# Bayes-FORL: Calibrated Bayesian Fusion of Forecast Priors and Offline Generative Beliefs for Non-Stationary Offline RL

Liz Lemma Future Detective

January 20, 2026

**Abstract**

Offline RL policies often fail at deployment when the observation function drifts non-stationarily across operational intervals (episodes), inducing partial observability even if the underlying dynamics and reward remain unchanged. FORL (Ada et al., NeurIPS 2025) addresses this by combining a zero-shot time-series forecaster for episodic offsets with a conditional diffusion model that proposes multimodal candidate states from within-episode action/effect history, fusing the two via a heuristic dimension-wise closest match (DCM). We push this line into a principled 2026-era interface between foundation forecasting and control: Bayes-FORL replaces DCM with a calibrated approximate Bayesian filter. We treat the forecaster as a multivariate prior over sensor shifts and the offline-trained generative belief model as a trajectory-consistency likelihood surrogate, yielding a posterior over offsets (and thus states) that preserves cross-dimensional correlations and quantifies uncertainty. Our main result is a modular value-loss bound for any frozen offline policy deployed on Bayes-FORL estimates: the performance degradation is upper bounded by a Lipschitz constant times the sum of forecast error, belief-model error, and inference approximation error; a matching lower bound shows this dependence is tight. We further specify an implementable particle-based algorithm and identify stress-test regimes (high-dimensional coupled offsets and tail-risk metrics) where Bayesian fusion should substantially reduce catastrophic state-estimation spikes compared to coordinate-wise fusion. Experiments on D4RL/OGBench augmented with correlated real-world time-series offsets would strengthen the empirical claim.

## Table of Contents

2. 2. Related Work: FORL; robust offline RL under perturbations; forecasting + control; Bayesian filtering / particle filters; diffusion/flow models as conditional priors.

3. 3. Formal Problem Setup: stationary training MDP; deployment as sequence of POMDPs with episodic offsets; access patterns (offline dataset, forecasting context, no test-time offset labels).

4. 4. Generative Belief Models as Likelihood Surrogates: conditional diffusion/flow belief model; from sampling-only models to energy/score surrogates; calibration and support constraints.

5. 5. Bayes-FORL Algorithm: posterior over offsets via product-of-experts; particle/importance sampling implementation; online updates; correlation-preserving estimation outputs; variants (MAP vs posterior mean; per-step vs aggregated likelihood).

6. 6. Theory I — Estimation: posterior concentration / mode selection with forecast priors and energy surrogates; finite-particle approximation bounds; conformal coverage integration.

7. 7. Theory II — Control: modular value-loss upper bound for frozen policies; matching lower bound constructions; discussion of tightness and required assumptions.

8. 8. Complexity and Practical Considerations: time/space; amortization; calibration costs; failure modes; recommended engineering choices.

9. 9. Experimental Protocol (Recommended): correlated-offset benchmark design; tail-risk metrics (max error, CVaR); ablations (univariate vs multivariate forecasting, DCM vs Bayes-FORL, calibration on/off).

10. 10. Discussion and Limitations: dependence on likelihood surrogate quality; multimodality and symmetry; extension to affine transforms and anchor observations; open problems.

# 1 Introduction and Motivation

We study a deployment mismatch that is ubiquitous in embodied control and industrial decision systems: the underlying Markovian dynamics and rewards remain stationary across time, yet the *observation function* is not. Concretely, we consider episodic deployment where, in each episode $j$, sensors incur an unknown additive shift $b_j \in \mathbb{R}^n$ that is constant within the episode but varies across episodes. Thus the deployed agent observes

$$o_t \;=\; s_t + b_j,$$

where $s_t$ denotes the true latent state (fully observed during offline data collection) and $o_t$ is the corrupted observation available online. Such offsets arise from sensor re-calibration, systematic bias, camera extrinsics drift, changes in reference frames, and imperfect zeroing procedures. While the additive form is simple, the resulting *control* problem is subtle because the deployed policy is often fixed: in offline RL, a policy $\pi$ is trained from a dataset collected under one observation convention and must be deployed without further interaction-driven policy updates. The only permissible online adaptation is then to *estimate* the latent state (or equivalently the offset) sufficiently well so that the frozen policy remains effective.

The difficulty is not merely that $b_j$ is unknown. Rather, $b_j$ is *episode-constant*, hence it is statistically identifiable only through within-episode temporal structure and action-conditioned evolution. If one were to treat each time step in isolation, $o_t$ does not determine $s_t$ or $b_j$. Consequently, any successful method must pool evidence across time without violating the online constraint of bounded computation. This suggests an inference problem: given a stream $(o_{0:t}, a_{0:t-1})$ in an episode, infer $b_j$ (and thus $s_t$) online. However, the inference must be compatible with the fact that the agent is actively choosing $a_t$ based on its current estimate, so the distribution of future observations depends on the estimator itself.

A common heuristic response to sensor shifts is *coordinate-wise fusion*: estimate each component $b_{j,i}$ independently by comparing observed coordinates to an expected reference range, or by applying per-dimension filters and then concatenating the resulting estimates. We emphasize that such procedures can fail catastrophically in high dimension, even when each coordinate appears individually "well-behaved." The reason is that the latent state typically satisfies strong cross-dimensional constraints induced by the physical system and by the data distribution implicit in the offline dataset. When one performs dimension-wise inference, one implicitly replaces the joint posterior $p(b \mid \text{data})$ by a product of marginals $\prod_i p(b_i \mid \text{data})$, thereby discarding correlations that may be essential for identifiability. In the extreme, the true posterior over $b$ may be multimodal along correlated subspaces (e.g., due to symmetries), and the coordinate-wise posterior mean or median can lie in a region of negligible joint probability. Then the induced state estimate $\tilde{s}_t = o_t - \hat{b}_t$ can be *inconsistent with any plausible trajectory* seen during training, and a frozen policy $\pi$ can respond arbitrarily poorly to such out-of-distribution inputs. This is not a measure-theoretic pathology: as $n$ grows, the probability mass of correlated distributions concentrates away from coordinate-wise "typical" points, so the mismatch between joint and marginal summaries becomes more severe.

The preceding issue is exacerbated by the online nature of control. A frozen policy $\pi$ is typically Lipschitz only locally on the support of training states, and it may be highly sensitive to perturbations in directions that were rarely explored offline. Thus a state estimator that produces small per-coordinate errors but violates the learned manifold can induce large action errors. Moreover, policy-induced feedback can amplify early estimation mistakes: if $\tilde{s}_t$ is biased, the selected $a_t$ changes the next state distribution, potentially reducing the informativeness of subsequent observations for correcting $b_j$. For this reason we seek estimators that (i) respect the joint structure of the state distribution learned offline, (ii) explicitly represent uncertainty over the episode-constant offset, and (iii) can integrate information sequentially with bounded per-step cost.

Our proposal, Bayes-FORL, is a modular procedure that instantiates these desiderata. We assume access to two deployment-time ingredients in addition to the frozen policy. First, we assume an offline-trained conditional belief model of state trajectories, summarized by an energy surrogate $E_\theta(s, \tau)$ that approximates the negative log-belief $-\log p(s_t = s \mid \tau)$ given a within-episode history window $\tau_{t,w}$. Importantly, we do *not* require a full generative simulator at test time; we only require fast evaluation of $E_\theta$ to score candidate latent states. Second, we assume that before episode $j$ begins, an external forecaster produces a prior $q_j(b)$ over the offset $b_j$ (optionally together with a conformal set $\mathcal{B}_j$ with marginal coverage). This forecaster may use any side information or long-horizon logs available at the system level; Bayes-FORL treats it as a black box and relies on calibration for reliability rather than on model correctness.

Given $q_j$ and $E_\theta$, Bayes-FORL maintains a weighted particle approximation to the posterior over $b_j$. Particles $b^{(m)}$ are sampled jointly in $\mathbb{R}^n$ and remain constant throughout the episode; only their weights evolve. Upon observing $o_t$, each particle proposes a latent state $s_t^{(m)} = o_t - b^{(m)}$ and receives a likelihood surrogate proportional to $\exp(-E_\theta(s_t^{(m)}, \tau_{t,w}))$. Weight updates aggregate evidence across time, so posterior mass concentrates on offsets that make the implied latent trajectory *jointly plausible* under the offline learned belief model. The estimator outputs either the posterior mean $\hat{b}_t$ or a MAP particle and forms $\tilde{s}_t = o_t - \hat{b}_t$, which is then fed into the frozen policy $a_t \sim \pi(\cdot \mid \tilde{s}_t)$. Because inference is carried out in the joint space of offsets, Bayes-FORL preserves cross-dimensional correlations and can represent multimodality, in contrast to coordinate-wise fusions.

A key modeling choice is that the belief model conditions on within-episode history through *delta observations* $\Delta o_t := o_t - o_{t-1}$. Under an episode-constant offset, $\Delta o_t = s_t - s_{t-1}$, so the deltas remove $b_j$ and provide offset-invariant information about dynamics. By incorporating $\tau_{t,w}$ containing $(\Delta o, a)$ pairs, the energy surrogate can learn rich action-conditioned priors over feasible state transitions while remaining robust to the unknown absolute reference frame. This furnishes Bayes-FORL with a principled mechanism to "explain away" the offset: offsets that render $o_t - b$ compatible with the learned dynamics receive higher posterior weight.

Our contributions are therefore conceptual and modular. We (i) formalize episodic additive shifts as a structured nonstationarity where online adaptation is possible via inference even when the policy is frozen; (ii) identify a concrete failure mode of dimension-wise offset correction in high dimension, stemming from the loss of joint structure; (iii) propose Bayes-FORL, which combines a forecast prior with an offline-trained energy-based belief model in an online particle filter; and (iv) provide a theory template linking estimation error to value degradation via Lipschitz regularity of $Q^\pi$ and $\pi$, thereby cleanly separating forecast error, belief-model approximation error, and finite-particle inference error. The remainder of the paper places Bayes-FORL in context, details the algorithmic design choices, and develops guarantees and stress-tests appropriate for deployment under sensor shifts.

## 2 Related Work

**Offline RL with distribution shift and robustness.** Offline reinforcement learning addresses the problem of learning a policy from a fixed dataset without online interaction **?**. A substantial line of work studies robustness to discrepancies between the data-generating distribution and the deployment distribution, including behavior-regularized objectives and conservative value estimation **??**. These approaches primarily mitigate *action-distribution* shift and compounding extrapolation error, and they typically intervene by *changing* the deployed policy (either explicitly through policy optimization or implicitly through pessimistic evaluation). In contrast, our setting isolates an *observation-model* mismatch—an episodic additive shift of the input coordinates—under the constraint that the policy is frozen at deployment. Classical robust MDP formulations **??** consider uncertainty sets over transitions and rewards and derive minimax-optimal policies; while conceptually related (uncertainty is resolved at decision time), they again presuppose the ability to optimize the policy against the uncertainty set. Our contribution is orthogonal: we treat robustness as an *inference* problem over an episode-constant nuisance parameter, and we quantify the resulting control degradation via modular Lipschitz bounds.

**FORL and state-estimation-centric deployment adaptation.** Recent work at the intersection of representation learning and offline RL emphasizes that a significant portion of deployment failures can be traced to mismatched state representations (sensor drift, calibration changes, partial observability) rather than to incorrect dynamics models. Approaches sometimes grouped under "filtering for RL" or "forward-model-based filtering" (we use the shorthand FORL to denote this general direction) combine an offline-learned predictive model with online inference over latent states, sometimes via recurrent architectures trained end-to-end on trajectories. Such methods are standard in partially observable RL, where recurrent policies approximate belief-state control. Our setting differs in two respects. First, during training the state is fully observed, and the nonstationarity enters only at deployment via an episodic observation offset. Second, we separate the deployed policy from the estimator: the estimator may use an offline-trained belief model, but the control law is a fixed map $\pi(\cdot \mid \tilde{s}_t)$ acting on the estimator output. This separation motivates guarantees that explicitly decompose performance loss into estimation error terms (forecast mis-specification, belief-model approximation, and finite-sample inference), rather than absorbing all errors into an end-to-end learned recurrent policy.

**Robustness to observation perturbations and test-time adaptation.** There is an extensive literature on robustness to observation noise and adversarial perturbations in RL, including robust policy learning under corrupted sensors and domain randomization in simulation-to-real transfer. These methods again typically rely on retraining or fine-tuning policies to become invariant to a family of corruptions. The episodic additive shift we study is a particularly structured corruption: it is constant within an episode and shared across all coordinates. This structure admits a low-dimensional latent variable $b_j$ whose posterior can, in principle, concentrate rapidly when one aggregates evidence across time. In this sense, our approach is closer to test-time system identification (estimating a latent environment parameter and conditioning the controller on it) than to worst-case adversarial robustness. The estimator-only adaptation constraint also places us nearer to classical sensor calibration and bias estimation than to policy-level robustification.

**Forecasting meets control and calibrated uncertainty.** Integrating probabilistic forecasts into decision-making is classical in operations research and control: predictive distributions over exogenous signals are propagated through stochastic control objectives, often via chance constraints, risk measures, or scenario optimization. In modern "predict-then-optimize" pipelines, a forecaster provides a distribution over future quantities and an optimizer consumes it. Our use of a forecaster is similar in spirit but differs in the object being forecast: we forecast an *episode-level offset* $b_j$ rather than future states. Moreover, we treat the forecaster as a black box and rely on calibration tools to guard against mis-specification. Conformal prediction provides finite-sample, distribution-free coverage guarantees under exchangeability **??**. Conformal sets for multivariate quantities can be built via scalar nonconformity scores (e.g., Mahalanobis or copula-based scores), yielding marginal coverage even when the forecaster is misspecified. In our framework, such sets $\mathcal{B}_j$ serve as an optional truncation region for posterior inference over $b_j$, providing a principled mechanism to bound tail risk stemming from forecast errors.

**Bayesian filtering and particle methods for static parameters.** Our deployment-time inference problem is a special case of filtering with a static latent parameter (here, an episode-constant bias) and a dynamic latent state. Sequential Monte Carlo (SMC) methods, including particle filters with sequential importance resampling, provide a standard tool for approximating filtering distributions in non-linear, non-Gaussian state-space models **?**. When the latent variable includes a static parameter, naive particle filtering can suffer from degeneracy, motivating parameter-learning variants, rejuvenation via MCMC moves, or Rao–Blackwellization when conditional structure is available **?**. Our procedure is best viewed as an SMC approximation to a posterior over $b_j$ with an analytically eliminated state variable given by $s_t = o_t - b$; the novelty is that the "likelihood" terms are supplied by a learned energy surrogate rather than by an explicit generative observation model. This combination places our method between classical model-based filters (where likelihoods are known) and amortized inference (where posteriors are learned directly): we amortize only the scoring function, retaining explicit Bayesian updating with a forecast prior.

**Energy-based models and learned conditional priors.** Energy-based models (EBMs) represent distributions via an unnormalized density proportional to $\exp(-E_\theta)$ and have been used as expressive priors and conditional models in vision and sequential modeling **??**. When EBMs are used for inference, sampling is often carried out by Langevin dynamics or other MCMC schemes, which can be computationally intensive at test time. In our setting, we avoid iterative sampling over the latent state trajectory by exploiting the additive observation structure: candidate offsets $b$ induce candidate states $s_t = o_t - b$, and the energy surrogate is evaluated pointwise on these candidates. Thus, online computation is dominated by $O(M)$ energy evaluations and weight updates, compatible with real-time constraints. We emphasize that our guarantees are phrased in terms of a uniform surrogate error $\varepsilon_G$, allowing the belief model to be trained by any suitable offline procedure (contrastive objectives, score matching, or supervised density surrogates) as long as the resulting energy approximates negative log-beliefs on the relevant compact set.

**Diffusion and flow models as conditional trajectory priors.** Diffusion models and normalizing flows provide alternative ways to represent conditional distributions over states or trajectories **??**. In offline RL, diffusion-based trajectory generation and planning have recently been explored as a way to model multi-modal behavior distributions and synthesize action sequences **?**. These methods underscore the importance of capturing joint correlations and multi-modality—precisely the failure mode of coordinate-wise fusion that motivates our approach. However, diffusion-based inference typically requires multiple denoising steps, which can be costly in an online control loop, and the interface to Bayesian updating with an external forecast prior is not immediate. Our method can be interpreted as using an EBM-style scoring function as a conditional prior over feasible states (given within-episode histories), while using the forecaster to supply the episode-level prior over offsets; this yields a plug-and-play Bayesian update whose online cost scales linearly with the number of particles.

**Summary.** Taken together, the above threads suggest a design point that is underrepresented in the literature: when policy adaptation is disallowed, robustness to structured observation nonstationarity should be pursued by (i) explicit episode-level latent-variable inference, (ii) calibrated forecast priors to encode cross-episode information, and (iii) joint (rather than coordinate-wise) posterior representations to preserve correlations. Our subsequent formal setup isolates these ingredients and makes the access pattern explicit so that both algorithmic and theoretical statements are unambiguous.

# 3 Formal Problem Setup

We formalize an episodic deployment setting in which the *control objective* remains stationary but the *observation map* is perturbed by an episode-constant additive offset. The key constraint is that the deployed policy is fixed; the only admissible adaptation at test time is via an online state estimator that preprocesses observations for the frozen policy.

**Training environment and offline data.** At training time we are given a fully observed discounted Markov decision process (MDP)

$$M_{\text{train}} := (\mathcal{S}, \mathcal{A}, T, r, \rho_0, \gamma),$$

where $\mathcal{S} \subseteq \mathbb{R}^n$ is a continuous state space, $\mathcal{A}$ is an action space, $T(\cdot \mid s, a)$ is the transition kernel, $r(s, a)$ is the reward function, $\rho_0$ is the initial-state distribution, and $\gamma \in (0, 1)$ is the discount factor. We assume that the training process produces an offline dataset

$$\mathcal{D} = \{(s_t, a_t, s_{t+1}, r_t)\}$$

collected in $M_{\text{train}}$ under some (possibly unknown) behavior policy. From $\mathcal{D}$ we train an *offline policy* $\pi$ (by any offline RL procedure) and then *freeze* $\pi$ for deployment. We emphasize that all quantities entering $\pi$ at training are functions of the *true* state $s_t$; the policy never observes offsets during training.

**Deployment as a sequence of offset-corrupted episodes.** Deployment consists of episodes indexed by $j \in \{1, 2, \dots\}$. Within episode $j$, there is a latent state sequence $(s_t)_{t=0}^{T}$ evolving under the same stationary dynamics $T$ and reward $r$ as in training, but the agent does not directly observe $s_t$. Instead it receives observations

$$o_t = s_t + b_j, \qquad b_j \in \mathbb{R}^n, \tag{1}$$

where the offset $b_j$ is *constant within the episode* and unknown to the agent. Equation (1) defines a family of partially observed processes, one per episode, with shared latent dynamics and an episode-level nuisance parameter $b_j$. (The analysis and algorithmic interface we develop are designed around this structured nonstationarity; we do not treat $b_j$ as i.i.d. per step noise.)

Because $b_j$ is constant, first differences remove it:

$$\Delta o_t := o_t - o_{t-1} = (s_t + b_j) - (s_{t-1} + b_j) = s_t - s_{t-1}.$$

Accordingly, $\Delta o_t$ is an *offset-invariant* signal that can be used to condition belief updates within an episode. We will use a finite history window size $w$ and define the within-episode context

$$\tau_{t,w} := \big[(\Delta o_{t-w+1}, a_{t-w}), \dots, (\Delta o_t, a_{t-1})\big],$$

with the natural truncation when $t < w$.

**Estimator–policy separation and the deployed control loop.** At deployment, the policy $\pi$ is evaluated on an *estimated* state $\tilde{s}_t$ rather than on $o_t$ or $s_t$. Concretely, we introduce an online estimator (or filter)

$$\mathcal{E} : \ (o_{0:t}, a_{0:t-1}, \tau_{t,w}, q_j) \ \mapsto \ \tilde{s}_t,$$

and actions are drawn according to the frozen policy

$$a_t \sim \pi(\cdot \mid \tilde{s}_t).$$

We regard this separation as a hard constraint: the estimator may maintain internal memory (e.g. a belief over $b_j$), but $\pi$ itself is not updated, fine-tuned, or re-optimized during evaluation. The resulting deployed value is denoted $V^{\pi \circ \mathcal{E}}$, emphasizing that control quality is mediated by estimation quality.

**Cross-episode forecasts and calibrated uncertainty for the offset.** The offset $b_j$ is episode-specific but not arbitrary: in many applications it reflects calibration drift or slow changes across episodes. We formalize the availability of cross-episode information by assuming that at the start of episode $j$ a probabilistic forecaster provides a *prior distribution* $q_j(b)$ over $b_j$. This prior may be misspecified; we therefore optionally augment it with a conformal prediction set $\mathcal{B}_j$ satisfying the marginal coverage guarantee

$$\mathbb{P}(b_j \in \mathcal{B}_j) \geq 1 - \delta,$$

under exchangeability assumptions on the calibration procedure. Operationally, $\mathcal{B}_j$ serves as a truncation or support constraint for inference, controlling tail behavior when $q_j$ assigns insufficient mass near the realized offset.

**Access patterns and what is *not* observed at test time.** The estimator has the following access pattern.

- *Offline (predeployment):* access to $\mathcal{D}$ and unlimited training compute to produce (i) the frozen policy $\pi$, and (ii) auxiliary models used by $\mathcal{E}$.
- *At episode start:* access to the forecast prior $q_j(b)$ and (optionally) $\mathcal{B}_j$.
- *Online during the episode:* sequential access to $o_t$ and the actions actually taken $a_t$ (which are generated by $\pi(\cdot \mid \tilde{s}_t)$). The offset $b_j$ is *not* revealed during evaluation; likewise, ground-truth states $s_t$ are not observed.

We also impose a bounded per-step compute budget: $\mathcal{E}$ must run online with fixed-cost updates, allowing (for example) a finite number of calls to a learned scoring function and simple particle-weight updates, but disallowing expensive test-time retraining or long-horizon planning.

**Objective: control loss relative to an oracle-offset agent.** The idealized benchmark is an oracle-offset agent that knows $b_j$ and therefore acts on the true state $s_t = o_t - b_j$, yielding value $V^{\pi \circ \text{oracle}}$ when using the same frozen policy $\pi$. Our primary objective is to minimize the deployment degradation

$$\Delta V := V^{\pi \circ \mathcal{E}} - V^{\pi \circ \text{oracle}},$$

and secondarily to control tail risk of estimation error within an episode (e.g. $\max_t \|\tilde{s}_t - s_t\|$ or a CVaR-type criterion). Since the policy is fixed, any improvement in $\Delta V$ must come through reducing the discrepancy between $\tilde{s}_t$ and $s_t$, while respecting the online access pattern above.

**Need for a likelihood surrogate from offline data.** To construct $\mathcal{E}$ we require a mechanism that scores candidate latent states $s$ given the within-episode context $\tau_{t,w}$ in a way that is compatible with Bayesian updating over $b_j$ through the relation $s = o_t - b$. We assume that offline training yields a conditional belief model summarized by an energy surrogate $E_\theta(s, \tau)$ intended to approximate $-\log p(s_t = s \mid \tau)$ (up to an additive, $\tau$-dependent constant) on the relevant compact domain. In the next section we specify how such conditional generative belief models are obtained, and how sampling-oriented models (diffusions or flows) can be converted into computationally efficient energy/score surrogates suitable for online filtering.

## 4 Generative Belief Models as Likelihood Surrogates

Bayes-FORL requires, for each time $t$, a mechanism for assigning relative plausibility to candidate latent states $s$ given the within-episode context $\tau_{t,w}$. Since the online estimator will compare hypotheses of the form $s = o_t - b$ across many candidate offsets $b$, we require a *fast* scoring rule in $s$ that is trained offline from $\mathcal{D}$ and can be evaluated online without inner-loop sampling. We therefore view offline learning as producing a *conditional belief model*

$$p_\star(s \mid \tau) \quad \text{(implicit or explicit)},$$

together with a computational surrogate $E_\theta(s, \tau)$ satisfying

$$E_\theta(s, \tau) \approx -\log p_\star(s \mid \tau) + c(\tau), \tag{2}$$

where $c(\tau)$ is an arbitrary additive normalization term that may depend on $\tau$ but not on $s$. Because our online posterior over offsets is computed only up to proportionality, such additive terms cancel and do not affect inference over $b$.

**Offline construction of conditional training pairs.** From trajectories in $\mathcal{D}$ we form supervised pairs $(s_t, \tau_{t,w})$ by computing $\Delta s_t = s_t - s_{t-1}$ and identifying $\Delta o_t$ with $\Delta s_t$ under the training observation model (no offset). In particular, we may set $\Delta o_t := s_t - s_{t-1}$ when training $E_\theta$, thereby matching the deployment statistic $\Delta o_t = o_t - o_{t-1}$. This produces a stationary conditional prediction problem: given a window of recent differences and actions, score the current state. We stress that the conditioning variable $\tau_{t,w}$ is chosen to be offset-invariant so that (2) remains meaningful under deployment shifts.

**Conditional normalizing flows: exact likelihoods and energies.** If we train a conditional normalizing flow $f_\psi(\cdot; \tau)$ such that $s = f_\psi(z; \tau)$ with

$z \sim \mathcal{N}(0, I)$, then the induced density $p_\psi(s \mid \tau)$ admits an exact change-of-variables likelihood:

$$\log p_\psi(s \mid \tau) = \log p_Z(f_\psi^{-1}(s;\tau)) + \log \left| \det \nabla_s f_\psi^{-1}(s;\tau) \right| .$$

In this case, an immediate choice is

$$E_\theta(s, \tau) := -\log p_\psi(s \mid \tau),$$

with $\theta = \psi$ and $\varepsilon_G$ determined by standard generalization error on held-out $(s, \tau)$ pairs. The flow case is algorithmically convenient: online we evaluate $E_\theta(o_t - b, \tau_{t,w})$ directly, and no further distillation is required.

**Conditional diffusion models: from sampling to scores/energies.** Diffusion models are typically trained to enable conditional sampling rather than direct likelihood evaluation. Concretely, for a noise schedule $\{\sigma_k\}_{k=1}^K$ one trains a denoiser or score network $u_\psi(s_k, k, \tau) \approx \nabla_{s_k} \log p(s_k \mid \tau)$ (or an equivalent $\varepsilon$-prediction parameterization), where $s_k = s + \sigma_k \epsilon$ and $\epsilon \sim \mathcal{N}(0, I)$. To use such models inside Bayes-FORL, we convert them into an online scoring surrogate by one of the following standard reductions.

*(i) Score-to-energy distillation.* On a compact domain $\mathcal{S}_0 \subset \mathbb{R}^n$ we may train an energy network $E_\theta(s, \tau)$ so that its gradient matches a target score field $\hat{g}_\psi(s, \tau)$ obtained from the diffusion model (e.g. via denoising score matching at a fixed low-noise level). A prototypical objective is

$$\min_\theta \; \mathbb{E}\Big[ \big\| \nabla_s E_\theta(s, \tau) + \hat{g}_\psi(s, \tau) \big\|^2 \Big],$$

where $(s, \tau)$ are drawn from the offline construction above and $\hat{g}_\psi$ is computed by a single network call. When the learned vector field is approximately conservative on $\mathcal{S}_0$, this yields an $E_\theta$ satisfying (2) up to an additive constant, with uniform error summarized by $\varepsilon_G$.

*(ii) Likelihood surrogates via ELBO/probability-flow.* Alternatively, one may approximate $-\log p(s \mid \tau)$ by an evidence lower bound (ELBO) computed from the diffusion training objective, or by integrating along the probability-flow ODE with Hutchinson trace estimators. While this can provide a closer approximation to likelihood, it may be too expensive online. We therefore treat such constructions primarily as *offline* tools to produce a cheaper $E_\theta$ by regression.

In both diffusion-derived approaches, we emphasize the operational requirement: online filtering evaluates $E_\theta$ at $M$ candidate states $s = o_t - b^{(m)}$ per time step, so $E_\theta$ must be a single forward pass (and ideally avoid iterative sampling).

**Calibration of energy scales for stable importance weights.** Even when (2) holds approximately, the scale of $E_\theta$ can be miscalibrated, causing importance weights to collapse. Since Bayes-FORL uses $\exp(-E_\theta)$ multiplicatively across time, small systematic scale errors can be amplified. We therefore optionally introduce a temperature parameter $\beta > 0$ and use the tempered surrogate $\exp(-\beta E_\theta)$, with $\beta$ selected on a held-out validation set to control effective sample size (ESS) or to minimize a predictive loss. Equivalently, we may perform an affine calibration

$$E_\theta^{\mathrm{cal}}(s, \tau) = \alpha E_\theta(s, \tau) + \kappa(\tau),$$

where $\kappa(\tau)$ is irrelevant for offset inference but may be used to stabilize numerics (e.g. centering by the minimum energy over a minibatch). This calibration is purely offline; online, we only evaluate $E_\theta^{\mathrm{cal}}$.

**Support constraints and robustness to forecast mis-specification.** The product form of the offset posterior combines the forecast prior $q_j(b)$ with within-episode evidence. If $q_j$ assigns negligible mass near the realized $b_j$, particle methods can fail regardless of the quality of $E_\theta$. We therefore incorporate explicit support constraints in two complementary ways.

First, if a conformal set $\mathcal{B}_j$ is available, we truncate the proposal and posterior to $\mathcal{B}_j$ by sampling $b^{(m)} \sim q_j(\cdot \mid b \in \mathcal{B}_j)$ (or rejecting samples outside $\mathcal{B}_j$). This enforces the marginal coverage guarantee and isolates the remaining error into a forecast term $\varepsilon_F$ measuring how much prior mass $q_j$ places near the true offset within $\mathcal{B}_j$.

Second, we may impose a *state-domain* constraint by restricting candidate states to a compact set $\mathcal{S}_0$ (e.g. the convex hull of offline states, possibly dilated). Operationally, when evaluating a particle $b^{(m)}$ we set $E_\theta(o_t - b^{(m)}, \tau_{t,w}) = +\infty$ (or a large constant) if $o_t - b^{(m)} \notin \mathcal{S}_0$. This prevents the filter from explaining observations using offsets that imply implausible states, and it aligns the analysis with the uniform approximation premise defining $\varepsilon_G$.

**Summary of the surrogate interface.** For the subsequent algorithmic development, the only required interface is the ability to compute $E_\theta(s, \tau)$ for arbitrary $s$ and $\tau$ at online time. Conditional flows provide this directly; conditional diffusions provide it after offline distillation into an energy/score surrogate. Calibration (via temperature or affine scaling) and explicit support constraints (via $\mathcal{B}_j$ and $\mathcal{S}_0$) are modular add-ons that improve stability without altering the estimator–policy separation. With this interface in hand, we can form a product-of-experts posterior over $b$ and implement it by particle/importance sampling, which we detail next.

# 5 Bayes-FORL: Filtering an Episode-Constant Offset

We now specify the online estimator $\mathcal{E}$ used to adapt a frozen policy $\pi$ to an episode $j$ in which observations satisfy $o_t = s_t + b_j$ for a fixed but unknown $b_j \in \mathbb{R}^n$. The estimator maintains an episode-level belief over $b$ and returns a point estimate $\hat{b}_t$ (and hence a state estimate $\tilde{s}_t = o_t - \hat{b}_t$) at each time $t$. The defining feature is that we do not attempt to re-plan in belief space; rather, we separate concerns by (i) performing approximate Bayesian inference over $b$ and (ii) feeding the resulting state estimate into the unchanged policy $\pi$.

**Product-of-experts posterior over offsets.** Fix episode $j$ and suppress the index $j$ for readability. Let $q(b)$ denote the forecaster-provided prior for $b$, optionally truncated to a conformal set $\mathcal{B}$ (and/or to those $b$ such that $o_t - b \in \mathcal{S}_0$ for all relevant $t$). Given a windowed, offset-invariant context $\tau_{t,w}$, the surrogate likelihood of the hypothesis $b$ at time $t$ is

$$\tilde{p}_t(o_t \mid b, \tau_{t,w}) \;\propto\; \exp\Big(-E_\theta(o_t - b, \tau_{t,w})\Big), \tag{3}$$

where proportionality absorbs any $b$-independent normalizers. Aggregating evidence up to time $t$ yields the approximate posterior

$$\tilde{p}_t(b \mid o_{0:t}, a_{0:t-1}) \;\propto\; q(b) \prod_{t'=0}^{t} \exp\Big(-E_\theta(o_{t'} - b, \tau_{t',w})\Big). \tag{4}$$

This has a product-of-experts form: the forecast prior contributes a global expert over $b$, while each time step contributes an expert favoring offsets that render the implied latent state $s = o_{t'} - b$ plausible under the offline belief model.

Two remarks are operationally important. First, since $b$ is episode-constant, (4) is not a standard state-space filtering recursion in the latent state; it is a static-parameter posterior updated sequentially as data arrive. Second, the surrogate likelihood (3) is evaluated at $s = o_t - b$, so inference over $b$ is intrinsically *joint* in $\mathbb{R}^n$; we do not decompose into independent coordinates, and thus preserve cross-dimensional correlations induced by both $q(b)$ and $E_\theta$.

**Sequential importance sampling for static parameters.** We approximate (4) with $M$ particles $\{b^{(m)}\}_{m=1}^{M}$ drawn at the beginning of the episode from the proposal $q$ (or $q(\cdot \mid b \in \mathcal{B})$). We attach weights $W_t^{(m)}$ that track the accumulated likelihood contributions. In its simplest form,

$$W_{-1}^{(m)} := \frac{1}{M}, \tag{5}$$

$$\widetilde{W}_t^{(m)} := W_{t-1}^{(m)} \cdot \exp\Big( - E_\theta(o_t - b^{(m)}, \tau_{t,w})\Big), \tag{6}$$

$$W_t^{(m)} := \frac{\widetilde{W}_t^{(m)}}{\sum_{\ell=1}^M \widetilde{W}_t^{(\ell)}}. \tag{7}$$

For numerical stability we implement (6) in log-space, optionally subtracting $\min_m E_\theta(o_t - b^{(m)}, \tau_{t,w})$ before exponentiation. When a temperature parameter $\beta$ is used, we replace $E_\theta$ by $\beta E_\theta$ in (6); this modifies the posterior approximation in a controlled manner and is primarily a variance-reduction device.

Because the particles do not move, the approximation quality hinges on the effective sample size

$$\mathrm{ESS}_t := \frac{1}{\sum_{m=1}^M (W_t^{(m)})^2}.$$

When $\mathrm{ESS}_t$ falls below a threshold (e.g. $M/2$), we perform resampling and reset weights to $1/M$. Since $b$ is static, repeated resampling can lead to impoverishment; we therefore treat resampling as optional and, when used, pair it with a mild rejuvenation kernel $b^{(m)} \leftarrow b^{(m)} + \xi^{(m)}$ for small $\xi^{(m)}$ (e.g. Gaussian with covariance tuned to the current weighted empirical covariance), followed by projection back to $\mathcal{B}$ if truncation is enforced.

**Outputs: posterior mean, MAP, and uncertainty.** At each time $t$ we may output either (i) a posterior mean estimate

$$\hat{b}_t^{\mathrm{mean}} := \sum_{m=1}^M W_t^{(m)} \, b^{(m)}, \tag{8}$$

or (ii) a particle-based MAP estimate

$$\hat{b}_t^{\mathrm{MAP}} := b^{(m_t^\star)} \qquad \text{where} \qquad m_t^\star \in \arg\max_m W_t^{(m)}. \tag{9}$$

The posterior mean is smoother and typically better under squared error; the MAP is often more robust when the posterior is multi-modal and we wish to commit to a single coherent hypothesis for control. Both preserve correlation structure because they are computed from joint particles in $\mathbb{R}^n$.

The state estimate passed to the policy is then

$$\tilde{s}_t := o_t - \hat{b}_t, \tag{10}$$

and the deployed action is sampled as $a_t \sim \pi(\cdot \mid \tilde{s}_t)$. For downstream risk control, we may also report uncertainty, e.g. the weighted covariance

$$\widehat{\mathrm{Cov}}_t(b) := \sum_{m=1}^{M} W_t^{(m)} \big(b^{(m)} - \hat{b}_t^{\mathrm{mean}}\big)\big(b^{(m)} - \hat{b}_t^{\mathrm{mean}}\big)^{\top},$$

or posterior samples of $b$ obtained by resampling particles according to $W_t$. This is useful for diagnosing imminent failure modes (e.g. posterior mass splitting across widely separated offsets).

**Per-step versus aggregated likelihood contributions.** Equation (4) uses the full product over $t' \le t$ and is the most direct approximation of the intended posterior under conditional independence assumptions implicit in the surrogate. In some environments, however, the surrogate $E_\theta(\cdot, \tau_{t,w})$ can over-count evidence due to overlapping windows. We therefore consider two variants.

*(i) Subsampled products.* We update weights only at times $t$ belonging to a subsequence (e.g. every $k$ steps) so that successive likelihood terms depend on more weakly overlapping contexts.

*(ii) Exponentially-forgotten products.* We replace the cumulative sum of energies by a discounted sum, equivalently

$$\widetilde{W}_t^{(m)} := \Big(W_{t-1}^{(m)}\Big)^{\lambda} \exp\Big( - E_\theta(o_t - b^{(m)}, \tau_{t,w})\Big),$$

with $\lambda \in (0, 1]$. This stabilizes weights when $E_\theta$ is slightly miscalibrated and can be interpreted as a robustness heuristic rather than a literal Bayesian update.

In all cases, the estimator remains an episode-level inference procedure over a static $b$ with online complexity linear in $M$.

**Computational and structural properties.** Bayes-FORL requires $M$ evaluations of $E_\theta$ per time step, plus bookkeeping for $\tau_{t,w}$ and (optionally) resampling. Memory is $O(Mn)$ for joint particles and weights. Crucially, because the inference variable is the full vector $b \in \mathbb{R}^n$, cross-dimensional correlations are retained; in contrast to dimension-wise fusion rules, the estimator can express and exploit structured forecast priors (e.g. correlated sensor offsets) as well as structured plausibility constraints induced by the belief model. These algorithmic choices are precisely those that enable the estimation and control guarantees developed next.

## 6 Theory I — Estimation: Concentration, Forecast Robustness, and Finite-$M$ Effects

We now formalize when Bayes-FORL selects (and remains near) the correct offset mode, and how forecast uncertainty, energy-surrogate error, and finite-particle inference contribute additively to the resulting state-estimation error. Throughout we fix a deployment episode and suppress the episode index. Recall that $\tilde{s}_t = o_t - \hat{b}_t$ and, under the additive-offset model $o_t = s_t + b$, we have the identity

$$\tilde{s}_t - s_t = b - \hat{b}_t, \tag{11}$$

so that it suffices to control $\|\hat{b}_t - b\|$.

**Energy-based posterior and the optimization landscape.** Let $\tau_{t,w}$ be the offset-invariant window used by the energy surrogate and define the (approximate) negative log-posterior potential

$$\Phi_t(b) \ := \ -\log q(b) \ + \ \sum_{t'=0}^{t} E_\theta(o_{t'} - b, \tau_{t',w}), \tag{12}$$

so that the Bayes-FORL target density is $\tilde{p}_t(b \mid o_{0:t}, a_{0:t-1}) \propto \exp(-\Phi_t(b))$ (cf. (4)). To state stability results, we compare $\Phi_t$ to the "ideal" potential obtained from the true conditional belief model,

$$\Phi_t^\star(b) \ := \ -\log q(b) + \sum_{t'=0}^{t} E^\star(o_{t'} - b, \tau_{t',w}), \qquad E^\star(s, \tau) := -\log p(s \mid \tau) + c(\tau),$$

where $c(\tau)$ is any additive normalizer independent of $s$ and hence irrelevant for inference over $b$.

Our standing regularity hypothesis is local identifiability of the offset through curvature of $\Phi_t^\star$. Concretely, for a neighborhood $U$ of the true offset $b$, we assume (for $t$ sufficiently large, or for all $t$ after an initial burn-in) that $\Phi_t^\star$ is $\mu$-strongly convex and $L$-smooth on $U$. This is a mode-separation condition: it excludes flat directions and indistinguishable offsets in $U$.

**Forecast misspecification enters as a local prior-mass condition.** The forecaster prior $q$ may be imperfect. For estimation, what matters is not global calibration but whether $q$ assigns non-negligible mass near the true $b$. We quantify this by requiring that on the same neighborhood $U$ one has

$$\inf_{b' \in U} q(b') \ \geq \ \exp(-\varepsilon_F), \tag{13}$$

for some $\varepsilon_F \geq 0$ (small when the forecast is locally accurate). When we truncate $q$ to a conformal set $\mathcal{B}$, (13) is interpreted as a condition on the truncated density (equivalently, we require $b \in \mathcal{B}$ and local mass within $\mathcal{B} \cap U$).

**MAP stability under surrogate energy error.** We assume a uniform approximation property of the energy surrogate on the compact domain of interest: for all $(s, \tau)$ in the domain,

$$\big|E_\theta(s,\tau) - E^\star(s,\tau)\big| \;\leq\; \varepsilon_G. \tag{14}$$

Under this condition, replacing $E^\star$ by $E_\theta$ perturbs $\Phi_t^\star$ by at most $(t+1)\varepsilon_G$ in value; more importantly, it perturbs gradients in a controlled manner when $E^\star$ is regular and the domain is compact. The following theorem records the resulting mode-selection guarantee.

**Theorem 6.1** (Posterior concentration / MAP offset error)**.** *Assume there exists a neighborhood $U$ of the true offset $b$ such that, with probability at least $1 - \eta$ over the episode trajectory, the potential $\Phi_t$ defined in* (12) *is $\mu$-strongly convex and $L$-smooth on $U$. Assume further that* (14) *holds and that the prior-mass condition* (13) *holds on $U$. Let $\hat{b}_t^{\mathrm{MAP}} \in \arg\min_{b'} \Phi_t(b')$ (ties broken arbitrarily). Then on the event above,*

$$\|\hat{b}_t^{\mathrm{MAP}} - b\| \;\leq\; \frac{1}{\mu}\Big(\varepsilon_F + C\,\varepsilon_G\Big) \;+\; \mathrm{Stat}_t, \tag{15}$$

*where $C$ depends only on local regularity constants (e.g. Lipschitz bounds for the gradients of $E^\star$ on $U$), and $\mathrm{Stat}_t$ is a term decreasing in $t$ that captures finite-sample fluctuation of the empirical sum $\sum_{t' \leq t} E^\star(o_{t'} - \cdot, \tau_{t',w})$ around its population counterpart (under mixing/boundedness assumptions).*

The proof is a perturbation argument for strongly convex objectives: strong convexity yields $\|\hat{b}_t^{\mathrm{MAP}} - b\| \leq \mu^{-1}\|\nabla\Phi_t(b) - \nabla\Phi_t(b^\star)\|$ for an appropriate reference point, while the replacement of $E^\star$ by $E_\theta$ and the use of an imperfect prior contribute additive gradient/value perturbations summarized by $\varepsilon_G$ and $\varepsilon_F$. The statistical term $\mathrm{Stat}_t$ is standard and can be made explicit given a concentration model for the trajectory-dependent energies (e.g. bounded differences or martingale concentration).

**From MAP to filtering: why we track a full posterior.** Theorem 6.1 is a *mode-selection* statement: once the posterior landscape becomes sufficiently curved around the true $b$, local optimization (or a particle approximation that maintains support in $U$) will remain near the correct mode. This is precisely why Bayes-FORL maintains a full posterior over $b$ rather than committing too early to a point estimate: if multiple modes are plausible initially, a particle representation can retain them until evidence (as measured by the accumulated energies) separates the modes.

**Finite-$M$ importance sampling error.** Bayes-FORL computes either the posterior mean (8) or a particle MAP (9). To isolate finite-$M$ effects for the posterior mean, we appeal to standard concentration for self-normalized importance sampling. Let

$$G_t(b) \;:=\; \exp\Big(-\sum_{t'=0}^{t} E_\theta(o_{t'} - b, \tau_{t',w})\Big),$$

so that the exact surrogate posterior mean is $\mathbb{E}_q[b\,G_t(b)]/\mathbb{E}_q[G_t(b)]$, and the particle estimator is its self-normalized Monte Carlo approximation. Under boundedness of $G_t$ on the truncated support (which can be enforced by compactness and mild clipping of energies), we obtain the usual $O(M^{-1/2})$ rate.

**Theorem 6.2** (Finite-particle error for the posterior mean)**.** *Assume* $0 < \underline{w} \le G_t(b) \le \overline{w} < \infty$ *for all* $b$ *in the sampling support. Let* $\hat{b}_t^{(M)}$ *be the self-normalized importance-sampling estimate of the posterior mean based on* $M$ *i.i.d. samples from* $q$*. Then for any* $\eta \in (0,1)$*, with probability at least* $1-\eta$*,*

$$\big\|\hat{b}_t^{(M)} - \mathbb{E}[b \mid o_{0:t}, a_{0:t-1}]\big\| \;\le\; C' \sqrt{\frac{\log(1/\eta)}{M}} \cdot \frac{\overline{w}}{\underline{w}}, \tag{16}$$

*for a constant* $C'$ *depending only on the diameter of the support.*

**Integrating conformal coverage.** When we truncate the prior to a conformal set $\mathcal{B}$, Theorem 1 guarantees $\mathbb{P}(b \in \mathcal{B}) \ge 1-\delta$ marginally. Conditioning on the event $\{b \in \mathcal{B}\}$, Theorems 6.1–6.2 apply with constants computed on $\mathcal{B}$. Unconditioning yields a two-level guarantee: with probability at least $1-\delta-\eta$, Bayes-FORL operates in the "well-specified support" regime and achieves estimation error controlled by $\varepsilon_F$, $\varepsilon_G$, and the inference term $\varepsilon_{\inf}(M) := O(\sqrt{\log(1/\eta)/M})$; on the remaining $\delta$ mass where $b \notin \mathcal{B}$, no nontrivial bound is possible without further assumptions. In view of (11), this directly implies corresponding bounds on $\|\tilde{s}_t - s_t\|$, which will be the only estimator-dependent quantity entering our control analysis in the next section.

# 7 Theory II — Control: Modular Value-Loss Bounds for Frozen Policies

We now translate state-estimation error into a bound on the deployment value degradation incurred by acting through the frozen policy $\pi$ on $\tilde{s}_t$ rather than on the true state $s_t$. The central point is that, once $\pi$ is fixed, *all* deploy-time adaptation enters exclusively through $\tilde{s}_t$; consequently, any control guarantee must be modular in the sense that it depends on the estimator only via a scalar summary of its state error.

**Setup and a compatible action metric.** Let $Q^\pi(s,a)$ be the action-value function corresponding to the fixed policy $\pi$ under the true latent MDP dynamics and rewards. We assume $Q^\pi$ is $L_Q$-Lipschitz in $(s,a)$ on a compact domain, i.e.,

$$\big|Q^\pi(s,a) - Q^\pi(s',a')\big| \;\le\; L_Q\big(\|s-s'\| + d_{\mathcal{A}}(a,a')\big),$$

where $d_{\mathcal{A}}$ is a metric on actions (for continuous actions, typically $\|a-a'\|$). Since $\pi$ is stochastic in general, we require a Lipschitz property mapping states to action *distributions*: there exists $L_\pi$ such that

$$d_{\mathcal{P}}\big(\pi(\cdot\mid s), \pi(\cdot\mid s')\big) \;\le\; L_\pi\|s-s'\|, \tag{17}$$

where $d_{\mathcal{P}}$ is a probability-metric compatible with $Q^\pi$, in the sense that for any fixed $s$ and any distributions $\nu, \nu'$ over actions,

$$\Big|\mathbb{E}_{a\sim\nu}Q^\pi(s,a) - \mathbb{E}_{a\sim\nu'}Q^\pi(s,a)\Big| \;\le\; L_Q\, d_{\mathcal{P}}(\nu,\nu'). \tag{18}$$

For example, (18) holds with $d_{\mathcal{P}}$ equal to 1-Wasserstein when $a \mapsto Q^\pi(s,a)$ is $L_Q$-Lipschitz in $a$, and also holds with $d_{\mathcal{P}}$ equal to total variation when $Q^\pi$ is bounded.

**A one-step deviation bound.** Fix a time $t$, and compare the oracle action distribution $\pi(\cdot\mid s_t)$ to the deployed one $\pi(\cdot\mid\tilde{s}_t)$. Using (18)–(17) and the Lipschitzness of $Q^\pi$ in $s$, we obtain

$$\Big|\mathbb{E}_{a\sim\pi(\cdot|\tilde{s}_t)}Q^\pi(s_t,a) - \mathbb{E}_{a\sim\pi(\cdot|s_t)}Q^\pi(s_t,a)\Big| \;\le\; L_Q\, d_{\mathcal{P}}\big(\pi(\cdot\mid\tilde{s}_t), \pi(\cdot\mid s_t)\big) \;\le\; L_Q L_\pi\|\tilde{s}_t - s_t\|, \tag{19}$$

$$\Big|\mathbb{E}_{a\sim\pi(\cdot|\tilde{s}_t)}Q^\pi(\tilde{s}_t,a) - \mathbb{E}_{a\sim\pi(\cdot|\tilde{s}_t)}Q^\pi(s_t,a)\Big| \;\le\; L_Q\|\tilde{s}_t - s_t\|. \tag{20}$$

Combining (19)–(20) yields the canonical $(1+L_\pi)$ factor: at time $t$ we pay once for evaluating $Q^\pi$ at the wrong state, and once more for sampling an action from the wrong conditional distribution.

**Modular discounted value-loss upper bound.** We lift the one-step bound to a discounted value statement by a standard Bellman telescoping argument. Let $V^{\pi\circ\mathcal{E}}(s_0)$ denote the value achieved when actions are sampled as $a_t \sim \pi(\cdot\mid\tilde{s}_t)$ with $\tilde{s}_t = \mathcal{E}(o_{0:t}, a_{0:t-1})$, and let $V^\pi(s_0)$ denote the oracle value when sampling $a_t \sim \pi(\cdot\mid s_t)$.

**Theorem 7.1** (Modular value-loss upper bound)**.** *Assume* (17)–(18) *and that $Q^\pi$ is $L_Q$-Lipschitz in $(s,a)$ on the relevant compact domain. Then, for any (possibly randomized) online estimator $\mathcal{E}$ producing $\tilde{s}_t$,*

$$\big|V^{\pi\circ\mathcal{E}}(s_0) - V^\pi(s_0)\big| \;\le\; \frac{L_Q(1+L_\pi)}{1-\gamma}\,\sup_{t\ge 0}\mathbb{E}\|\tilde{s}_t - s_t\|. \tag{21}$$

*In particular, under the additive-offset model $o_t = s_t + b$ we have $\tilde{s}_t - s_t = b - \hat{b}_t$ and the right-hand side depends only on the offset-estimation error.*

*Proof sketch.* We write the value difference as a discounted sum of per-step $Q^\pi$ differences under a coupling that aligns the latent trajectory while comparing actions sampled from $\pi(\cdot \mid \tilde{s}_t)$ and $\pi(\cdot \mid s_t)$. Applying (19)–(20) at each time step yields

$$\Big|\mathbb{E}\big[Q^\pi(\tilde{s}_t, a_t) - Q^\pi(s_t, a_t^\star)\big]\Big| \;\le\; L_Q(1 + L_\pi)\,\mathbb{E}\|\tilde{s}_t - s_t\|,$$

and summing with weights $\gamma^t$ gives (21) by the geometric series bound $\sum_{t\ge 0}\gamma^t \le (1-\gamma)^{-1}$. □

**Instantiation for Bayes-FORL and separation of error sources.** Theorem 7.1 reduces control to filtering. Combining (21) with the estimation statements from Section 6 yields an immediate decomposition: whenever Bayes-FORL operates in the "well-specified support" regime (e.g. $b \in \mathcal{B}$ under conformal truncation), we may substitute a bound of the form

$$\sup_t \mathbb{E}\|\tilde{s}_t - s_t\| \;\lesssim\; \varepsilon_F + \varepsilon_G + \varepsilon_{\inf}(M) + \sup_t \mathrm{Stat}_t,$$

and thus obtain an explicit end-to-end guarantee on value loss whose dependence on forecast misspecification, surrogate error, and finite-$M$ inference is additive up to constants. Importantly, no property of $\pi$ beyond Lipschitz regularity enters; in particular, we do not require $\pi$ to be optimal, nor do we require any special structure of the MDP beyond existence and regularity of $Q^\pi$ on the visited domain.

**Matching lower bound: tightness of the Lipschitz dependence.** We now record that the linear dependence on $\sup_t \mathbb{E}\|\tilde{s}_t - s_t\|$ and the factor $(1-\gamma)^{-1}$ are unavoidable without additional structure. The proof proceeds by constructing an MDP family in which rewards are directly sensitive to one coordinate of the state, and the frozen policy $\pi$ maps that coordinate (Lipschitzly) into actions which linearly control reward. In such a case, any estimator error induces an essentially proportional reward shortfall.

**Theorem 7.2** (Matching lower bound)**.** *Fix $\gamma \in (0,1)$ and any estimator $\mathcal{E}$ that produces actions by sampling $a_t \sim \pi(\cdot \mid \tilde{s}_t)$. There exists a family of stationary deterministic MDPs with additive observation offsets and a policy $\pi$ such that $Q^\pi$ is $L$-Lipschitz and, for an absolute constant $c > 0$,*

$$V^\pi(s_0) - V^{\pi\circ\mathcal{E}}(s_0) \;\ge\; c\,\frac{L}{1-\gamma}\,\sup_{t\ge 0}\mathbb{E}\|\tilde{s}_t - s_t\|. \tag{22}$$

*Construction sketch.* We embed a one-step decision problem into a discounted MDP by making the next state absorbing. Let $s \in [-1, 1]$ be scalar and let the reward be $r(s, a) = -|a - s|$, with dynamics $s_{t+1} = s_t$ (absorbing). For a fixed stochastic policy $\pi(\cdot \mid s)$ that concentrates around $a = s$ and is Lipschitz in $s$ (e.g. a narrow Gaussian centered at $s$), the induced $Q^\pi$ is Lipschitz in $s$ and $a$. If the deployed agent acts on $\tilde{s}$ rather than $s$, then $\mathbb{E}[|a - s|]$ increases by a constant fraction of $|\tilde{s} - s|$ under mild regularity of $\pi$, yielding a one-step value gap proportional to $|\tilde{s} - s|$; the absorbing structure amplifies it by $(1 - \gamma)^{-1}$. □

**Discussion: what would be required to beat the bound.** Theorems 7.1–7.2 identify the estimator error as the unique lever in the frozen-policy regime and show that Lipschitz regularity alone cannot yield sublinear dependence on $\|\tilde{s}_t - s_t\|$. Any improvement must therefore exploit additional structure beyond generic Lipschitzness, such as (i) flat directions of $Q^\pi$ aligned with offset ambiguity, (ii) robust policies whose action distribution is locally invariant to certain state perturbations, or (iii) richer observation models that make $b$ identifiable faster (reducing $\sup_t \mathbb{E}\|\tilde{s}_t - s_t\|$). In the absence of such structure, the appropriate goal is not to seek a sharper control inequality, but to engineer an estimator that minimizes the relevant error summaries (including tail-risk versions) under the forecast and computational constraints.

## 8 Complexity and Practical Considerations

We collect here the implementation-level considerations that determine whether the Bayes-FORL filter is a viable deploy-time module under bounded compute. Since the policy $\pi$ is frozen, the only online degrees of freedom are (i) the number of particles $M$, (ii) the manner in which we maintain and update the episode-level belief over $b_j$, and (iii) simple numerical safeguards that prevent weight collapse and out-of-support behavior.

**Online time and space complexity.** At time $t$, Bayes-FORL performs $M$ evaluations of the energy surrogate $E_\theta(s, \tau_{t,w})$ at the candidate states $s = o_t - b^{(m)}$. Writing $C_E$ for the cost of a single energy evaluation and $C_\pi$ for sampling from $\pi(\cdot \mid \tilde{s}_t)$, the per-step time is

$$O\big(MC_E + C_\pi + C_{\text{buf}}\big),$$

where $C_{\text{buf}}$ is the constant-time update for the history buffer $\tau_{t,w}$ (typically $O(w)$ but implemented as a ring buffer so that the amortized per-step cost is constant). Resampling, when triggered, adds an $O(M)$ step. Over a horizon $T$, the per-episode complexity is thus $O(TMC_E)$, and the memory footprint is

$$O(Mn) \quad \text{for the particles } \{b^{(m)}\}_{m=1}^{M} \subset \mathbb{R}^n, \qquad \text{plus} \qquad O\big(w(n+|a|)\big) \text{ for } \tau_{t,w}.$$

We stress that this complexity is independent of any planning step: the control component remains a single policy query at $\tilde{s}_t$.

**Numerical stability: log-weights and effective sample size.** Since the posterior weights are multiplicative over time, naive weight updates underflow quickly. We therefore maintain log-weights

$$\log \bar{w}_t^{(m)} \;=\; \log \bar{w}_{t-1}^{(m)} \;-\; E_\theta(o_t - b^{(m)}, \tau_{t,w}),$$

followed by normalization via a log-sum-exp. A standard diagnostic for degeneracy is the effective sample size

$$\mathrm{ESS}_t \;:=\; \frac{\big(\sum_{m=1}^{M} w_t^{(m)}\big)^2}{\sum_{m=1}^{M} (w_t^{(m)})^2} \;\in [1, M],$$

and we trigger resampling when $\mathrm{ESS}_t/M \;\leq\; \alpha$ for some $\alpha \in (0,1)$ (e.g. $\alpha = 0.3$). In an episodic constant-offset model, resampling is typically sufficient; however, when the posterior is multimodal, resampling alone can prematurely eliminate viable modes. We return to this failure mode below.

**Amortization and batching.** Although the update is conceptually sequential, it is vectorizable: the $M$ candidate states $s_t^{(m)} = o_t - b^{(m)}$ can be stacked and passed through $E_\theta$ in a single batched forward call. This typically yields near-linear speedups on accelerators, and it also simplifies memory locality on CPU. When $\tau_{t,w}$ is represented by a fixed-size tensor (deltas and actions), the energy network can be structured to reuse an embedding of $\tau_{t,w}$ across all $m$ at time $t$, so that only the $s$-dependent path is replicated over particles. Concretely, if $E_\theta(s, \tau)$ factors as $g_\theta(h_\theta(\tau), s)$, we compute $z_t = h_\theta(\tau_{t,w})$ once per step and evaluate $g_\theta(z_t, s_t^{(m)})$ for all $m$. This reduces the effective constant in $MC_E$ without changing the asymptotic bound.

**Calibration and conformal truncation costs.** If we employ conformal sets $\mathcal{B}_j$ to guarantee $\mathbb{P}(b_j \in \mathcal{B}_j) \geq 1 - \delta$, the cost is paid primarily offline. In the common split-conformal pattern, we compute nonconformity scores on a calibration log of past offsets and forecaster outputs, then select a quantile threshold. For multivariate $b$, a practical choice is a Mahalanobis-type score $\alpha(b) = \|(b - \mu)/\Sigma^{1/2}\|$, where $(\mu, \Sigma)$ are taken from the forecaster or estimated from residuals; the only nontrivial offline expense is a covariance estimation/inversion, which is negligible relative to training $E_\theta$. Online, conformal truncation is essentially free: we either (i) reject-sample particles until $b^{(m)} \in \mathcal{B}_j$ (acceptable when $\mathbb{P}_{q_j}(b \in \mathcal{B}_j)$ is not too small), or (ii) sample once and clamp/transport particles to the set by a deterministic projection (useful for ellipsoidal $\mathcal{B}_j$). We emphasize that conformal calibration does not correct a wrong energy model; rather, it bounds prior support error by ensuring the true $b_j$ lies in a region we do not discard with high probability.

**Recommended engineering choices (default settings).** In deployments where $n$ is moderate and $E_\theta$ is a neural surrogate, we have found the following choices to be robust:

- *Use $\Delta o_t$ in $\tau_{t,w}$ whenever possible.* Under constant $b_j$, deltas remove the offset and supply identifiability signal that is invariant to $b_j$; this reduces the burden on the forecaster and improves the conditioning of the posterior.
- *Prefer posterior mean for control, MAP for diagnostics.* The posterior mean $\hat{b}_t = \sum_m w_t^{(m)} b^{(m)}$ is stable when the posterior is unimodal, while the MAP particle is useful for detecting multimodality (large mean–MAP discrepancy).
- *Use tempering when the energy scale is uncertain.* If the magnitude of $E_\theta$ is miscalibrated, weights may collapse. A simple remedy is to introduce an inverse-temperature $\beta \in (0, 1]$ and update with $\exp(-\beta E_\theta)$; $\beta$ can be chosen by maintaining a target ESS range.
- *Monitor* $\mathrm{ESS}_t$ *and posterior dispersion.* Low ESS, rapidly decreasing posterior variance, or abrupt shifts in $\hat{b}_t$ are all actionable signals for triggering resampling, tempering, or a fallback estimator.

**Failure modes and mitigations.** We distinguish four common pathologies.

1. *Forecast misspecification (support failure).* If $q_j$ assigns negligible mass near the true $b_j$, importance sampling fails regardless of $M$. Conformal truncation addresses the opposite error (over-pruning) but does not create missing mass. A practical mitigation is to use heavy-tailed priors (e.g. Gaussian scale mixtures) or to mix the forecaster with a broad "safety" component $q_j^{\text{mix}} = (1 - \lambda)q_j + \lambda q_0$.
2. *Multimodal or symmetric posteriors over $b$.* When the likelihood induced by $E_\theta$ is approximately symmetric in $b$, the posterior may have separated modes. Particle resampling can then lock onto an arbitrary mode. Remedies include stratified initialization (draw from a mixture covering plausible modes), rejuvenation steps (e.g. a small Gaussian

random-walk move on $b$ followed by Metropolis correction using the energy surrogate), or delayed commitment (control on a mixture-averaged $\tilde{s}_t$ while deferring mode selection until sufficient evidence accumulates).

3. *Energy surrogate error out of distribution.* If $o_t - b^{(m)}$ exits the compact domain on which $E_\theta$ approximates the negative log-belief, the filter can assign spurious likelihood. This is best handled by explicit domain checks (clipping to a trusted set, or inflating energy outside-range), and by training-time augmentation over plausible offsets so that $E_\theta$ is well-behaved on the states induced by $o_t - b$ with $b \in \mathcal{B}_j$.

4. *Particle impoverishment at long horizons.* Even with correct models, repeated resampling without rejuvenation reduces particle diversity. Since $b_j$ is constant, rejuvenation can be implemented cheaply (a few move steps every $K$ time steps) without changing the per-step asymptotics, and often stabilizes late-episode behavior.

**Compute–risk tradeoffs and choosing $M$.** Theorems controlling $\varepsilon_{\text{inf}}(M)$ suggest the usual $M^{-1/2}$ improvement, but in practice the relevant criterion is avoiding catastrophic weight collapse in the early time steps, since control errors compound through the discount factor only linearly. We therefore recommend sizing $M$ by a stress-test that measures the worst-case (over episodes) early-time ESS decay and the tail of $\|\hat{b}_t - b_j\|$, rather than by average loss. When compute is severely constrained, a hybrid approach is viable: run Bayes-FORL with small $M$ to obtain a coarse $\hat{b}_t$, and switch to a deterministic local optimizer for $\Phi_t(b)$ (warm-started at $\hat{b}_t$) once the posterior is evidently unimodal, thereby replacing the $O(M)$ particle update with a small fixed number of gradient steps.

**Summary.** Bayes-FORL is designed so that all substantial costs are shifted offline (training $E_\theta$ and, optionally, calibrating $\mathcal{B}_j$), while the online loop consists of a batched energy evaluation and a light-weight SMC update. The dominant practical risks are not asymptotic, but rather mode ambiguity and support mismatch; these are addressable by conservative prior mixing, ESS-driven tempering/resampling, and lightweight rejuvenation. This positions the method for an empirical evaluation focused on correlated offsets and tail-risk behavior, which we specify next.

# 9 Experimental Protocol (Recommended)

We recommend an evaluation protocol whose purpose is not merely to improve mean return, but to (i) expose the failure modes predicted by the modular analysis (support mismatch, multimodality, finite-$M$ collapse), and

(ii) measure *tail* state-estimation risk, since the value gap bound in Theorem 4 is controlled by worst-case (or supremal) state error rather than an average-case proxy.

**Training–deployment split and ground-truth access.** We assume access to an offline dataset $D$ collected in $M_{\text{train}}$ with fully observed states. All components are trained offline: (i) the frozen policy $\pi$ (any offline RL method), and (ii) the energy surrogate $E_\theta(s, \tau_{t,w})$ trained to approximate $-\log p(s_t \mid \tau_{t,w})$ on the training distribution induced by $D$ (optionally augmented; cf. Section 8). The forecaster that produces $q_j(b)$ is trained on a log of offsets from past episodes (or a simulator-generated offset process); crucially, at *deployment-time evaluation* we do not reveal $b_j$ to the estimator, but we do record it for metrics.

**Correlated-offset benchmark design.** To test the specific advantage of a joint belief over $b_j \in \mathbb{R}^n$, we recommend benchmarks in which $b_j$ has *nontrivial cross-dimensional correlation* and exhibits across-episode temporal structure. A simple, controllable family is

$$b_j \;=\; \mu \;+\; A(b_{j-1} - \mu) \;+\; \xi_j, \qquad \xi_j \sim \mathcal{N}(0, \Sigma), \tag{23}$$

with $A$ chosen to be stable (e.g. $A = \rho I$ for $\rho \in (0, 1)$, or a low-rank perturbation) and with $\Sigma$ dense. To generate *structured* correlations at scale, we also recommend a factor model

$$b_j \;=\; \mu \;+\; W z_j \;+\; \epsilon_j, \qquad z_j \sim \mathcal{N}(0, I_k),\; \epsilon_j \sim \mathcal{N}(0, \sigma^2 I_n), \tag{24}$$

where $k \ll n$ controls the intrinsic correlation dimension. One may combine (24) with an AR(1) process on $z_j$ to induce persistent episode-to-episode drift while keeping $b_j$ constant *within* each episode. This benchmark directly probes whether an estimator that treats coordinates independently can recover offsets that lie near a low-dimensional correlated manifold.

We propose sweeping difficulty along at least three axes: (a) correlation strength (via $\|W\|$ or off-diagonal mass in $\Sigma$), (b) forecastability (via $\rho$ in (23) or SNR in (24)), and (c) out-of-support rate (by occasionally sampling $b_j$ from a broader "shock" distribution, e.g. a mixture with a heavy-tailed component). The latter is necessary to evaluate support failure and the practical value of mixing priors.

**Forecaster baselines (univariate vs. multivariate).** Because Bayes-FORL is modular in $q_j$, we recommend two forecaster classes:

1. *Univariate forecaster:* fit $n$ independent predictors to each coordinate, yielding a factorized prior $q_j^{\text{uni}}(b) = \prod_{i=1}^{n} q_{j,i}(b_i)$ (e.g. ARIMA per coordinate, or per-dimension quantile regression).

2. *Multivariate forecaster:* fit a joint model (e.g. VAR, low-rank state-space, or a neural multivariate probabilistic model) to obtain a full-covariance prior $q_j^{\mathrm{multi}}(b)$.

The comparison isolates whether joint correlation modeling matters *even when the filter itself is unchanged.* For both classes, we recommend reporting a scalar mis-specification diagnostic on held-out offsets, such as negative log-likelihood or calibration error for marginal quantiles, but we stress that these do not replace control-centric evaluation.

**Estimator baselines: DCM vs. Bayes-FORL (and oracles).** We recommend at least the following estimators: (i) *No correction:* $\tilde{s}_t = o_t$. (ii) *Oracle offset:* $\tilde{s}_t = s_t$ or $\tilde{s}_t = o_t - b_j$ (upper bound). (iii) *DCM (dimension-wise correction):* any estimator that forms $\hat{b}_t$ by coordinate-wise fusion of independent beliefs or scores (the precise implementation may vary, but the defining restriction is per-coordinate updating without maintaining a joint posterior over $\mathbb{R}^n$). (iv) *Bayes-FORL:* the joint particle filter described earlier, with fixed compute budget $M$ and optional conformal truncation to $\mathcal{B}_j$. To ensure fairness, we recommend matching wall-clock or matching the number of $E_\theta$ evaluations per step across methods; for DCM variants, this typically means equating total forward passes through $E_\theta$.

**Tail-risk metrics for state estimation.** Since $b_j$ is constant within an episode, state-estimation error and offset-estimation error coincide: $\tilde{s}_t - s_t = -(\hat{b}_t - b_j)$. We recommend reporting both pointwise and tail metrics:

$$e_{j,t} := \|\tilde{s}_{j,t} - s_{j,t}\|, \tag{25}$$

$$e_j^{\max} := \max_{0 \le t \le T} e_{j,t}, \tag{26}$$

$$\mathrm{CVaR}_\alpha(e^{\max}) := \inf_{u \in \mathbb{R}} \left\{ u + \frac{1}{1-\alpha} \mathbb{E}\big[(e^{\max} - u)_+\big] \right\}, \tag{27}$$

with $\alpha \in \{0.9, 0.95\}$. We additionally recommend a discounted error proxy aligned with value bounds,

$$e_j^\gamma \;:=\; (1-\gamma) \sum_{t=0}^{T} \gamma^t e_{j,t}, \tag{28}$$

but we treat $e_j^{\max}$ and $\mathrm{CVaR}_\alpha$ as primary, since they are most sensitive to early catastrophic mistakes and mode-locking.

**Control metrics.** We recommend reporting (i) mean return $\mathbb{E}[\sum_{t=0}^{T} \gamma^t r_t]$, (ii) return tail risk (e.g. $\mathrm{CVaR}_\alpha$ of negative return), and (iii) the value gap to oracle,

$$\Delta V \;:=\; V^{\pi \circ \mathcal{E}} - V^{\pi \circ \mathrm{oracle}}. \tag{29}$$

When environment stochasticity is substantial, we recommend multiple rollouts per episode-offset pair to separate offset-induced error from transition noise.

**Calibration ablations (conformal on/off).** To evaluate the effect of conformal truncation, we recommend three conditions: (a) no truncation (particles from $q_j$), (b) truncation to $\mathcal{B}_j$ at level $1-\delta$, and (c) an intentionally mis-set $\delta$ (e.g. overly aggressive truncation) to show the support-risk tradeoff. We recommend reporting empirical coverage $\widehat{\mathbb{P}}(b_j \in \mathcal{B}_j)$ and relating failures to spikes in $e_j^{\max}$.

**Protocol-level ablations.** Finally, we recommend ablations that isolate the sources of robustness: (i) history choice (using $\Delta o_t$ versus raw $o_t$ in $\tau_{t,w}$), (ii) window size $w$, (iii) particle count $M$ and resampling/tempering on/off, and (iv) prior mixing $q_j^{\text{mix}} = (1-\lambda)q_j + \lambda q_0$ with a broad $q_0$. The outcome of these ablations should be summarized not only by means but by *rankings under tail criteria* (e.g. which method minimizes $\text{CVaR}_{0.95}(e^{\max})$), since this is the regime in which correlated offsets and multimodality most clearly separate joint from coordinate-wise approaches.

# 10 Discussion and Limitations

**Dependence on likelihood-surrogate quality.** Our online inference procedure reduces, in effect, to importance-weighting candidate offsets by the surrogate likelihood contributions

$$\ell_t(b) \;\propto\; \exp\bigl(-E_\theta(o_t - b, \tau_{t,w})\bigr), \qquad p_\theta(b \mid o_{0:t}, a_{0:t-1}) \;\propto\; q_j(b) \prod_{t' \le t} \ell_{t'}(b).$$

Consequently, deployment performance hinges on whether $E_\theta(\cdot,\cdot)$ assigns *relative* energies that are faithful to the true conditional density of $s_t$ given history. The uniform approximation condition $\|E_\theta - E^*\|_\infty \le \varepsilon_G$ is convenient for analysis but does not by itself preclude harmful pathologies: (i) localized regions where the surrogate is overly sharp, causing weight collapse and premature mode-locking, and (ii) regions where the surrogate is too flat, producing poor discrimination among offsets and slow concentration. Both effects can occur even when average predictive metrics (e.g. held-out negative log-likelihood) appear satisfactory, because the filter compounds errors multiplicatively over time.

A practical corollary is that one should treat the *dynamic range* of $E_\theta$ as a first-class object: if $E_\theta$ is miscalibrated by a multiplicative temperature, then $\prod_t \exp(-E_\theta)$ may concentrate too quickly (or not at all). While we have presented conformal truncation and prior mixing as safeguards on the *support* of $b$, they do not correct a systematically misshapen surrogate. Designing calibration procedures that preserve the product structure while controlling effective sample size (ESS) remains an important implementation detail; our theoretical bounds implicitly assume that the induced importance weights remain within a manageable range.

**Multimodality, symmetry, and non-identifiability.** Theorem-style guarantees for offset recovery require some form of local curvature or mode separation of the (negative) log-posterior

$$\Phi_t(b) \;:=\; -\log q_j(b) \;+\; \sum_{t' \le t} E_\theta(o_{t'} - b, \tau_{t',w}).$$

This is unavoidable: if $\Phi_t$ has multiple well-separated minima with comparable values, then neither MAP estimation nor finite-$M$ sampling can be expected to recover the true $b_j$ reliably without additional information. In our setting, multimodality may arise from at least two sources.

First, the belief model itself may be ambiguous: if the learned conditional $p_\theta(s_t \mid \tau_{t,w})$ is multimodal (e.g. due to partial observability induced by using only a short window), then $E_\theta$ may exhibit multiple low-energy regions corresponding to distinct plausible states, which translate into multiple plausible offsets $b = o_t - s$. Second, and more fundamentally, the environment may possess *symmetries* that render offsets intrinsically unidentifiable. For instance, if the true dynamics and rewards are invariant under translations along some coordinates (or approximately so on the relevant domain), then distinct offsets can generate indistinguishable trajectories under the frozen policy, implying that no estimator can concentrate beyond the symmetry class. In such cases, one should expect persistent posterior uncertainty over $b_j$ and therefore a non-vanishing value gap whenever $\pi$ is sensitive to those coordinates (cf. the Lipschitz lower bound phenomenon).

These observations suggest two diagnostics we regard as essential: (i) explicit checks for symmetry-induced flat directions by probing whether $E_\theta(s, \tau)$ changes materially under candidate shifts in $s$, and (ii) stress tests in which the forecaster prior is intentionally broadened to reveal whether the filter remains stable under genuine ambiguity or instead collapses arbitrarily due to finite-$M$ effects. Put differently, multimodality is not merely a sampling nuisance; it is an identifiability question that directly controls tail risk.

**Finite-$M$ collapse and the limits of importance sampling.** The particle implementation inherits the standard failure modes of self-normalized importance sampling: weight degeneracy, sensitivity to proposal mismatch, and path dependence through resampling. In our episodic-constant setting, the weights are updated repeatedly while particle locations remain fixed, which can lead to rapid ESS decay when the accumulated likelihood ratio is large. While resampling stabilizes numerical behavior, it can also irreversibly discard minority modes that later become favored as new evidence arrives. Remedies such as tempering (using $\exp(-\beta E_\theta)$ with $\beta \uparrow 1$), rejuvenation moves (e.g. local MCMC steps on $b$), or maintaining a mixture of proposals are compatible with our modular framing, but they increase online compute and complicate guarantees. Our present analysis captures finite-$M$ error via coarse concentration terms; obtaining sharp, pathwise tail bounds under resampling remains open.

**Extensions beyond additive offsets.** A natural extension replaces the observation model $o_t = s_t + b_j$ by an episode-specific affine transform,

$$o_t \;=\; A_j s_t + b_j, \tag{30}$$

with $A_j \in \mathbb{R}^{n\times n}$ unknown (or structured, e.g. diagonal gains) and $b_j \in \mathbb{R}^n$. In principle, our approach extends by treating $\theta_j := (A_j, b_j)$ as the episode-constant latent and weighting particles using $s_t^{(m)} = A^{(m)-1}(o_t - b^{(m)})$ (when invertible), or more generally by scoring candidate $(A, b)$ through $E_\theta(\cdot, \tau)$. However, identifiability becomes substantially more delicate: even in deterministic settings, $(A, b)$ may not be recoverable without excitation, and the strong convexity conditions required for concentration may fail unless we impose structure (e.g. $A_j$ near identity, low-dimensional parametrization, or priors that exclude degenerate transformations). Moreover, errors in $A_j$ couple multiplicatively into state error, potentially amplifying value loss relative to the additive case. We therefore view (30) as feasible but requiring explicit structural assumptions and careful experimental validation.

**Anchor observations and intermittent ground truth.** Many applications provide occasional "anchors": absolute-position fixes, trusted proprioceptive measurements, or external calibration signals that partially reveal $s_t$ (or $b_j$) at a sparse set of times. Such information can be incorporated by adding anchor likelihood terms to $\Phi_t(b)$, equivalently multiplying the particle weights by an additional factor at anchor steps. The resulting filter can break symmetries and collapse multimodality dramatically, but it also highlights a limitation of our current exposition: we have not quantified how anchor frequency and noise interact with the energy-surrogate error $\varepsilon_G$ and the forecast mis-specification level $\varepsilon_F$. Deriving explicit tradeoffs between anchor rate, particle budget $M$, and tail estimation risk is a concrete direction for strengthening the framework.

**Open problems.** We close by isolating several questions that remain unresolved in our modular treatment. (i) *Training for filtering robustness:* how should one train $E_\theta$ so that the induced product-of-experts posterior is well-behaved under sequential multiplication, rather than merely accurate in one-step prediction? (ii) *Non-asymptotic tail guarantees:* can one obtain high-probability bounds on $\max_t \|\hat{b}_t - b_j\|$ under realistic resampling schemes and mild mixing assumptions? (iii) *Diagnosing and handling symmetries:* can one algorithmically detect non-identifiable subspaces online and report calibrated uncertainty in a way that correlates with control risk? (iv) *Beyond constant offsets:* slowly drifting within-episode biases (e.g. $b_{t+1} = b_t + \zeta_t$) interpolate between our episodic model and general POMDP filtering; understanding the minimal modifications that preserve tractability, while retaining conformal safeguards, is an important step toward broader applicability.