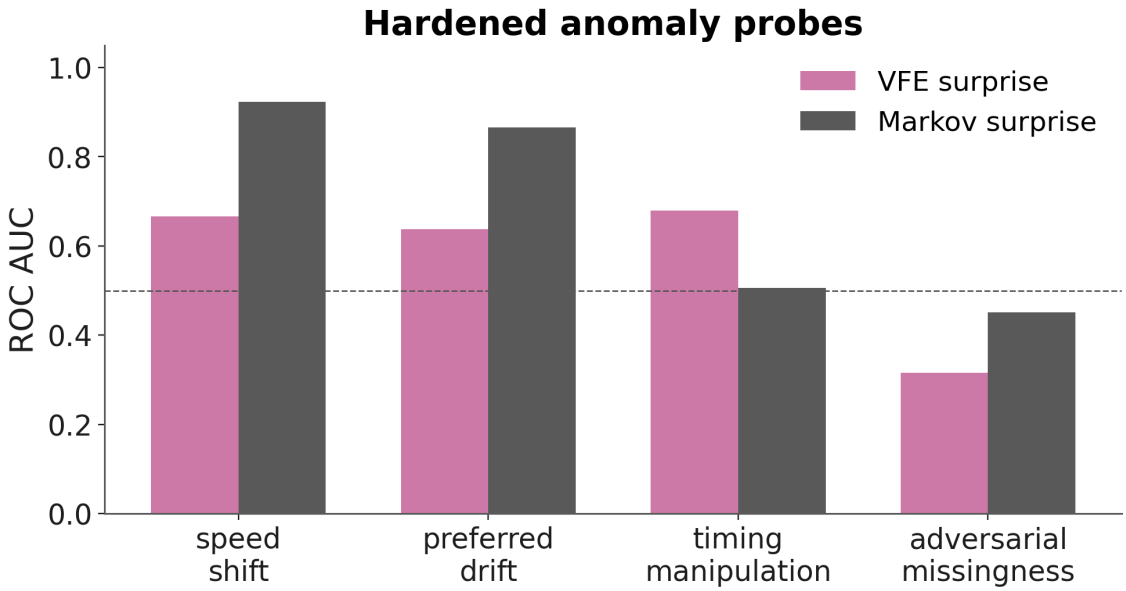


Figure 47: Source hardened_anomaly_analysis.json re-tests VFE surprise against an online Markov-surprise baseline on every hardened synthetic perturbation present in the artifact, including speed-regime shifts, preferred-set drift, timing manipulation, and adversarial missingness. Operating-point metrics live in the artifact. The figure strengthens the routine-break diagnostic while keeping anomaly claims inside synthetic software evaluation. Evidence: harder synthetic routine-break checks against an online Markov-surprise baseline. Boundary: deployment-ready anomaly monitoring.

The integrity boundary here is read from the regenerated artifacts rather than from a fixed narrative. On the hardened-anomaly artifact (fig. 47), distinct from the poisoning gate above, the timing-manipulation probe is not a null on this fixture: its VFE surprise shows a modest above-chance separation while the online Markov baseline stays near chance; that artifact carries no positive-control field, so no positive-control claim is attached to the timing probe, and the separation is reported as a synthetic perturbation-response diagnostic rather than a detector. On the poisoning gate, the transition-aware comparison is the negative one reported above: at the mid-rate gate neither teleport nor label-flip separates from the online Markov baseline (gross teleport poisoning detected no), with label-flip clearing only the above-chance floor, so no poisoning mode currently carries signal above the transition-aware null. Timing-channel integrity as a designed capability would still require a generative model in which report timing is itself informative.

The external stochastic lane status is: completed on this host with the pinned upstream build. That status is derived from the presence of the retained pinned-build metrics artifact (`output/runs/fractalrabbit_external/metrics.json`) on this host, not from the current full-study invocation. The current-run forward-gate and external poisoning artifacts record their own lane status independently and can simultaneously report the external lane as not completed in this run (sec. 14.4), in which case every external number in this paragraph describes that retained prior pinned-build run rather than same-invocation evidence. From that retained artifact, the pinned simulator build generated 5737 waypoint rows from 8 travellers over 16 latent cells, with mean predictive loss 2.920 nats, mean posterior entropy 0.128 nats, holdout mean predictive loss 3.029 nats, and a pymdp matched-prior agreement of unavailable. The external sensitivity lane resimulates waypoints for each of its 10 runs, so its replicate family reflects genuine simulator stochasticity: the replicate spread in mean predictive loss is 0.783 nats. The simulator's ground-truth place stream is also ingested: the external run reports 165 unique places with revisit rate 0.976 and 34.770 reports per place on average, and external-lane recovery (reference mean mass 0.972) is scored the same way as the fixture lane, with the same objective decomposition applied (external reference joint-loss skill 0.561 against its own uniform-chance floor). These numbers exercise the full stochastic simulator; they remain synthetic-data software evidence, not empirical mobility validation.

15.4 The Privacy-Utility Frontier Shows What Survives Minimization

The privacy frontier of sec. 15.3 measures only the *risk* side of minimization. Its dual is the governance question that matters for a minimization policy: as location is coarsened or withheld, how much useful next-cell prediction *survives*, and is that surviving utility bought back in privacy? fig. 48 pairs the two. Predictive utility is reported on a resolution-fair, shuffle-corrected axis — bits of next-location pinned beyond chance (`log2(n_states)` minus the loss in bits), minus a label-shuffle static-occupancy floor — because two confounds otherwise make a naive utility axis dishonest: a coarser grid is intrinsically easier to predict (target cardinality), and even a temporally-shuffled predictor scores on a peaked occupancy marginal. Controlling both, the endpoint comparison moves from 3.933 bits and 0.389 event-uniqueness risk at the finest grid to 1.747 bits and 0 risk at the coarsest grid, while withholding location entirely drives surviving location utility to zero by construction. On the larger canonical fixture, the full-grid ladder is monotone and clears the graceful-degradation gate; the minimum-collection operating point is grid_2x2, with 1.747 excess utility bits and 0 event-uniqueness risk against the configured 0.100 utility floor and 0.200 risk ceiling. The reading is the honest governance one: this is a source-cleared synthetic minimization frontier, not a free lunch and not a transferable real-mobility or operational claim. The risk axis is an oversight-positive minimization diagnostic over synthetic identifiers, never association capability over real people, k-anonymity, differential privacy, or a trajectory-privacy certification; the utility axis remains task-specific synthetic utility, not evidence of downstream real-life usefulness [Sweeney, 2002, Dwork, 2006, Buchholz et al., 2024, Deng et al., 2025, Kapp and Mihaljevic, 2023].

Reading the tradeoff as a **minimum-collection operating point** sharpens the governance question into a falsifiable one: what is the coarsest grid — the least location disclosed — that still clears a 0.100-bit predictive-utility floor while holding re-identification event-uniqueness below 0.200? On this larger synthetic generator the answer is source-positive (operating point found: yes): grid_2x2 preserves 1.747 excess utility bits while reducing event-uniqueness risk to 0. The binding caveat is scope, not threshold failure. This is a precise statement about this generated fixture, this grid ladder, and this next-cell prediction task; it is not evidence that the same minimization policy would preserve utility or privacy under real mobility, different tasks, temporal coarsening, suppression, k-anonymity, or deployment constraints [European Parliament and Council, 2016, National Institute of Standards and Technology, 2020, Sweeney, 2002]. The thresholds remain deliberately strict so future regenerated artifacts can fail rather than inherit this positive.

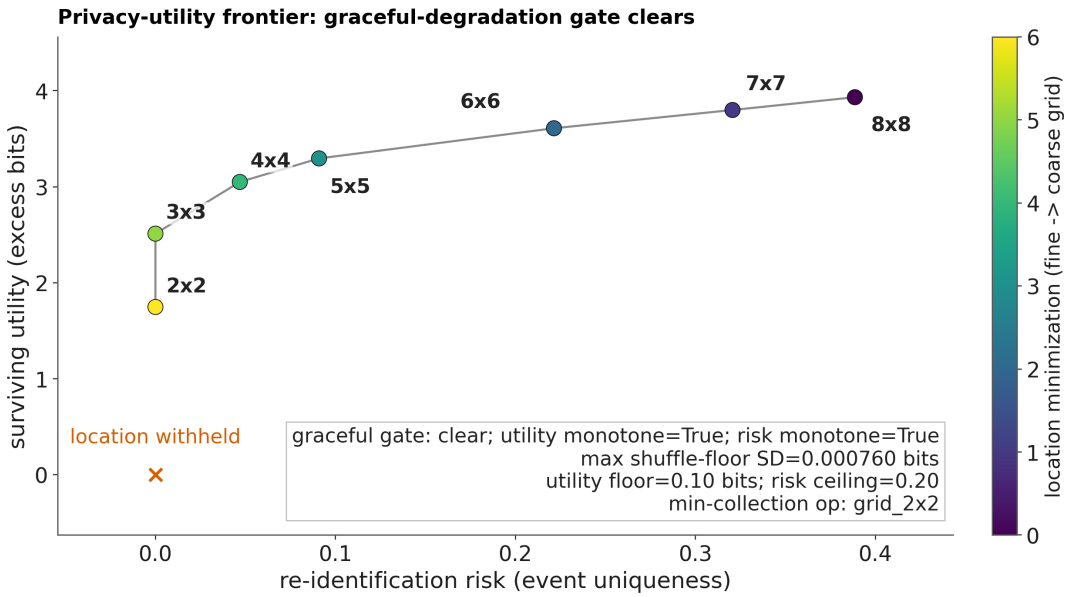


Figure 48: Source privacy_utility_frontier.json pairs surviving next-cell utility with per-event uniqueness risk across location-minimization regimes. Utility is measured as bits beyond chance and corrected for the shuffled static-occupancy floor, making grids comparable. The current artifact reports a monotone graceful-degradation gate and minimum-collection operating point grid_2x2. Evidence: a synthetic governance audit showing surviving predictive utility, re-identification risk, a source-cleared monotone graceful-degradation gate, and the current minimum-collection operating point as location is minimized. Boundary: a state-of-the-art claim, association over real people, or a real-mobility claim.

The repeated-draw audit in fig. 49 keeps this privacy-utility reading honest across available generated observations. It records 11 completed repeat rows and 0 unavailable variance-realization rows. Unavailable rows are not imputed and do not count as failures or successes; they block repeat-stability wording until the source observations exist and the regenerated gate clears. Completing every row does not by itself clear the claim: the audit content-hashes each completed source and flags any realization byte-identical to the canonical fixture as a determinism rerun rather than an independent draw, and a repeated-draw claim requires more than one genuinely distinct draw among the completed rows. When the completed rows are almost entirely byte-identical reruns of the canonical fixture, the repeat-stability wording stays blocked for that reason — lack of independent draws — even with zero unavailable rows, and the figure attests determinism rather than independent-sample stability in that case.

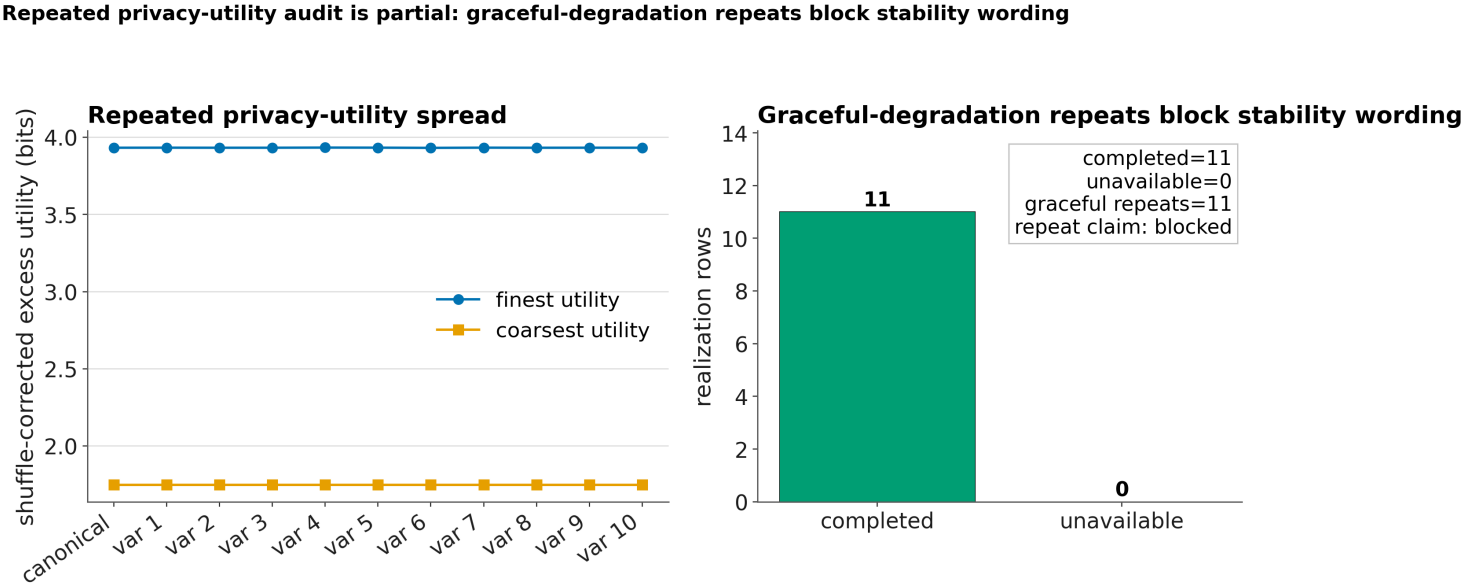
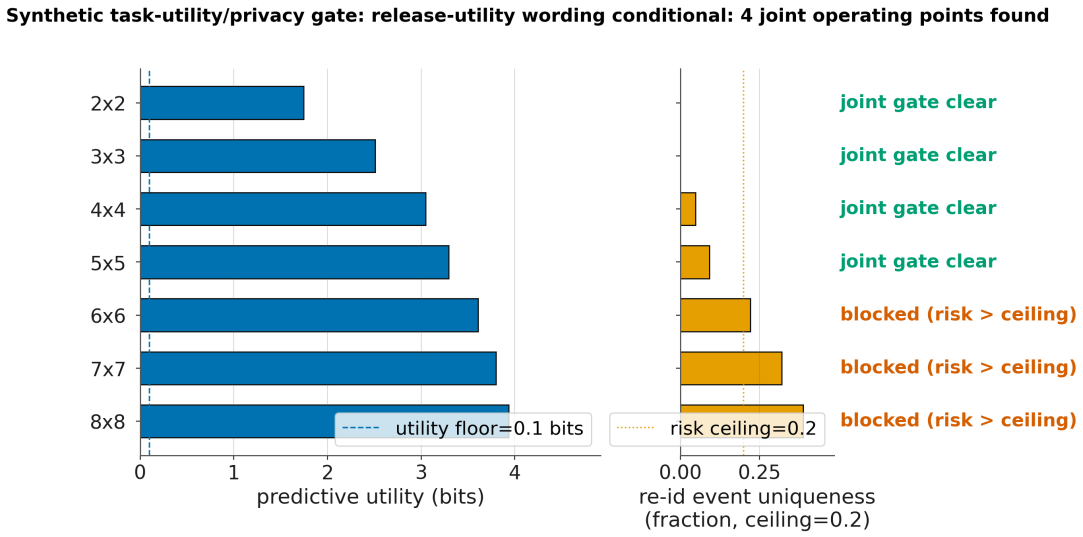


Figure 49: Source privacy_utility_frontier_repeats.json reruns the privacy-utility frontier over the canonical fixture and any generated variance-realization observation files that are present. Completed rows show finest and coarsest shuffle-corrected utility spread; unavailable realization rows are counted explicitly rather than imputed. The figure supports only a repeated-draw audit surface: when required rows are unavailable or the full repeated gate does not clear, stability and graceful-degradation wording remain blocked. Evidence: a repeated synthetic privacy-utility audit with explicit unavailable-row accounting. Boundary: repeat-stability wording when required realization rows are missing, empirical privacy evidence, or real-mobility claims.

The task-utility/privacy screen in fig. 50 makes release-utility language thresholded rather than rhetorical. It joins surviving predictive utility with re-identification event uniqueness, co-travel, recoverability, and linkage-minimization context. Current joint operating points: 4; release-utility wording allowed: yes. This remains a synthetic governance screen, not a privacy guarantee or operational usefulness claim.



status=release_utility_operating_point_found utility floor=0.10 bits risk ceiling=0.20 best minimized regime=grid_2x2 joint gate=cleared on selected rows

Figure 50: Source synthetic_task_utility_privacy.json joins predictive utility, event uniqueness, co-travel coincidence, recoverability, and linkage-minimization status. Bars compare utility with the configured utility floor and risk with the risk ceiling; each regime states whether both thresholds define a joint operating point. The synthetic artifact contains such points, so utility wording remains conditional on joint clearance. Evidence: a thresholded synthetic task-utility/privacy screen that blocks release-utility wording unless utility and privacy thresholds jointly clear. Boundary: release utility, operational usefulness, privacy guarantees, or real-world utility claims when the joint gate is blocked.

16 Supplementary Confound Taxonomy Keeps Apparent Wins Auditable

This supplement records the confound taxonomy behind the main-text claim discipline. Across development, apparent model wins were raised and then checked against live artifacts, null models, and cross-vendor audit. The resulting taxonomy is included here as an audit trail rather than as a main-line performance result. fig. 51 maps 7 catalogued apparent advantages onto 7 confound classes: data density, spatial confinement, in-sample selection, knob or measurement asymmetry, no signal, family misclassification, and instrumentation bug — the last added when a reporting-artifact miscomputation, not any model-level effect, was found to have manufactured an apparent regime clearance. The taxonomy reports 7 currently verified entries out of 8, with any missing-artifact cases left explicitly unverified rather than treated as dissolved. It also marks the fair-by-construction partial-observability positive as the surviving non-confound, because it varies only marginalize-versus-commit under a matched emission and grants no oracle access.

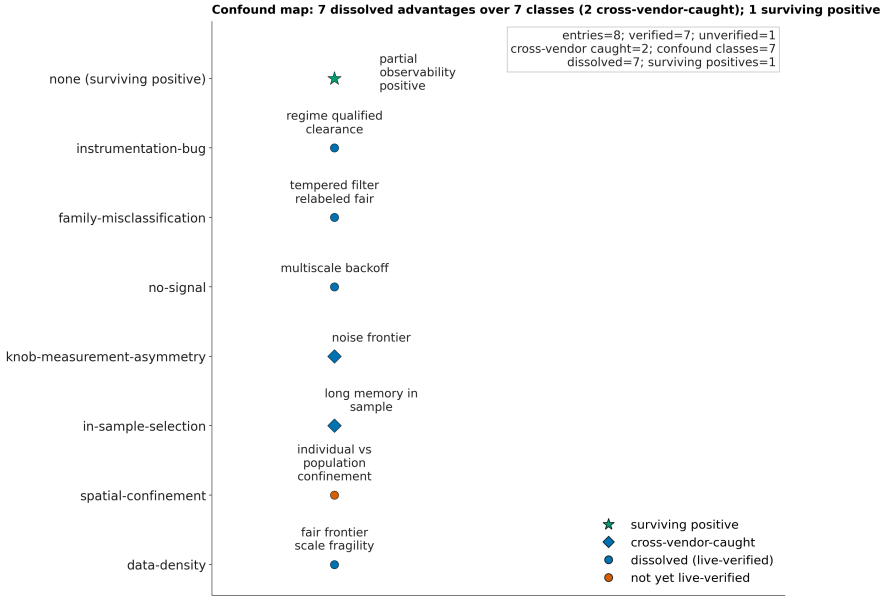


Figure 51: Source confound_taxonomy.json maps every apparent active-inference advantage this synthetic benchmark raised and then checked onto the confound class that explains it, with each entry marked as live-verified, documentation-verified, or explicitly unverified when the current external trace needed by the hook is absent. Markers sit on confound-class rows; diamonds flag confounds caught only by cross-vendor audit, and a star marks the single fair-by-construction surviving positive. The panel turns the project's pattern of dissolved naive wins into an auditable map rather than a performance result. Evidence: an auditable map of which confound class explains each dissolved active-inference advantage and which catches required cross-vendor review. Boundary: a state-of-the-art claim, an operational capability, or a real-mobility claim.

The taxonomy is bound to current repository state rather than asserted from memory. Each entry carries a verification hook that re-derives its status from an artifact, a code invariant, a gate result, an explicitly weaker documentation check for historical events, or an explicit unverified marker when the required current artifact is absent. Read with the multiplicity correction in sec. 3.2.1.1, the supplement explains why apparent active-inference wins are not promoted unless they survive the same evidence contract as the partial-observability result. It does not support an oracle-beating claim, a global state-of-the-art claim, an operational claim, or a real-mobility claim.

16.1 Where, How Much, and Why: A Per-Regime Mechanism Map

The confound taxonomy above and the multiplicity ledger in sec. 3.2.1.1 each answer part of a single question this benchmark keeps returning to: not just whether active inference beats a fair baseline, but *where* it does so, *how much* it wins or loses by, and *why*. fig. 52 joins them, adding a diagnostic that `aif_regime_searc h.py` itself never ran: this project's own confound-control battery (dominant-cell, pooled-Markov, return-boosted-soft, and oracle-previous-cell baselines), applied per regime rather than only on the default fixture. Across 5 synthetic regimes, no directional signal (the traveller-clustered confidence interval straddles zero) is found in 3 regime(s), and active inference loses outright to the strongest implemented fair baseline in 1 regime(s). The one negative-control regime (zero-noise emission) is trivial by construction, as designed.

The mechanism map's finding is more precise than a blanket "AIF reduces to a confound." 3 of the 4 non-trivial regimes see the best AIF candidate beat every confound-battery baseline (dominant-cell, pooled-Markov, and return-boosted-soft, all scored against the same noisy observations AIF itself has to work from), so an apparent AIF signal on those regimes is not reducible to those simpler naive or pooled effects. What still blocks a promoted claim differs by regime, not by one shared mechanism: on the symmetric-noise and burst-reporting-gap regimes the AIF-vs-fair-baseline advantage is nominally positive but the traveller-clustered confidence interval straddles zero, so no direction is established; on the latent-regime-switching lane, AIF clears the confound battery yet still loses outright to `online_base_rate`, a fair baseline whose adaptive tracking fits this regime's switching structure better than AIF's fixed generative model does — a loss to a stronger adaptive competitor, not to a naive confound. The remaining regime fails the confound battery outright, beating neither the pooled-Markov nor the return-boosted-soft baseline. A regime can therefore clear this project's own confound battery and still not earn a promoted claim, for structurally different reasons (statistical insufficiency versus a loss to a stronger adaptive competitor outside the battery); collapsing that distinction into a single verdict would have hidden exactly the mechanism this map exists to surface. Every gate at once (fair-baseline win, confound-battery clearance, and Holm-survival) is cleared by 0 regime(s) in the current grid — the mechanism-map methodology's bar for a claim this benchmark would call a robust, regime-local, still-synthetic win.

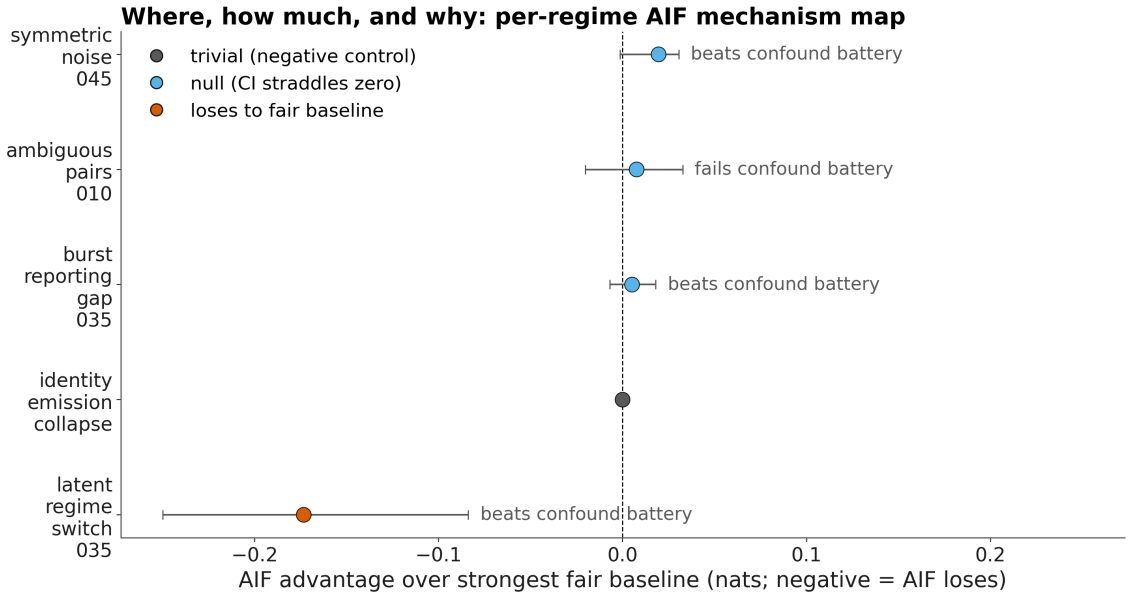


Figure 52: Source regime_mechanism_synthesis.json joins three checks per regime: AIF-versus-strongest-fair magnitude and traveller-clustered interval from aif_regime_search.json, Holm survival from multiplicity_ledger.json, and the dominant-cell, pooled-Markov, return-boosted-soft, and oracle-previous-cell confound battery. Dot and whisker encode magnitude and interval; color encodes the mechanism verdict; the margin labels confound clearance and spatial confinement. Each row names its current blocker, keeping directional loss, confounding, and multiplicity as separate reasons. Evidence: a per-regime join of AIF-vs-fair magnitude and CI, multiplicity-survival status, and confound-battery/mobility standing into one mechanism verdict per regime. Boundary: a state-of-the-art claim, a global or operational capability, or a real-mobility claim.

This map is a diagnostic layered on top of the canonical regime-search gate in sec. 3.2.1.1, not a replacement for it: it uses a single generator realization and a smaller bootstrap resample count than `aif_regime_search.json`'s own multi-realization average, so its confound-battery and mobility columns are directional, not the headline gate's own publication-grade estimate. Like every other synthetic result in this manuscript, it supports no state-of-the-art claim, no global or operational capability claim, and no real-mobility claim.

16.2 A Structural Diagnostic: Where Within the Trajectory the Loss Sits

The mechanism map above answers WHERE (`latent_regime_switch_035`) and HOW MUCH (a pooled loss to `online_base_rate`), but not WHY within that one trajectory. This regime is the only one in the grid whose generator has a named, 3-segment movement-mode schedule per traveller — a low-variance, high-return "commute" segment, a high-variance, low-return "explore" segment (the movement scale jumps sixfold at this boundary), and an intermediate "return_stabilization" segment — so it is the one regime where a within-trajectory structural breakdown is possible without inventing a new generator. fig. 53 re-slices the exact same per-step losses `aif_regime_search.py` already computes for its best AIF candidate and the strongest fair baseline, this time grouped by movement-mode segment instead of pooled, with a traveller-clustered bootstrap CI computed separately within each segment.

The one-sentence version going in would have been a guess: does AIF's disadvantage concentrate in the volatile "explore" segment, where a fixed generative model might be expected to struggle most? The regenerated artifact answers without forcing that story. Its `uniform_disadvantage` flag is False; the worst segment is return_stabilization, the best is commute, and their magnitude ratio is n/a. The ratio deliberately renders `n/a` whenever segment intervals do not support a common disadvantage sign, rather than dividing incomparable effects. The current non-uniform result localizes the loss: the adaptive online-base-rate comparator's advantage is strongest in the return-stabilization phase, present in exploration, and unresolved in commute. Online transition adaptation improves the fixed-transition AIF ablation, but does not erase that phase-structured deficit or beat the adaptive fair baseline. This is evidence against both blanket explanations — neither "AIF fails only during volatility" nor "AIF is uniformly worse everywhere" survives the phase decomposition.

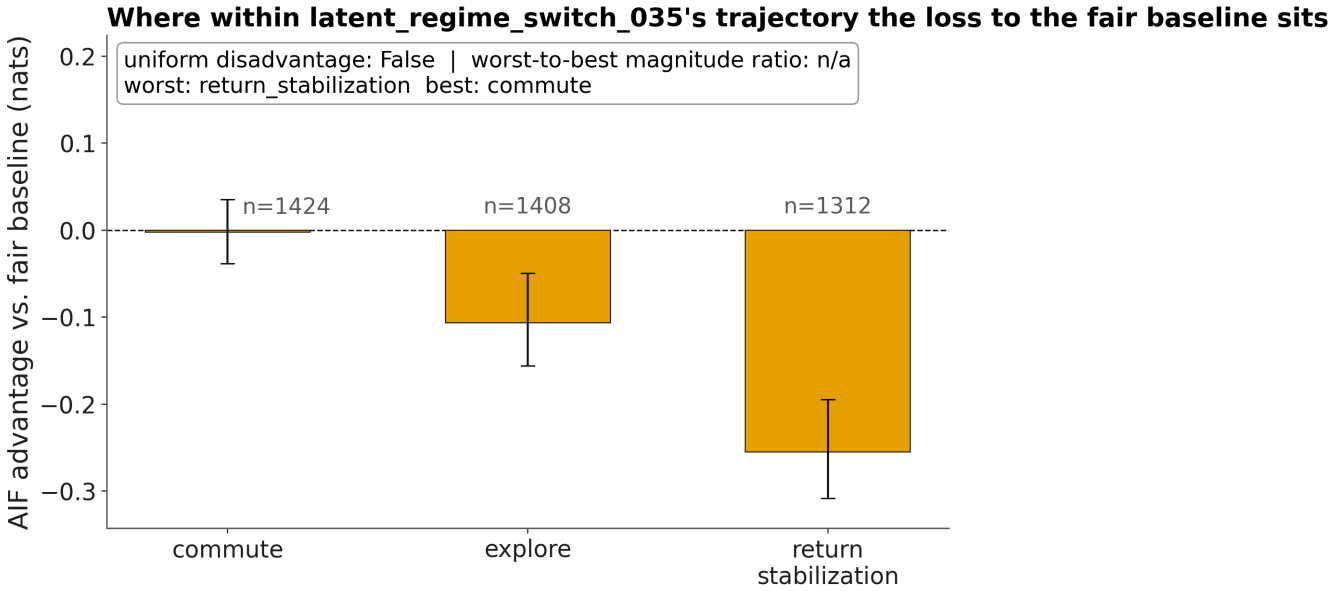


Figure 53: Source regime_switch_phase_diagnostic.json partitions the AIF-versus-strongest-fair per-step losses from latent_regime_switch_035, the one synthetic regime with named commute, explore, and return_stabilization segments. Bars report AIF advantage with traveller-clustered bootstrap intervals. CI-separated negative segments: explore, return_stabilization; zero-crossing unresolved segments: commute; uniform disadvantage: False; worst-to-best magnitude ratio: n/a. Evidence: a structural, per-movement-phase breakdown of an AIF-vs-fair-baseline loss on one regime, showing whether the loss is uniform or concentrated across phase segments. Boundary: a state-of-the-art claim, an operational capability, or a real-mobility claim.

This diagnostic is scoped to one regime and one generator realization, with a smaller bootstrap resample count than the canonical gate. It supports a structural, per-phase account of where within this one synthetic trajectory a fair-baseline loss sits; it does not support global capability, state-of-the-art standing, deployment, surveillance readiness, operational use, or real mobility.

References

Miguel Aguilera, Beren Millidge, Alexander Tschantz, and Christopher L. Buckley. How particular is the physics of the free energy principle? *Physics of Life Reviews*, 40:24–50, 2022. doi: 10.1016/j.plrev.2021.11.001. URL https://arxiv.org/abs/2105.11203.

M. Sanjeev Arulampalam, Simon Maskell, Neil Gordon, and Tim Clapp. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2):174–188, 2002. doi: 10.1109/78.978374.

Hikaru Asano, Hiroki Ouchi, Akira Kasuga, and Ryo Yonetani. MobQA: A benchmark dataset for semantic understanding of human mobility data through question answering, 2025. URL https://arxiv.org/abs/2508.11163.

Albert-Laszlo Barabasi. The origin of bursts and heavy tails in human dynamics. *Nature*, 435:207–211, 2005. doi: 10.1038/nature03459.

Hugo Barbosa, Marc Barthelemy, Gourab Ghoshal, Charlotte R. James, Maxime Lenormand, Thomas Louail, Ronaldo Menezes, Jose J. Ramasco, Filippo Simini, and Marcello Tomasini. Human mobility: Models and applications. *Physics Reports*, 734:1–74, 2018. doi: 10.1016/j.physrep.2018.01.001. URL https://arxiv.org/abs/1710.00004.

Thomas Bayes and Richard Price. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53: 370–418, 1763. doi: 10.1098/rstl.1763.0053. URL https://doi.org/10.1098/rstl.1763.0053.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x. URL https://academic.oup.com/jrsssb/article/57/1/289/7035855.

Jeremy Bentham. *Panopticon; Or, The Inspection-House*. T. Payne, 1791. URL https://archive.org/details/10059224bsb.

George Berkeley. *An Essay towards a New Theory of Vision*. Aaron Rhames, 1709. URL https://archive.org/details/bim_eighteenth-century_an-essay-towards-a-new-t_1709.

Daniel Bernoulli. Specimen theoriae novae de mensura sortis. *Commentarii Academiae Scientiarum Imperialis Petropolitanae*, 5:175–192, 1738. URL https://archive.org/details/SpecimenTheoriaeNovaeDeMensuraSortis.

Jakob Bernoulli. *Ars conjectandi*. Thurnisiorum, 1713. URL https://archive.org/details/bub_gb_kz9nvk99EWoC.

William Blackstone. *Commentaries on the Laws of England*, volume 1. Clarendon Press, 1765. URL https://archive.org/details/bim_eighteenth-century_commentaries-on-the-laws_blackstone-william-sir_1765_1.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518): 859–877, 2017. doi: 10.1080/01621459.2017.1285773.

Stavros Bouras, Ioannis Kontopoulos, Chiara Pugliese, Francesco Lettich, Emanuele Carlini, Hanna Kavalionak, Chiara Renso, and Konstantinos Tserpes. Privacy evaluation of generative models for trajectory generation, 2026. URL https://arxiv.org/abs/2605.15246.

Robert Boyle. *The Sceptical Chymist*. Internet Archive, 1661. URL https://archive.org/details/scepticalchymist0000boyl_i6r3. Reprint of the 1661 edition.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950. doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Erik Buchholz, Alsharif Abuadbba, Shuo Wang, Surya Nepal, and Salil S. Kanhere. SoK: Can trajectory generation combine privacy and utility? *Proceedings on Privacy Enhancing Technologies*, 2024(3):75–93, 2024. doi: 10.56553/popets-2024-0068. URL https://doi.org/10.56553/popets-2024-0068.

Erik Buchholz, Natasha Fernandes, David D. Nguyen, Alsharif Abuadbba, Surya Nepal, and Salil S. Kanhere. What is the cost of differential privacy for deep learning-based trajectory generation?, 2025. URL https://arxiv.org/abs/2506.09312.

A. Colin Cameron and Douglas L. Miller. A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, 50(2):317–372, 2015. doi: 10.3368/jhr.50.2.317. URL https://jhr.uwpress.org/content/50/2/317.

Ozan Catal, Tim Verbelen, Toon Van de Maele, Bart Dhoedt, and Adam Safron. Robot navigation as hierarchical active inference. *Neural Networks*, 142:192–204, 2021. doi: 10.1016/j.neunet.2021.05.010.

Theophile Champion, Howard Bowman, Dimitrije Markovic, and Marek Grzes. Reframing the expected free energy: Four formulations and a unification, 2024. URL https://arxiv.org/abs/2402.14460.

Linyao Chen, Qinlao Zhao, Zechen Li, Mingming Li, Likun Ni, Jinyu Chen, Yuhao Yao, Xuan Song, Noboru Koshizuka, and Hiroki Kobayashi. Towards efficient and evidence-grounded mobility prediction with LLM-driven agent, 2026. URL https://arxiv.org/abs/2606.05130.

Aya Cherigui, Florent Guepin, Arnaud Legendre, and Jean-Francois Couchot. A dual perspective on synthetic trajectory generators: Utility framework and privacy vulnerabilities, 2026. URL https://arxiv.org/abs/2604.19653.

William S. Cleveland and Robert McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387):531–554, 1984. doi: 10.2307/2288400.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2 edition, 2006. ISBN 9780471241959. doi: 10.1002/047174882X.

Lancelot Da Costa, Thomas Parr, Noor Sajid, Sebastijan Veselic, Victorita Neacsu, and Karl Friston. Active inference on discrete state-spaces: A synthesis. *Journal of Mathematical Psychology*, 99:102447, 2020. doi: 10.1016/j.jmp.2020.102447. URL https://arxiv.org/abs/2001.07203.

Lancelot Da Costa, Karl Friston, Conor Heins, and Grigorios A. Pavliotis. Bayesian mechanics for stationary processes. *Proceedings of the Royal Society A*, 477 (2256):20210518, 2021. doi: 10.1098/rspa.2021.0518. URL https://royalsocietypublishing.org/doi/10.1098/rspa.2021.0518.

R. W. R. Darling. Retro-preferential stochastic mobility models on random fractals under sporadic observations, 2018. URL https://www.researchgate.net/publication/340741639_Retro-preferential_Stochastic_Mobility_Models_on_Random_Fractals_Under_Sporadic_Observations.

A. Philip Dawid. Present position and potential developments: Some personal views: Statistical theory: The prequential approach. *Journal of the Royal Statistical Society. Series A*, 147(2):278–292, 1984. doi: 10.2307/2981683.

Abraham de Moivre. *The Doctrine of Chances: Or, A Method of Calculating the Probability of Events in Play*. W. Pearson, 1718. URL https://archive.org/details/bim_eighteenth-century_the-doctrine-of-chances_moivre-abraham-de_1718.

Yves-Alexandre de Montjoye, Cesar A. Hidalgo, Michel Verleysen, and Vincent D. Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific Reports*, 3:1376, 2013. doi: 10.1038/srep01376.

Bangchao Deng, Xin Jing, Tianyue Yang, Bingqing Qu, Dingqi Yang, and Philippe Cudre-Mauroux. Revisiting synthetic human trajectories: Imitative generation and benchmarks beyond datasaurus. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2025. doi: 10.1145/3690624.3709180. URL https://arxiv.org/abs/2409.13790.

Francis X. Diebold and Roberto S. Mariano. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263, 1995. doi: 10.1080/07350015.1995.10524599.

Simon Dirmeier, Ye Hong, and Fernando Perez-Cruz. Synthetic location trajectory generation using categorical diffusion models, 2024. URL https://arxiv.org/abs/2402.12242.

Arnaud Doucet, Nando de Freitas, and Neil Gordon, editors. *Sequential Monte Carlo Methods in Practice*. Springer, 2001. ISBN 9781475734379. doi: 10.1007/978-1-4757-3437-9.

Cynthia Dwork. Differential privacy. In *Automata, Languages and Programming*, volume 4052 of *Lecture Notes in Computer Science*, pages 1–12. Springer, 2006. doi: 10.1007/11787006_1.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, 1994. ISBN 9780412042317. doi: 10.1201/9780429246593.

European Parliament and Council. Regulation (eu) 2016/679 (general data protection regulation), arts. 5(1)(c), 25, 89. Official Journal of the European Union, 2016. URL http://data.europa.eu/eli/reg/2016/679/oj. Data minimisation; data protection by design and by default; research safeguards.

Tianye Fang, Xuanshu Luo, and Martin Werner. Human mobility prediction with multi-task curriculum training. In *Proceedings of the 1st International Workshop on the Human Mobility Prediction Challenge*, 2025. doi: 10.1145/3748636.3771316. URL https://www.bgd.ed.tum.de/pdf/2025_short_SIGSPATIAL_giscup.pdf.

Federal Trade Commission. FTC finalizes order with X-Mode and successor Outlogic prohibiting it from sharing or selling sensitive location data, 2024. URL https://www.ftc.gov/news-events/news/press-releases/2024/04/ftc-finalizes-order-x-mode-successor-outlogic-prohibiting-it-sharing-or-selling-sensitive-location.

Federal Trade Commission. FTC finalizes order settling allegations that GM and OnStar collected and sold geolocation data without consumers' informed consent, 2026. URL https://www.ftc.gov/news-events/news/press-releases/2026/01/ftc-finalizes-order-settling-allegations-gm-onstar-collected-sold-geolocation-data-without-consumers.

C. A. Field and A. H. Welsh. Bootstrapping clustered data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(3):369–390, 2007. doi: 10.1111/j.1467-9868.2007.00593.x.

Karl Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010. doi: 10.1038/nrn2787.

Karl Friston. A free energy principle for a particular physics, 2019. URL https://arxiv.org/abs/1906.10184.

Karl Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, and Giovanni Pezzulo. Active inference: A process theory. *Neural Computation*, 29 (1):1–49, 2017. doi: 10.1162/NECO_a_00912. URL https://activeinference.github.io/papers/process_theory.pdf.

Karl Friston, Lancelot Da Costa, Noor Sajid, Conor Heins, Kai Ueltzhoeffer, Grigorios A. Pavliotis, and Thomas Parr. The free energy principle made simpler but not too simple. *Physics Reports*, 1024:1–29, 2023a. doi: 10.1016/j.physrep.2023.07.001. URL https://arxiv.org/abs/2201.06387.

Karl Friston, Lancelot Da Costa, Dalton A. R. Sakthivadivel, Conor Heins, Grigorios A. Pavliotis, Maxwell Ramstead, and Thomas Parr. Path integrals, particular kinds, and strange things. *Physics of Life Reviews*, 47:35–62, 2023b. doi: 10.1016/j.plrev.2023.08.016. URL https://doi.org/10.1016/j.plrev.2023.08.016.

Qiang Gao, Jinyu Hong, Xovee Xu, Ping Kuang, Fan Zhou, and Goce Trajcevski. Predicting human mobility via self-supervised disentanglement learning, 2022. URL https://arxiv.org/abs/2211.09625.

Andrew Gelman, Cristian Pasarica, and Rahul Dodhia. Let's practice what we preach: Turning tables into graphs. *The American Statistician*, 56(2):121–130, 2002. doi: 10.1198/000313002317572790.

Tilmann Gneiting and Matthias Katzfuss. Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1:125–151, 2014. doi: 10.1146/annurev-statistics-062713-085831.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477): 359–378, 2007. doi: 10.1198/016214506000001437.

Kwang-Il Goh and Albert-Laszlo Barabasi. Burstiness and memory in complex systems. *Europhysics Letters*, 81(4):48002, 2008. doi: 10.1209/0295-5075/81/48002.

Marta C. Gonzalez, Cesar A. Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453:779–782, 2008. doi: 10.1038/nature06958.

John Graunt. *Natural and Political Observations Mentioned in a Following Index, and Made upon the Bills of Mortality*. National Library of Medicine Digital Collections, 1662. URL https://archive.org/details/2356015R.nlm.nih.gov.

Florent Guepin, Cheick Tidiani Cisse, Denis Renaud, Francois Bidet, and Arnaud Legendre. diffGHOST: Diffusion based generative hedged oblivious synthetic trajectories, 2026. URL https://arxiv.org/abs/2605.10647.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Kevin D. Haggerty and Richard V. Ericson. The surveillant assemblage. *British Journal of Sociology*, 51(4):605–622, 2000. doi: 10.1080/00071310020015280. URL https://doi.org/10.1080/00071310020015280.

Edmond Halley. An estimate of the degrees of the mortality of mankind, drawn from curious tables of the births and funerals at the city of breslaw. *Philosophical Transactions of the Royal Society of London*, 17:596–610, 1693. doi: 10.1098/rstl.1693.0007. URL https://doi.org/10.1098/rstl.1693.0007.

Chonghua Han, Yuan Yuan, Jingtao Ding, Jie Feng, Fanjin Meng, and Yong Li. MoveGPT: Scaling mobility foundation models with spatially-aware mixture of experts, 2025. URL https://arxiv.org/abs/2505.18670.

Ammar Haydari, Dongjie Chen, Zhengfeng Lai, Michael Zhang, and Chen-Nee Chuah. MobilityGPT: Enhanced human mobility modeling with a GPT model, 2024. URL https://arxiv.org/abs/2402.03264.

Conor Heins, Beren Millidge, Daphne Demekas, Brennan Klein, Karl Friston, Iain D. Couzin, and Alexander Tschantz. pymdp: A python library for active inference in discrete state spaces. *Journal of Open Source Software*, 7(73):4098, 2022. doi: 10.21105/joss.04098. URL https://doi.org/10.21105/joss.04098.

Richards J. Heuer. *Psychology of Intelligence Analysis*. Center for the Study of Intelligence, Central Intelligence Agency, 1999. URL https://www.cia.gov/resources/csi/static/Pyschology-of-Intelligence-Analysis.pdf.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Michael C. Horowitz and Lauren Kahn. Bending the automation bias curve: A study of human and AI-based decision making in national security contexts. *International Studies Quarterly*, 68(2):sqae020, 2024. doi: 10.1093/isq/sqae020. URL https://doi.org/10.1093/isq/sqae020.

Jessica Hullman. Why authors don't visualize uncertainty. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):130–139, 2020. doi: 10.1109/TVCG.2019.2934287.

David Hume. *An Enquiry Concerning Human Understanding*. Project Gutenberg, 1748. URL https://www.gutenberg.org/ebooks/9662. Project Gutenberg edition.

Christiaan Huygens. *De ratiociniis in aleae ludo*. Johannes Elsevier, 1657. URL https://archive.org/details/ned-kbn-all-00003375-001. Included in Frans van Schooten, Exercitationum mathematicarum libri quinque.

Edin Lind Ikanovic and Anders Mollgaard. An alternative approach to the limits of predictability in human mobility, 2016. URL https://arxiv.org/abs/1608.06419. Predictability estimates depend on the prediction target and temporal aggregation.

infer-actively. pymdp documentation, 2026a. URL https://pymdp-rtd.readthedocs.io/en/latest/. JAX-backend documentation.

infer-actively. pymdp numpy/legacy to jax migration guide, 2026b. URL https://pymdp-rtd.readthedocs.io/en/latest/migration/numpy-to-jax/. Migration documentation.

Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999. doi: 10.1023/A:1007665907178.

Bayrem Kaabachi, Jeremie Despraz, Thierry Meurers, Karen Otte, Mehmed Halilovic, Bogdan Kulynych, Fabian Prasser, and Jean Louis Raisaro. A scoping review of privacy and utility metrics in medical synthetic data. *npj Digital Medicine*, 8, 2025. doi: 10.1038/s41746-024-01359-3. URL https://www.nature.com/articles/s41746-024-01359-3.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1–2):99–134, 1998. doi: 10.1016/S0004-3702(98)00023-X.

Immanuel Kant. *Critique of Pure Reason*. Project Gutenberg, 1781. URL https://www.gutenberg.org/ebooks/4280. Project Gutenberg translation.

Raphael Kaplan and Karl J. Friston. Planning and navigation as active inference. *Biological Cybernetics*, 112(4):323–343, 2018. doi: 10.1007/s00422-018-0753-2.

Alexandra Kapp and Helena Mihaljevic. Reconsidering utility: Unveiling the limitations of synthetic mobility data generation algorithms in real-life scenarios. In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, pages 1–12, 2023. doi: 10.1145/3589132.3625661. URL https://doi.org/10.1145/3589132.3625661.

Alexandra Kapp, Julia Hansmeyer, and Helena Mihaljevic. Generative models for synthetic urban mobility data: A systematic literature review. *ACM Computing Surveys*, 56(4):1–37, 2023. doi: 10.1145/3610224. URL https://doi.org/10.1145/3610224.

Eun-Kyeong Kim and Hang-Hyun Jo. Measuring burstiness for finite event sequences. *Physical Review E*, 94(3):032311, 2016. doi: 10.1103/PhysRevE.94.032311.

Michael D. Kirchhoff, Thomas Parr, Ensor Palacios, Karl Friston, and Julian Kiverstein. The markov blankets of life: autonomy, active inference and the free energy principle. *Journal of The Royal Society Interface*, 15(138):20170792, 2018. doi: 10.1098/rsif.2017.0792. URL https://royalsocietypublishing.org/doi/10.1098/rsif.2017.0792.

Daniel Kondor, Behrooz Hashemian, Yves-Alexandre de Montjoye, and Carlo Ratti. Towards matching user mobility traces in large-scale datasets. *IEEE Transactions on Big Data*, 6(4):714–726, 2020. doi: 10.1109/TBDATA.2018.2871693. URL https://arxiv.org/abs/1709.05772.

Vaibhav Kulkarni, Natasa Tagasovska, Thibault Vatter, and Benoit Garbinato. Generative models for simulating mobility trajectories, 2018. URL https://arxiv.org/abs/1811.12801.

Pierre-Simon Laplace. Memoire sur la probabilite des causes par les evenements. *Memoires de mathematique et de physique, presentes a l'Academie royale des sciences*, 1774. URL https://sites.mathdoc.fr/cgi-bin/oetoc?id=OE_LAPLACE___8. Reprinted in Oeuvres completes de Laplace, volume 8.

Timothy LaRock, Chen Zhang, and Jurgen Hackl. Higher-order network analysis of human mobility data, 2026. URL https://arxiv.org/abs/2606.00733.

Siyu Li, Toan Tran, Lingyi Zhao, Khurram Shafique, and Li Xiong. TrajGenAgent: A hierarchical LLM agent for human mobility trajectory generation, 2026. URL https://arxiv.org/abs/2606.12657.

Leonard Christopher Limanjaya and Dae-Ki Kang. Breaking the fog with SIGHT: Attention-guided state prediction for partially observable reinforcement learning. *ACM Transactions on Intelligent Systems and Technology*, 2026. doi: 10.1145/3787973. URL https://dl.acm.org/doi/10.1145/3787973.

Haiqun Lin, Daniel O. Scharfstein, and Robert A. Rosenheck. Analysis of longitudinal data with irregular, outcome-dependent follow-up. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 66(3):791–813, 2004. doi: 10.1111/j.1467-9868.2004.b5543.x. URL https://doi.org/10.1111/j.1467-9868.2004.b5543.x.

Bo Liu, Tong Li, Zhu Xiao, Ruihui Li, Geyong Min, Zhuo Tang, and Kenli Li. All cities are equal: A unified human mobility generation model enabled by LLMs, 2026a. URL https://arxiv.org/abs/2602.19694.

Qingxiang Liu, Anqi Liang, Zhuoyang Jiang, Yutian Jiang, Sisuo Lyu, Yu Ji, Haomin Wen, and Yuxuan Liang. Think before you act: Intention-guided reasoning for LLM-based location prediction, 2026b. URL https://arxiv.org/abs/2606.08122.

Shuai Liu, Ning Cao, Yile Chen, Yue Jiang, George Rosario Jagadeesh, and Gao Cong. NextLocLLM: Location semantics modeling and coordinate-based next location prediction with LLMs, 2024. URL https://arxiv.org/abs/2410.09129.

John Locke. *An Essay Concerning Humane Understanding*. Project Gutenberg, 1690. URL https://www.gutenberg.org/ebooks/10615. Project Gutenberg edition.

Qingyue Long, Yuan Yuan, and Yong Li. UniMob: A universal model for human mobility prediction, 2024. URL https://arxiv.org/abs/2412.15294.

Massimiliano Luca, Gianni Barlacchi, Bruno Lepri, and Luca Pappalardo. A survey on deep learning for human mobility. *ACM Computing Surveys*, 55(1):1–44, 2021. doi: 10.1145/3485125. URL https://arxiv.org/abs/2012.02825.

Massimiliano Luca, Luca Pappalardo, Bruno Lepri, and Gianni Barlacchi. Trajectory test-train overlap in next-location prediction datasets. *Machine Learning*, 112:4597–4634, 2023. doi: 10.1007/s10994-023-06386-x. URL https://doi.org/10.1007/s10994-023-06386-x.

Domenico Maisto, Francesco Gregoretti, Karl J. Friston, and Giovanni Pezzulo. Active inference tree search in large POMDPs. *Neurocomputing*, 623:129319, 2025. doi: 10.1016/j.neucom.2024.129319. URL https://doi.org/10.1016/j.neucom.2024.129319.

Jesse Merhi, Erik Buchholz, and Salil S. Kanhere. Synthetic trajectory generation through convolutional neural networks, 2024. URL https://arxiv.org/abs/2407.16938. Introduces a reversible trajectory-to-CNN transformation.

Beren Millidge, Alexander Tschantz, and Christopher L. Buckley. Whence the expected free energy? *Neural Computation*, 33(2):447–482, 2021. doi: 10.1162/neco_a_01354. URL https://arxiv.org/abs/2004.08128.

Abhishek Kumar Mishra, Mathieu Cunche, and Heber H. Arcolezi. How tough is location anonymization? re-identifying 100k real-user trajectories in japan, 2025. URL https://arxiv.org/abs/2506.05611.

Tamara Munzner. *Visualization Analysis and Design*. A K Peters/CRC Press, 2014. doi: 10.1201/b17511.

Ran Nathan, Wayne M. Getz, Eloy Revilla, Marcel Holyoak, Ronen Kadmon, David Saltz, and Peter E. Smouse. A movement ecology paradigm for unifying organismal movement research. *Proceedings of the National Academy of Sciences*, 105(49):19052–19059, 2008. doi: 10.1073/pnas.0800375105.

National Institute of Standards and Technology. Nist privacy framework: A tool for improving privacy through enterprise risk management, version 1.0, 2020. URL https://doi.org/10.6028/NIST.CSWP.01162020. Privacy as an organizational risk-management problem.

National Institute of Standards and Technology. Artificial intelligence risk management framework (AI RMF 1.0), 2023. URL https://doi.org/10.6028/NIST.AI.100-1.

National Institute of Standards and Technology. The NIST cybersecurity framework (CSF 2.0), 2024. URL https://doi.org/10.6028/NIST.CSWP.29.

National Security Agency. FRACTALRABBIT: Simulate realistic trajectory data seen through sporadic reporting, 2026a. URL https://github.com/NationalSecurityAgency/fractalrabbit/tree/9933449c4f4fe1b26b6ac7bfdeeac76583085df5. GitHub repository pinned to commit 9933449c4f4fe1b26b6ac7bfdeeac76583085df5; no upstream releases were available, accessed 2026-06-10.

National Security Agency. FRACTALRABBIT mainclassfr output contract, 2026b. URL https://raw.githubusercontent.com/NationalSecurityAgency/fractalrabbit/9933449c4f4fe1b26b6ac7bfdeeac76583085df5/src/main/java/fractalRabbitGenerator/MainClassFR.java. Upstream Java entry point pinned to commit 9933449c4f4fe1b26b6ac7bfdeeac76583085df5, accessed 2026-06-10.

National Security Agency. FRACTALRABBIT parameters.csv, 2026c. URL https://raw.githubusercontent.com/NationalSecurityAgency/fractalrabbit/9933449c4f4fe1b26b6ac7bfdeeac76583085df5/resources/parameters.csv. Upstream parameter contract pinned to commit 9933449c4f4fe1b26b6ac7bfdeeac76583085df5, accessed 2026-06-10.

Victorita Neacsu, Laura Convertino, and Karl J. Friston. Synthetic spatial foraging with active inference in a geocaching task. *Frontiers in Neuroscience*, 16: 802396, 2022. doi: 10.3389/fnins.2022.802396.

Isaac Newton. *The Mathematical Principles of Natural Philosophy*. Benjamin Motte, 1729. URL https://archive.org/details/bub_gb_6EqxPav3vIsC. English translation of the Principia.

Viet Dung Nguyen, Zhizhuo Yang, Christopher L. Buckley, and Alexander Ororbia. R-AIF: Solving sparse-reward robotic tasks from pixels with active inference and world models, 2024. URL https://arxiv.org/abs/2409.14216.

Jeremy Nixon, Michael Dusenberry, Ghassen Jerfel, Timothy Nguyen, Jeremiah Liu, Linchuan Zhang, and Dustin Tran. Measuring calibration in deep learning, 2019. URL https://arxiv.org/abs/1904.01685. Expected Calibration Error is binning-sensitive; a descriptive diagnostic, not a definitive metric.

Office of the Director of National Intelligence. Panel on commercially available information: Report to the director of national intelligence, 2022. URL https://www.dni.gov/files/ODNI/documents/assessments/ODNI-Declassified-Report-on-CAI-January2022.pdf. Declassified report released by ODNI.

Office of the Director of National Intelligence. Intelligence community policy framework for commercially available information, 2024. URL https://www.dni.gov/files/ODNI/documents/CAI/Commercially-Available-Information-Framework-May2024.pdf.

Jun'ichi Ozaki, Ryosuke Susuta, Takuhiro Moriyama, and Yohei Shida. Privacy-preserving synthetic dataset of individual daily trajectories for city-scale mobility analytics. In *2025 IEEE International Conference on Big Data (BigData)*, pages 2619–2626, 2025. doi: 10.1109/BigData66926.2025.11401071. URL https://doi.org/10.1109/BigData66926.2025.11401071.

Thomas Parr, Lancelot Da Costa, and Karl J. Friston. Markov blankets, information geometry and stochastic thermodynamics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 378(2164):20190159, 2020. doi: 10.1098/rsta.2019.0159. URL https://royalsocietypublishing.org/doi/10.1098/rsta.2019.0159.

Thomas Parr, Giovanni Pezzulo, and Karl J. Friston. *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. MIT Press, 2022. ISBN 9780262045353. URL https://mitpress.mit.edu/9780262045353/active-inference/.

Toby A. Patterson, Len Thomas, Chris Wilcox, Otso Ovaskainen, and Jason Matthiopoulos. State-space models of individual animal movement. *Trends in Ecology and Evolution*, 23(2):87–94, 2008. doi: 10.1016/j.tree.2007.10.009.

Toby A. Patterson, Alison Parton, Roland Langrock, Paul G. Blackwell, Len Thomas, and Ruth King. Statistical modelling of individual animal movement: an overview of key methods and a discussion of practical challenges. *AStA Advances in Statistical Analysis*, 101(4):399–438, 2017. doi: 10.1007/s10182-017-0302-7. URL https://arxiv.org/abs/1603.07511.

Marco A. F. Pimentel, David A. Clifton, Lei Clifton, and Lionel Tarassenko. A review of novelty detection. *Signal Processing*, 99:215–249, 2014. doi: 10.1016/j.sigpro.2013.12.026.

Zhenlin Qin, Leizhen Wang, Yancheng Ling, Francisco Camara Pereira, and Zhenliang Ma. A foundational individual mobility prediction model based on open-source large language models. *Transportation Research Part C: Emerging Technologies*, 185:105562, 2026. doi: 10.1016/j.trc.2026.105562. URL https://doi.org/10.1016/j.trc.2026.105562.

Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989. doi: 10.1109/5.18626.

Maxwell J. D. Ramstead, Dalton A. R. Sakthivadivel, Conor Heins, Magnus Koudahl, Beren Millidge, Lancelot Da Costa, Brennan Klein, and Karl J. Friston. On bayesian mechanics: a physics of and by beliefs. *Interface Focus*, 13(3):20220029, 2023. doi: 10.1098/rsfs.2022.0029. URL https://royalsocietypublishing.org/doi/10.1098/rsfs.2022.0029.

Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. doi: 10.1093/biomet/63.3.581. URL https://doi.org/10.1093/biomet/63.3.581.

Noor Sajid, Philip J. Ball, Thomas Parr, and Karl J. Friston. Active inference: Demystified and compared. *Neural Computation*, 33(3):674–712, 2021a. doi: 10.1162/neco_a_01357. URL https://arxiv.org/abs/1909.10863.

Noor Sajid, Lancelot Da Costa, Thomas Parr, and Karl Friston. Active inference, bayesian optimal design, and expected utility, 2021b. URL https://arxiv.org/abs/2110.04074.

Geir Kjetil Sandve, Anton Nekrutenko, James Taylor, and Eivind Hovig. Ten simple rules for reproducible computational research. *PLOS Computational Biology*, 9(10):e1003285, 2013. doi: 10.1371/journal.pcbi.1003285.

Gustavo H. Santos, Aline Carneiro Viana, and Thiago H. Silva. When plausible is not realistic: Evaluating human mobility in LLM-based urban simulation, 2026. URL https://arxiv.org/abs/2606.13835.

Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3–4):379–423, 623–656, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.

Liushuai Shi, Le Wang, Sanping Zhou, Wei Tang, and Gang Hua. Sparse trajectory prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):2610–2627, 2026. doi: 10.1109/TPAMI.2025.3626815. URL https://doi.org/10.1109/TPAMI.2025.3626815.

John Sinclair, editor. *The Statistical Account of Scotland*. William Creech, 1791. URL https://archive.org/details/b21365799_018. Parish accounts compiled from ministers' communications.

Arfon M. Smith, Daniel S. Katz, Kyle E. Niemeyer, and FORCE11 Software Citation Working Group. Software citation principles. *PeerJ Computer Science*, 2: e86, 2016. doi: 10.7717/peerj-cs.86. URL https://doi.org/10.7717/peerj-cs.86.

Ryan Smith, Karl J. Friston, and Christopher J. Whyte. A step-by-step tutorial on active inference and its application to empirical data. *Journal of Mathematical Psychology*, 107:102632, 2022. doi: 10.1016/j.jmp.2021.102632. URL https://discovery.ucl.ac.uk/id/eprint/10143770/.

Aivin V. Solatorio. GeoFormer: Predicting human mobility using generative pre-trained transformer (GPT), 2023. URL https://arxiv.org/abs/2311.05092.

Chaoming Song, Tal Koren, Pu Wang, and Albert-Laszlo Barabasi. Modelling the scaling properties of human mobility. *Nature Physics*, 6:818–823, 2010a. doi: 10.1038/nphys1760.

Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-Laszlo Barabasi. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010b. doi: 10.1126/science.1177170.

David Spiegelhalter, Mike Pearson, and Ian Short. Visualizing uncertainty about the future. *Science*, 333(6048):1393–1400, 2011. doi: 10.1126/science.1191181.

Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570, 2002. doi: 10.1142/S0218488502001648. URL https://dl.acm.org/doi/10.1142/S0218488502001648.

Patrick Sweeney, Jaime Ruiz-Serra, and Michael S. Harre. Decision, inference, and information: Formal equivalences under active inference. *Entropy*, 28(1):1, 2026. doi: 10.3390/e28010001. URL https://www.mdpi.com/1099-4300/28/1/1.

Peizhi Tang, Chuang Yang, Tong Xing, Xiaohang Xu, Jiayi Xu, Renhe Jiang, and Kaoru Sezaki. Llama-Mob: Instruction-tuning Llama-3-8B excels in city-scale mobility prediction, 2024. URL https://arxiv.org/abs/2410.23692.

United States Census Bureau. 1790 census, 1790. URL https://www.census.gov/programs-surveys/decennial-census/decade.1790.html. Official Census Bureau historical overview.

U.S. Department of Defense. Implementing responsible artificial intelligence in the department of defense, 2021. URL https://media.defense.gov/2021/May/27/2002730593/-1/-1/0/IMPLEMENTING-RESPONSIBLE-ARTIFICIAL-INTELLIGENCE-IN-THE-DEPARTMENT-OF-DEFENSE.PDF.

Jesse van Oostrum, Carlotta Langer, and Nihat Ay. A concise mathematical description of active inference in discrete time, 2024. URL https://arxiv.org/abs/2406.07726.

Huandong Wang, Changzheng Gao, Yuchen Wu, Depeng Jin, Lina Yao, and Yong Li. Pategail: A privacy-preserving mobility trajectory generator with imitation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14539–14547, 2023. doi: 10.1609/aaai.v37i12.26700. URL https://ojs.aaai.org/index.php/AAAI/article/view/26700. Issue 12.

Hua Wei, Chacha Chen, Chang Liu, Guanjie Zheng, and Zhenhui Li. Learning to simulate on sparse trajectory data, 2021. URL https://arxiv.org/abs/2103.11845.

Greg Wilson, Jennifer Bryan, Karen Cranston, Justin Kitzes, Lex Nederbragt, and Tracy K. Teal. Good enough practices in scientific computing. *PLOS Computational Biology*, 13(6):e1005510, 2017. doi: 10.1371/journal.pcbi.1005510.

Xinhua Wu, Haoyu He, Yanchao Wang, and Qi Wang. Pretrained mobility transformer: A foundation model for human mobility, 2024. URL https://arxiv.org/abs/2406.02578.

Yuanyuan Wu, Zhenlin Qin, and Zhenliang Ma. A comprehensive evaluation framework for synthetic trip data generation in public transport, 2025. URL https://arxiv.org/abs/2510.24375.

Yi Xu, Ruining Yang, Yitian Zhang, Jianglin Lu, Mingyuan Zhang, Yizhou Wang, Lili Su, and Yun Fu. Trajectory prediction meets large language models: A survey, 2025. URL https://arxiv.org/abs/2506.03408.

Sourabh Yadav, Chenyang Yu, Xinpeng Xie, Yan Huang, and Chenxi Qiu. Protecting vehicle location privacy with contextually-driven synthetic location generation. In *Proceedings of the 32nd ACM International Conference on Advances in Geographic Information Systems*, 2024. doi: 10.1145/3678717.3691211. URL https://doi.org/10.1145/3678717.3691211.

Xiao-Yong Yan, Wen-Xu Wang, Zi-You Gao, and Ying-Cheng Lai. Universal model of individual and population mobility on diverse spatial scales. *Nature Communications*, 8(1):1639, 2017. doi: 10.1038/s41467-017-01892-8. URL https://doi.org/10.1038/s41467-017-01892-8.

Yuan Yuan, Yukun Liu, Chonghua Han, Jie Feng, and Yong Li. Breaking data silos: Towards open and scalable mobility foundation models via generative continual learning, 2025. URL https://arxiv.org/abs/2506.06694.

Yuan Yuan, Yuheng Zhang, Jingtao Ding, and Yong Li. WorldMove, a global open data for human mobility. *Scientific Data*, 13, 2026. doi: 10.1038/s41597-026-06555-2. URL https://www.nature.com/articles/s41597-026-06555-2.

Walter Zucchini, Iain L. MacDonald, and Roland Langrock. *Hidden Markov Models for Time Series: An Introduction Using R*. CRC Press, 2 edition, 2016. ISBN 9781482253832. doi: 10.1201/b20790.