# CONSIDERING CONTEXT: WHEN WORLD MODELS NEED CONTEXT ENCODERS

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

## ABSTRACT

Methods for generalization in model-based reinforcement learning typically assume that an agent cannot recover the latent context governing the environment dynamics from its own experience, and therefore supplies it externally. We formalize and test this assumption with *predictive sufficiency*, which quantifies what access to the context adds to next-step prediction under the visitation distribution an agent induces, and separates that quantity into a history-recoverable part, a residual requiring the true context, and the deficit added by a finite model. We classify context-aware algorithms by the predictive risk their conditioning set can target and demonstrate across environments of increasing identification difficulty that the headroom does not follow the MDP class. The same task under different priors leaves predictive headroom in one setting and nothing distinguishable from zero in another, where the agent's behavior implicitly identifies the context and any benefit of such a mechanism cannot be attributed to missing information. Where headroom persists, the learned state exposes it only partially, and adding the true context still lowers the risk. Our contribution is a practical criterion for matching contextual mechanisms to the information available to them, estimated from the ordinary trained agent without a reference policy.

## 1 INTRODUCTION

In model-based reinforcement learning (RL), even minor changes in environment dynamics, e.g., the mass of a body or the friction of a floor in robotics, can degrade next-state predictions and, through compounding rollout errors, the policies built upon them (Lee et al., 2020). Settings with such variations are often formalized as contextual MDPs (Hallak et al., 2015) or hidden parameter MDPs (Doshi-Velez & Konidaris, 2016). The latent parameters may be drawn i.i.d. from a shared prior (Benjamins et al., 2023), evolve according to a background process, as in dynamic parameter MDPs (Xie et al., 2021), or vary without structure, folding the problem into a general partially observable MDP. They can also change between episodes, during an episode, or even at every timestep (Dulac-Arnold et al., 2021).

In these settings, a predictor that only sees the current state and action has to average over the contexts that are still plausible at that point, resulting in a model mixture and reduced prediction accuracy. A common solution is to infer the missing information online and provide it to the agent, where existing approaches include supervised system identification, meta-learned adaptation, latent context encoders, or conditioning on ground-truth context (Duan et al., 2016; Nagabandi et al., 2019; Lee et al., 2020; Zintgraf et al., 2020; Evans et al., 2022; Röder et al., 2025; Prasanna et al., 2024). Those methods aim for improving how a finite model uses experience without adding new observations. However, the agent's trajectories can contain sufficient evidence about the active context, and a model that uses history already has access to this information. What a contextual mechanism can still add is exactly the one-step advantage that observing the context grants a predictor under the data distribution the agent itself induces.

Figure 1 illustrates these effects. Pooling distinct contexts yields an averaged kernel that governs no particular episode, while trajectories reveal or conceal the context depending on where they lead, identifying the dynamics before any dedicated mechanism infers them.

We formalize this question through *predictive sufficiency*, quantifying the Bayes one-step advantage of context for memoryless and history-aware predictors. We distinguish prior-averaged structural



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

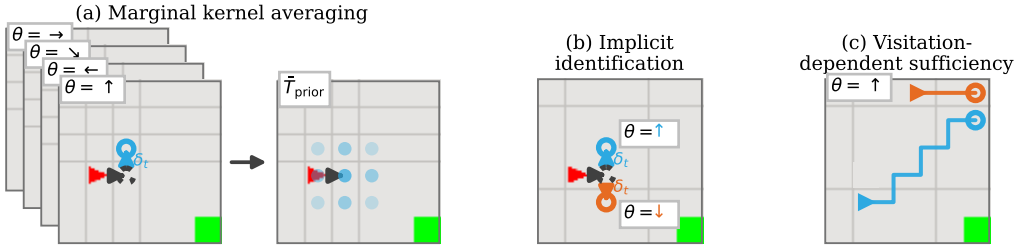


Figure 1: Contextual variation in MiniGrid, where a hidden wind variable probabilistically displaces the agent after each move, and walls block the displacement. (a) A model trained on pooled experiences with different winds learns the averaged kernel, which has no net displacement and inflated spread and governs no individual episode. (b) Under the same policy the trajectories separate mid-episode, so the state alone is consistent with only one wind and identifies it implicitly, rendering contextual mechanisms partially redundant. (c) Under one and the same context, the trajectory of an agent crossing the open grid (blue) bends downwind and betrays the context, while an agent moving along the downwind border (orange) shows no drift at all, since the obstacle blocks it. What remains for context mechanisms to gain depends on what the agent's own trajectories already reveal, and on what they cover.

variation from *predictive headroom*, the potential improvement in next-state prediction from observing the true context under the agent's observed trajectories. We split headroom into information recoverable from history and a residual requiring context, while separating the learned predictor's finite-model deficit. Reading context-aware methods through the information they access turns the decomposition into class-level predictions: an encoder has no missing information to supply where visitation alone makes the context identifiable, short windows can recover headroom carried by recent evidence, and persistent representations can exploit information spanning episode boundaries.

Our contribution is threefold: i) we analyze visitation-mediated implicit identification, where the agent's behavior concentrates the posterior over the context until nothing is left for another mechanism to supply; ii) we formulate predictive sufficiency, which asks a representation to preserve the transition information the context carries rather than the context itself, and decompose the resulting gap exactly across the inputs a method conditions on; and iii) we propose a principled criterion, confirmed in experiments across MDP classes of increasing identification difficulty, that determines which conditioning sets have predictive information to add and whether linear decoding of a learned state recovers it.

## 2 Related Work

**MDPs with latent context variables.** Contextual MDP formulation (Hallak et al., 2015) also appears as *hidden-parameter* (Doshi-Velez & Konidaris, 2016), *latent* (Kwon et al., 2021), and is closely related to *Bayes-adaptive* (Duff, 2002) MDPs, which consider a belief over the unknown parameters. Work in these settings has studied exploration (Henaff et al., 2023), adaptation, regret, and zero-shot generalization (Kirk et al., 2023). Common benchmarks in gaming settings (Chevalier-Boisvert et al., 2023; Cobbe et al., 2020; Hafner, 2022) typically vary per-episode initial state distribution rather than dynamics, while meta-RL tasks (Rakelly et al., 2019) mix reward, goal, and dynamics variation. The Real-World Reinforcement Learning suite (Dulac-Arnold et al., 2021) includes dynamics randomization, stochastic disturbances, and degradations in continuous control, varied between episodes. CARL (Benjamins et al., 2023) turns continuous environment parameters into explicit context variables and defines protocols on contextualized control tasks. LILAC (Xie et al., 2021) instead considers structured evolution of an episode-level latent across a sequence of tasks. We use these settings to study transition prediction and ask whether and how predictive differences carry to downstream control.

**Context inference and conditioning.** In meta-RL, PEARL (Rakelly et al., 2019) infers a latent task variable with a permutation-invariant encoder and VariBAD (Zintgraf et al., 2020) maintains a Bayes-adaptive belief over task parameters, separating information shared across tasks from information inferred within one, in the spirit of the empirical-Bayes recasting of meta-learning (Grant

et al., 2018). For dynamics, UP-OSI (Yu et al., 2017) learns a supervised map from recent state–action histories to physical parameters that condition a universal policy, GrBAL and ReBAL (Nagabandi et al., 2019) meta-train forward models that adapt from recent trajectory segments through gradient-based and recurrent updates respectively, and CaDM (Lee et al., 2020) and IIDA (Evans et al., 2022) infer a latent dynamics context from recent transitions without observing the true parameters. Within Dreamer, cRSSM (Prasanna et al., 2024) conditions the world model and policy on observed context, while DALI (Röder et al., 2025) infers it with a separate history encoder and forward-dynamics objective. While these methods offer inductive biases for inference from experience, we ask a prior question, namely what information their inputs can supply under a given visitation, which our framework makes precise.

## 3 Theoretical Framework

**Preliminaries.** We formalize the environment as a contextual MDP $(S, A, T_\theta, R_\theta, \gamma)$, where both the transition kernel $T_\theta$ and the reward function $R_\theta$ are parameterized by a latent context variable $\theta$. We denote an episode trajectory by $\tau = (s_0, a_0, s_1, a_1, \ldots, s_T)$ and its prefix up to time $t$ by $\tau_{\leq t}$. The context is drawn at the start of each episode, either independently from a prior $p(\theta \mid \alpha)$ with shared hyperparameters $\alpha$, or from an autonomous Markov chain $p(\theta_{i+1} \mid \theta_i)$ across episodes. We restrict attention to contexts fixed within an episode: time-varying context requires an augmented latent process and remains outside the present framework. Our analysis is concerned with the transition kernel, and its predictive quantities require no reward labels.

**Pooled and prior-averaged prediction.** Let $p^\pi(\theta, s, a, s')$ denote the joint distribution induced by drawing $\theta \sim p(\theta \mid \alpha)$, rolling out policy $\pi$ under $T_\theta$, and sampling a transition according to the evaluation weighting. In the realizable infinite-data limit, a memoryless world model trained by maximum likelihood on the pooled transitions converges to

$$\bar{T}_{\mathrm{emp}}(s' \mid s, a) = p^\pi(s' \mid s, a) = \mathbb{E}_{\theta \sim p^\pi(\theta \mid s,a)}\big[T_\theta(s' \mid s, a)\big]. \tag{1}$$

By Bayes' rule, $p^\pi(\theta \mid s, a) \propto p(\theta \mid \alpha) d^\pi_\theta(s, a)$, where $d^\pi_\theta$ is the corresponding state–action occupancy induced by $T_\theta$ and $\pi$. Thus it need not equal $p(\theta \mid \alpha)$, making $\bar{T}_{\mathrm{emp}}$ an observational mixture and the Bayes-optimal log-loss predictor conditioned only on $(s, a)$. At a fixed query $(s, a)$, averaging the conditional kernels under the context prior gives

$$\bar{T}_{\mathrm{prior}}(s' \mid s, a) = \mathbb{E}_{\theta \sim p(\theta \mid \alpha)}\big[T_\theta(s' \mid s, a)\big], \tag{2}$$

the Bayesian model average (Hoeting et al., 1999). We refer to this mixture's predictive loss relative to the context-conditional kernels as the *marginal averaging problem*. This standardized reference compares kernels at a shared $(s, a)$ before visitation reweights the context prior. If $\theta \perp (s, a)$ under $p^\pi$, it equals $\bar{T}_{\mathrm{emp}}$. For a $\theta$-independent reference distribution $\mu(s, a)$, define

$$\mathcal{I}_{\mathrm{struct}} = \mathbb{E}_{(s,a)\sim\mu}\, \mathbb{E}_{\theta \sim p(\theta \mid \alpha)}\left[D_{\mathrm{KL}}\big(T_\theta(\cdot \mid s, a) \,\big\|\, \bar{T}_{\mathrm{prior}}(\cdot \mid s, a)\big)\right]. \tag{3}$$

$\mathcal{I}_{\mathrm{struct}}$ is conditional mutual information under the constructed joint $\mu(s, a)p(\theta \mid \alpha)T_\theta(s' \mid s, a)$ and averages the corresponding Jensen gap across shared queries (Appendix A). This quantity measures kernel separation rather than the training loss of a pooled predictor.

**Predictive headroom under visitation.** Consider the memoryless predictor as per Equation 1 and an Oracle predictor that additionally observes the realized context, so that its target is $T_\theta$. The difference between their minimum expected log losses is

$$\mathcal{I}_{\mathrm{emp}} = I(\theta; s_{t+1} \mid s_t, a_t) = \mathbb{E}_{(s,a)\sim p^\pi}\, \mathbb{E}_{\theta \sim p^\pi(\theta \mid s,a)}\left[D_{\mathrm{KL}}\big(T_\theta(\cdot \mid s, a) \,\big\|\, \bar{T}_{\mathrm{emp}}(\cdot \mid s, a)\big)\right]. \tag{4}$$

We call $\mathcal{I}_{\mathrm{emp}}$ the *predictive headroom*. It is the exact reduction in one-step log loss from revealing $\theta$ to a predictor conditioned on $(s_t, a_t)$. A finite memoryless model can incur additional estimation, approximation, or optimization error.

$\mathcal{I}_{\mathrm{emp}}$ uses the query distribution and context weights induced by data collection, whereas $\mathcal{I}_{\mathrm{struct}}$ averages over $\mu$. The two quantities therefore coincide when $\mu(s, a) = p^\pi(s, a)$ and the independence condition above holds, but neither bounds the other in general. The difference $\mathcal{I}_{\mathrm{struct}} - \mathcal{I}_{\mathrm{emp}}$ is, therefore, descriptive, and may be negative. If $p^\pi(\theta \mid s, a)$ is concentrated, the current $(s, a)$ already identifies much of the context, so large structural variation can coexist with little predictive

headroom. We call this *visitation-mediated implicit identification*. However, context remains informative for prediction when visitation overlaps across contexts and the posterior over $\theta$ stays diffuse. In practice, measuring how well $\theta$ can be decoded from the current $(s, a)$ reveals to what extent the state and action already identify the context.

In discrete environments with repeated contexts, $\mathcal{I}_{\mathrm{emp}}$ can be estimated from context-labelled transition counts, whereas continuous settings require mutual information estimation. Small values leave little advantage even for a memoryless Oracle, while large values expose one-step errors that can compound in model rollouts. Proposition 2 in Appendix A makes this connection explicit: with bounded rewards, the value error under any stationary Markov policy is bounded by $\gamma R_{\max}/(1-\gamma)^2$ times the worst-case transition error (Kearns & Singh, 2002; Brafman & Tennenholtz, 2002), motivating separate evaluation of predictive headroom and return.

**History, Oracle context, and finite models.** Assume the transition process is Markov conditional on $\theta$, so that $s_{t+1} \perp \tau_{<t} \mid (s_t, a_t, \theta)$. The Bayes-optimal history-aware target is

$$\bar{T}_{\mathrm{hist}}(s' \mid \tau_{\leq t}, a_t) = \mathbb{E}_{\theta \sim p^{\pi}(\theta \mid \tau_{\leq t}, a_t)}\big[T_\theta(s' \mid s_t, a_t)\big]. \tag{5}$$

Let $\mathcal{R}_{\mathrm{M}}$, $\mathcal{R}_{\mathrm{H}}$, and $\mathcal{R}_{\mathrm{O}}$ denote the minimum expected one-step log loss when the predictor observes $(s_t, a_t)$, $(\tau_{\leq t}, a_t)$, and $(s_t, a_t, \theta)$, respectively. The residual Oracle headroom after observing history is

$$\mathcal{I}_{\mathrm{hist}} = I(\theta; s_{t+1} \mid \tau_{\leq t}, a_t) = \mathcal{R}_{\mathrm{H}} - \mathcal{R}_{\mathrm{O}}. \tag{6}$$

**Proposition 1** (Predictive headroom decomposition)**.** *Under the conditional-Markov assumption and a common transition distribution,*

$$\mathcal{I}_{\mathrm{emp}} = \mathcal{R}_{\mathrm{M}} - \mathcal{R}_{\mathrm{O}} = \underbrace{\mathcal{R}_{\mathrm{M}} - \mathcal{R}_{\mathrm{H}}}_{\textit{headroom closed by history}} + \underbrace{\mathcal{I}_{\mathrm{hist}}}_{\textit{residual Oracle headroom}} \tag{7}$$

*Consequently,* $0 \leq \mathcal{I}_{\mathrm{hist}} \leq \mathcal{I}_{\mathrm{emp}}$.

The proof appears in Appendix A. The first term measures predictive information recovered from trajectory history beyond $(s_t, a_t)$. The second is the improvement still available from revealing $\theta$ to an ideal history-aware predictor. When $\mathcal{I}_{\mathrm{hist}} = 0$, context provides no further advantage after history. For any finite recurrent model $q$, the remaining Oracle gap is

$$\mathcal{R}(q) - \mathcal{R}_{\mathrm{O}} = \big[\mathcal{R}(q) - \mathcal{R}_{\mathrm{H}}\big] + \mathcal{I}_{\mathrm{hist}}, \tag{8}$$

where the bracketed term is estimation, approximation, or optimization error. Neither a context pathway computed from the same within-episode history nor a different recurrent core can improve on $\mathcal{R}_{\mathrm{H}}$ in principle, but either may help a finite model approach it. The proposition applies to a conditionally Markov state; in a partially observed process, the Oracle risk must also condition on history.

**Contexts correlated across episodes.** The decomposition above takes $\theta$ independent across episodes, so histories end at episode boundaries. When contexts instead evolve as a Markov chain, $\theta_{i+1} \sim p(\theta_{i+1} \mid \theta_i)$, completed episodes carry information about the current context (Choi et al., 2001; Xie et al., 2021). For an ergodic chain, the stationary distribution plays the role of the prior. Let $\tau^{+}_{\leq t}$ denote the cross-episode history concatenating completed episodes with the current prefix, and $\mathcal{R}_{\mathrm{H}^+}$ the corresponding minimum risk. Since $\tau_{\leq t} \subseteq \tau^{+}_{\leq t}$,

$$\mathcal{I}_{\mathrm{emp}} = \underbrace{\mathcal{R}_{\mathrm{M}} - \mathcal{R}_{\mathrm{H}}}_{\text{within-episode}} + \underbrace{\mathcal{R}_{\mathrm{H}} - \mathcal{R}_{\mathrm{H}^+}}_{\text{cross-episode}} + \underbrace{I\big(\theta;\, s_{t+1} \mid \tau^{+}_{\leq t}, a_t\big)}_{\text{residual}}, \tag{9}$$

where the middle term vanishes under independent draws, which motivates episode-start evaluation before within-episode evidence accumulates. The in-principle bound of Equation 8 is relative to the information interface: $\mathcal{R}_{\mathrm{H}}$ bounds any predictor computed from the current episode, while a recurrent state carried across resets or an explicit model of the context chain targets $\mathcal{R}_{\mathrm{H}^+} \leq \mathcal{R}_{\mathrm{H}}$.

**Predictive sufficiency.** Predictive sufficiency asks a context representation to preserve the transition information carried by $\theta$ without requiring recovery of $\theta$ itself.

**Definition 1** ($\varepsilon$-predictive sufficiency)**.** *For a data-collection policy $\pi$, let $\tau$ be generated under episode context $\theta \sim p(\theta \mid \alpha)$. A representation $c_t$ computed from $\tau_{\leq t}$ induces the Bayes-optimal predictor*

$$\bar{T}_c(s' \mid s_t, a_t, c_t) = \mathbb{E}_{\theta' \sim p^{\pi}(\theta' \mid s_t, a_t, c_t)}\big[T_{\theta'}(s' \mid s_t, a_t)\big]. \tag{10}$$

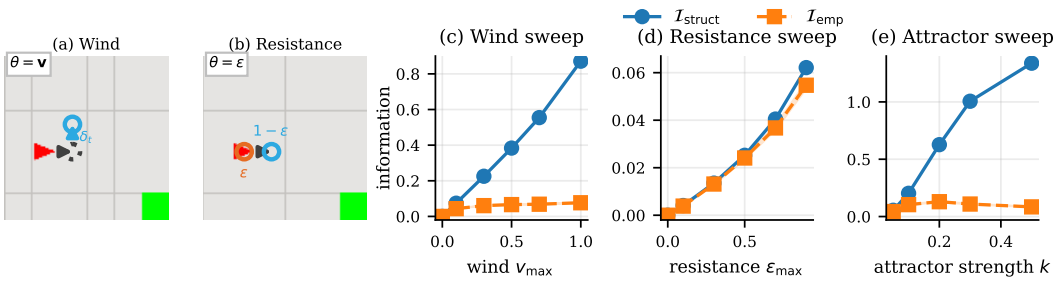


Figure 2: Controlled diagnostics. Wind applies a sampled drift after the native action (a), whereas Resistance either commits or voids the action (b). Across perturbations (c–e), $\mathcal{I}_{\text{struct}}$ separates from predictive headroom on Wind and Point-Mass-Attractor but remains close to it on Resistance.

*The representation is $\varepsilon$-predictively sufficient under $\pi$ if*

$$\mathbb{E}_{\theta\sim p(\theta|\alpha)}\,\mathbb{E}_{\tau\sim\pi,T_\theta}\left[\frac{1}{T}\sum_{t=0}^{T-1} D_{\mathrm{KL}}\big(T_\theta(\cdot\mid s_t,a_t)\,\big\|\,\bar{T}_c(\cdot\mid s_t,a_t,c_t)\big)\right]\le\varepsilon. \tag{11}$$

A constant $c_t$ induces $\bar{T}_{\text{emp}}$ and attains the sufficiency gap $\varepsilon = \mathcal{I}_{\text{emp}}$. Preserving the full posterior-predictive information in $\tau_{\le t}$ attains $\varepsilon = \mathcal{I}_{\text{hist}}$, while $\varepsilon = 0$ requires identifying the realized transition kernel only on the visited support. Contexts inducing the same kernel there are predictively equivalent and may share a representation. Predictive sufficiency retains trajectory information that affects transitions and may discard other context components. In the correlated-episode setting, $c_t$ may be computed from $\tau^{+}_{\le t}$. The definition then applies with the expectation taken over the stationary chain, and preserving the full cross-episode posterior attains the residual of Equation 9.

The decomposition identifies the part of an Oracle gap accessible to architectural intervention. Its control effect remains an empirical question: return also depends on horizon and reward sensitivity, and the definition is relative to the state–action distribution induced by $\pi$, so policy improvement requires coverage of the new distribution or a fresh estimate of $\varepsilon$.

## 4 Controlled Diagnostic

The framework distinguishes variation in a transition family from predictive headroom under the distribution generated by an agent. We test this distinction in tractable i.i.d. settings where the context-conditional kernels are known. MiniGrid (Chevalier-Boisvert et al., 2023) provides a finite state space in which both $\mathcal{I}_{\text{struct}}$ and $\mathcal{I}_{\text{emp}}$ can be evaluated from the exact kernels at the logged continuous contexts. The Point-Mass-Attractor environment provides a continuous-state counterpart with known noisy one-step kernels.

**Diagnostic design.** We implement two MiniGrid transition variants: MiniGrid-Wind samples a drift that changes the realized displacement after each action, while MiniGrid-Resistance samples an action-failure probability. Point-Mass-Attractor (Figure 3, Appendix B) samples one of the attractor locations, causing trajectories to curve toward the selected center. In all cases, the diagnostic varies only an unobserved episode-level transition parameter. We collect trajectories under uniform random actions and record context for estimation while withholding it from the predictive information set. Both continuous and discrete settings exploit the known kernels rather than estimating next-state distributions from samples. Environment definitions and protocols appear in Appendix B.

Alongside the fixed $\theta$-independent reference, we evaluate $\mathcal{I}^{\pi}_{\text{struct}}$ by setting $\mu(s,a) = p^{\pi}(s,a)$ while retaining prior context weights. Comparing $\mathcal{I}^{\pi}_{\text{struct}}$ with $\mathcal{I}_{\text{emp}}$ holds the query distribution fixed and changes only the context weights. All information quantities use natural logarithms.

**Structural variation and predictive headroom.** Panels c–e in Figure 2 depict two regimes. Wind develops substantial sensitivity while its estimated predictive headroom is 8.8% of $\mathcal{I}_{\text{struct}}$. At $v_{\max} = 1$, matching the query distribution reduces the reference by 29%, and the data-conditional context weights account for a further 88% reduction. Resistance changes little across the three

Table 1: Representative MiniGrid and Point-Mass-Attractor settings. The matched column sets $\mu(s, a) = p^{\pi}(s, a)$ while retaining prior context weights. CIs are 95% episode-level bootstrap. All estimates are conservatively biased upward, so finite-sample error can inflate headroom.

| State | Setting | $\mathcal{I}_{\mathrm{struct}}$ | $\mathcal{I}^{\pi}_{\mathrm{struct}}$ | $\mathcal{I}_{\mathrm{emp}}$ |
|---|---|---|---|---|
| Discrete | Wind ($v_{\max} = 1$) | 0.871 | 0.620 | $0.077 \pm 0.002$ |
| Discrete | Resistance ($\epsilon_{\max} = 0.9$) | 0.062 | 0.059 | $0.055 \pm 0.001$ |
| Continuous | Attractor ($k = 0.3$) | 1.006 | 1.006 | $0.109 \pm 0.002$ |

columns of Table 1; its low headroom follows from modest kernel separation rather than state-mediated identification. Point-Mass-Attractor reproduces the Wind regime in continuous state, where its kernel separation is invariant to the query distribution on the interior support, but an estimated 11% remains under observed visitation.

**Intervening on visitation.** To demonstrate the dependency, we reset selected state components during collection without resampling the episode context. Resetting MiniGrid position and direction every ten steps raises $\mathcal{I}_{\mathrm{emp}}$ by 430%, and resets every two steps raise it by 833%. In Point-Mass-Attractor, resetting position and velocity every ten steps raises the estimate by 324%; preserving velocity raises it by only 28%. Hence, the control identifies velocity as the principal carrier of context in the continuous system. Appendix B reports the full sweeps and complementary state-projection checks.

Together, these diagnostics show why contextual variation alone does not warrant contextual modeling. Wind and Point-Mass-Attractor have strongly separated kernels, but their induced states leave little one-step headroom, whereas MiniGrid-Resistance reaches similar headroom because its kernels differ little.

## 5 Experimental Validation

Prior work has proposed a range of mechanisms for latent context MDPs, each targeting a specific level in a hierarchy of settings and covering the levels below it. Based on the theoretical analysis, we predict that predictive sufficiency under a given visitation determines when a mechanism has additional information to exploit: headroom must remain, and conditioning on its inputs must lower the corresponding optimal risk. The resulting bounds separate the available information from how effectively the inductive bias uses it, which depends on the information structure rather than the class the mechanism was designed for.

Based on those insights, we formulate four falsifiable predictions:

- **(P1)** Implicit identification renders observed context redundant for one-step prediction irrespective of the MDP class. When visitation makes the context identifiable from the current state and action, the agent's own behavior has already performed the identification that context mechanisms are built for, making the predictive headroom vanish. Benefits of the mechanism therefore cannot be attributed to missing predictive information.
- **(P2)** Predictive sufficiency separates information failure from finite-model error. When headroom persists but the episode's history reaches the observed-context risk, any remaining predictive deficit for a model with access to that history is due to estimation, approximation, or optimization.
- **(P3)** The sufficiency optimum is relative to its conditioning set. For methods that do not observe context, only mechanisms that carry information across episode boundaries can improve on the within-episode optimum when consecutive contexts are dependent. Their advantage should be largest at episode start and vanish under independent draws.
- **(P4)** For a fixed visitation distribution, the reduction in expected log loss obtained by adding a context representation to the current state and action equals the difference between predictive headroom and that representation's minimum sufficiency gap. Under implicit identification, both the reduction and the tolerance vanish.

**Method classes.** Table 2 groups representative methods by their conditioning information, ordered by the predictive risk each class can target. The window class adds a dedicated context encoder over

Table 2: Method classes ordered by the predictive risk their conditioning set can target. Bold denotes the evaluated representatives. The memoryless class is instantiated by the measurement protocol itself.

| Class | Conditioning set (bound) | Sufficient when | Methods |
|---|---|---|---|
| Memoryless | $(s_t, a_t)$ $(\mathcal{R}_{\mathrm{M}})$ | $\theta$ decodable from $(s_t, a_t)$ | **Base probes** |
| Transition window | last $K$ transitions $(\geq \mathcal{R}_{\mathrm{H}})$ | recent evidence | **DALI-S**, CaDM, UP-OSI, GrBAL, PEARL, IIDA |
| Recurrent | within-episode summary $(\geq \mathcal{R}_{\mathrm{H}})$ | evidence within the episode | **DreamerV3**, ReBAL, VariBAD |
| Cross-episode chain | cross-episode summary $(\geq \mathcal{R}_{\mathrm{H}+})$ | episodes correlated | **Carried state**, **LILAC**, $\mathrm{RL}^2$, L2RL |
| Ground-truth context | realized $\theta$ $(\mathcal{R}_{\mathrm{O}})$ | always | **cRSSM-D**, cRSSM-C |

Table 3: Environment settings, ordered by the level at which the context becomes identifiable.

| Setting | Environment | Context process | Difficulty level | Tests |
|---|---|---|---|---|
| Calibration | Point mass | persistent two-state chain, i.i.d. and short-episode variants | within/across episodes | P3 (control) |
| Identifiable | CARL Walker | i.i.d. per episode | current state and action | P1, P4 |
| Hard | CARL Walker | i.i.d. per episode | episode history | P2, P4 |
| Drift | RWRL Walker | non-i.i.d. Markov chain across episodes | across episodes | P3 |

the last $K$ transitions, ordered or unordered. We take DALI-S (Röder et al., 2025) as the ordered representative, since none of our settings makes transition order informative and the framework therefore cannot separate the variants. For the cross-episode class, we evaluate a Markov chain-latent agent similar to LILAC (Xie et al., 2021) and a carried-state variant that preserves recurrent state across resets, the mechanism of $\mathrm{RL}^2$ (Duan et al., 2016) and L2RL (Wang et al., 2016), and the smallest member of this family, used as an ablation in the non-i.i.d. setting only. Finally, the observed-context class provides the privileged reference that separates recoverable headroom from task difficulty, represented by cRSSM-D (Prasanna et al., 2024). All evaluated methods share the same DreamerV3 backbone (Hafner et al., 2025). Implementation, training, and evaluation details are provided in Appendix C.

**Environments.** Common benchmarks often vary initial states rather than the latent dynamics, while meta-RL suites mix reward with dynamics variation, and perturbation studies confound predictive headroom with task difficulty. To address this, we build on CARL (Benjamins et al., 2023) and RWRL (Dulac-Arnold et al., 2021), which isolate dynamics variation. As with the methods, we select representative environments for the classes of contextual MDPs studied in prior work, ordered by the difficulty of identifying the context.

The two i.i.d. settings are based on DMC Walker from the CARL suite, repeating the contrast of Section 4 at benchmark scale: (i) an *identifiable* setting, where a latent horizontal wind, sampled uniformly per episode, exerts force through nonzero fluid density, probe headroom collapses while decodability stays high, and return remains competitive (P1, P4), and (ii) a *hard* setting, where a latent actuator-strength scale is sampled uniformly under fixed gravity and headroom persists above the permutation null (P2). The settings separate collapsed and persistent headroom under baseline visitation. Each method is then profiled under its own. In the non-i.i.d. drift setting, a single Walker parameter follows an RWRL perturbation schedule, performing a random walk across episodes at three rates on identical frozen realizations (P3).

Additionally, we include a simple linear-Gaussian point mass environment whose hidden context evolves across episodes as a Markov chain with high self-transition probability, so completed episodes carry predictive value about the current one. There, known transition kernels and disturbance filters provide analytical references for calibrating the Gaussian-score diagnostics (Appendix D). A short-episode variant limits within-episode evidence, and an independent-draw variant controls for the information supplied by cross-episode disturbance history (P3). One difficulty class remains uncovered: to our knowledge, no standard benchmark makes the context identifiable only through deliberately informative actions, so every passive class is predicted equivalent there. Table 3 summarizes the settings with their context process and difficulty level.

**Measurements.** To estimate headroom, history recovery, and the sufficiency gap, we compare held-out next-state log losses from MLP and recurrent probes with Gaussian likelihood heads, fitted independently of the agent on each method's evaluation episodes. Their inputs are (i) current state and action, (ii) adding the true context, (iii) adding a linear readout $c_t$ trained to predict $\theta$ from the

Table 4: Computed quantities over i.i.d. environments. Identifiable and Hard settings report return, probe headroom, and the residual context gain of the linear context readout. Entries are three-seed IQM with 95% bootstrap confidence intervals, method differences use paired seed resampling.

| | Identifiable | | | Hard | | |
|---|---|---|---|---|---|---|
| Method | Return ↑ | Headroom | Resid. gain | Return ↑ | Headroom | Resid. gain |
| DreamerV3 | $712 \pm 10$ | $-0.00 \pm 0.12$ | $-0.13 \pm 0.11$ | $946 \pm 6$ | $2.39 \pm 0.70$ | $0.38 \pm 0.08$ |
| DALI-S | $756 \pm 72$ | $0.03 \pm 0.13$ | $-0.15 \pm 0.13$ | $941 \pm 14$ | $2.10 \pm 1.76$ | $0.44 \pm 0.19$ |
| LILAC-style | $589 \pm 88$ | $-0.09 \pm 0.08$ | $0.00 \pm 0.09$ | $\mathbf{953 \pm 4}$ | $3.49 \pm 0.68$ | $0.73 \pm 0.29$ |
| cRSSM-D | $\mathbf{862 \pm 17}$ | $-0.14 \pm 0.16$ | $-0.06 \pm 0.09$ | $948 \pm 6$ | $3.64 \pm 1.71$ | $-0.22 \pm 0.80$ |

Table 5: Computed quantities over the non-i.i.d. environment: start transient and start prior loss at three drift rates. Entries are means ± sample stdev across seeds.

| | Start transient ↓ | | | Start prior loss ↓ | | |
|---|---|---|---|---|---|---|
| Method | Low | Mid | High | Low | Mid | High |
| DreamerV3 | $\mathbf{2.64 \pm 0.45}$ | $2.80 \pm 0.15$ | $2.79 \pm 0.86$ | $\mathbf{10.15 \pm 0.82}$ | $10.77 \pm 0.59$ | $11.24 \pm 0.83$ |
| Carried state | $4.06 \pm 0.36$ | $3.92 \pm 0.69$ | $3.93 \pm 0.51$ | $13.30 \pm 0.56$ | $12.51 \pm 0.75$ | $12.75 \pm 0.77$ |
| DALI-S | $3.12 \pm 0.19$ | $\mathbf{2.52 \pm 0.13}$ | $3.05 \pm 0.89$ | $10.68 \pm 0.05$ | $\mathbf{10.45 \pm 1.23}$ | $11.37 \pm 0.33$ |
| LILAC-style | $2.94 \pm 0.92$ | $3.19 \pm 0.86$ | $\mathbf{2.46 \pm 0.61}$ | $11.14 \pm 0.93$ | $11.54 \pm 0.90$ | $\mathbf{10.22 \pm 0.91}$ |
| cRSSM-D | $2.84 \pm 0.35$ | $3.05 \pm 0.41$ | $2.55 \pm 0.71$ | $10.48 \pm 0.31$ | $10.68 \pm 1.05$ | $10.38 \pm 0.33$ |

learned recurrent state, (iv) adding both, (v) adding a fixed-size history window, and (vi) adding both window and readout. A separate decodability probe predicts context from state and action; a permutation null shuffles contexts across episodes. Comparisons use fixed training budgets and separate episodes for fitting, validation, and scoring.

Because the probes are computed under each method's visitation, they are considered as within-method evidence, following the policy dependence of Definition 1. Cross-method comparisons rely on return over paired evaluation contexts and the class-level predictions. On the drift setting the primary metric is the start transient: early-window prior loss minus the cached-prefix reference loss.

**Evaluation design.** We use differences in held-out probe risks as empirical diagnostics of the predictive quantities in Section 3. Let $\widehat{\mathcal{R}}_X$ denote the next-state log loss of a fitted probe with inputs $X$. Probe headroom is defined as $\widehat{G}_\theta = \widehat{\mathcal{R}}_{\mathrm{M}} - \widehat{\mathcal{R}}_{\mathrm{O}}$. Under the assumptions of Section 3, its population counterpart satisfies $G_\theta = \mathcal{I}_{\mathrm{emp}} + e_{\mathrm{M}} - e_{\mathrm{O}}$, where $e_X$ includes approximation and fitting error above the corresponding Bayes risk. The finite-family reference $\widehat{\mathcal{R}}_{\mathrm{ref}}$ is the minimum risk over the window, the learned-state readout, their combination, and a recurrent probe. For readout $c_t$, we define $\widehat{\Delta}_{\mathrm{FM}}(c) = \widehat{\mathcal{R}}_c - \widehat{\mathcal{R}}_{\mathrm{ref}}$ on the same transitions. When probe headroom separates from the permutation null, we report normalized gain $(\widehat{\mathcal{R}}_{\mathrm{M}} - \widehat{\mathcal{R}}_c)/\widehat{G}_\theta$. Residual context gain is $\widehat{\mathcal{R}}_c - \widehat{\mathcal{R}}_{c,\theta}$. For Bayes-optimal predictors, the corresponding risk difference under the averaging law of Definition 1 equals the smallest admissible $\varepsilon$. We also report differences over the first $k{=}25$ transitions of each episode.

## 6 Results

**Settings with i.i.d. contexts.** Table 4 confirms the intended contrast across selected methods. In the hard setting, estimated headroom remains open and varies for every method, while returns stay within 2% of the baseline. There, successful control coexists with unresolved predictive context. Residual context gain remains positive for the inference-based methods and vanishes only for cRSSM-D, whose linear readout carries the ground-truth context (P4). In the identifiable setting, headroom collapses under each profiled visitation while the context stays decodable from the state with $R^2 \geq 0.93$, consistent with P1. cRSSM-D still leads the returns there, despite negligible predictive headroom.

**Settings with non-i.i.d. contexts.** As reported in Table 5, the context-blind baseline achieves the lowest mean startup losses at the low drift rate. However, at the mid rate, the transition-window agent already reduces the start transient by approximately $10\%$ and start prior loss by $3\%$ relative to the baseline. At the high rate, the chain-latent agent's advantage reaches approximately $12\%$ and $9\%$, respectively, with start prior loss also about $2\%$ below the observed-context variant. The same methods lead in absolute start prior loss at their respective rates. Carried state has the largest transient across rates, separating episode-level latent prediction from simply continuing the previous episode's state through the reset. Together, these profiles support regime-related benefits of the conditioning mechanisms under each method's own visitation (P3). Figure 7 in Appendix E depicts the considered metrics across the methods and drift rates.

## 7 Discussion

The results demonstrate that the value of context mechanisms is governed by the information structure of the setting in addition to their design, consistent with the theoretical analysis. First, the probes tracked exact targets on the calibration setting, where history recovery reached $0.89$ against the exact $0.99$ (P2), providing a reference for interpreting their estimates. Next, in the identifiable i.i.d. setting, the headroom collapsed under every profiled visitation while the context stayed decodable, so return differences reflect optimization rather than information. In the hard setting, estimated headroom stayed open across the evaluated methods despite comparable returns, showing that successful control does not have to resolve predictive context. Finally, the non-i.i.d. results connect predictive benefits to both the drift regime and mechanism design. The contrast between chain-latent inference and carried state demonstrates the distinction between predicting episode-level context and preserving recurrent state, consistent with Equation 9.

Our findings carry two implications: i) when Bayes predictive headroom is zero, any predictive gain under the same visitation reflects a reduction in finite-model error, and ii) for benchmark construction, the class of a setting alone does not determine its difficulty. The same task yielded headroom well above the null in one setting but not in another, which suggests reporting estimated headroom alongside an MDP's class and its benchmark definition.

**Limitations.** In the present analysis all estimated quantities are properties of a closed loop: each method's visitation shapes the information available to it, and that information shapes the method in turn. We deliberately do not compare under a shared reference policy, since no policy is neutral, random behavior collapses the visited region and competent behavior encodes the context it has already identified, so any fixed choice would characterize that policy rather than the learning problem. The policy dependence is instead the mechanism under study: implicit identification is produced by the agent's own behavior, and removing it would remove the phenomenon. It is also the conditioning that matters in practice, since a deployed mechanism operates under the visitation its training produces, rather than under a standardized one.

Extending comparison across visitations, through paired counterfactual resets where simulators permit them, and constructing settings whose context yields only to deliberately informative actions, are the future directions we consider most fruitful.

## 8 Conclusion

Motivated by real-world control, where the dynamics an agent faces vary and are rarely observed directly, a growing number of methods target MDPs with latent dynamics, but nothing tells a practitioner when any of them is needed: in practice the choice often follows the environment's class and the method's design intent. However, we show that both can be unreliable guides and propose a principled replacement. Our analysis establishes, and our experiments support, that the predictive opportunity available to a context mechanism is set by where the latent information resides under the trained agent's visitation, yielding a criterion assessed from the ordinary agent alone. Behind it lies a simple observation: an agent that acts in a world also performs inference about it, and whatever its behavior identifies along the way is already available to the optimal predictor. What remains for such additions is confined to specific regimes that measurement can identify, allowing it to inform design.

### AI USE STATEMENT

Generative AI tools were used to assist with software implementation, including environmental benchmarks and analysis scripts, and to edit the manuscript for language, wording, and phrasing to improve readability. AI was also used to translate the proposition and its proof into a formal notation for verification by a theorem prover. The authors reviewed all code, analyses, and text and take full responsibility for the final content of this work.

### REPRODUCIBILITY STATEMENT

Appendix B details the MiniGrid and Point-Mass-Attractor diagnostics, Appendix C reports the benchmark definitions, evaluation distributions, implementations, and hyperparameters, while Appendix D reports the calibration protocol. The reference codebase[1] includes all source code, along with configuration scripts and protocols.

## REFERENCES

Carolin Benjamins, Theresa Eimer, Frederik Schubert, Aditya Mohan, Sebastian Döhler, André Biedenkapp, Bodo Rosenhahn, Frank Hutter, and Marius Lindauer. Contextualize me—the case for context in reinforcement learning. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=y42xVBQusn.

Ronen I. Brafman and Moshe Tennenholtz. R-MAX—a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3:213–231, 2002. URL https://www.jmlr.org/papers/v3/brafman02a.html.

Maxime Chevalier-Boisvert, Bolun Dai, Mark Towers, Rodrigo Perez-Vicente, Lucas Willems, Salem Lahlou, Suman Pal, Pablo Samuel Castro, and J. Terry. MiniGrid & MiniWorld: Modular & customizable reinforcement learning environments for goal-oriented tasks. In *Advances in Neural Information Processing Systems*, volume 36, pp. 73383–73394, 2023. doi: 10.52202/075280-3209.

Samuel PM Choi, Dit-Yan Yeung, and Nevin L Zhang. Hidden-mode Markov decision processes for nonstationary sequential decision making. In *Sequence learning: paradigms, algorithms, and applications*, pp. 264–287. Springer, 2001.

Karl Cobbe, Chris Hesse, Jacob Hilton, and John Schulman. Leveraging procedural generation to benchmark reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2048–2056. PMLR, 2020. URL https://proceedings.mlr.press/v119/cobbe20a.html.

Finale Doshi-Velez and George Konidaris. Hidden parameter Markov decision processes: A semiparametric regression approach for discovering latent task parametrizations. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pp. 1432–1440. IJCAI/AAAI Press, 2016. URL https://www.ijcai.org/Proceedings/16/Papers/206.pdf.

Yan Duan, John Schulman, Xi Chen, Peter L Bartlett, Ilya Sutskever, and Pieter Abbeel. $RL^2$: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*, 2016.

Michael O'Gordon Duff. *Optimal Learning: Computational Procedures for Bayes-Adaptive Markov Decision Processes*. PhD thesis, University of Massachusetts Amherst, Amherst, Massachusetts, 2002.

Gabriel Dulac-Arnold, Nir Levine, Daniel J Mankowitz, Jerry Li, Cosmin Paduraru, Sven Gowal, and Todd Hester. Challenges of real-world reinforcement learning: definitions, benchmarks and analysis. *Machine Learning*, 110(9):2419–2468, 2021.

[1] https://anonymous.4open.science/r/considering_context/

Ben Evans, Abitha Thankaraj, and Lerrel Pinto. Context is everything: Implicit identification for dynamics adaptation. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 2642–2648. IEEE, 2022.

Erin Grant, Chelsea Finn, Sergey Levine, Trevor Darrell, and Thomas L. Griffiths. Recasting gradient-based meta-learning as hierarchical Bayes. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=BJ_UL-k0b.

Danijar Hafner. Benchmarking the spectrum of agent capabilities. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=1W0z96MFEoH.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025. doi: 10.1038/s41586-025-08744-2.

Assaf Hallak, Dotan Di Castro, and Shie Mannor. Contextual Markov decision processes, 2015. URL https://arxiv.org/abs/1502.02259.

Mikael Henaff, Minqi Jiang, and Roberta Raileanu. A study of global and episodic bonuses for exploration in contextual MDPs. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 12972–12999. PMLR, 2023. URL https://proceedings.mlr.press/v202/henaff23a.html.

Jennifer A. Hoeting, David Madigan, Adrian E. Raftery, and Chris T. Volinsky. Bayesian model averaging: A tutorial. *Statistical Science*, 14(4):382–417, 1999. doi: 10.1214/ss/1009212519.

Aapo Hyvärinen, Hiroaki Sasaki, and Richard E. Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 859–868. PMLR, 2019. URL https://proceedings.mlr.press/v89/hyvarinen19a.html.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49(2–3):209–232, 2002. doi: 10.1023/A:1017984413808.

Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2207–2217. PMLR, 2020. URL https://proceedings.mlr.press/v108/khemakhem20a.html.

Robert Kirk, Amy Zhang, Edward Grefenstette, and Tim Rocktäschel. A survey of zero-shot generalisation in deep reinforcement learning. *Journal of Artificial Intelligence Research*, 76:201–264, 2023.

Jeongyeol Kwon, Yonathan Efroni, Constantine Caramanis, and Shie Mannor. RL for latent MDPs: Regret guarantees and a lower bound. In *Advances in Neural Information Processing Systems*, volume 34, pp. 24523–24534, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/c9037f06ebb3d1135d7228e3e381aa48-Abstract.html.

Kimin Lee, Younggyo Seo, Seunghyun Lee, Honglak Lee, and Jinwoo Shin. Context-aware dynamics model for generalization in model-based reinforcement learning. In *International Conference on Machine Learning*, pp. 5757–5766. PMLR, 2020.

Anusha Nagabandi, Ignasi Clavera, Simin Liu, Ronald S Fearing, Pieter Abbeel, Sergey Levine, and Chelsea Finn. Learning to adapt in dynamic, real-world environments through meta-reinforcement learning. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HyztsoC5Y7.

Sai Prasanna, Karim Farid, Raghu Rajan, and André Biedenkapp. Dreaming of many worlds: Learning contextual world models aids zero-shot generalization. *Reinforcement Learning Journal*, 3:1317–1350, 2024. URL https://rlj.cs.umass.edu/2024/papers/Paper167.html.

Kate Rakelly, Aurick Zhou, Chelsea Finn, Sergey Levine, and Deirdre Quillen. Efficient off-policy meta-reinforcement learning via probabilistic context variables. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5331–5340. PMLR, 2019. URL https://proceedings.mlr.press/v97/rakelly19a.html.

Frank Röder, Jan Benad, Manfred Eppe, and Pradeep Banerjee. Dynamics-aligned latent imagination in contextual world models for zero-shot generalization. In *Advances in Neural Information Processing Systems*, volume 38, pp. 80885–80920, 2025. doi: 10.52202/085713-2438.

Jane X Wang, Zeb Kurth-Nelson, Dhruva Tirumala, Hubert Soyer, Joel Z Leibo, Remi Munos, Charles Blundell, Dharshan Kumaran, and Matt Botvinick. Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*, 2016.

Annie Xie, James Harrison, and Chelsea Finn. Deep reinforcement learning amidst continual structured non-stationarity. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11393–11403. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/xie21c.html.

Wenhao Yu, Jie Tan, C. Karen Liu, and Greg Turk. Preparing for the unknown: Learning a universal policy with online system identification. In Nancy M. Amato, Siddhartha S. Srinivasa, Nora Ayanian, and Scott Kuindersma (eds.), *Robotics: Science and Systems XIII, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA, July 12-16, 2017*, 2017. doi: 10.15607/RSS.2017.XIII.048. URL http://www.roboticsproceedings.org/rss13/p48.html.

Luisa M. Zintgraf, Kyriacos Shiarlis, Maximilian Igl, Sebastian Schulze, Yarin Gal, Katja Hofmann, and Shimon Whiteson. VariBAD: A very good method for Bayes-adaptive deep RL via meta-learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Hkl9JlBYvr.

# A Proofs and Technical Remarks

**Proposition** (Predictive headroom decomposition)**.** *Under the conditional-Markov assumption and a common transition distribution,*

$$\mathcal{I}_{\mathrm{emp}} = \mathcal{R}_{\mathrm{M}} - \mathcal{R}_{\mathrm{O}} = \underbrace{\mathcal{R}_{\mathrm{M}} - \mathcal{R}_{\mathrm{H}}}_{\textit{headroom closed by history}} + \underbrace{\mathcal{I}_{\mathrm{hist}}}_{\textit{residual Oracle headroom}}.$$

*Consequently,* $0 \le \mathcal{I}_{\mathrm{hist}} \le \mathcal{I}_{\mathrm{emp}}$.

*Proof.* $\mathcal{R}_{\mathrm{M}} = H(s_{t+1} \mid s_t, a_t)$, $\mathcal{R}_{\mathrm{H}} = H(s_{t+1} \mid \tau_{<t}, a_t)$, and the conditional Markov property gives $\mathcal{R}_{\mathrm{O}} = H(s_{t+1} \mid s_t, a_t, \theta) = H(s_{t+1} \mid \tau_{\le t}, a_t, \theta)$. Subtracting these risks yields Equation 7. □

**Proposition 2** (Value sensitivity to transition error)**.** *Let $T_\theta$ and $\widehat{T}$ be transition kernels over a shared state–action space, with common bounded reward $r \in [0, R_{\max}]$ and discount $\gamma \in [0, 1)$. For any stationary Markov policy $\pi$ and any state $s$,*

$$\left|V^\pi_{T_\theta}(s) - V^\pi_{\widehat{T}}(s)\right| \;\le\; \frac{\gamma R_{\max}}{(1-\gamma)^2} \sup_{s,a} \sqrt{\tfrac{1}{2}\, D_{\mathrm{KL}}\big(T_\theta(\cdot \mid s,a) \,\|\, \widehat{T}(\cdot \mid s,a)\big)}.$$

*Proof.* Write $V_\theta = V^\pi_{T_\theta}$, $\widehat{V} = V^\pi_{\widehat{T}}$, and $T^\pi(\cdot \mid s) = \sum_a \pi(a \mid s)\, T(\cdot \mid s, a)$ for the policy-induced state kernel. Subtracting the two Bellman equations, whose shared reward cancels, and inserting $\pm\, \gamma\, T^\pi_\theta \widehat{V}$:

$$V_\theta(s) - \widehat{V}(s) = \gamma\, T^\pi_\theta(\cdot \mid s)\big[V_\theta - \widehat{V}\big] + \gamma\, \big[T^\pi_\theta - \widehat{T}^\pi\big](\cdot \mid s)\, \widehat{V}.$$

We bound the two terms. The first is a sup-norm non-expansion:

$$\left|T^\pi_\theta(\cdot \mid s)\big[V_\theta - \widehat{V}\big]\right| \;\le\; \|V_\theta - \widehat{V}\|_\infty \qquad (T^\pi_\theta(\cdot \mid s) \text{ is a probability measure}).$$

For the second, $T^\pi_\theta - \widehat{T}^\pi$ has zero total mass, so

$$\begin{aligned}
\left|[T^\pi_\theta - \widehat{T}^\pi]\, \widehat{V}\right| &= \left|[T^\pi_\theta - \widehat{T}^\pi]\, (\widehat{V} - c)\right| && \text{(zero mass: subtract any constant } c) \\
&\le \|\widehat{V} - c\|_\infty \left\|T^\pi_\theta - \widehat{T}^\pi\right\|_1 && \text{(Hölder's ineq.)} \\
&\le \frac{R_{\max}}{2(1-\gamma)} \cdot 2\, D_{\mathrm{TV}}(T^\pi_\theta, \widehat{T}^\pi) && (c \text{ the midrange of } \widehat{V} \in [0, \tfrac{R_{\max}}{1-\gamma}]) \\
&\le \frac{R_{\max}}{1-\gamma} \sup_{s,a} D_{\mathrm{TV}}(T_\theta, \widehat{T}) && \text{(triangle ineq. for TV under the shared policy mixture).}
\end{aligned}$$

Taking absolute values in the identity, applying the triangle inequality with the two bounds, and taking a supremum over $s$:

$$\begin{aligned}
\|V_\theta - \widehat{V}\|_\infty &\le \gamma\, \|V_\theta - \widehat{V}\|_\infty + \frac{\gamma R_{\max}}{1-\gamma} \sup_{s,a} D_{\mathrm{TV}}(T_\theta, \widehat{T}) \\
&\le \frac{\gamma R_{\max}}{(1-\gamma)^2} \sup_{s,a} D_{\mathrm{TV}}(T_\theta, \widehat{T}) && \text{(rearrange; } \gamma < 1,\ \|V_\theta - \widehat{V}\|_\infty < \infty) \\
&\le \frac{\gamma R_{\max}}{(1-\gamma)^2} \sup_{s,a} \sqrt{\tfrac{1}{2}\, D_{\mathrm{KL}}\big(T_\theta \,\|\, \widehat{T}\big)} && \text{(Pinsker's ineq.)}
\end{aligned}$$

Since $|V_\theta(s) - \widehat{V}(s)| \le \|V_\theta - \widehat{V}\|_\infty$ for every $s$, the claim follows. □

**Jensen-gap identity.** At each fixed $(s, a)$, the integrand of Equation 3 satisfies

$$\mathbb{E}_\theta D_{\mathrm{KL}}(T_\theta \| \bar{T}_{\mathrm{prior}}) = H(\bar{T}_{\mathrm{prior}}) - \mathbb{E}_\theta H(T_\theta), \tag{12}$$

with the $(s, a)$ arguments suppressed for clarity. The entropy and expected-log-loss identities assume finite terms under a common dominating measure.

**Remark on recovering the context.** Tasks requiring $\theta$ itself need the stronger condition $p(\theta \mid \tau_{\le t}) = p(\theta \mid c_t)$ and trajectory-level information such as $I(\theta; \tau)$. Establishing this condition from observational trajectories requires assumptions beyond episode grouping (Hyvärinen et al., 2019; Khemakhem et al., 2020).

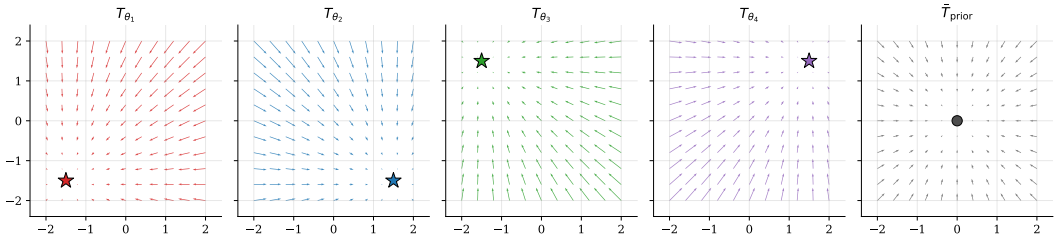


Figure 3: Point-Mass-Attractor transition fields. Each episode uses one hidden attractor $\theta_i$, inducing a distinct field $T_{\theta_i}$; prior averaging produces a center-seeking field $\bar{T}_{\text{prior}}$ that governs no episode.




Figure 4: Point-Mass-Attractor trajectories under uniform random actions. The hidden attractor is fixed within an episode, causing visited states to become informative about context.

## B Controlled Diagnostic Details

The controlled diagnostic uses two state spaces to test the same distinction. MiniGrid makes every $(s, a)$ pair enumerable, so $\mathcal{I}_{\text{struct}}$ can be evaluated directly and visitation can be manipulated by setting. Point-Mass-Attractor retains known one-step kernels in continuous state, where the empirical quantity requires estimating only the occupancy-induced context posterior. MiniGrid-Resistance supplies a contrast with low $\mathcal{I}_{\text{struct}}$ within the discrete setting.

**Shared evaluation protocol.** Context is sampled once per episode and omitted from the predictive information set, but retained in the logged data used to estimate conditional information. Because every context-conditional kernel is known in closed form, no next-state distribution is estimated from samples: empirical values score each logged transition's exact kernel against a cross-fitted reference mixture, so every transition contributes and no count-based filtering occurs. All pooled transitions receive equal weight, while confidence intervals resample episodes. The standardized $\mathcal{I}_{\text{struct}}$ uses a fixed $\theta$-independent query distribution.

**Discrete state: MiniGrid.** Both MiniGrid families use the fully observed Markov state $(x, y, d)$ on the `Empty-8x8` layout and leave the goal and reward unchanged. In MiniGrid-Wind, the episode context is $\mathbf{v} = (v_x, v_y)$ with independent coordinates sampled from $\mathcal{U}[-v_{\max}, v_{\max}]$. After the native action, each axis receives drift $\text{sign}(v_j)$ with probability $|v_j|$ and zero drift otherwise. Invalid displacement into a wall is blocked. MiniGrid-Resistance samples $\epsilon \sim \mathcal{U}[0, \epsilon_{\max}]$; an action takes effect with probability $1 - \epsilon$ and otherwise leaves the state unchanged. Zero perturbation recovers the base environment in both families.

For the standardized reference, $\mu$ is uniform over interior position, direction, and the seven MiniGrid actions. The matched-query reference instead weights those same local calculations by $(s, a)$ frequencies in the collected trajectories. Figure 2 reports the full $v_{\max}$ and $\epsilon_{\max}$ sweeps.

Each setting uses 5,000 uniform-random episodes of at most 256 steps. The $\mathcal{I}_{\text{struct}}$ calculation enumerates $\approx$ 1,000 interior $(s, a)$ queries and integrates the continuous context prior by quadrature. The empirical estimator is a known-kernel occupancy plug-in: each logged transition contributes the exact Kullback–Leibler divergence between its context-conditional kernel and a kernel mixture over the transitions observed at the same $(s, a)$ cell. The reported value cross-fits the mixture leave-one-episode-out with one pseudo-count of uniform smoothing on the cell's outcome support, which makes it a conservative upper estimate of the true headroom in expectation; the within-sample

mixture value is always within 2% of it. Percentile CIs come from 200 episode-level bootstrap resamples with full re-estimation of all mixtures. The sweeps use $v_{\max} \in \{0, 0.1, 0.3, 0.5, 0.7, 1.0\}$ and $\epsilon_{\max} \in \{0, 0.1, 0.3, 0.5, 0.7, 0.9\}$.

**Continuous state: Point-Mass-Attractor.** Point-Mass-Attractor is a unit-mass two-dimensional system integrated at $\Delta t = 0.05$. The hidden attractor is selected uniformly from the four corners $(\pm 1.5, \pm 1.5)$ and held fixed for the episode. At each step, the agent applies $\mathbf{a}_t \in [-1, 1]^2$ and the attractor contributes force $-k(\mathbf{p}_t - \theta)$. Independent Euler–Maruyama increments add variance $\sigma^2 \Delta t$ to each velocity coordinate and $\sigma^2 \Delta t^3$ to each position coordinate. We use $k = 0.3$ and $\sigma = 0.1$ for the representative setting.

Episodes begin with $\mathbf{p}_0 \sim \mathcal{U}[-0.5, 0.5]^2$ and $\mathbf{v}_0 = \mathbf{0}$, and run for 50 steps (Figure 4). Positions are confined to $[-2.5, 2.5]^2$ by elastic reflection with velocity damping 0.5. Away from the boundary, the conditional kernel is diagonal Gaussian, with four means separated by a query-independent context displacement. Thus, the uniform and matched-query values of $\mathcal{I}_{\text{struct}}$ in the interior coincide exactly. Per coordinate, the context sign defines a binary-input Gaussian channel, so $\mathcal{I}_{\text{struct}}$ can be computed by one-dimensional Gauss–Hermite quadrature. A standard bound shows that boundary reflection has a negligible effect at the occupancies used here.

Each $k$ setting uses $5{,}000$ episodes, with $k \in \{0.05, 0.1, 0.2, 0.3, 0.5\}$. The empirical estimate scores every logged transition's exact context-conditional Gaussian kernel, including the exact single-reflection boundary densities, against the four-kernel mixture weighted by a cross-fitted $k$-NN context posterior (episode-level folds; the number of neighbors is selected by cross-validated log loss). For each candidate neighbor count the fitted mixture is a valid state-measurable predictor, so the estimate is biased upward in expectation up to a small model-selection effect. Percentile CIs come from $1{,}000$ episode-level bootstrap resamples of per-episode contributions. As an internal certificate, transitions immediately after a zero-velocity reset in the intervention datasets below have an exactly known uniform context posterior; the fitted estimator reproduces the exact stratum contribution within about 1.5% (for example, 0.4845 versus 0.4834 at reset interval two).

**Locating the visitation effect.** Two checks vary the information carried by the current state. First, state projection removes coordinates from the conditioning variable while retaining the full next state. MiniGrid uses $(x, y, d)$, direction alone, or no state; Point-Mass-Attractor uses $(\mathbf{p}, \mathbf{v})$, position, a $5 \times 5$ position bin, or no state. Coarsening therefore changes the estimand. The MiniGrid projections are evaluated at a declared discretization of ten bins per context coordinate, with both the projected conditional and marginal formed as occupancy-weighted exact-kernel mixtures, cross-fitted leave-one-episode-out. The Point-Mass-Attractor projections use a cross-fitted classifier contrast: with four known context values, the estimand equals the difference between the context entropy given the projected state and given the projected state, action, and next state, each estimated by a cross-fitted $k$-NN class posterior (for the no-state projection the first term is the exact design prior $\ln 4$, making the estimate conservative). Every projection continues to condition on action.

Second, resets intervene on visitation while preserving the episode context. MiniGrid resets position and direction before the current state is recorded, at intervals of ten, five, or two steps. Point-Mass-Attractor resets position at intervals of 25, ten, five, or two steps. Its control either resets velocity to zero or preserves it across the position reset. All subsequent transitions are recorded from the post-reset state.

The projection results, depicted in Figure 5, locate context information already present in the observed state. Removing position from MiniGrid raises estimated $\mathcal{I}_{\text{emp}}$ by about 2,300%. In Point-Mass-Attractor, progressively removing velocity and position raises it by about 980%. The resets partially restore headroom without imposing the standardized query distribution. Relative to baseline, estimated MiniGrid headroom increases by 430%, 658%, and 833% as the reset interval decreases. In Point-Mass-Attractor, resetting both position and velocity raises headroom by as much as 698%, increasing monotonically as the reset interval shrinks, whereas preserving velocity keeps it within 35% of baseline.

## C Experimental Details

**Benchmark environments.** The i.i.d. settings use CARL DMC Walker Walk (Benjamins et al., 2023). The identifiable setting samples a latent horizontal wind uniformly per episode, with nonzero

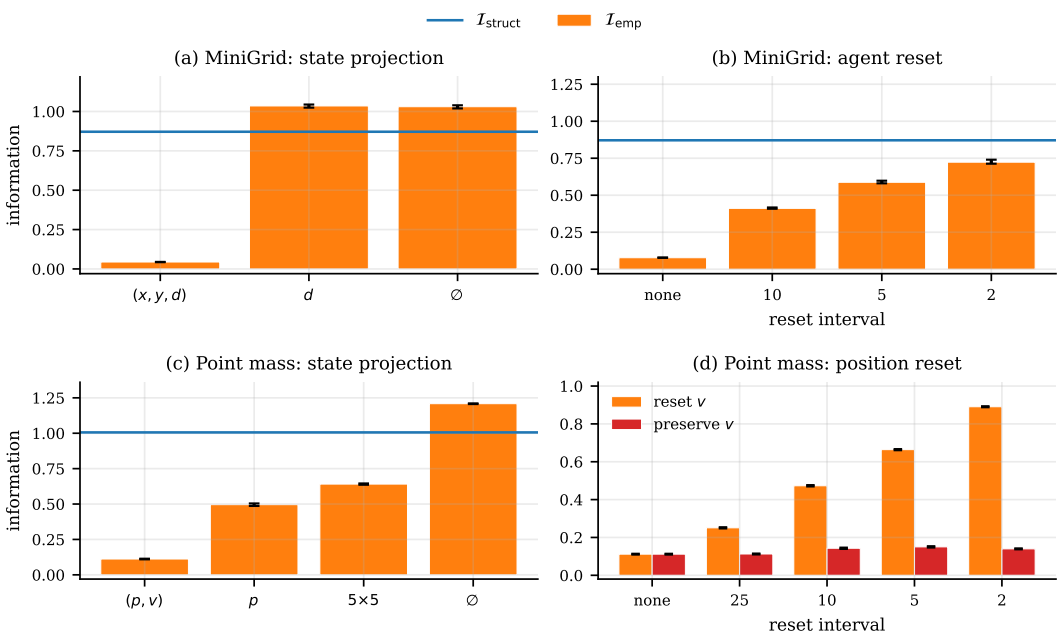


Figure 5: Parallel checks of state-mediated identification. State projection removes current-state information in MiniGrid (a) and Point-Mass-Attractor (c). State resets weaken context-dependent visitation in MiniGrid (b). In Point-Mass-Attractor (d), headroom rises when velocity is reset and remains near baseline when velocity is preserved. Error bars denote 95% bootstrap CIs. Where shown, blue lines are full-state structural references, not bounds after the information set or query distribution changes.

fluid density so the wind exerts force on the body. The hard setting samples a latent actuator-strength scale uniformly around nominal, with gravity fixed at nominal. An 84-context benchmark panel is held out for the final evaluation. The drift setting follows the registered RWRL protocol (Dulac-Arnold et al., 2021): one Walker parameter performs a random walk across episodes at three step sizes on frozen realizations shared by every method, with the step sizes chosen so that consecutive-episode autocorrelation stays above 0.5 at the highest rate.

**Agents and conditioning interfaces.** All agents use the common DreamerV3 hyperparameters (Hafner et al., 2025) in Table 6. Every method shares this backbone and differs only in its context configuration, which selects one information path while leaving the training stack unchanged: the context-blind baseline receives no context, DALI-S infers one from a recent window, cRSSM-D receives the true context through deep conditioning, the carried-state variant preserves the recurrent state across episode boundaries, and the chain-latent agent maintains a cross-episode latent inspired by LILAC (Xie et al., 2021). The true-context method receives the context as a separate normalized observation with explicit bounds, and the ordinary observation encoder always excludes it. DreamerV3 already carries a recurrent within-episode state, so its comparison with DALI isolates the value of a dedicated recent-history pathway while holding that recurrent backbone fixed.

DALI-S (Röder et al., 2025) uses a causal 64-step state–action encoder with one head and width 256, an 8-dimensional context, and a two-layer width-128 forward model, with unit loss scale and context learning rate $10^{-4}$. The inferred context is detached before it enters the world model, its heads, the actor, and the critic, and the cross-modal loss is omitted. cRSSM-D (Prasanna et al., 2024) receives true normalized context through the transition and prediction/control heads and adds no learned context encoder. The chain-latent baseline uses an 8-dimensional context, a two-layer width-128 episode encoder, and a width-64 GRU over the preceding 128 episode latents, trained at $10^{-4}$ with chain and consistency loss scales 1 and 0.1. All comparators are implemented within the same DreamerV3 codebase using the conditioning interfaces described above.

**Training and evaluation.** Return entries are IQM over paired evaluation contexts after within-seed aggregation, with CARL return normalized by the maximum episodic return of 1000. Confidence

Table 6: Agent and probe hyperparameters.

| Agent setting | Value |
|---|---|
| Backbone | DreamerV3, block-GRU |
| Parameters (Walker baseline) | 9.7M (4.3M transition) |
| Deterministic state | 2048 |
| Stochastic state | $32 \times 16$ categorical |
| RSSM hidden width | 256 |
| Encoder / decoder | 3 layers, width 256 |
| Actor / critic | 3 layers, width 256 |
| Activation / normalization | SiLU / layer norm |
| Batch / sequence | $16/64$ |
| Replay capacity / ratio | $5 \times 10^5/512$ |
| Optimizer / learning rate | LaProp / $4 \times 10^{-5}$ |
| Betas / warmup | $(0.9, 0.999)/1000$ |
| Weight decay | none |
| Gradient clipping | adaptive, 0.3 |
| KL free threshold | 1.0 |
| Imagination horizon | 15 |
| Discount / $\lambda$ | $1 - 1/333$ / 0.95 |
| Actor entropy scale | $3 \times 10^{-4}$ |
| Action repeat | 1 |

| Probe setting | Value |
|---|---|
| Feedforward architecture | 2-layer MLP, width 256 |
| Recurrent architecture | 1-layer GRU, width 128 |
| Activation | SiLU (feedforward) |
| Predictive output | diagonal Gaussian |
| Optimizer / learning rate | AdamW / $10^{-3}$ |
| Training updates | 4,000 |
| Batch | 1,024 transitions / 32 episodes |
| Feedforward weight decay | $\{10^{-4}, 3 \times 10^{-2}, 1\}$ |
| Recurrent weight decay | $\{10^{-4}, 3 \times 10^{-2}\}$ |
| Cross-fitting | 5 episode-level folds |
| History span | 16 steps / full episode |
| Episode-start window | first 25 steps |
| Selection / calibration | validation split |

intervals use 10,000 paired, seed-level bootstrap samples. Contexts are not treated as independent units. On the drift setting the primary endpoint is the start transient over the first 25 steps, computed post hoc from final checkpoints. Drift entries are means of seed means with seed ranges. Return and per-seed collapse counts are secondary.

Runs use one 24 GB NVIDIA L4. A $10^6$-step run takes about 9 hours, with 6–7 GB peak allocated memory. Checkpoint evaluation freezes the model and policy and performs no learning updates, while only explicitly carried recurrent state persists across episode resets.

**Probe implementation.** Probe architectures and training settings are reported beside the agent settings in Table 6. Scoring is cross-fitted at episode level, so every transition is scored by a probe that never trained on its episode. The permutation null re-pairs episodes with contexts drawn from other episodes, and the decodability readout uses the feedforward architecture with a squared-error objective to predict the context from the current state and action; its cross-fitted $R^2$ is reported. Learned-state inputs are extracted before the target transition is observed. The shared Gaussian likelihood defines the predictive risk used in continuous benchmark settings. Episode-start estimates restrict every quantity to the beginning of each episode. During training, a frozen panel of open-loop trajectories, one per evaluation context with autocorrelated exploration actions of length 64, tracks prior prediction loss at every evaluation point.

## D Analytical Calibration Details

The considered benchmark settings do not offer exact information targets, while the diagnostic environments have known kernels but no analytic probe targets, since the proposed probes have closed-form equivalents only under linear-Gaussian dynamics. To this end, we calibrate the measurements using the simplest such system.

**Environment and analytical quantities.** Figure 6 shows the calibration task: a two-dimensional stabilization point mass whose state $(x, v, d)$ is fully observed and whose hidden regime $z \in \{-1, +1\}$ follows a slow-switching Markov chain across episodes with stay probability 0.9. The disturbance follows $d_{t+1} = \rho R(z\omega)d_t + \sigma\sqrt{1-\rho^2}\epsilon_t$, with rotation $R$, $\rho = 0.9$, $\omega = \pi/4$, $\sigma = 0.3$, and independent $\epsilon_t \sim \mathcal{N}(0, I_2)$. Velocity and position update as $v_{t+1} = 0.9v_t + 0.5a_t + d_{t+1}$ and $x_{t+1} = x_t + v_{t+1}$; the reward is $-(\|x\|^2 + 0.01\|a\|^2)$. At reset, $x_0 = v_0 = 0$ and $d_0 \sim \mathcal{N}(0, \sigma^2 I_2)$ independently of the regime. The first regime is drawn uniformly and episodes last 200 steps. Conditional disturbance densities and disturbance-history posteriors are available analytically: their averaged scores are evaluated by numerical integration or sample averages.

**Evaluation protocol.** We evaluate DreamerV3, DALI-S, carried state, the LILAC-inspired chain-latent agent, and cRSSM-D using the common agent and probe settings in Table 6. Each method uses seeds 0–2, trains for $10^5$ environment steps, and is evaluated on 10 episodes. The primary

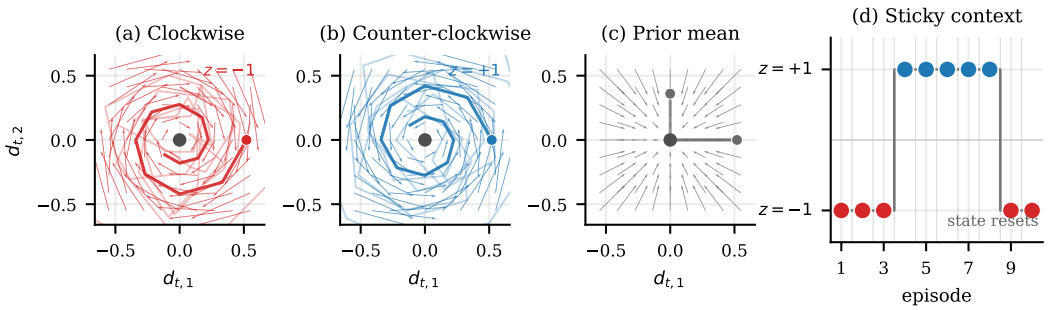


Figure 6: Analytical calibration environment. The hidden regime determines the rotation of the autoregressive disturbance. The physical state resets between episodes while the regime follows a Markov chain. The prior mixture averages the two conditional disturbance fields but is not a realized regime.

setting uses the sticky context chain. An independent-draw control sets the stay probability to 0.5, and a short-episode variant truncates episodes at 12 steps to leave a nonzero residual.

**Score-based targets and information ceilings.** The probe targets are ideal *diagonal-Gaussian score gaps*, rather than mutual information. Given $(S_t, A_t)$, the six observed next-state coordinates share one two-dimensional innovation $d_{t+1}$ across the position, velocity, and disturbance blocks, so the conditional covariance has rank two and no six-dimensional density exists; the coordinatewise diagonal-Gaussian score is the well-defined reference, and the three identical blocks contribute the factor of three in its gap. With uniform regime weights the population gap is

$$G_0 = \tfrac{3}{2}\,\mathbb{E}\sum_{j=1}^{2}\log\Big[1+\tfrac{(\mu_{+}-\mu_{-})_j^2}{4\tau^2}\Big] = 2.518 \quad (\tau^2=\sigma^2(1-\rho^2)),$$

evaluated by quadrature, whereas the sampled evaluation in Table 7 gives $2.51 \pm 0.05$. Because $w(1-w) \le \frac{1}{4}$ for any regime probability $w$, $G_0$ upper-bounds the state-conditional score gap. The corresponding *information* quantities are much smaller: the binary regime carries at most $\ln 2 \approx 0.693$ nats, and the stationary disturbance-mixture information is $I(Z; D_{t+1} \mid D_t) = 0.507$, which upper-bounds the memoryless context information $I(Z; S_{t+1} \mid S_t, A_t)$ here because $(S_t, A_t)$ determines $D_t$ and $D_{t+1}$ is independent of $(S_t, A_t)$ given $(Z, D_t)$.

**Calibration results.** The estimated score gap in Table 7 tracks the exact target at a stable ratio in the primary and independent-draw settings, which characterizes its attenuation on the calibration environment. The unadjusted reference upper-bounds the population context gain for moment-matched diagonal-Gaussian predictors conditioned on state and action, by the argument above. The shuffled null stays near zero, while history recovery reaches $0.89 \pm 0.03$ of the 0.99 reference computed from the exact disturbance-history filter, giving a recovered fraction for the nested disturbance-only information sets rather than a certified bound on the full on-policy history optimum. The estimated sufficiency gap is near zero for the observed-context agent, concentrates at episode start for every inference method, and is largest for carried state, whose recurrent state continues the previous episode through the reset. The chain classes have no advantage at episode start, consistent with the small cross-episode increment in the closed form.

The short-episode variant yields an exact per-transition residual of 0.24; since the scoring window covers all twelve steps of this variant, start-window and full-episode averages coincide there, and the probes reproduce the residual as a sufficiency gap that persists for the context-blind model and nearly vanishes for the observed-context one. Under the independent-draw control the exact cross-episode increment is zero and the chain classes remain indistinguishable from within-episode ones, consistent with P3. The characterized ratio reproduces at $0.51 \pm 0.04$.

These controls support using the probes as comparative diagnostics at the operating episode length; the attenuation ratio is characterized on this environment rather than assumed to transfer elsewhere.

## E Additional Results

Table 7: Probe calibration on the analytical point-mass environment. The context score gap is a diagonal-Gaussian log-score difference: the exact column is a sampled evaluation of the 2.518-nat population gap, and the estimate column reports the estimate-to-target ratio. History recovery reports a recovered fraction and the shuffled null a probe-risk difference. Entries are IQM $\pm$ the larger 95% bootstrap CI half-width.

*(a) Probe calibration*

| Quantity | Exact | Est. |
|---|---|---|
| Context score gap | $2.51 \pm 0.05$ | $0.48 \pm 0.03$ |
| History recovery | $0.99$ | $0.89 \pm 0.03$ |
| Shuffled null | $0$ | $-0.04 \pm 0.02$ |

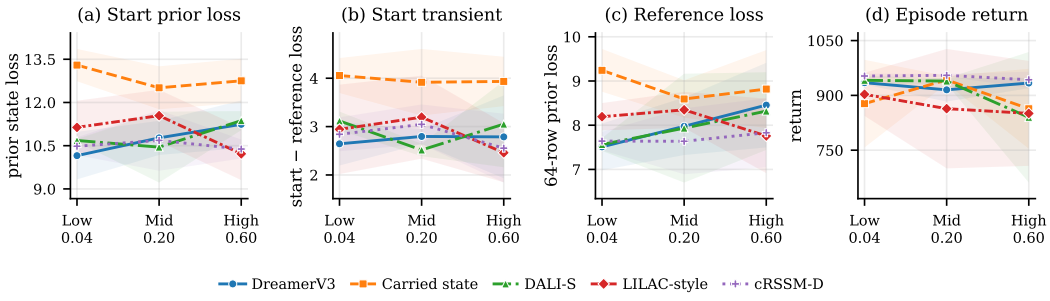


Figure 7: Drift metrics at the three rates on frozen realizations shared across methods. Lower is better in (a)–(c), higher in (d); lines are seed means and bands show sample stdev. The lowest mean startup losses shift from the context-blind baseline to the transition-window and chain-latent agents as drift increases.