Likelihood-Free Shifts-Aware Reward Learning for Implicit World Models via f-Divergence Duality

Liz Lemma Future Detective

January 20, 2026

Abstract

Model-based offline RL mitigates data scarcity by training policies in learned dynamics, but suffers from distribution shift due to model bias and policy shift. Recent work (e.g., SAR/SAMBO) corrects this shift with reward terms involving log-likelihood ratios $\log \frac{p}{q}$ and $\log \frac{\pi}{\pi_c}$, estimated by classifiers. In 2026-era pipelines, however, the best world models are often implicit (diffusion, autoregressive latent dynamics) where calibrated transition likelihoods are unavailable or meaningless. We propose a likelihood-free generalization of shifts-aware rewards that replaces log ratios with critic scores obtained from variational f-divergence duals in a learned latent space. The result is a discriminator-only correction that can be applied to diffusion/transformer world models, stabilizing imagination rollouts and policy updates without requiring tractable densities. We formalize a clean offline setting with implicit models and provide (i) a variational surrogate objective that recovers SAR when the divergence is KL and critics are optimal, (ii) finite-sample performance bounds that explicitly separate representation mismatch, critic duality gaps, and rollout-horizon compounding, and (iii) lower bounds showing that without overlap, no likelihood-free correction can guarantee improvement. Experiments (recommended) should validate robustness on narrow-coverage benchmarks (NeoRL) and new OOD model-mismatch suites targeting implicit world models.

Table of Contents

3. 3. Problem setup: offline dataset + implicit dynamics sampler; mixed real/model training distribution; formalize model bias and policy shift in trajectory weighting; define representation-space setting.

4. 4. Likelihood-free shift-aware rewards: derive corrected reward from f-divergence variational duals for transition shift and action/policy shift; specialize to KL to connect to SAR; discuss reward positivity/log transform and utility alternatives.

5. 5. Algorithm (LF-SAR): training loop with implicit world model rollouts, latent encoder, transition critic, action critic; stabilization (spectral norm, gradient penalties, logit clipping/calibration); practical notes for diffusion/transformer models.

6. 6. Theory I (surrogate objective): prove an ELBO-like lower bound on a utility of return under optimal critics; quantify degradation under critic duality gaps; horizon dependence.

7. 7. Theory II (policy performance): translate surrogate optimization error into environment return suboptimality using occupancy measures/PEVI-style arguments in representation space; explicit additive bound in terms of divergence, critic error, representation error, and rollout length.

8. 8. Lower bounds / impossibility: prove that without overlap (absolute continuity), any offline algorithm (including likelihood-free SAR) can fail; sample-complexity lower bounds for ratio estimation imply unavoidable value error.

9. 9. Experiments (recommended for full paper): NeoRL + D4RL; implicit world model ablations; OOD mismatch suite; calibration and stability diagnostics; compare to SAMBO/MOPO/MOBILE and to likelihood-based discriminator baselines.

10. 10. Discussion and limitations: when representation-space divergences are meaningful; dependence on encoder; negative/zero rewards; long-horizon imagination; broader implications for foundation world models.

1 Introduction

Offline reinforcement learning aims to synthesize a policy using only a fixed dataset of transitions collected by an unknown behavior policy, without further interaction with the environment. In the model-based variant, we additionally fit a dynamics model from the same dataset and then optimize a policy by rolling out the learned model. This paradigm promises improved generalization and sample efficiency, but it is constrained by a central obstruction: the optimization procedure induces *distribution shift* relative to the data-generating distribution. In offline settings, this shift is not merely a statistical nuisance; it can create systematic overestimation and compounding error because the policy is optimized on regions where neither the dataset nor the learned model is reliable.

It is useful to separate two distinct sources of shift. First, *dynamics shift* occurs because the learned dynamics model does not coincide with the true environment kernel. Even when the model is accurate on the dataset support, model rollouts under the learned policy can quickly drift toward states and transitions that were rarely observed, at which point multi-step prediction errors compound over the effective horizon. Second, *policy shift* arises because the optimized policy may differ substantially from the unknown behavior policy that generated the dataset. This mismatch can cause the value estimates to be extrapolated to actions not represented in the data, which is known to produce brittle behavior in purely offline control. Model-based offline methods thus face a coupled failure mode: policy optimization pushes toward actions that look beneficial under the model, and the resulting state-action distribution is precisely the one for which the model is least trustworthy.

A principled family of approaches addresses these shifts by reweighting or correcting objectives using likelihood ratios, often derived from a change-of-measure argument along trajectories. Representative instances include shift-aware reward (or advantage) correction schemes that use terms of the form $\log \frac{p(s'|s,a)}{m(s'|s,a)}$ to penalize model transitions that are implausible under the environment, as well as policy regularization terms such as $\log \frac{\pi_b(a|s)}{\pi(a|s)}$ to discourage deviation from the behavior distribution. Such corrections may be understood as constructing an evidence lower bound (ELBO)-like surrogate objective, whose maximization yields a conservative policy update when the corrections are accurate. When both the environment likelihood and the model likelihood are tractable (or estimable in a calibrated way), these likelihood-based shift corrections admit clean interpretations and, in some regimes, provable safety guarantees.

However, modern world models used in high-dimensional domains are increasingly *implicit*. Diffusion models, autoregressive transformers with stochastic decoding, and other implicit generative mechanisms can provide

high-quality samples $s' \sim m_\theta(\cdot \mid s, a)$, but typically do not provide a normalized and tractable density $m_\theta(s' \mid s, a)$, nor a computable $\log m_\theta(s' \mid s, a)$. In continuous state spaces, even models that define densities in principle may make likelihood evaluation prohibitively expensive or numerically unstable. Consequently, likelihood-based shift-aware reward correction is not directly applicable precisely in the regimes where model-based planning and rollouts are most attractive.

This work develops a likelihood-free alternative that preserves the conceptual structure of shift-aware correction while requiring only sampling access to the learned dynamics. The key observation is that the corrections needed for conservatism can be formulated in terms of *distributional discrepancy* between environment transitions and model transitions, as well as between behavior actions and policy actions. Such discrepancies can be measured by an f-divergence, and crucially, f-divergences admit variational dual representations that depend only on expectations under the relevant distributions. Expectations are estimable from samples, hence they remain available in the implicit-model setting. We thus replace explicit log-likelihood ratios by learned *witness functions* (critics) obtained from a discriminator-style training objective.

Concretely, we operate in a representation space $z = f(s)$, both for statistical efficiency and to express sufficiency assumptions that are natural in high-dimensional observation domains. We consider the pushforward transition distributions induced by the environment and the model, namely $p_f(z' \mid s, a)$ and $m_{\theta,f}(z' \mid s, a)$. We train a *transition critic* $T_\phi(z, a, z')$ via the dual objective of a chosen f-divergence, using samples (z, a, z') derived from real transitions in the offline dataset and from model rollouts. In parallel, we train an *action critic* $U_\psi(z, a)$ to detect policy shift by contrasting action-state pairs generated by the current policy with those present in the dataset. The outputs of these critics are then used as additive corrections to a base reward, yielding a corrected reward $\tilde{r}_{\phi,\psi}$ that can be optimized by any convergent off-policy algorithm on a mixture of real and model-generated transitions.

Our development is guided by two requirements. First, the method must be *likelihood-free*: all learning signals for the corrections must be computable from samples, not from explicit densities. Second, the method must permit *end-to-end performance control*: we wish to relate the true environment performance of the final policy to quantities that measure model mismatch and critic error, and we wish to expose the dependence on the effective horizon $H \asymp (1-\gamma)^{-1}$. The latter is essential because multi-step compounding is the primary failure mode in model-based optimization, and because offline learning without overlap is information-theoretically impossible.

The contributions are as follows.

- We formulate a shift-aware corrected objective for model-based offline

RL with implicit dynamics models, using variational f-divergence critics in representation space to replace likelihood ratios. The resulting algorithm requires only (i) i.i.d. minibatches from the offline dataset and (ii) sampling access to the learned model.

- We show that, under standard overlap assumptions (absolute continuity of the relevant pushforward measures) and bounded critic duality gaps, the learned transition critic provides a principled proxy for likelihood-based dynamics correction. For KL in particular, the optimal critic recovers the familiar log-density ratio up to an additive constant, while our implementation remains valid without computing any likelihoods.

- We establish a surrogate lower bound relating the corrected objective to a utility-transformed version of the true return. For log-reward variants (requiring $r_{\min} > 0$), this yields a direct lower bound on $\log J_{\mathcal{M}}(\pi)$ up to additive terms controlled by critic error and representation sufficiency.

- We provide an end-to-end additive performance bound for the learned policy $\hat{\pi}$, with explicit dependence on (i) representation-space model mismatch measured by an f-divergence, (ii) policy shift relative to the dataset, (iii) critic duality gaps and representation error, and (iv) optimization error of the policy improvement routine. The bound scales linearly with the effective horizon, matching known lower-bound phenomena in offline RL.

The central message is that likelihood-based shift-aware correction is not intrinsically tied to tractable densities; rather, it is tied to *distinguishability* between distributions. By translating the correction terms into variational divergences, we obtain a method that is compatible with implicit world models while retaining a conservative interpretation: the corrected reward penalizes transitions and actions that are distinguishable as "out-of-distribution" relative to the offline dataset. The theoretical development makes explicit which assumptions are needed for this interpretation to hold—notably overlap in the representation space and critic learnability—and quantifies how violations or approximation errors propagate into value suboptimality.

Finally, we emphasize the scope of what can and cannot be guaranteed. Our results do not circumvent the impossibility of offline RL without coverage: if the optimal policy relies on actions or transitions unsupported by the dataset (or, here, unsupported by model samples in the relevant representation), then no algorithm can provide a uniform performance guarantee. Instead, our framework clarifies the precise role of overlap and supplies a likelihood-free mechanism for enforcing conservatism in the presence of both model bias and policy shift, thereby aligning modern implicit generative modeling with the requirements of reliable offline control.

2 Background and related work

Likelihood-based shift-aware correction. A line of work formalizes model-based policy optimization as a change-of-measure problem along trajectories, leading to objectives that incorporate explicit likelihood ratios between the environment dynamics and the learned model, and between the behavior policy and the learned policy. In its simplest form, one obtains trajectory weights of the type

$$w(\tau) \;\propto\; \prod_{t\geq 0} \frac{p(s_{t+1} \mid s_t, a_t)}{m(s_{t+1} \mid s_t, a_t)} \cdot \frac{\pi_b(a_t \mid s_t)}{\pi(a_t \mid s_t)},$$

or additive reward corrections involving $\log \frac{p}{m}$ and $\log \frac{\pi_b}{\pi}$, which yield an ELBO-like conservative surrogate when combined with concave utility transforms (notably the log transform when $r_{\min} > 0$). Shift-Aware Reward (SAR) and closely related formulations (including SAMBO and variants thereof) make this structure explicit and provide guarantees under overlap and boundedness conditions; the key technical ingredient is that, for KL-based corrections, the optimal witness function is a log density ratio and Jensen-type arguments convert likelihood ratios into lower bounds on the desired return or on a monotone transform thereof (**??**). Our contribution may be viewed as retaining this conservative change-of-measure template while removing the assumption that $m(\cdot \mid s, a)$ admits tractable likelihoods.

Off-dynamics reward correction (DARC). A complementary approach corrects for dynamics mismatch by modifying rewards to counteract discrepancies between the transition distributions induced by two Markov kernels. Off-Dynamics RL and DARC-like methods estimate a correction term that, informally, encourages the policy to prefer transitions that are more plausible under the target dynamics than under the source dynamics (**?**). These techniques are naturally connected to density-ratio estimation between transition measures (or state-action-next-state triples) and can be interpreted as constructing a shaped reward whose value function under the learned kernel approximates the value function under the true kernel. In offline settings, the practical difficulty is that the required ratios can be high variance and, when implemented using explicit likelihoods, become unavailable for implicit models. The likelihood-free critics we use occupy the same conceptual position as DARC's correction term, but are learned via variational divergence objectives from samples rather than from explicit density evaluations.

Model-based offline RL via pessimism and uncertainty penalties. A broad set of model-based offline algorithms addresses distribution shift by discouraging rollout regions where the learned model is unreliable. MOPO-style methods penalize rewards by an epistemic uncertainty estimate of the

dynamics model, typically implemented via ensembles and predictive variance, so that planning in the model becomes pessimistic away from the data support (**?**). Related methods (e.g., MOBILE, COMBO, and other hybrid schemes) combine short-horizon model rollouts with conservative value estimation or conservative Q-learning penalties to prevent exploitation of model errors (**??**). These approaches differ in (i) whether conservatism is enforced by an explicit uncertainty penalty, a lower confidence bound, or a conservative value regularizer, and (ii) whether rollouts are performed in the learned model, in a learned latent model, or in a mixture with real transitions. Our framework is compatible with these design choices but emphasizes a different axis: we treat conservatism as a divergence-controlled correction between *distributions* (environment versus model, policy versus behavior), learned by classification-style objectives that remain well-defined with sampling-only models.

Density-ratio estimation and classifier-based correction. Estimating ratios such as $\frac{dP}{dQ}$ from samples is a classical problem that appears in off-policy evaluation, covariate shift correction, and model bias correction (**??**). In high dimensions, direct density estimation is typically avoided in favor of *classification-based* ratio estimation: given samples from P and Q, train a discriminator D to distinguish the two; under suitable losses, the discriminator recovers a monotone function of the density ratio, e.g.,

$$\log \frac{dP}{dQ}(x) \;=\; \log \frac{D(x)}{1 - D(x)} \;+\; \text{const}$$

for logistic regression at optimum. This perspective is operationally attractive in our setting because both $p_f(\cdot \mid s, a)$ and $m_{\theta,f}(\cdot \mid s, a)$ are accessible through samples (from D_{env} and from model rollouts) even when neither density is tractable. We use this idea twice: first to compare transition triples (z, a, z') from environment versus model, and second to compare action choices (z, a) from the current policy versus the dataset. The latter is closely related to techniques that regularize policy improvement by constraining divergences from the behavior distribution, as in many conservative or behavior-regularized offline RL algorithms (**???**).

Variational f-divergences, f-GAN, and NCE. Our treatment of discrepancy is phrased in terms of f-divergences because they admit a variational dual of the form

$$D_f(P\|Q) \;=\; \sup_T \; \mathbb{E}_P[T] \;-\; \mathbb{E}_Q[f^*(T)],$$

where f^* is the convex conjugate of f (**?**). This dual representation is the basis of f-GANs and encompasses familiar objectives: KL corresponds to

log-ratio recovery, Jensen–Shannon corresponds to logistic classification, and χ^2-type divergences yield quadratic witnesses. Noise-contrastive estimation (NCE) and related contrastive objectives also fit this template, providing consistent estimators of density ratios or unnormalized models by discriminating data from noise (**?**). We exploit precisely the property that the dual depends only on expectations, not on likelihood evaluations. In our algorithmic instantiation, the transition critic T_ϕ and action critic U_ψ are trained by such variational objectives, and their outputs are used as additive corrections in a surrogate reward. From the theoretical side, we track the effect of imperfect critic optimization via duality gaps, which is standard in variational divergence estimation.

Implicit generative modeling for dynamics. Modern world models increasingly rely on implicit generative mechanisms. Diffusion models, autoregressive sequence models with stochastic decoding, and latent-variable models with intractable marginals can produce high-fidelity samples yet lack tractable normalized likelihoods $m_\theta(s' \mid s, a)$, or make likelihood computation computationally infeasible at training time (**??**). In reinforcement learning, such models are used for imagination-based policy learning and planning, often by rolling out predicted future states and optimizing expected returns under the model (**??**). In online settings, model bias can be corrected by environment interaction; in offline settings, the same bias becomes a dominant failure mode. The central methodological point for us is that, with implicit dynamics, any approach requiring $\log m_\theta(s' \mid s, a)$ is structurally incompatible with the model class. This motivates our emphasis on objectives that require only conditional sampling from m_θ.

Representation learning and latent-space comparisons. Many successful model-based RL methods operate in a learned latent space in which dynamics are simpler and prediction is easier (**???**). In offline regimes, representation learning also serves as a variance-reduction device for discrepancy estimation: discriminators trained on raw observations may overfit or focus on irrelevant features, whereas a suitably learned $f : \mathcal{S} \to \mathcal{Z}$ can expose the task-relevant components of the transition while suppressing nuisance variation. This connects to sufficiency and bisimulation-style representation results, where value functions and optimal policies can be approximated as functions of latent variables under appropriate invariances (**?**). Our subsequent setup therefore measures model mismatch and overlap in the pushforward transition measures $p_f(\cdot \mid s, a)$ and $m_{\theta,f}(\cdot \mid s, a)$, making explicit the role of representation sufficiency error ϵ_{rep} in end-to-end guarantees.

Summary and positioning. The above threads collectively suggest a unifying view: conservatism in model-based offline RL can be enforced by

penalizing *distinguishability* between (i) environment and model transitions and (ii) behavior and learned actions, and this distinguishability can be learned by variational critics from samples. In the next section we formalize the offline learning problem under an implicit dynamics sampler, define the mixed real/model training distribution induced by rollouts, and state overlap and representation assumptions in a form tailored to likelihood-free f-divergence correction.

3 Problem setup

We study offline policy learning in a discounted MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, r, \mu_0, \gamma)$ with reward function $r : \mathcal{S} \times \mathcal{A} \to [r_{\min}, r_{\max}]$ and discount $\gamma \in (0, 1)$. The return of a trajectory $\tau = (s_0, a_0, s_1, a_1, \dots)$ is

$$R(\tau) \;=\; \sum_{t \geq 0} \gamma^t r(s_t, a_t), \qquad J_{\mathcal{M}}(\pi) \;=\; \mathbb{E}_{\tau \sim p^\pi}\big[R(\tau)\big],$$

where p^π denotes the trajectory measure induced by μ_0, the policy $\pi(\cdot \mid s)$, and the true transition kernel $p(\cdot \mid s, a)$.

Offline data and implicit dynamics. Our only access to the environment is an offline dataset

$$D_{\text{env}} \;=\; \{(s_i, a_i, r_i, s_i')\}_{i=1}^n,$$

collected by an unknown behavior policy π_b (or mixture thereof) under p. In addition, we are given a learned dynamics model $m_\theta(\cdot \mid s, a)$ trained from D_{env} that is *implicit*: we can sample $s^+ \sim m_\theta(\cdot \mid s, a)$ for any queried (s, a), but we do not assume the conditional likelihood $m_\theta(s' \mid s, a)$ is tractable or even well-defined in closed form. Accordingly, any method requiring $\log m_\theta(s' \mid s, a)$ is not admissible in our computational model; the primitive we use is conditional sampling.

Model rollouts and mixed training distribution. Training proceeds by interleaving (i) minibatches from D_{env} and (ii) short rollouts in m_θ under the current policy π. Concretely, for a rollout horizon h we sample a seed state s_0 from an empirical distribution supported on states in D_{env} (e.g., uniform over dataset states), then iterate

$$a_t \sim \pi(\cdot \mid s_t), \qquad s_{t+1} \sim m_\theta(\cdot \mid s_t, a_t), \qquad t = 0, \dots, h-1,$$

and store the resulting triples in a model-generated buffer $D_m = \{(s_t, a_t, s_{t+1})\}$. The induced training data for subsequent critic/policy updates is then drawn from a mixture of real and model transitions. For analysis it is convenient to view this mixture as defining an auxiliary Markov chain whose one-step

transitions equal p with some probability and m_θ otherwise, or more simply as an algorithm-dependent sampling distribution over (s, a, s') that we denote by ξ^π. The precise mixing scheme is not essential; what matters is that ξ^π is supported on (a) the dataset transitions and (b) the transitions produced by m_θ along π-rollouts starting from dataset states, and that its effective horizon is controlled by h to limit compounding model error.

Trajectory measures and shift decomposition. The main difficulty is that performance is evaluated under the true environment dynamics p, while training uses a distribution ξ^π influenced by D_{env} and m_θ. To make explicit the sources of mismatch, we write the environment trajectory density under π as

$$p^\pi(\tau) \;=\; \mu_0(s_0) \prod_{t \geq 0} \pi(a_t \mid s_t)\, p(s_{t+1} \mid s_t, a_t),$$

and the corresponding model trajectory density as

$$m_\theta^\pi(\tau) \;=\; \mu_0(s_0) \prod_{t \geq 0} \pi(a_t \mid s_t)\, m_\theta(s_{t+1} \mid s_t, a_t),$$

understanding these as measures when densities do not exist. If likelihoods were available, we could express a change of measure between p^π and m_θ^π through the multiplicative product of one-step ratios. In particular, whenever $p(\cdot \mid s, a) \ll m_\theta(\cdot \mid s, a)$ we have the Radon–Nikodym derivative

$$\frac{dp^\pi}{dm_\theta^\pi}(\tau) \;=\; \prod_{t \geq 0} \frac{dp(\cdot \mid s_t, a_t)}{dm_\theta(\cdot \mid s_t, a_t)}(s_{t+1}),$$

which isolates *model bias* as a transition-kernel mismatch.

A second shift arises because D_{env} reflects π_b rather than π. If we define the behavior-induced trajectory measure

$$p^{\pi_b}(\tau) \;=\; \mu_0(s_0) \prod_{t \geq 0} \pi_b(a_t \mid s_t)\, p(s_{t+1} \mid s_t, a_t),$$

then, on the event that $\pi(\cdot \mid s) \ll \pi_b(\cdot \mid s)$, the policy shift admits a trajectory derivative

$$\frac{dp^\pi}{dp^{\pi_b}}(\tau) \;=\; \prod_{t \geq 0} \frac{\pi(a_t \mid s_t)}{\pi_b(a_t \mid s_t)}.$$

In offline learning π_b is unknown, and in our setting m_θ is implicit; thus neither derivative is directly computable. Nevertheless, these identities dictate the *structure* of a conservative objective: we must control deviations between (p, π) and the training sources (m_θ, π_b) in a way that remains meaningful with sampling-only access.

Representation-space formulation. Since both transition comparison and action comparison are high-dimensional when s is complex (e.g., images), we introduce a representation map $f : \mathcal{S} \to \mathcal{Z}$ and write $z = f(s)$. We measure mismatch in $\mathcal{Z}$ via pushforward transition measures

$$p_f(z' \mid s, a) \text{ the law of } z' = f(s') \text{ when } s' \sim p(\cdot \mid s, a), \qquad m_{\theta,f}(z' \mid s, a) \text{ the law of } z' = f(s') \text{ when}$$

We emphasize that $p_f(\cdot \mid s, a)$ is observable through $(s, a, s') \in D_{\text{env}}$ by mapping $s' \mapsto f(s')$, and $m_{\theta,f}(\cdot \mid s, a)$ is observable by sampling $s^+ \sim m_\theta(\cdot \mid s, a)$ and mapping $s^+ \mapsto f(s^+)$. Hence, although p and m_θ may be intractable in $\mathcal{S}$, their induced distributions in $\mathcal{Z}$ are accessible *through samples*.

We will quantify transition mismatch by an f-divergence $D_f(p_f(\cdot \mid s, a) \| m_{\theta,f}(\cdot \mid s, a))$ and, analogously, policy shift by a divergence between the action distributions $\pi(\cdot \mid s)$ and the (implicit) behavior action distribution induced by D_{env} at state s. We record the overlap conditions needed for such comparisons in the representation space:

$$p_f(\cdot \mid s, a) \ll m_{\theta,f}(\cdot \mid s, a) \quad \text{for relevant } (s, a), \qquad \pi(\cdot \mid s) \ll \pi_b(\cdot \mid s) \quad \text{for relevant } s,$$

where "relevant" refers to the occupancy induced by the training pipeline (dataset seeding and h-step model rollouts). These absolute continuity requirements are the minimal conditions under which any change-of-measure or divergence-based control can be well-posed.

Value sufficiency in latent space. The purpose of f is not only statistical (variance reduction for discrepancy estimation) but also semantic: we require that planning and evaluation can be carried out using z with limited loss. Formally, we assume a representation sufficiency condition: there exists a value function class $\mathcal{V}$ on $\mathcal{Z}$ such that the optimal value function in $\mathcal{S}$ is approximable by $V(z)$ with error at most ϵ_{rep} in the sense relevant to our Bellman backups and occupancy measures. This condition allows us to relate (i) mismatch measured on p_f versus $m_{\theta,f}$ to (ii) errors in value estimates and policy gradients computed from mixed real/model data.

Effective horizon and short-rollout regime. Throughout, we use the standard effective horizon notation $H \asymp (1 - \gamma)^{-1}$. Algorithmically, we further cap model rollouts at a finite h to avoid unbounded accumulation of model error; analytically, this produces bounds that scale with either H or $\min\{H, h\}$ depending on the quantity being controlled. The role of the subsequent correction terms is to ensure that, even within this short-rollout regime, the policy does not exploit systematic bias in m_θ nor drift excessively away from the behavioral support.

Preparation for likelihood-free correction. In the likelihood-based setting, the preceding change-of-measure identities suggest additive corrections of the form $\log \frac{p}{m_\theta}$ and $\log \frac{\pi_b}{\pi}$ along trajectories, often coupled with a concave utility such as $\log r$ (hence our standing assumption $r_{\min} > 0$). Since explicit likelihoods are unavailable here, we will instead construct *sample-based* surrogates for these corrections by learning witness functions in the variational dual of an f-divergence, separately for (i) transition shift between samples from p_f and $m_{\theta,f}$ and (ii) action shift between policy actions and dataset actions. The next section carries out this derivation and defines the corrected reward used for offline actor–critic updates on the mixed buffer $D_{\text{env}} \cup D_m$.

4 Likelihood-free shift-aware rewards

Our objective is to construct an additive reward correction that plays the role of the intractable log-ratio terms suggested by the change-of-measure identities, while requiring only samples from D_{env} and conditional samples from the implicit model m_θ. We do so by learning *witness functions* in the variational dual of an f-divergence, separately for (i) transition shift between $p_f(\cdot \mid s, a)$ and $m_{\theta,f}(\cdot \mid s, a)$ and (ii) action shift between dataset actions and actions proposed by π at dataset states.

Variational dual for transition shift in latent space. Fix a convex function $f : (0, \infty) \to \mathbb{R}$ with $f(1) = 0$, and recall the associated f-divergence

$$D_f(P\|Q) \;=\; \mathbb{E}_{x\sim Q}\Big[f\Big(\frac{dP}{dQ}(x)\Big)\Big], \qquad P \ll Q.$$

A standard variational representation (Fenchel dual) states that, for suitable function classes,

$$D_f(P\|Q) \;=\; \sup_T \Big\{\mathbb{E}_{x\sim P}[T(x)] - \mathbb{E}_{x\sim Q}[f^*(T(x))]\Big\}, \tag{1}$$

where f^* is the convex conjugate $f^*(t) = \sup_{u>0}\{ut - f(u)\}$. We apply (11) conditionally at each (s, a) with

$$P \equiv p_f(\cdot \mid s, a), \qquad Q \equiv m_{\theta,f}(\cdot \mid s, a), \qquad x \equiv z' \in \mathcal{Z},$$

and we allow the witness to depend on (z, a, z') through a parametric critic $T_\phi(z, a, z')$. Concretely, we train T_ϕ using (i) real transitions $(s, a, s') \in D_{\text{env}}$ mapped to (z, a, z') and (ii) model transitions (s, a, s^+) with $s^+ \sim m_\theta(\cdot \mid s, a)$ mapped to (z, a, z^+). Writing expectations with respect to the corresponding empirical/sampling procedures, the conditional dual objective reads

$$\max_\phi \;\; \mathbb{E}_{(s,a,s')\sim D_{\text{env}}}\big[T_\phi(f(s), a, f(s'))\big] - \mathbb{E}_{(s,a)\sim\nu,\; s^+\sim m_\theta(\cdot|s,a)}\big[f^*(T_\phi(f(s), a, f(s^+)))\big], \tag{2}$$

where ν is the state–action sampling distribution used to query the model (typically induced by dataset seeding and short rollouts). When T_ϕ is optimized and the function class is rich enough, (2) provides a likelihood-free estimate of the discrepancy between p_f and $m_{\theta,f}$ on the region of interest.

KL specialization and connection to log-ratio corrections. The case most directly aligned with SAR is the forward KL divergence. Taking

$$f(u) = u \log u, \qquad f^*(t) = \exp(t-1),$$

the conditional dual (11) becomes

$$\mathrm{KL}(P\|Q) \;=\; \sup_T \Big\{ \mathbb{E}_P[T] - \mathbb{E}_Q[\exp(T-1)] \Big\}. \tag{3}$$

Moreover, the pointwise optimizer satisfies $T^*(x) = 1 + \log \frac{dP}{dQ}(x)$ (any additive constant is immaterial up to normalization conventions). In our conditional setting this yields, formally,

$$T^*(z, a, z') \;=\; 1 + \log \frac{dp_f(\cdot \mid s, a)}{dm_{\theta,f}(\cdot \mid s, a)}(z'), \qquad z = f(s),\ z' = f(s').$$

Thus, once trained, T_ϕ can be interpreted (up to an additive constant and optimization error) as a proxy for the unavailable per-transition log-ratio $\log \frac{p_f}{m_{\theta,f}}$, and hence as a substitute for $\log \frac{p}{m_\theta}$ in the SAR-style correction. Importantly, this interpretation is *likelihood-free*: the critic is learned from samples without ever evaluating $m_\theta(s' \mid s, a)$.

Variational dual for action (policy) shift. We next construct an analogous witness for policy shift relative to the unknown behavior policy. Since $\pi_b(\cdot \mid s)$ is not available, we use the dataset itself to provide samples from the behavior action distribution at visited states. A convenient conditional comparison is: sample s from the dataset state marginal, then compare actions $a \sim \pi(\cdot \mid s)$ to dataset actions paired with the same (or nearby) state. In representation space we write $z = f(s)$ and train an action critic $U_\psi(z, a)$ via an f-divergence dual between two distributions on (z, a):

$$P_{za} \;\equiv\; \text{law of } (f(s), a) \text{ when } (s, a) \sim D_{\mathrm{env}}, \qquad Q_{za} \;\equiv\; \text{law of } (f(s), a) \text{ when } s \sim D_{\mathrm{env}},\ a \sim \pi(\cdot \mid s).$$

Applying (11) with $x = (z, a)$ gives the objective

$$\max_\psi \;\; \mathbb{E}_{(s,a)\sim D_{\mathrm{env}}}\big[U_\psi(f(s), a)\big] \;-\; \mathbb{E}_{s\sim D_{\mathrm{env}},\, a\sim\pi(\cdot|s)}\big[f^*(U_\psi(f(s), a))\big]. \tag{4}$$

Under the KL choice, the optimizer satisfies

$$U^*(z, a) \;=\; 1 + \log \frac{dP_{za}}{dQ_{za}}(z, a).$$

If the state sampling in P_{za} and Q_{za} is matched (both use $s \sim D_{\text{env}}$), then the state-marginal ratio cancels and we obtain the conditional form

$$U^*(f(s), a) \;=\; 1 + \log \frac{\pi_b(a \mid s)}{\pi(a \mid s)},$$

again up to an additive constant. This is exactly the per-step policy-shift term that appears in likelihood-based conservative objectives, but realized through a sample-based discriminator.

Corrected reward as a surrogate for SAR-style objectives. We now combine the base reward transformation with the learned witnesses. Since the likelihood-based derivations typically yield additive *log* corrections along trajectories, we adopt the log-reward transform and define per-transition corrected rewards of the form

$$\tilde{r}_{\phi,\psi}(s, a, s') \;=\; \log r(s, a) \;+\; \alpha\, T_\phi(f(s), a, f(s')) \;+\; \beta\, U_\psi(f(s), a), \qquad (5)$$

with weights $\alpha, \beta \geq 0$ controlling the strength of dynamics and policy-shift corrections. In practice, we may apply the transition correction only to model-generated transitions (where it is intended to compensate model bias) and apply the action correction primarily on dataset-seeded states (where the behavior distribution is well-defined). The precise gating is algorithmic; the analytic role is that αT_ϕ penalizes regions where $m_{\theta,f}$ deviates from p_f, and βU_ψ penalizes actions that depart from the behavioral support.

Under the KL specialization and assuming optimal critics, (5) recovers the canonical SAR structure up to constants:

$$\tilde{r}(s, a, s') \;\approx\; \log r(s, a) \;+\; \alpha \log \frac{p_f(z' \mid s, a)}{m_{\theta,f}(z' \mid s, a)} \;+\; \beta \log \frac{\pi_b(a \mid s)}{\pi(a \mid s)}.$$

Constants (including the "+1" in the KL optimizer) can be dropped or absorbed into a baseline since they shift values by at most an additive constant and do not affect the optimal policy under standard entropy-regularized objectives; nonetheless, for numerical stability we will later enforce boundedness of T_ϕ and U_ψ via clipping or calibration.

On reward positivity and utility alternatives. The use of $\log r(s, a)$ requires $r_{\min} > 0$, which we assume throughout. This is not merely technical: the log transform is the mechanism by which multiplicative change-of-measure factors become additive trajectory sums, enabling Jensen-type lower bounds and ELBO-like surrogates. When the environment reward may be zero or negative, one may (i) shift and scale rewards to enforce positivity (noting that such transformations alter the objective unless one simultaneously adjusts the performance criterion), or (ii) replace the log transform

by a different concave utility u and analyze $J_u(\pi) = \mathbb{E}[u(R(\tau))]$ instead of $J_{\mathcal{M}}(\pi)$. For general f-divergences, this utility view is natural: the variational form (11) provides a family of inequalities in which f^* determines the penalty applied to samples from the reference distribution, and the corresponding surrogate objective controls a utility-transformed return whose curvature matches the chosen divergence. In all cases, the guiding principle is unchanged: we construct additive, sample-based witness functions that (a) discourage exploitation of model bias and (b) discourage unsupported policy deviation, while remaining implementable with an implicit dynamics sampler and offline data alone.

5 Algorithm (LF-SAR): practical training loop with implicit models

We now make the preceding construction operational in the sampling-oracle setting, i.e., when $m_\theta(\cdot \mid s, a)$ can be queried for samples but does not admit tractable likelihood evaluation. The resulting method, which we refer to as LF-SAR, alternates between (i) short-horizon rollouts in the implicit model under the current policy, (ii) training variational critics for transition shift and action shift, and (iii) updating the policy by off-policy RL on a mixture of real and model-generated transitions with corrected rewards.

Data structures and seeding. We maintain two replay buffers: the fixed offline buffer D_{env} and a growing model buffer D_m containing tuples (s, a, r, s') where $s' \sim m_\theta(\cdot \mid s, a)$ and (s, a) are generated by rolling out the current policy from dataset seed states. In each iteration we sample a minibatch of seed states s_1 from the empirical state marginal of D_{env} (or from the s-components of transitions), and we run h-step synthetic rollouts

$$a_t \sim \pi(\cdot \mid s_t), \qquad s_{t+1} \sim m_\theta(\cdot \mid s_t, a_t), \qquad t = 1, \ldots, h,$$

storing the resulting transitions in D_m. We emphasize that h is not intended to approximate the full discounted horizon; rather, it controls the extent to which model bias can compound before we re-anchor to the dataset. In practice we treat h as a conservative hyperparameter (small to moderate) and increase it only when the learned model is demonstrably accurate on the relevant latent features.

Representation and caching. Because both critics operate in latent space, we either (a) learn an encoder f jointly with the critics and policy, or (b) pretrain f on D_{env} and subsequently freeze it. The latter frequently improves stability by preventing nonstationarity in the discriminator features. In either case, we precompute and cache $z = f(s)$ for states in D_{env} when

feasible; for D_m we compute $f(s)$ on the fly (or cache in parallel) to avoid repeated encoder passes during critic updates. When m_θ itself is defined in latent space (e.g., a latent diffusion model), we identify s with a decoded observation and set z to the model's internal latent; when doing so we must ensure that the same f is applied consistently to both real and model transitions to avoid spurious detectability unrelated to dynamics mismatch.

Transition-shift critic updates. Given minibatches of real latent transitions (z, a, z') from D_{env} and synthetic latent transitions (z, a, z^+) from D_m, we update T_ϕ by stochastic ascent on the dual objective already specified in (2). For implicit models, the second term is computed by sampling $s^+ \sim m_\theta(\cdot \mid s, a)$ and mapping $z^+ = f(s^+)$. We may optionally balance the two terms by reweighting minibatches so that the marginal distribution of (z, a) is comparable across real and synthetic samples; this reduces the burden on T_ϕ to separate distributions using spurious covariate shift in (z, a) rather than genuine discrepancy in z' conditional on (z, a).

Action-shift critic updates. We update U_ψ by stochastic ascent on (4), contrasting dataset action pairs (z, a) with policy-proposed pairs (z, a^π) where $z = f(s)$ and $a^\pi \sim \pi(\cdot \mid s)$ for s drawn from the dataset state marginal. The use of matched state sampling is essential: it makes U_ψ primarily a witness for conditional action shift rather than a confounder for state visitation shift. When the behavior policy is a mixture (as is typical in offline benchmarks), U_ψ implicitly estimates shift relative to the mixture induced by D_{env}, which is the relevant reference for support constraints.

Corrected reward and gating on real vs. synthetic data. We use the template $\tilde{r}_{\phi,\psi}$ from (5) but implement it with explicit gating, reflecting the distinct roles of the two witnesses:

- For *real* transitions from D_{env}, we set

$$\tilde{r} \;=\; \log r(s, a) \;+\; \beta\, U_\psi(f(s), a),$$

  since real transitions do not require correction for model bias, while the action witness remains meaningful as a conservative penalty on deviation from behavioral support.

- For *synthetic* transitions from D_m, we set

$$\tilde{r} \;=\; \log r(s, a) \;+\; \alpha\, T_\phi(f(s), a, f(s')),$$

  since the primary concern is exploitation of dynamics errors; depending on the application we may also include the βU_ψ term on synthetic data, but we view this as optional and typically use it only when the actor rapidly leaves the dataset action support.

This separation ensures that each correction term is used where it is most interpretable, while retaining a unified off-policy learning interface.

Policy and value updates on a mixed buffer. We update π (and any auxiliary value critics used by the chosen off-policy algorithm) on mini-batches drawn from a mixture distribution over $D_{\text{env}} \cup D_m$. We typically control the mixture by a parameter $\lambda \in [0, 1]$ specifying the fraction of real transitions per update, and we either fix λ or schedule it (e.g., start with mostly real data, then increase synthetic usage as T_ϕ improves). Any convergent off-policy method can be used; in continuous control, we instantiate this step with entropy-regularized actor–critic updates (e.g., SAC) using $\tilde{r}$ in place of r. We treat α, β as Lagrange-like weights: increasing α discourages reliance on model regions where T_ϕ signals mismatch, while increasing β discourages unsupported actions.

Stabilization of variational critics. Because common divergences yield rapidly growing $f^*(\cdot)$ (notably $f^*(t) = \exp(t - 1)$ for KL), unconstrained optimization of (2)–(4) can produce large logits and unstable gradients. We therefore enforce boundedness and smoothness of T_ϕ and U_ψ via a combination of:

1. *Spectral normalization* on critic layers, yielding a global Lipschitz constraint that empirically prevents discriminator collapse.

2. *Gradient penalties* on interpolations between real and synthetic samples (in the style of WGAN-GP), applied to the critic inputs (z, a, z') or (z, a); this is especially helpful when f is learned jointly and the feature distribution drifts.

3. *Logit clipping or calibration*: we replace T_ϕ by $\text{clip}(T_\phi, [-c_T, c_T])$ and similarly for U_ψ, or subtract a running mean baseline so that the critics remain centered and their additive constants do not induce large reward shifts.

4. *Penalty regularization*: we add ℓ_2 penalties on critic outputs or enforce trust regions on critic updates, which controls variance of the corrected reward signal seen by the actor.

These interventions do not change the conceptual role of the witnesses; rather, they ensure that the corrected reward remains within a bounded range compatible with stable Bellman backups.

Notes for diffusion and transformer world models. When m_θ is a diffusion model, each conditional sample $s' \sim m_\theta(\cdot \mid s, a)$ may require multiple denoising steps; the dominant cost in LF-SAR is then model rollout

sampling rather than policy optimization. We therefore (i) keep h small, (ii) reuse synthetic rollouts across multiple gradient steps (i.e., amortize D_m), and (iii) optionally reduce diffusion steps during training rollouts while reserving higher-fidelity sampling for evaluation or for periodic refresh of D_m. When m_θ is an autoregressive transformer, we similarly amortize rollouts by caching predicted next states and by truncating rollouts when the critic T_ϕ indicates severe mismatch (an implicit early-termination criterion). In both cases, the key requirement is only that the sampler be conditionally callable; likelihood evaluation is never used.

Summary of the implementable loop. At a high level, LF-SAR repeatedly (i) expands D_m by short model rollouts under π, (ii) tightens the two variational witnesses T_ϕ and U_ψ using only samples from D_{env} and m_θ, and (iii) improves π by standard off-policy RL on corrected rewards. This completes the algorithmic layer; in the next section we formalize the sense in which the resulting surrogate objective constitutes an ELBO-like lower bound (under optimal witnesses) and how critic duality gaps degrade the bound with explicit horizon dependence.

6 Theory I (surrogate objective): an ELBO-like lower bound and degradation by critic gaps

In this section we isolate the purely variational component of LF-SAR: we show that, under optimal critics, the corrected reward defines a surrogate objective that lower-bounds a suitable utility of the true return, in direct analogy with an evidence lower bound (ELBO). We then quantify how imperfect critics degrade the bound, with an explicit dependence on the effective horizon.

6.1 A change-of-measure inequality for positive rewards

Fix a policy π and write $p^\pi(\tau)$ for the trajectory law in the environment,

$$p^\pi(\tau) \;=\; \mu_0(s_0)\prod_{t\geq 0}\pi(a_t \mid s_t)\,p(s_{t+1} \mid s_t, a_t), \qquad \tau = (s_0, a_0, s_1, a_1, \dots).$$

Likewise define the model-induced trajectory law

$$m^\pi(\tau) \;=\; \mu_0(s_0)\prod_{t\geq 0}\pi(a_t \mid s_t)\,m_\theta(s_{t+1} \mid s_t, a_t).$$

Under the overlap condition $p_f(\cdot \mid s,a) \ll m_{\theta,f}(\cdot \mid s,a)$ on the relevant support, the Radon–Nikodym derivative $\frac{dp^\pi}{dm^\pi}(\tau)$ is well-defined and factorizes

in the usual way:

$$\log \frac{dp^\pi}{dm^\pi}(\tau) \;=\; \sum_{t\geq 0} \log \frac{p(s_{t+1} \mid s_t, a_t)}{m_\theta(s_{t+1} \mid s_t, a_t)}. \tag{6}$$

Since $r_{\min} > 0$, we may relate the discounted sum return $R(\tau) = \sum_{t\geq 0} \gamma^t r(s_t, a_t)$ to a discounted average of $\log r$ via the log-sum inequality. Let $w_t = (1-\gamma)\gamma^t$ so that $\sum_{t\geq 0} w_t = 1$. Then for every trajectory,

$$\log\big((1-\gamma)R(\tau)\big) = \log\Big(\sum_{t\geq 0} w_t\, r(s_t, a_t)\Big) \;\geq\; \sum_{t\geq 0} w_t\, \log r(s_t, a_t) - \sum_{t\geq 0} w_t \log w_t, \tag{7}$$

where the final term depends only on γ. Denoting the constant

$$C_\gamma \;:=\; -\log(1-\gamma) \;-\; \sum_{t\geq 0} w_t \log w_t,$$

we may rewrite (7) as

$$\log R(\tau) \;\geq\; (1-\gamma) \sum_{t\geq 0} \gamma^t \log r(s_t, a_t) \;-\; C_\gamma. \tag{8}$$

Now we apply Jensen after changing measure from p^π to m^π:

$$\begin{aligned} \log J_{\mathcal{M}}(\pi) = \log \mathbb{E}_{\tau\sim p^\pi}\big[R(\tau)\big] &= \log \mathbb{E}_{\tau\sim m^\pi}\Big[R(\tau)\, \frac{dp^\pi}{dm^\pi}(\tau)\Big] \\ &\geq \mathbb{E}_{\tau\sim m^\pi}\Big[\log R(\tau) + \log \tfrac{dp^\pi}{dm^\pi}(\tau)\Big], \end{aligned} \tag{9}$$

where we used concavity of log. Combining (9) with (8) yields the "ideal" (likelihood-based) lower bound

$$\log J_{\mathcal{M}}(\pi) \;\geq\; (1-\gamma)\, \mathbb{E}_{\tau\sim m^\pi}\Big[\sum_{t\geq 0} \gamma^t \log r(s_t, a_t)\Big] + \mathbb{E}_{\tau\sim m^\pi}\Big[\sum_{t\geq 0} \log \tfrac{p}{m_\theta}(s_{t+1} \mid s_t, a_t)\Big] - C_\gamma. \tag{10}$$

The obstacle is that the per-step log-ratios in (10) are not available when m_θ is implicit.

6.2 Replacing log-ratios by variational witnesses in latent space

We now show how to replace the inaccessible log-ratio term by a learned witness T_ϕ operating in latent space. We work with pushforward conditional laws $p_f(z' \mid s, a)$ and $m_{\theta,f}(z' \mid s, a)$, where $z = f(s)$ and $z' = f(s')$. For a chosen f-divergence, we have the variational representation

$$D_f(P\|Q) \;=\; \sup_T \Big\{\mathbb{E}_{x\sim P}[T(x)] - \mathbb{E}_{x\sim Q}[f^*(T(x))]\Big\}. \tag{11}$$

We interpret $T_\phi(z, a, z')$ as an approximate maximizer of (11) for $P = p_f(\cdot \mid s, a)$ and $Q = m_{\theta,f}(\cdot \mid s, a)$ (or the reverse direction, depending on the divergence and which direction yields the desired correction). In the KL case, $f(u) = u \log u$ and $f^*(t) = \exp(t-1)$, and the optimizer satisfies

$$T^*(z, a, z') \;=\; 1 + \log \frac{dp_f(\cdot \mid s, a)}{dm_{\theta,f}(\cdot \mid s, a)}(z'). \tag{12}$$

Thus, up to an additive constant, the optimal critic recovers the latent log density ratio. Because additive constants do not affect policy optimization (they shift all rewards uniformly), we may treat T^* as a likelihood-free proxy for $\log \frac{p_f}{m_{\theta,f}}$.

For general f, we do not obtain a literal log-ratio; instead the witness T^* is the Fenchel dual optimizer controlling mismatch between p_f and $m_{\theta,f}$. The key point for LF-SAR is that (11) allows us to inject T_ϕ into an ELBO-like argument without ever evaluating p or m_θ. Concretely, for any measurable T and any (s, a) on-support,

$$\mathbb{E}_{z' \sim p_f(\cdot \mid s,a)}[T(z, a, z')] - \mathbb{E}_{z' \sim m_{\theta,f}(\cdot \mid s,a)}[f^*(T(z, a, z'))] \;\le\; D_f\big(p_f(\cdot \mid s, a) \,\|\, m_{\theta,f}(\cdot \mid s, a)\big), \tag{13}$$

with equality at $T = T^*$. When T_ϕ is learned with duality gap $\epsilon_T(s, a)$, we have

$$D_f\big(p_f(\cdot \mid s, a) \,\|\, m_{\theta,f}(\cdot \mid s, a)\big) - \Big(\mathbb{E}_{p_f}[T_\phi] - \mathbb{E}_{m_{\theta,f}}[f^*(T_\phi)]\Big) \;\le\; \epsilon_T(s, a). \tag{14}$$

Analogously, we learn an action-shift witness $U_\psi(z, a)$ for a divergence between conditional action laws $\pi(\cdot \mid s)$ and the (unknown) behavior actions implicit in D_{env}; in the KL case, the optimal $U^*(z, a)$ again recovers $\log \frac{\pi(a|s)}{\pi_b(a|s)}$ up to a constant.

6.3 An ELBO-like surrogate and explicit horizon dependence of slack

We now state the bound at the level needed for subsequent performance analysis. Define the *surrogate return functional*

$$\mathcal{L}_{\phi,\psi}(\pi) \;:=\; \mathbb{E}\Big[\sum_{t \ge 0} \gamma^t \, \tilde{r}_{\phi,\psi}(s_t, a_t, s_{t+1})\Big], \tag{15}$$

where the expectation is taken under the training distribution induced by the LF-SAR pipeline (mixtures of real transitions from D_{env} and synthetic transitions produced by m_θ under π), and $\tilde{r}_{\phi,\psi}$ is the corrected reward used by the algorithm (with gating between T_ϕ and U_ψ depending on whether the transition is synthetic or real). For the purpose of the bound, it is helpful to interpret $\mathcal{L}_{\phi,\psi}(\pi)$ as an empirical proxy for the model-rollout expectation in (10), plus a separate correction accounting for action shift on real data.

Under optimal critics and KL, substituting (12) into (10) and rewriting the (undiscounted) sum of log-ratios in terms of discounted occupancies yields, up to constants depending only on $(\gamma, r_{\min}, r_{\max})$,

$$\log J_{\mathcal{M}}(\pi) \;\geq\; (1-\gamma)\,\mathcal{L}_*(\pi) \;-\; C_\gamma, \tag{16}$$

where $\mathcal{L}_*(\pi)$ is the surrogate obtained by replacing T_ϕ, U_ψ with the corresponding optimal log-ratio witnesses (and absorbing additive constants into C_γ). The representation sufficiency assumption allows us to replace the state-level ratios by latent ratios without changing the bound beyond an additive ϵ_{rep} term, since values and thus the relevant occupancy weighting depend on s only through $z = f(s)$ up to ϵ_{rep}.

When critics are imperfect, we quantify the degradation by combining (14) (and its action analogue) with a telescoping occupancy argument. Writing $\bar{\epsilon}_T$ and $\bar{\epsilon}_U$ for suitable discounted averages of $\epsilon_T(s_t, a_t)$ and $\epsilon_U(s_t, a_t)$ along the training rollouts, we obtain a bound of the form

$$\log J_{\mathcal{M}}(\pi) \;\geq\; (1-\gamma)\,\mathcal{L}_{\phi,\psi}(\pi) - \underbrace{\frac{1}{1-\gamma}\big(\bar{\epsilon}_T + \bar{\epsilon}_U\big)}_{\text{critic duality gaps}} - \underbrace{\epsilon_{\text{rep}}}_{\text{representation mismatch}} - C_\gamma, \tag{17}$$

where $\frac{1}{1-\gamma} \asymp H$ is the effective horizon. The appearance of H is unavoidable: even if $\epsilon_T(s,a)$ is small per step, its contribution accumulates over $\Theta(H)$ effective steps under discounting. More explicitly, if $\epsilon_T(s,a) \leq \epsilon_T$ and $\epsilon_U(s,a) \leq \epsilon_U$ uniformly on the relevant support, then

$$\frac{1}{1-\gamma}\big(\bar{\epsilon}_T + \bar{\epsilon}_U\big) \;\leq\; \frac{1}{1-\gamma}\,(\epsilon_T + \epsilon_U) \;=\; \mathcal{O}(H(\epsilon_T + \epsilon_U)). \tag{18}$$

For non-KL f-divergences, the same reasoning yields an ELBO-like bound for a *utility-transformed* objective rather than $\log J_{\mathcal{M}}(\pi)$. Concretely, the Fenchel–Young inequality underlying (11) implies that there exists a monotone utility u_f (depending on f and on the scaling used in the corrected reward) such that

$$\log \mathbb{E}_{\tau\sim p^\pi}\big[u_f(R(\tau))\big] \;\geq\; (1-\gamma)\,\mathcal{L}_{\phi,\psi}(\pi) \;-\; \mathcal{O}(H(\bar{\epsilon}_T + \bar{\epsilon}_U) + \epsilon_{\text{rep}}) \;-\; C_{\gamma,f}, \tag{19}$$

with a constant $C_{\gamma,f}$ absorbing the reward-aggregation term (cf. (8)) and any additive normalizations of the critics. In particular, KL recovers $u_f(x) = x$ (equivalently, $\log J_{\mathcal{M}}(\pi)$ on the left-hand side as in (17)), while other choices interpolate toward risk-sensitive utilities.

6.4 Remarks on clipping, rollout truncation, and the role of h

In practice we clip or calibrate the critic outputs to stabilize learning. Clipping converts T_ϕ and U_ψ into biased witnesses; in the above bounds this can

be modeled as an additional contribution to ϵ_T, ϵ_U (or, equivalently, as a restriction of the dual function class in (11)). The horizon dependence remains linear in H as long as the clipped corrections remain uniformly bounded.

Finally, LF-SAR uses h-step model rollouts rather than unbounded model trajectories. This does not alter the variational nature of the bound; it changes only the reference distribution under which $\mathcal{L}_{\phi,\psi}(\pi)$ is estimated. Intuitively, shorter rollouts reduce the compounding of model bias in the data-generation process, which improves critic learnability and reduces the empirical $\bar{\epsilon}_T$, but the translation from surrogate improvement to true return (addressed next) still incurs an $\mathcal{O}(H)$ factor due to the environment's effective horizon.

The inequalities (17)–(19) provide the promised ELBO-like justification for optimizing corrected rewards: maximizing $\mathcal{L}_{\phi,\psi}(\pi)$ improves a lower bound on a utility of the true return, and the degradation due to imperfect critics and representations is explicit and horizon-controlled. In the next section we convert near-optimality for the surrogate into a direct suboptimality bound for $J_{\mathcal{M}}(\pi)$ via occupancy-measure arguments in representation space.

7 Theory II (policy performance): from surrogate near-optimality to environment return

We now translate near-optimality of the learned policy for the corrected-reward surrogate into a bound on the true environment return. The argument has two components: (i) an occupancy-measure comparison that controls how transition and action shift affect value under p versus the distributions induced during training; and (ii) a representation-space simulation lemma that propagates one-step mismatch into $\mathcal{O}(H)$ value error, in the style of pessimistic evaluation/value iteration (PEVI).

7.1 Discounted occupancies and truncation-aware training distributions

For any policy π, we write the normalized discounted state–action occupancy under environment dynamics p as

$$d_p^\pi(s,a) \;:=\; (1-\gamma)\sum_{t\geq 0}\gamma^t\,\mathbb{P}_{p,\pi}(s_t=s,a_t=a), \qquad d_{p,f}^\pi(z,a) \;:=\; (1-\gamma)\sum_{t\geq 0}\gamma^t\,\mathbb{P}_{p,\pi}(z_t=z,a_t=a),$$

where $z_t = f(s_t)$. Analogously define d_m^π and $d_{m,f}^\pi$ for rollouts under m_θ.

Because LF-SAR uses h-step synthetic rollouts, it is convenient to isolate the amount of discounted mass carried by model-generated transitions:

$$H_h \;:=\; \sum_{t=0}^{h-1}\gamma^t \;=\; \frac{1-\gamma^h}{1-\gamma}.$$

In particular, when synthetic data are generated by repeatedly restarting from seed states sampled from D_{env} and rolling out h steps, any expectation over model transitions appearing in the training objective is naturally weighted by at most H_h rather than $H \asymp (1-\gamma)^{-1}$. This is the only point at which h enters the analysis: the conversion from one-step mismatch into return error remains $\Theta(H)$, but the *exposure* of the algorithm to model bias through synthetic data is $\Theta(H_h)$.

7.2 A representation-space simulation lemma for implicit models

We first state a generic value perturbation bound in latent space. Let V_p^π denote the true value function under p for reward r, and similarly V_m^π under m_θ. We use only the boundedness of rewards and the overlap assumption in representation space.

Lemma 7.1 (Latent simulation bound)**.** *Assume $r \in [0, r_{\max}]$. Let $f : \mathcal{S} \to \mathcal{Z}$ be any representation and suppose that, for all (s,a) on the relevant support, $p_f(\cdot \mid s,a) \ll m_{\theta,f}(\cdot \mid s,a)$. Then for any policy π,*

$$\left|J_{\mathcal{M}}(\pi) - J_{m_\theta}(\pi)\right| \;\le\; \frac{\gamma}{(1-\gamma)^2}\, r_{\max} \cdot \mathbb{E}_{(s,a)\sim d_p^\pi}\Big[\big\|p_f(\cdot \mid s,a) - m_{\theta,f}(\cdot \mid s,a)\big\|_{\text{TV}}\Big] + \mathcal{O}(\epsilon_{\text{rep}}),$$

where the $\mathcal{O}(\epsilon_{\text{rep}})$ term accounts for representation sufficiency (i.e., the loss incurred by evaluating/bootstrapping values as functions of z rather than s). Moreover, if the divergence control is expressed via an f-divergence, then there exists $c_f > 0$ such that

$$\big\|p_f(\cdot \mid s,a) - m_{\theta,f}(\cdot \mid s,a)\big\|_{\text{TV}} \;\le\; c_f \cdot \sqrt{D_f\big(p_f(\cdot \mid s,a) \,\|\, m_{\theta,f}(\cdot \mid s,a)\big)},$$

and hence the value difference is controlled by an occupancy-weighted $\sqrt{D_f}$ term.

Proof sketch. We apply the standard resolvent identity

$$V_p^\pi - V_m^\pi \;=\; \gamma(I - \gamma P_p^\pi)^{-1}(P_p^\pi - P_m^\pi)V_m^\pi,$$

where P_p^π is the Markov operator induced by (p,π). Taking sup-norms yields $\|V_p^\pi - V_m^\pi\|_\infty \le \frac{\gamma}{1-\gamma} \sup_{s,a} \big|\mathbb{E}_p[V_m^\pi] - \mathbb{E}_m[V_m^\pi]\big|$. Since $V_m^\pi \in [0, r_{\max}/(1-\gamma)]$, the difference of expectations is bounded by $\frac{r_{\max}}{1-\gamma}\|p(\cdot \mid s,a) - m(\cdot \mid s,a)\|_{\text{TV}}$, and we translate to the representation space via pushforward and sufficiency, incurring ϵ_{rep}. The f-divergence to total-variation relation follows from standard inequalities for divergences with f twice differentiable at 1 (constants absorbed into c_f). □

Lemma 7.1 isolates the dynamics-mismatch contribution. The remaining terms in the final performance bound arise because the policy is learned

from an *offline* dataset with unknown behavior actions and from *synthetic* data produced by m_θ, so the learned policy's occupancy can differ from the data-generating occupancy unless explicitly regularized.

7.3 Occupancy control for behavior/policy shift

Let d_{env} denote the (unknown) discounted occupancy induced by the data-collection process (e.g., a mixture of behavior policies). We encode policy shift by an occupancy-weighted divergence

$$\epsilon_{\text{policy}}(\pi) \;:=\; \mathbb{E}_{s\sim d_p^\pi}\Big[D_f\big(\pi(\cdot \mid s) \,\|\, \pi_b(\cdot \mid s)\big)\Big],$$

where π_b is any conditional action law consistent with D_{env} (the bound is stated in terms of ϵ_{policy} rather than π_b itself, which is unobserved). Under overlap $\pi(\cdot \mid s) \ll \pi_b(\cdot \mid s)$ and the same divergence–TV conversion as above, we may control the discrepancy between expectations under $(s,a) \sim d_p^\pi$ and expectations under dataset actions at the same states. This is the usual step in offline RL analyses where off-support actions lead to unavoidable error; here it is made explicit through the learned action critic U_ψ and the term ϵ_{policy}.

7.4 Main bound: surrogate near-optimality implies environment near-optimality

We now combine: (i) the ELBO-like surrogate comparison from Theory I (which turns critic duality gaps into additive slack), (ii) the latent simulation bound (which turns representation-space mismatch into $\mathcal{O}(H)$ value error), and (iii) an optimization error term measuring how well the actor optimizes the surrogate.

Theorem 7.2 (From surrogate optimization to true return)**.** *Assume overlap and representation sufficiency as stated in the enclosing scope, and suppose the actor returns $\hat{\pi} \in \Pi$ satisfying*

$$\mathcal{L}_{\phi,\psi}(\hat{\pi}) \;\geq\; \sup_{\pi\in\Pi} \mathcal{L}_{\phi,\psi}(\pi) \;-\; \epsilon_{\text{opt}}.$$

Let $\bar{\epsilon}_T, \bar{\epsilon}_U$ denote discounted averages of the transition-critic and action-critic duality gaps along the training distribution. Define a dynamics mismatch level in representation space by

$$\epsilon_{\text{model},f} \;:=\; \sup_{(s,a)\ \text{relevant}} D_f\big(p_f(\cdot \mid s,a) \,\|\, m_{\theta,f}(\cdot \mid s,a)\big),$$

and let $\epsilon_{\text{policy}} := \epsilon_{\text{policy}}(\hat{\pi})$. Then there exist constants $C_1, C_2, C_3, C_4 > 0$ depending only on $(r_{\min}, r_{\max}, \gamma)$ and on the choice/scaling of the divergence

such that, with high probability over the sampling noise in D_{env} and model rollouts,

$$J_{\mathcal{M}}(\pi^*) - J_{\mathcal{M}}(\hat{\pi}) \;\le\; C_1\, H_h\, \epsilon_{\mathrm{model},f} + C_2\, H\, \epsilon_{\mathrm{policy}} + C_3\, H\, (\bar{\epsilon}_T + \bar{\epsilon}_U + \epsilon_{\mathrm{rep}}) + C_4\, \epsilon_{\mathrm{opt}}.$$

In particular, if $\epsilon_{\mathrm{model},f}$ and the critic gaps are uniformly small, the suboptimality scales linearly with the effective horizon and, for fixed γ, improves monotonically as h decreases through the factor $H_h = (1-\gamma^h)/(1-\gamma)$ in the model-mismatch term.

Proof sketch. We proceed by a three-step comparison.

Step 1 (critic substitution). By the duality-gap definitions for T_ϕ and U_ψ, replacing optimal witnesses by learned critics in the corrected reward perturbs the surrogate objective by at most $\mathcal{O}(H(\bar{\epsilon}_T + \bar{\epsilon}_U))$ in the same occupancy-weighted manner as in Theory I. This step is purely variational and does not require likelihoods.

Step 2 (surrogate optimality to an idealized latent objective). We introduce an "ideal" latent objective in which (a) synthetic transitions are generated from the true latent kernel p_f rather than $m_{\theta,f}$ and (b) behavior actions are drawn from π_b. The gap between the LF-SAR training objective and this ideal objective is controlled by two change-of-measure terms: a transition-shift term governed by $\epsilon_{\mathrm{model},f}$ and an action-shift term governed by $\epsilon_{\mathrm{policy}}$. The transition-shift contribution is weighted by at most H_h because only h-step synthetic rollouts are used in the pipeline, whereas the action-shift term affects evaluation over the full horizon and retains an $\mathcal{O}(H)$ factor.

Step 3 (latent objective to true return). Finally we translate the ideal latent objective back to the true return using Lemma 7.1 (for transition mismatch) and the representation sufficiency assumption (for the z-dependence error), yielding an $\mathcal{O}(H\epsilon_{\mathrm{rep}})$ contribution. Combining the three steps and inserting the actor suboptimality ϵ_{opt} yields the stated inequality after collecting constants. □

Theorem 7.2 is the desired translation: optimizing the corrected-reward surrogate is sufficient to guarantee small true-return regret provided (i) representation-space dynamics mismatch is controlled on the occupancy relevant to training, (ii) the learned policy does not deviate too far from the behavior support, and (iii) the variational critics are learned to small duality gap. The next section shows that the overlap assumptions implicit in these conditions are not merely technical: without them, no offline procedure can guarantee nontrivial performance in general.

8 Lower bounds and impossibility results

We formalize the sense in which the overlap assumptions in Theorem 7.2 are necessary. The statements below are not specific to LF-SAR; they apply

to *any* offline algorithm that has access only to a fixed dataset D_{env} and (possibly) to a learned implicit model sampler m_θ trained on that same dataset. The implicit-model oracle cannot create information about regions that are absent from the data distribution, and therefore cannot circumvent the standard offline RL hardness.

8.1 Impossibility without absolute continuity (coverage)

We consider an algorithm Alg that maps $(D_{\text{env}}, m_\theta)$ (and any internal randomness) to a policy $\hat{\pi}$. We show that if the optimal policy can place nonzero occupancy on state–action pairs outside the support of the data-collection distribution (or outside the support of the model rollouts induced by the training pipeline), then Alg can be forced to be arbitrarily suboptimal on some instance consistent with its observations.

Theorem 8.1 (No offline guarantee without overlap)**.** *Fix $\gamma \in (0,1)$ and $r_{\max} > 0$. For any offline algorithm Alg that, given an offline dataset D_{env} and an implicit model sampler m_θ trained from D_{env}, outputs a policy $\hat{\pi}$, there exist two discounted MDPs $\mathcal{M}_0, \mathcal{M}_1$ with the same $(\mathcal{S}, \mathcal{A}, \mu_0, \gamma)$ and bounded rewards in $[0, r_{\max}]$ such that:*

1. *the joint distribution of $(D_{\text{env}}, m_\theta)$ is identical under $\mathcal{M}_0$ and $\mathcal{M}_1$ (in particular, all observed offline transitions and all model samples produced by m_θ during training are equal in law); yet*
2. *the optimal values differ by a constant: $J_{\mathcal{M}_0}(\pi^*) - J_{\mathcal{M}_0}(\hat{\pi}) \geq c$ or $J_{\mathcal{M}_1}(\pi^*) - J_{\mathcal{M}_1}(\hat{\pi}) \geq c$ for some $c = \Omega\big(\frac{r_{\max}}{1-\gamma}\big)$.*

Moreover, the construction can be chosen so that the failure is driven by a single state–action pair $(s^\dagger, a^\dagger)$ with $\pi^*(a^\dagger \mid s^\dagger) = 1$, but with zero support under the data-collection process (and hence no absolute continuity $\pi^* \not\ll \pi_b$ at $s^\dagger$).

Proof sketch. We use the standard "two indistinguishable instances" argument. Let $\mathcal{S} = \{s_0, s_1\}$ and $\mathcal{A} = \{a_0, a_1\}$, with $\mu_0 = \delta_{s_0}$. The behavior/data-collection policy is taken to be $\pi_b(a_0 \mid s_0) = 1$, so the offline dataset contains only the action a_0 at s_0 (and never a_1). Define $\mathcal{M}_0$ and $\mathcal{M}_1$ to agree on the observed transition: under either MDP, taking a_0 at s_0 transitions deterministically to s_1 with reward 0, and s_1 is absorbing with reward 0 thereafter. The two MDPs differ only on the unobserved action a_1 at s_0: in $\mathcal{M}_0$ it yields reward 0 and transitions to s_1, while in $\mathcal{M}_1$ it yields reward $r_{\max}$ and transitions to s_1.

Because a_1 is never taken in D_{env}, the offline data distribution is identical under $\mathcal{M}_0$ and $\mathcal{M}_1$. Furthermore, since m_θ is trained only from D_{env} and is queried only on state–action pairs the training pipeline encounters, we may

define $m_\theta(\cdot \mid s_0, a_0)$ to match the observed transition and define its behavior arbitrarily on (s_0, a_1) without affecting the algorithm's observations—indeed, by arranging that the training pipeline never queries (s_0, a_1) (which holds whenever the learned policy stays within the data support, or when conservative regularization prevents selection of a_1), we can ensure that all model samples observed by Alg are also identical in law across $\mathcal{M}_0, \mathcal{M}_1$. Consequently $\hat{\pi}$ is identically distributed under the two instances.

However, the optimal policy in $\mathcal{M}_1$ takes a_1 at s_0 and obtains value $r_{\max}$ at $t = 0$, whereas any policy that takes a_0 obtains 0. Thus $J_{\mathcal{M}_1}(\pi^*) = r_{\max}$ and $J_{\mathcal{M}_1}(\hat{\pi}) \approx 0$ unless $\hat{\pi}$ selects a_1 at s_0. Since $\hat{\pi}$ cannot depend on unobserved differences between $\mathcal{M}_0$ and $\mathcal{M}_1$, it must fail on at least one of the two instances. Taking c to be a constant fraction of $r_{\max}$ yields the claim; by modifying the construction to make s_0 recur with probability γ, one obtains $c = \Omega(\frac{r_{\max}}{1-\gamma})$, matching the horizon scaling. □

Theorem 8.1 captures the necessity of the absolute continuity requirements $\pi(\cdot \mid s) \ll \pi_b(\cdot \mid s)$ and (for model-based training) $p_f(\cdot \mid s, a) \ll m_{\theta,f}(\cdot \mid s, a)$ on the occupancy induced by the learning procedure. If a policy can visit (s, a) where the data (or the model rollouts) provide no information, then any algorithm is forced to extrapolate, and the above indistinguishability construction makes such extrapolation arbitrarily wrong.

8.2 Lower bounds matching the linear-in-H dependence

We also record a horizon dependence statement: even when two MDPs are extremely close on the data distribution, the value difference can remain $\Theta(H)$ due to compounding. This explains why Theorem 7.2 exhibits linear dependence on H.

Proposition 8.2 (Indistinguishability implies $\Omega(H)$ value error)**.** *For any $H \asymp (1-\gamma)^{-1}$ and any $\delta > 0$, there exist two MDPs $\mathcal{M}_0, \mathcal{M}_1$ such that their induced distributions over D_{env} (and over any model-generated rollouts constrained to the dataset support) have total variation distance at most δ, but $\sup_\pi |J_{\mathcal{M}_0}(\pi) - J_{\mathcal{M}_1}(\pi)| \geq c\, H\, \delta$ for a universal constant $c > 0$.*

Proof sketch. One may construct a chain MDP in which the two instances differ by an additive $\pm\delta$ perturbation in a rare transition that is not reliably observed from n samples; the resulting occupancy difference accumulates over H steps, yielding a value separation proportional to $H\delta$. This is a standard Le Cam two-point method: the data distributions remain close, hence any test (and thus any algorithm) has error at least $\Omega(\delta)$ in identifying the instance, which translates into $\Omega(H\delta)$ value error. □

8.3 Information-theoretic limits for likelihood-free ratio estimation

Finally, we connect the critic-error terms $\bar{\epsilon}_T, \bar{\epsilon}_U$ to minimax lower bounds for density-ratio estimation and hypothesis testing. Even if overlap holds, learning the variational witnesses in the f-divergence dual is statistically nontrivial. Since T_ϕ (resp. U_ψ) is trained by distinguishing samples from two distributions (env vs model transitions in latent space; policy actions vs dataset actions), the problem contains binary classification as a special case.

Theorem 8.3 (Critic estimation lower bound (schematic))**.** *Let P, Q be two distributions over a common measurable space and suppose we observe n i.i.d. samples from each. Consider any estimator $\widehat{T}$ (measurable function of the samples) intended to approximate an optimal dual witness T^* for $D_f(P\|Q)$, or equivalently (for KL) the log density ratio $\log \frac{dP}{dQ}$ on the relevant support. Then there exists a pair (P, Q) with $\mathrm{TV}(P, Q) \asymp \eta$ such that, for any such estimator,*

$$\inf_{\widehat{T}} \sup_{(P,Q)} \mathbb{E}\Big[\big(\mathbb{E}_P[\widehat{T}] - \mathbb{E}_Q[f^*(\widehat{T})]\big) - D_f(P\|Q)\Big] \;\geq\; c\,\frac{\eta}{\sqrt{n}},$$

for a constant $c > 0$ depending only on the divergence family (and mild regularity conditions).

Interpretation. Theorem 8.3 states that the duality gap cannot, in general, be driven below $\Omega(n^{-1/2})$ uniformly over problem instances: distinguishing P from Q with small advantage requires $\Theta(1/\eta^2)$ samples. Since our end-to-end performance bound in Theorem 7.2 multiplies critic error by an effective horizon factor, this implies an unavoidable contribution of order $\Omega(H/\sqrt{n})$ in worst case (up to problem-dependent constants and occupancy weighting). In particular, likelihood-free access to samples does not eliminate statistical hardness: it merely changes *how* we estimate the correction (via variational classification rather than explicit likelihood ratios), but not the fundamental information requirements.

Taken together, the preceding results justify the structure of our guarantee. The overlap assumptions are not optional: without them, offline learning is information-theoretically ill-posed. Even with overlap, the critic terms represent irreducible estimation error unless $|D_{\mathrm{env}}|$ (and the number of informative model rollouts) grows, and the resulting value uncertainty must scale at least linearly in the effective horizon.

9 Experiments

We evaluate LF-SAR along five axes: (i) end-to-end performance on standard offline continuous-control benchmarks; (ii) sensitivity to the choice of implicit

world model and to the degree of model utilization; (iii) robustness under controlled out-of-distribution (OOD) model mismatch; (iv) calibration and stability of the learned variational critics that define the reward correction; and (v) comparisons to existing model-based offline RL methods and to likelihood-based ratio/discriminator variants when likelihoods are available.

9.1 Benchmarks, protocols, and metrics

We consider two families of benchmarks. First, we use D4RL-style continuous-control datasets, reporting normalized scores when available and raw discounted returns otherwise. These tasks span (a) dense-reward locomotion with varying data quality (e.g., "random", "medium", "medium-replay", "medium-expert" regimes), and (b) sparse-reward navigation/manipulation settings where extrapolation and compounding error are pronounced. Second, we use the NeoRL benchmark suite, which provides multiple offline datasets per underlying environment with controlled variation in coverage and reward noise; we use these as a convenient testbed for evaluating how the proposed correction behaves as overlap degrades.

Our training protocol follows the algorithmic access pattern assumed in the main development. We train an implicit dynamics model m_θ from D_{env} only, and then run the mixed offline+model training loop for a fixed number of iterations. Model rollouts are initiated from states sampled from D_{env}; this ensures that any benefit of imagination is due to composing m_θ rather than due to an optimistic initial-state distribution. We evaluate the learned policy $\hat{\pi}$ in the true environment dynamics p (with no model access) and average over 3–10 random seeds depending on the benchmark variance. Hyperparameters (critic learning rates, clipping thresholds, rollout horizon h, and mixture proportions of D_{env} vs. D_m) are selected via a fixed validation protocol based on held-out dataset splits or a small set of environment tasks, without per-task tuning.

9.2 Implicit world model: architecture and ablations

Since our guarantee and algorithm only require a *sampler* $s' \sim m_\theta(\cdot \mid s, a)$, we explicitly vary the modeling family to test whether the proposed correction is genuinely likelihood-free and not an artifact of a particular model class. Concretely, we instantiate m_θ as (i) a diffusion-style conditional generator (iterative denoising in state space or latent space), (ii) an autoregressive transformer over discretized state tokens, and (iii) a deterministic-plus-noise baseline (e.g., an ensemble regressor with Gaussian residual sampling). For each class we match overall parameter count and training compute as closely as possible.

We then ablate three aspects of model usage. First, we sweep the rollout horizon $h \in \{1, 5, 10, 25\}$ to probe the long-horizon imagination regime

where compounding mismatch is expected. Second, we vary the number of model samples per real transition used to populate D_m, which controls how strongly the policy optimizer is exposed to model-generated data. Third, we compare single-step model sampling (freshly seeded from D_{env} at each step) against multi-step chained rollouts, which isolates whether LF-SAR primarily corrects one-step shift or can tolerate accumulation over several steps.

In each ablation we report not only policy performance but also a direct proxy for model mismatch in representation space: an empirical estimate of the learned f-divergence dual objective evaluated on a held-out split of (z, a, z') tuples. The intended qualitative prediction is that longer horizons and heavier reliance on D_m increase mismatch, and that the learned transition critic T_ϕ should respond by reducing the effective utility of model transitions via its correction term.

9.3 OOD mismatch suite

To stress-test the shift correction beyond the nominal i.i.d. regime, we introduce controlled OOD conditions in which m_θ is deliberately biased relative to p while keeping the offline dataset fixed. We consider three mismatch mechanisms.

Data-induced mismatch. We train m_θ on strict subsets of D_{env} (e.g., removing high-velocity or near-terminal transitions), yielding a model that is accurate on a restricted region but systematically wrong elsewhere. This simulates limited model coverage and is directly relevant to the overlap assumptions in our theory.

Perturbation-induced mismatch. We post-compose the model sampler with known perturbations (e.g., additive disturbances in selected state dimensions, or action scaling) that preserve marginal plausibility but shift conditional dynamics. This creates a regime where naive model-based rollouts can be harmful even though one-step predictions appear reasonable.

Representation-induced mismatch. We vary the encoder f (random features, contrastive learning on D_{env}, and jointly learned end-to-end encoders) to test whether mismatch is more or less detectable in certain latent spaces. Since our correction operates on pushforward measures p_f and $m_{\theta,f}$, this isolates the dependence of the method on representation sufficiency.

For each mismatch setting we compare LF-SAR to an *uncorrected* model-based pipeline that uses the same m_θ and the same policy optimization algorithm but with $\alpha = \beta = 0$. The key measurement is the performance degradation under increasing mismatch. We additionally report a diagnostic "mismatch–penalty curve": the empirical correlation between (a) critic scores

$T_\phi(z, a, z')$ on model transitions and (b) realized downstream return loss attributable to those transitions, estimated by reweighting trajectories by their proportion of model-originated steps.

9.4 Calibration and stability diagnostics

Since LF-SAR relies on critic outputs inside a corrected reward, stability is a first-order concern. We therefore track: (i) the distribution of T_ϕ and U_ψ logits over training, (ii) the frequency of clipping events when we enforce bounded corrections, (iii) gradient norms and the presence/absence of critic collapse (e.g., constant outputs), and (iv) the sensitivity of the learned policy to small changes in α, β.

We also perform calibration checks. In small tabular or low-dimensional continuous tasks where density ratios can be approximated by kernel methods or discretization, we compare T_ϕ to a ground-truth proxy for $\log \frac{dp_f}{dm_{\theta,f}}$ (in the KL case) or to the corresponding optimal f-divergence witness. In larger tasks, where ground truth is unavailable, we instead evaluate *held-out* discrimination performance (env vs model for T_ϕ; dataset actions vs policy actions for U_ψ) and compute reliability-style curves relating critic score quantiles to empirical classification odds. While such diagnostics do not prove correctness of the dual optimizer, they identify failure modes where the correction becomes numerically unstable or uninformative.

Finally, we report an "effective pessimism" statistic: the expected corrected reward gap between environment transitions and model transitions, $\mathbb{E}[\tilde{r} \mid D_{\text{env}}] - \mathbb{E}[\tilde{r} \mid D_m]$, which should increase when mismatch grows. This is the operational analogue of the theoretical intuition that the correction discourages reliance on unreliable model rollouts.

9.5 Baselines and likelihood-based variants

We compare to model-based offline RL methods designed to mitigate model bias, including MOPO, MOBILE, and SAMBO, instantiated with their recommended hyperparameters and (when applicable) with matched world-model capacity and rollout budgets. In addition, we include a likelihood-free discriminator baseline that uses a standard GAN-style classifier loss on (z, a, z') but does *not* correspond to an f-divergence dual with calibrated conjugates; this tests whether the specific variational form matters.

To isolate the benefit of likelihood-free training (as opposed to correction per se), we also construct likelihood-based counterparts in settings where tractable model likelihoods can be computed (e.g., Gaussian predictive models or normalizing-flow dynamics). In these cases we replace T_ϕ by the explicit log ratio $\log m_\theta(s' \mid s, a)$ (and, when possible, an estimate of $\log p(s' \mid s, a)$ on held-out data) to emulate SAR-like corrections. Comparing these variants to LF-SAR clarifies whether the learned variational witnesses

provide comparable behavior in regimes where explicit likelihood access is feasible.

Across all comparisons we keep the policy optimizer fixed (e.g., SAC-style off-policy updates) and vary only (i) how model-generated transitions are produced and (ii) how rewards are corrected. This isolates the empirical contribution of the proposed likelihood-free shift-aware correction.

10 Discussion and limitations

We discuss the regimes in which representation-space shift correction is meaningful, and we enumerate the principal limitations suggested by the theory and by the algorithmic design.

When are representation-space divergences informative? Our correction operates on pushforward transition measures $p_f(\cdot \mid s, a)$ and $m_{\theta,f}(\cdot \mid s, a)$ rather than on raw-state conditionals. This is advantageous when s is high-dimensional and the model is an implicit sampler, but it introduces an identifiability question: even if $D_f(p_f \| m_{\theta,f})$ is small, the true-state mismatch $D_f(p \| m_\theta)$ may be large if f forgets task-relevant aspects of the dynamics. Conversely, $D_f(p_f \| m_{\theta,f})$ may be large even when the induced values are insensitive to those differences. Thus, representation-space divergences are meaningful precisely insofar as f preserves *value-relevant* structure: informally, f should collapse nuisance variation while retaining the Markovian information needed for control. Our representation sufficiency assumption makes this explicit; absent such an assumption, the learned critic T_ϕ may penalize mismatches that are irrelevant to return, or fail to detect mismatches that are catastrophic for planning.

This tension suggests a practical guideline: the purpose of T_ϕ is not to estimate a physically faithful likelihood ratio, but rather to provide a witness for *model unreliability as it matters to the policy optimization pipeline.* In particular, if f is learned jointly with π and the critics, there is a risk of degenerate solutions in which f collapses information to make the discrimination task artificially easy or artificially hard (depending on optimization pressures), thereby distorting the correction. Mitigating such pathologies requires architectural and optimization choices (e.g., stop-gradient paths, auxiliary reconstruction/contrastive losses, or explicit regularization of f) that are not captured by the core guarantee.

Dependence on the encoder and "shortcut" features. The encoder dependence is not merely statistical; it is also geometric. The variational dual for an f-divergence is sensitive to the choice of feature space in which the critic T_ϕ operates. If f exposes features that strongly separate D_{env} from D_m for incidental reasons (e.g., artifacts of the model sampler, discretization

effects, or differences in preprocessing), then T_ϕ may learn a high-confidence discriminator that yields large corrections without corresponding semantic mismatch. Since these corrections enter the reward, this can lead to excessive pessimism and reduced model utilization. Conversely, if f hides the discriminative signal (e.g., by being too low-dimensional or by enforcing invariances that remove causal variables), the critic may be underpowered and the correction ineffective.

We view this as an instance of a more general issue in likelihood-free ratio estimation: one estimates density ratios *relative to a chosen σ-algebra*. The most robust strategy is therefore to (i) treat f as part of the hypothesis class whose adequacy must be validated, and (ii) incorporate diagnostics that explicitly test whether critic scores correlate with downstream performance degradation under model rollouts. Such diagnostics cannot fully certify correctness, but they can detect failure modes in which the correction becomes detached from control-relevant errors.

Reward sign, magnitude, and the log-reward variant. Several of our bounds and the corrected reward definition use $\log r(s,a)$, which requires $r_{\min} > 0$. This is a real restriction: many control benchmarks have rewards that are zero, negative, or shaped with additive constants. In practice, one may shift and scale rewards to enforce positivity, e.g.,

$$r^+(s,a) \;=\; \max\{r(s,a) - c,\; \varepsilon\},$$

with c chosen so that typical rewards are positive and $\varepsilon > 0$ for numerical stability, and then apply the log transform to r^+. However, such transformations change the control objective unless the downstream algorithm is invariant to monotone utilities (which standard RL objectives are not, in general). Alternatively, one can avoid the log transform and incorporate the critic terms as additive penalties to the original reward,

$$\tilde{r} \;=\; r(s,a) + \alpha\, T_\phi(\cdot) + \beta\, U_\psi(\cdot),$$

interpreting the correction as a regularizer rather than as an ELBO-like decomposition. This variant sacrifices the clean multiplicative/rationing interpretation available in the KL/SAR case, and the corresponding guarantees require different constants and, typically, boundedness assumptions on T_ϕ, U_ψ enforced by clipping. We therefore regard strict positivity as a limitation of the most direct theoretical instantiation, and an invitation to develop utility-robust versions where the chosen divergence naturally matches a utility function u without requiring $r > 0$.

Long-horizon imagination and multi-step mismatch accumulation. Our transition correction is learned from one-step discrimination between (z, a, z') drawn from D_{env} and from model rollouts. When we chain the

model for h steps, the distribution of (s_t, a_t) itself drifts, and the induced mismatch is no longer captured solely by a one-step divergence at the dataset state-action marginals. In effect, we face a compounding-error phenomenon: even if $D_f(p_f(\cdot \mid s, a) \| m_{\theta,f}(\cdot \mid s, a))$ is small on-support, the rollouts may visit regions where this divergence is large and where critic estimation is statistically weak. Our bounds reflect this through horizon factors (linear in an effective horizon H) and through occupancy-weighted mismatch terms; nevertheless, these bounds are pessimistic in regimes where errors cancel and optimistic in regimes where rare but severe model failures dominate return.

Algorithmically, this indicates two competing design choices. Short rollouts reduce compounding error but limit the ability to propagate rewards and discover long-term consequences in sparse settings. Long rollouts provide more synthetic data but require stronger overlap and more reliable critics. The practical implication is that h should be treated as a primary knob, ideally adapted during training based on measured mismatch (e.g., shrinking h when T_ϕ indicates large divergence). A more principled approach would learn *state-dependent* rollout truncation or uncertainty-aware branching, but this lies beyond our current scope.

Critic calibration, boundedness, and optimization coupling. The variational critics enter the reward; thus any instability in T_ϕ or U_ψ can directly destabilize actor-critic training. This necessitates clipping, normalization, or temperature scaling of critic outputs. Such interventions are practically essential but introduce an additional approximation layer not fully modeled by the idealized variational statement. Moreover, the critics and the policy are coupled: as π changes, the distribution of model rollouts changes, which changes the discrimination task and thus the reward shaping experienced by π. This feedback loop can, in principle, induce oscillations or exploitation of critic weaknesses (analogous to adversarial training pathologies). While our analysis accounts for critic error via additive ϵ_T, ϵ_U terms, it does not by itself guarantee stability of the joint learning dynamics.

Broader implications for foundation world models. A motivating use case is a large implicit "foundation" dynamics model—for instance, a diffusion or transformer generator trained broadly across tasks—used as a simulator for downstream offline RL. In this regime, likelihood-free correction is not merely convenient; it may be necessary because exact likelihoods are unavailable or meaningless due to latent-variable structure and approximate inference. Our perspective suggests that one should treat such a world model as providing *plausible rollouts* rather than a calibrated probabilistic model, and rely on discriminative critics to modulate trust in those rollouts relative to the offline evidence.

At the same time, the foundation-model setting exacerbates our limita-

tions: (i) the encoder f may be inherited from pretraining and misaligned with control, (ii) the model may generate artifacts that are easy to discriminate but not value-relevant, and (iii) the mismatch may be highly non-uniform across the state space. These considerations point toward hybrid systems in which representation learning, discriminative shift estimation, and conservative policy optimization are co-designed, rather than treated as modular components. Our contribution is a step in this direction, but we emphasize that strong performance in the foundation-model regime will likely require additional structure (e.g., task-conditioned representations, calibrated uncertainty estimates, or explicit constraints on rollout support) beyond the present likelihood-free divergence correction alone.