UG-SRDP: Calibrated Uncertainty-Gated Diffusion Policies for Offline RL under State Distribution Shift

Liz Lemma Future Detective

January 20, 2026

Abstract

Diffusion policies have emerged as an expressive policy class for offline reinforcement learning (RL), but Q-guided diffusion methods can become brittle under state distribution shift: in out-of-distribution (OOD) states, miscalibrated critics may drive the diffusion sampler toward catastrophic actions. Building on SRDP (State Reconstruction for Diffusion Policies), which improves OOD generalization by injecting a state-reconstruction objective at every diffusion timestep, we propose UG-SRDP: a calibrated uncertainty-gated diffusion policy framework that uses reconstruction-derived OOD scores to modulate (i) critic guidance strength, (ii) imitation regularization weight, and optionally (iii) sampling budget. We formalize OOD-robust offline control as maximizing return while controlling OOD-induced risk, and prove bounds in bandit/linear-MDP settings showing that state-dependent gating converts unbounded OOD critic error into a bounded additive performance penalty proportional to the probability of encountering far-OOD states, with matching impossibility lower bounds absent coverage assumptions. We further provide finite-sample calibration guarantees for the gating rule via conformal prediction on reconstruction residuals. Experiments (recommended) on missing-data maze navigation, controlled region-removal benchmarks, and real-robot multimodal manipulation would validate that UG-SRDP improves stability and safety over Diffusion-QL and SRDP by automatically reducing reliance on critic guidance when states fall outside the dataset support.

Table of Contents

sampling procedure.

3. 3. Problem Formulation (OOD-safe offline control): define covariate shift in state distributions, risk/catastrophe modeling, and desiderata (improve near-support, avoid catastrophic OOD actions).

4. 4. UG-SRDP Algorithm: (i) learn SRDP with probabilistic/ensemble reconstruction, (ii) define OOD score $u(s)$, (iii) calibrate threshold(s) via conformal methods, (iv) define gating schedules for $\eta(s), \lambda(s)$ and optional sampler switching/budgeting.

5. 5. Theory I (Upper Bounds): performance decomposition into in-support improvement and far-OOD penalty; show gating scales critic-error contribution by $\mathbb{E}[\eta(s)]$ and yields an additive term $\propto \mathbb{P}[u(s) \geq \tau]$; special-case contextual bandits and linear MDPs.

6. 6. Theory II (Lower Bounds / Impossibility): unidentifiability of OOD values in offline RL; show any method must incur $\Omega(\mathbb{P}[\text{far-OOD}])$ loss without coverage; interpret gating as optimal (up to constants) among safe methods.

7. 7. Calibration Guarantees: split conformal thresholds for reconstruction residuals; bounds on false-positive and abstention rates under $d_{\mathcal{D}}$; discussion of autoencoder-OOD paradox and how probabilistic decoders/ensembles help.

8. 8. Complexity & Systems: compute cost vs. safety; expected sampling steps with gated sampler switching; practical considerations for real-time robotics.

9. 9. Experimental Plan (recommended to strengthen): missing-goal Maze2D, synthetic region removal across D4RL tasks, critic-miscalibration stress tests, real-robot 'forbidden zones'; ablations for calibration, gating shape, and decoder type.

10. 10. Discussion & Limitations: when gating helps/hurts; dependence on calibration validity; interaction with partial observability and vision; relation to pessimism/conservative offline RL.

11. 11. Conclusion: summary of theoretical and practical contributions; future directions (vision, trajectory diffusion, consistency distillation).

1 Introduction

Offline reinforcement learning seeks to compute a policy from a fixed dataset $\mathcal{D}$ of transitions collected by an unknown behavior policy π_β, without further interaction with the underlying Markov decision process $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \rho_0, \gamma)$. The central obstruction is support mismatch: at deployment, a learned policy may visit states and select actions that are poorly represented in $\mathcal{D}$, while the learned value function (or critic) is only constrained by the data on that limited support. When function approximation and bootstrapping are combined, such mismatch can produce uncontrolled value extrapolation and, consequently, unstable policy improvement. Since rewards are bounded ($|r(s, a)| \leq R_{\max}$), the performance degradation can be small if the policy remains near the data distribution, but it can become significant when the policy is induced to act in regions where the critic error is unconstrained.

Diffusion policies have recently emerged as a convenient class of expressive, state-conditional action generators in offline RL. In this view, for each state s the policy $\pi_\theta(\cdot \mid s)$ is implemented by a denoising diffusion model over actions, with a forward corruption process producing noisy actions a_t and a learned reverse process predicting noise (or an equivalent denoising signal) over T timesteps. Such policies inherit two distinct modes of training: a behavior cloning component that matches the conditional action distribution observed in $\mathcal{D}$, and a value-driven component that adjusts sampling or training via a learned critic Q_ϕ so as to prefer higher-value actions. The second component is the source of both potential improvement and potential failure: if Q_ϕ is accurate on the states and actions under consideration, guidance can improve upon imitation; if Q_ϕ is inaccurate off-support, guidance can amplify error by systematically steering samples into actions whose high predicted value is an artifact of extrapolation.

A natural response is to regularize the policy toward the dataset; however, uniform regularization does not address the fact that the degree of support mismatch is state dependent. In particular, the same policy update that is beneficial in well-covered regions may be harmful in states that are rare, novel, or absent from $\mathcal{D}$. We therefore seek a mechanism that decides, from the observed state alone, when critic guidance should be trusted and when it should be suppressed. This mechanism must be computable strictly offline, must be compatible with diffusion sampling, and must admit a quantitative guarantee that isolates the unavoidable price of encountering truly out-of-distribution states at test time.

We build on a dual-head architecture, SRDP, in which a shared representation module f_ϕ feeds (i) a diffusion head f_θ used to model $\pi_\theta(\cdot \mid s)$ and (ii) a reconstruction head f_ψ trained to predict the state (or a distribution over the state) from the same representation. The reconstruction head is not used to act directly; rather, it provides a diagnostic of whether the current state resembles those present in $\mathcal{D}$. Concretely, we define an OOD score

$u_\psi(s)$ via a residual, a negative log-likelihood, or an ensemble uncertainty measure derived from f_ψ. This score is then mapped to state-dependent coefficients $(\eta(s), \lambda(s))$ that modulate, respectively, critic guidance and (optionally) behavior regularization in the diffusion policy update and/or sampling procedure. The key structural constraint is monotonicity: as $u_\psi(s)$ increases, we do not increase reliance on the critic, and beyond a threshold τ we set $\eta(s) = 0$ (a hard gate), thereby reverting to a behavior-cloned diffusion policy in those states.

The contribution of this design is that it converts an uncalibrated notion of "being out of distribution" into an explicit control signal with a finite-sample calibration guarantee. We select τ using split conformal calibration on a held-out subset of $\mathcal{D}$ at miscoverage level α, so that under the data distribution $d_\mathcal{D}$ the event $u_\psi(s) \leq \tau$ holds with probability at least $1 - \alpha$ for a fresh in-distribution state. This calibration does not require parametric assumptions on the score distribution and yields a principled bound on false-positive gating (unnecessary disabling of guidance) on in-distribution data. The remaining failure mode is then concentrated on states that genuinely lie outside the support of $\mathcal{D}$, for which offline identifiability is impossible without additional assumptions.

Our thesis is that reconstruction-based gating yields a robust form of "safe improvement": on in-support states, critic guidance can be used while its effect is controlled by the guidance coefficient, and on far-OOD states guidance is disabled so that critic error cannot be adversarially amplified. The resulting performance bounds decompose into (i) a term proportional to the critic error on the gated in-support set, scaled by $\mathbb{E}[\eta(s)]$, and (ii) an additive term proportional to the probability of encountering states that trigger the gate, scaled by $R_{\max}$ (and by $(1 - \gamma)^{-1}$ in MDPs). Moreover, we emphasize that an additive dependence on the mass of far-OOD states is information-theoretically unavoidable in offline learning: when test-time contexts or states have no coverage in $\mathcal{D}$, one can construct indistinguishable instances of the environment whose optimal actions disagree there, implying a matching $\Omega(p_{\mathrm{OOD}})$ lower bound on achievable regret.

Beyond the theoretical motivation, the gating mechanism is operationally useful for diffusion sampling. Since sampling cost scales with the number of denoising steps, the same state-dependent signal that modulates $\eta(s)$ may also select a cheaper fallback sampler (e.g., pure behavior cloning diffusion, or fewer denoising steps) when $u_\psi(s)$ indicates novelty. Thus, the method targets both value stability and predictable compute: we spend guidance and sampling budget where the data support makes it meaningful, and we avoid expensive, brittle extrapolation where it is not.

The remainder of the paper formalizes these claims. We first review offline RL, diffusion policies, and the critic-guided diffusion objective, and we describe the SRDP dual-head parameterization used to produce both actions and OOD scores. We then define the uncertainty-gated policy family and present bounds that separate in-support critic error from the probability of encountering gated states, together with the conformal calibration guarantee for the threshold τ and a matching lower bound capturing offline unidentifiability under support shift.

2 Preliminaries

Offline reinforcement learning. We work in a discounted MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \rho_0, \gamma)$ with $\gamma \in (0,1)$ and bounded rewards $|r(s,a)| \le R_{\max}$. An offline dataset $\mathcal{D} = \{(s_i, a_i, r_i, s_i')\}_{i=1}^N$ is collected by an unknown behavior policy π_β and fixed thereafter. Our goal is to learn a policy π maximizing the discounted return

$$J(\pi) \;=\; \mathbb{E}\Big[\sum_{t\ge 0} \gamma^t r(s_t, a_t)\Big], \qquad s_0 \sim \rho_0,\; a_t \sim \pi(\cdot \mid s_t),\; s_{t+1} \sim P(\cdot \mid s_t, a_t),$$

using only samples from $\mathcal{D}$. Throughout, d_π denotes the discounted state occupancy induced by π. Since $\mathcal{D}$ provides information only on the state–action support visited by π_β, our analysis and algorithmic choices explicitly track which parts of $\mathcal{S}$ are well supported by the data and which are not.

Diffusion policies over actions. A diffusion policy specifies a conditional generative model over actions given a state. Fix a number of diffusion steps $T \in \mathbb{N}$. The forward diffusion (corruption) process constructs a sequence of noisy actions $(a_t)_{t=0}^T$ via

$$q(a_t \mid a_{t-1}) = \mathcal{N}\Big(\sqrt{1-\beta_t}\, a_{t-1},\; \beta_t I\Big), \qquad t = 1, \dots, T,$$

with a variance schedule $\{\beta_t\}$ and a_0 interpreted as the clean action. Equivalently, $a_t = \sqrt{\bar\alpha_t}\, a_0 + \sqrt{1-\bar\alpha_t}\, \epsilon$ where $\epsilon \sim \mathcal{N}(0, I)$ and $\bar\alpha_t = \prod_{k=1}^t (1-\beta_k)$. The reverse process defines the policy via a learned denoiser predicting the forward noise, $\hat\epsilon = f_\theta(s, a_t, t)$, yielding a parameterized transition $p_\theta(a_{t-1} \mid a_t, s)$. Sampling an action from $\pi_\theta(\cdot \mid s)$ proceeds by drawing $a_T \sim \mathcal{N}(0, I)$ and iterating the reverse transitions to obtain a_0, which is then executed.

Training typically uses the diffusion denoising score matching objective (a conditional DDPM loss). For $(s, a_0) \sim \mathcal{D}$, $t \sim \mathrm{Unif}(\{1, \dots, T\})$, and $\epsilon \sim \mathcal{N}(0, I)$, define $a_t = \sqrt{\bar\alpha_t} a_0 + \sqrt{1-\bar\alpha_t}\epsilon$, and minimize

$$L_{\mathrm{DP}}(\theta) \;=\; \mathbb{E}\big[\|\epsilon - f_\theta(s, a_t, t)\|_2^2\big],$$

which corresponds to behavior cloning in the sense that the induced $\pi_\theta(\cdot \mid s)$ matches the dataset conditional action distribution.

Critic learning and Diffusion-QL style guidance. To obtain improvements over imitation, we learn a critic $Q_\phi(s,a)$ from $\mathcal{D}$ by minimizing a Bellman error. Concretely, with a target network $\bar{\phi}$ and (optionally) double critics, we form targets using actions sampled from the current diffusion policy:

$$y \;=\; r(s,a)+\gamma\,\mathbb{E}_{a'\sim\pi_\theta(\cdot|s')}\big[Q_{\bar{\phi}}(s',a')\big], \qquad L_Q(\phi)=\mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\big[(Q_\phi(s,a)-y)^2\big].$$

This update is purely offline: the only distributional dependence beyond $\mathcal{D}$ enters through policy-sampled actions $a' \sim \pi_\theta(\cdot \mid s')$, which are generated by the learned diffusion model.

Diffusion-QL and related methods then incorporate the critic into the policy update or into the sampling rule. We adopt an abstract formulation in which the diffusion loss is augmented by a value-seeking term applied to the sampled clean action a_0 produced by the reverse process. Writing $a_0 = a_0(s;\theta)$ for a stochastic sample from the denoiser chain, a representative objective is

$$L_\pi(\theta) \;=\; L_{\mathrm{DP}}(\theta) \;-\; \eta_0\,\widetilde{\mathbb{E}}_{s\sim\mathcal{D},\,a_0\sim\pi_\theta(\cdot|s)}\big[Q_\phi(s,a_0)\big],$$

where $\eta_0 \geq 0$ is a guidance strength and $\widetilde{\mathbb{E}}$ indicates that the expectation may be estimated using the same noise and timestep samples as in L_{DP}. This term biases the diffusion model toward generating actions assigned higher value by the critic, while retaining the denoising objective that anchors the policy to the data.

SRDP dual-head architecture and OOD scoring. We use a dual-head parameterization in which a shared representation module (the "trunk") f_ϕ maps states to a latent representation $z = f_\phi(s)$. The diffusion head consumes (z, a_t, t) to predict noise, $f_\theta(z,a_t,t) \approx \epsilon$, implementing π_θ. In parallel, a reconstruction head f_ψ predicts the state (or a distribution over it) from the same representation, e.g.,

$$\hat{s} = f_\psi(z) \qquad \text{or} \qquad p_\psi(s \mid z),$$

trained by a reconstruction loss $L_{\mathrm{R}}(\psi)$ such as $\|s-\hat{s}\|_2^2$ or a negative log-likelihood. The reconstruction head does not act; it induces an OOD score $u_\psi(s)$ used later for gating. Typical instantiations include (i) a deterministic residual $u_\psi(s) = \|s - f_\psi(f_\phi(s))\|$, (ii) a probabilistic score $u_\psi(s) = -\log p_\psi(s \mid f_\phi(s))$, or (iii) an ensemble-based uncertainty estimate computed from multiple decoders sharing the trunk.

Critic-guided diffusion sampling. At deployment, we may incorporate Q_ϕ directly into the reverse diffusion steps. Abstractly, each reverse transition produces a mean action update $\mu_\theta(s,a_t,t)$ and noise scale σ_t. A critic-guided sampler modifies the denoising direction by a term proportional to

$\nabla_{a_t} Q_\phi(s, a_t)$ (or $\nabla_{a_0} Q_\phi(s, a_0)$ via the reparameterization between a_t and a_0), yielding an update of the schematic form

$$a_{t-1} \;=\; \mu_\theta(s, a_t, t) \;+\; \sigma_t \xi \;+\; \eta(s)\, g_t \,\nabla_{a_t} Q_\phi(s, a_t), \qquad \xi \sim \mathcal{N}(0, I),$$

where g_t is a known scale factor depending on the diffusion schedule and $\eta(s)$ is a state-dependent guidance coefficient. This sampling rule is the mechanism by which critic errors can be amplified off-support: when Q_ϕ extrapolates, the gradient term can steer the chain into regions of $\mathcal{A}$ that were never anchored by L_{DP} on the relevant states. The next section formalizes this failure mode under covariate shift and motivates our statewise gating of $\eta(s)$ using the reconstruction-derived score $u_\psi(s)$.

3 Problem Formulation: OOD-safe Offline Control

We consider strictly offline learning in $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \rho_0, \gamma)$ from a fixed dataset $\mathcal{D}$ collected by an unknown behavior policy π_β. The central difficulty is that the distribution of states encountered at deployment is generally not the same as the marginal state distribution implicit in $\mathcal{D}$. In particular, once a learned policy π is executed in the environment, its discounted occupancy d_π is determined by both the induced action choices and the transition kernel P, and can place nontrivial mass on states that are rare or absent in $\mathcal{D}$. This mismatch is the offline analogue of covariate shift, and it is precisely the regime in which critic-guided generative policies are vulnerable to harmful extrapolation.

State-distribution shift and support mismatch. Let $d_{\mathcal{D}}$ denote the (discounted) state distribution associated with the dataset, e.g., the empirical state marginal of $\mathcal{D}$ viewed as a proxy for the occupancy of π_β. At evaluation time, we allow a test-time state distribution d_{test} (or, in the sequential setting, the endogenous d_π induced by the deployed policy) that may differ from $d_{\mathcal{D}}$. We use "out-of-distribution" (OOD) to refer to those states that are insufficiently supported by $\mathcal{D}$ in the sense relevant to our learned models (representation, critic, and diffusion policy). Since exact support is not observable from finite data in high dimension, we do not assume access to an indicator of membership in $\text{supp}(d_{\mathcal{D}})$; rather, we assume we can compute a real-valued score $u_\psi : \mathcal{S} \to \mathbb{R}$ from the SRDP reconstruction head, with larger values indicating greater novelty relative to $\mathcal{D}$. For a threshold τ, we define the induced in-support set

$$\mathcal{S}_{\text{in}}(\tau) \;=\; \{s \in \mathcal{S} : u_\psi(s) < \tau\},$$

and interpret its complement as the region in which both value estimation and critic-guided sampling are unreliable.

Risk and catastrophe modeling. Beyond return maximization, we explicitly track OOD risk. We allow two equivalent ways to encode unsafe behavior. First, we may specify a catastrophe set $\mathcal{C} \subseteq \mathcal{S} \times \mathcal{A}$ containing state–action pairs that should be avoided, such as entering forbidden zones, violating joint limits, or commanding destabilizing torques. Second, we may specify a bounded cost $c : \mathcal{S} \times \mathcal{A} \to [0, 1]$ capturing graded notions of risk. In either case, we measure risk under the deployed policy π via quantities such as

$$\mathbb{P}_\pi\big[(s_t, a_t) \in \mathcal{C} \text{ for some } t \geq 0\big] \qquad \text{or} \qquad \mathbb{E}_\pi\Big[\sum_{t\geq 0} \gamma^t c(s_t, a_t)\Big].$$

We emphasize that, in the strict offline setting, these risks cannot be directly optimized against the true environment; instead, we require a mechanism that mitigates the specific failure mode of critic-guided diffusion when the policy visits states outside the effective support of $\mathcal{D}$.

Guided diffusion as a source of OOD amplification. Our starting point is the empirical observation (and the theoretical fact in stylized settings) that a learned critic Q_ϕ may be accurate on $\mathcal{S}_{\text{in}}(\tau)$ yet arbitrarily wrong on its complement. When such a critic is used to guide a generative policy—either by augmenting the training objective with a value term or by adding gradient-based guidance during reverse diffusion—the resulting policy can be driven toward actions that are high under Q_ϕ but poor under Q^*. This is benign when Q_ϕ is accurate, but it can be catastrophic when the state is far OOD, since the guidance term can overwhelm the behavior-cloning anchor and produce actions that the dataset provides no evidence for. Accordingly, our problem is not merely to learn an accurate critic, but to deploy the critic selectively based on a statewise estimate of whether the current state is within the reliable region of the offline data.

OOD-safe objective and desiderata. We seek a deployable policy π together with a gating rule g that converts the OOD score $u_\psi(s)$ into state-dependent coefficients $(\eta(s), \lambda(s))$ controlling, respectively, the strength of critic guidance and the strength of reconstruction/behavior regularization (where λ may be used in training variants or to choose a conservative sampler at test time). The core desiderata are: (i) *near-support improvement*: on states s that are effectively in-distribution, the learned policy should be allowed to deviate from pure imitation and exploit critic information to improve return, ideally achieving $J(\pi) > J(\pi_{\text{BC}})$ when the critic is sufficiently accurate; (ii) *far-OOD safety*: on states with large $u_\psi(s)$, the policy should revert to a behavior-regularized action generator so that critic extrapolation cannot induce unbounded degradation; and (iii) *graceful degradation under shift*: any unavoidable loss due to visiting OOD states should enter performance guarantees additively through the probability of encountering such states, rather than through uncontrolled critic error terms.

Formally, we restrict attention to gating rules satisfying the monotonicity constraint that $\eta(s)$ is non-increasing in $u_\psi(s)$, and we impose the hard-safety condition that $\eta(s) = 0$ whenever $u_\psi(s) \geq \tau$. Under this structure, the impact of critic error is confined to $\mathcal{S}_{\text{in}}(\tau)$, while the complement is handled by a safe fallback (e.g., the unguided diffusion policy π_{BC} or a cheaper conservative sampler).

Calibration as an offline control of false triggers. A remaining issue is that the threshold τ must be set without online interaction. We therefore require that the gate be calibratable from a held-out split of $\mathcal{D}$ so that, under $s \sim d_{\mathcal{D}}$, the probability of incorrectly declaring an in-distribution state as OOD is controlled at a user-specified level α. Concretely, we will choose τ by split conformal calibration applied to the scores $\{u_\psi(s)\}$ computed on the calibration split, thereby ensuring marginal coverage guarantees of the form $\mathbb{P}_{s \sim d_{\mathcal{D}}}[u_\psi(s) \leq \tau] \geq 1 - \alpha$. This provides an explicit tradeoff knob: smaller α reduces false-positive gating (and thus preserves near-support improvement), while larger α yields more conservative behavior under suspected shift.

These requirements define our OOD-safe offline control problem: maximize $J(\pi)$ under state-distribution shift while controlling risk, by using reconstruction-derived uncertainty to gate critic guidance. The next section instantiates this formulation as UG-SRDP by specifying the score construction, calibration procedure, and concrete schedules for $(\eta(s), \lambda(s))$, including optional sampler switching to respect a bounded inference budget.

4 UG-SRDP Algorithm: Uncertainty-Gated SRDP for Offline Control

We now specify UG-SRDP as a concrete instantiation of the preceding formulation. The algorithm has four components: (i) learn an SRDP-style diffusion policy together with a reconstruction model that yields a state-wise novelty score, (ii) convert reconstruction residual/uncertainty into an OOD score $u_\psi(s)$, (iii) calibrate a gate threshold τ from a held-out split via split conformal, and (iv) define a deployment-time rule that maps $u_\psi(s)$ to state-dependent guidance and regularization coefficients (and, optionally, a sampler switch and denoising-step budget).

(i) SRDP backbone with probabilistic/ensemble reconstruction. We use a diffusion policy $\pi_\theta(\cdot \mid s)$ parameterized by a shared trunk f_ϕ and a diffusion head f_θ predicting noise. In forward diffusion, we form noisy actions $\{a_t\}_{t=1}^{T}$ from dataset actions by injecting Gaussian noise; in reverse diffusion, we denoise from a_T to a_0 using $f_\theta(f_\phi(s), a_t, t)$. In parallel, we train a reconstruction head f_ψ sharing f_ϕ to predict (or model) the state. Concretely, we allow either a deterministic decoder $\hat{s} = f_\psi(f_\phi(s))$ trained by an ℓ_2 loss, or a probabilistic decoder $p_\psi(s \mid z)$ with $z = f_\phi(s)$ trained by negative log-likelihood. To increase sensitivity to epistemic uncertainty, we may use an ensemble $\{f_{\psi^{(e)}}\}_{e=1}^{E}$ and/or an ensemble of critics; this changes only constant factors in cost and does not alter the gate definition below.

Policy learning follows the SRDP/Diffusion-QL template: we combine a diffusion behavior cloning objective $L_{\mathrm{DP}}(\theta, \phi)$ (e.g., denoising score matching) with reconstruction regularization $L_{\mathrm{R}}(\psi, \phi)$ and critic guidance through a learned Q_ϕ (in practice two critics to reduce overestimation). Writing $\widetilde{\mathbb{E}}$ for an empirical minibatch average from $\mathcal{D}$ and $a_0 \sim \pi_\theta(\cdot \mid s)$ for the denoised action, a representative objective is

$$L_{\mathrm{BC}}(\theta, \phi, \psi) \;=\; L_{\mathrm{DP}}(\theta, \phi) + \lambda_0 \, L_{\mathrm{R}}(\psi, \phi) - \eta_0 \, \widetilde{\mathbb{E}}[Q_\phi(s, a_0)]\,,$$

where (η_0, λ_0) are base coefficients used during training. The critic(s) are trained by offline Bellman regression using $\mathcal{D}$ and target actions sampled from the current diffusion policy (as in Diffusion-QL), and the policy is updated by stochastic gradients of L_{BC}.

(ii) OOD score from reconstruction residual/uncertainty. After training, we define an OOD score $u_\psi : \mathcal{S} \to \mathbb{R}$ from the reconstruction head. The intended property is monotonicity with novelty: larger $u_\psi(s)$ should indicate that s is less consistent with the data manifold learned from $\mathcal{D}$. We support several interchangeable constructions:

1. *Deterministic residual:* $u_\psi(s) = \|s - \hat{s}\|$, where $\hat{s} = f_\psi(f_\phi(s))$.
2. *Likelihood score:* $u_\psi(s) = -\log p_\psi(s \mid f_\phi(s))$ when f_ψ outputs a conditional density.
3. *Ensemble dispersion:* if $\hat{s}^{(e)}$ are ensemble reconstructions, set

   $$u_\psi(s) = \frac{1}{E} \sum_{e=1}^{E} \|s - \hat{s}^{(e)}\| \;+\; \kappa \cdot \mathrm{Tr}\Big(\widehat{\mathrm{Var}}_e[\hat{s}^{(e)}]\Big)\,,$$

   with a tunable $\kappa \geq 0$.

We emphasize that UG-SRDP uses u_ψ only to gate reliance on the critic; we do not assume that u_ψ is a perfect support indicator, only that it correlates with regions where critic extrapolation is unreliable.

(iii) Split conformal calibration of the gate threshold. We set the hard-gating threshold τ using a calibration split from $\mathcal{D}$. Specifically, we partition $\mathcal{D} = \mathcal{D}_{\mathrm{train}} \cup \mathcal{D}_{\mathrm{cal}}$, train (ϕ, θ, ψ) on $\mathcal{D}_{\mathrm{train}}$, compute scores $\{u_\psi(s_i)\}_{i=1}^{m}$ on the states in $\mathcal{D}_{\mathrm{cal}}$, and choose

$$\tau \;=\; u_{(\lceil (m+1)(1-\alpha) \rceil)},$$

the corresponding order statistic. By the standard split conformal guarantee, for a fresh $s \sim d_{\mathcal{D}}$ independent of $\mathcal{D}_{\text{cal}}$ we have $\mathbb{P}[u_\psi(s) \le \tau] \ge 1 - \alpha$. Thus α directly controls the tolerated false-positive rate of declaring an in-distribution state as OOD (under $d_{\mathcal{D}}$), without requiring parametric assumptions on u_ψ.

(iv) Deployment rule: gating schedules and sampler switching. At test time, given a state s, we compute $u = u_\psi(s)$ and apply a gating rule g producing state-dependent coefficients. The default hard gate sets

$$\eta(s) = \eta_0\, \mathbf{1}[u < \tau], \qquad \lambda(s) = \lambda_0,$$

so critic guidance is used only on $\mathcal{S}_{\text{in}}(\tau) = \{s : u_\psi(s) < \tau\}$. We also allow a soft monotone schedule to avoid discontinuities, e.g.,

$$\eta(s) = \eta_0\, h\!\left(\frac{\tau - u}{\sigma}\right), \quad \text{with} \quad h(x) \in [0, 1] \text{ nondecreasing},$$

and a corresponding conservative regularization schedule such as $\lambda(s) = \lambda_0 + k\, u$ to bias the policy toward imitation as novelty grows. The monotonicity constraint $(u_1 \le u_2 \Rightarrow \eta(u_1) \ge \eta(u_2))$ is enforced by construction.

Finally, we optionally couple gating to the sampling procedure to respect bounded inference budgets and to provide an explicit fallback. If $u \ge \tau$, we either (a) disable guidance within the same sampler (set $\eta = 0$ during reverse diffusion), or (b) switch to a cheaper conservative sampler, e.g., an unguided behavior-cloning diffusion policy π_{BC} and/or a reduced number of denoising steps $T_{\text{BC}} < T_{\text{guided}}$. This yields a state-dependent compute profile

$$\mathbb{E}[T(s)] = T_{\text{guided}}\, \mathbb{P}[u < \tau] + T_{\text{BC}}\, \mathbb{P}[u \ge \tau],$$

while ensuring that far-OOD states do not activate critic-driven extrapolation. The resulting deployed policy is denoted π_{UG} and is fully specified by the learned SRDP components, the calibrated threshold τ, and the chosen schedules for $(\eta(s), \lambda(s))$.

5 Theory I (Upper Bounds): Gated Performance Decomposition

We formalize the intended effect of uncertainty gating as a decomposition of performance into (i) an *in-support* term in which critic guidance can improve over behavior cloning up to a controlled critic-error penalty, and (ii) a *far-OOD* term in which we assume no reliable information is available and thus bound loss only by reward boundedness. Throughout, we write $\mathcal{S}_{\text{in}}(\tau) = \{s : u_\psi(s) < \tau\}$ and $\mathcal{S}_{\text{out}}(\tau) = \mathcal{S} \setminus \mathcal{S}_{\text{in}}(\tau)$, and we consider gating rules with $\eta(s) = 0$ on $\mathcal{S}_{\text{out}}(\tau)$ and $\eta(\cdot)$ non-increasing in $u_\psi(\cdot)$.

A generic decomposition principle. Let π_{UG} denote the deployed uncertainty-gated diffusion policy, and let π_{BC} denote its unguided (pure imitation) counterpart. By the performance difference lemma, for any pair of policies π, π',

$$J(\pi)-J(\pi') = \frac{1}{1-\gamma}\,\mathbb{E}_{s\sim d_\pi}\Big[\mathbb{E}_{a\sim\pi(\cdot|s)}\big[A^{\pi'}(s,a)\big]\Big], \qquad A^{\pi'}(s,a) := Q^{\pi'}(s,a)-V^{\pi'}(s).$$

We apply this with $(\pi, \pi') = (\pi_{\mathrm{BC}}, \pi_{\mathrm{UG}})$ or $(\pi, \pi') = (\pi_{\mathrm{UG}}, \pi_{\mathrm{BC}})$ depending on which direction yields a convenient upper bound, and then split the resulting expectation over $\mathcal{S}_{\mathrm{in}}(\tau)$ and $\mathcal{S}_{\mathrm{out}}(\tau)$. On $\mathcal{S}_{\mathrm{out}}(\tau)$ we do not attempt to compare π_{UG} to any optimal policy: instead we invoke bounded rewards to obtain a worst-case value gap. Concretely, for any s and any policies π, π',

$$\big|V^{\pi}(s) - V^{\pi'}(s)\big| \le \frac{2R_{\max}}{1-\gamma},$$

and hence any contribution from $\mathcal{S}_{\mathrm{out}}(\tau)$ can be upper bounded by $(2R_{\max}/(1-\gamma)) \cdot \mathbb{P}_{s\sim d_\pi}[s \in \mathcal{S}_{\mathrm{out}}(\tau)]$.

In-support improvement with critic-error control. We now characterize how the guidance signal introduces a dependence on the critic error only through the *effective* guidance magnitude. Under (H1), on $\mathcal{S}_{\mathrm{in}}(\tau)$ we have $\sup_a |Q^*(s,a) - Q_\phi(s,a)| \le \varepsilon$. Under (H2), the SRDP-style diffusion update is L-Lipschitz with respect to the guidance coefficient in the sense that the induced change in the statewise expected true value is linear in $|\eta(s)|$. A convenient abstract form is: there exists C_{lip} such that for any $s \in \mathcal{S}_{\mathrm{in}}(\tau)$,

$$\Big|\mathbb{E}_{a\sim\pi_\eta(\cdot|s)}[Q^*(s,a)] - \mathbb{E}_{a\sim\pi_{\eta'}(\cdot|s)}[Q^*(s,a)]\Big| \le C_{\mathrm{lip}}\,|\eta(s) - \eta'(s)|.$$

Taking $\eta'(s) = 0$ and using that the guided policy is computed using Q_ϕ rather than Q^*, standard approximate-greediness arguments yield an additive error on $\mathcal{S}_{\mathrm{in}}(\tau)$ of order $C_{\mathrm{lip}}\eta(s)\varepsilon$ (constants depend on the precise parameterization of the guidance step and, in multi-step diffusion, on how guidance is aggregated across denoising steps). Consequently, the in-support contribution takes the form

$$\text{(in-support loss)} \;\lesssim\; C\,\mathbb{E}_{s\sim d_{\pi_{\mathrm{UG}}}}\big[\eta(s)\,\mathbf{1}\{s\in\mathcal{S}_{\mathrm{in}}(\tau)\}\big]\,\varepsilon \;\le\; C\,\mathbb{E}_{s\sim d_{\pi_{\mathrm{UG}}}}\big[\eta(s)\big]\,\varepsilon,$$

which isolates the dependence on critic error through the *average activated guidance* $\mathbb{E}_{d_{\pi_{\mathrm{UG}}}}[\eta(s)]$. In particular, hard gating gives $\mathbb{E}[\eta(s)] = \eta_0 \cdot \mathbb{P}[u_\psi(s) < \tau]$, while soft schedules interpolate continuously.

Contextual bandits as a one-step special case. In the contextual bandit setting (equivalently $\gamma = 0$), the occupancy reduces to the test context distribution d_{test}, and the above decomposition becomes particularly transparent. Writing $p_{\text{OOD}} := \mathbb{P}_{s \sim d_{\text{test}}}[u_\psi(s) \geq \tau]$, we obtain a bound of the schematic form

$$\mathbb{E}_{s \sim d_{\text{test}}}\big[V^*(s) - V^{\pi_{\text{UG}}}(s)\big] \;\leq\; C_1\, \eta_0\, \varepsilon \;+\; 2R_{\max}\, p_{\text{OOD}},$$

where the first term is the *critic-error amplification* term suppressed by gating (through η_0 and the Lipschitz sensitivity), and the second term is the *unavoidable* price of encountering contexts outside the calibrated support indicator.

Discounted MDP upper bound and interpretation. Combining the in-support and out-of-support contributions yields the bound

$$J(\pi_{\text{UG}}) \;\geq\; J(\pi_{\text{BC}}) - C_2\, \varepsilon\, \mathbb{E}_{s \sim d_{\pi_{\text{UG}}}}[\eta(s)] - \frac{2R_{\max}}{1-\gamma}\, \mathbb{P}_{s \sim d_{\pi_{\text{UG}}}}\big[u_\psi(s) \geq \tau\big],$$

with $C_2 = O((1-\gamma)^{-1})$ arising from the performance difference lemma and the Lipschitz stability constant. The key qualitative point is that hard gating removes any dependence on $\sup_{s \in \mathcal{S}_{\text{out}}(\tau)} |Q^*(s,a) - Q_\phi(s,a)|$, replacing it by a term proportional only to the probability of visiting $\mathcal{S}_{\text{out}}(\tau)$ under the deployed policy.

Linear MDP instantiation (optional). If we further specialize to a linear MDP (or linear contextual bandit) in which $Q^*(s,a) = \langle w^*, \phi(s,a) \rangle$ for a known feature map ϕ of dimension d, then standard offline least-squares analysis can yield an explicit in-support critic error bound $\varepsilon = \tilde{O}(\sqrt{d/m})$ on $\mathcal{S}_{\text{in}}(\tau)$ under a suitable restricted eigenvalue/coverage condition on the feature covariance within $\mathcal{S}_{\text{in}}(\tau)$. Substituting this ε into the preceding inequalities makes the tradeoff explicit: gating converts the dependence on ill-conditioned or unsupported regions into the occupancy-weighted quantity $\mathbb{P}_{d_{\pi_{\text{UG}}}}[u_\psi(s) \geq \tau]$, while preserving the usual $\tilde{O}(\sqrt{d/m})$ statistical rate on the calibrated in-support region.

6 Theory II (Lower Bounds / Impossibility): Unidentifiability of Far-OOD Values

The preceding guarantees necessarily contain an additive term proportional to the probability of visiting states outside the calibrated in-support set. We now justify that, without additional assumptions (coverage, realizability with known structure, smoothness across the state manifold, etc.), *no strictly-offline method can remove such a dependence.* The obstruction is information-theoretic: the offline dataset does not identify rewards and transitions on regions that are never (or essentially never) visited under the behavior distribution, hence the optimal action on those regions is not learnable.

A minimal statement of the obstacle. Fix any offline algorithm $\mathcal{A}$ mapping a dataset $\mathcal{D}$ to a deployed policy $\pi = \mathcal{A}(\mathcal{D})$ (possibly randomized). Let $\mathsf{Law}_{\mathcal{M}}(\mathcal{D})$ denote the distribution of datasets induced by running the unknown behavior policy in $\mathcal{M}$. If there exist two MDPs $\mathcal{M}_0, \mathcal{M}_1$ such that

$$\mathsf{Law}_{\mathcal{M}_0}(\mathcal{D}) = \mathsf{Law}_{\mathcal{M}_1}(\mathcal{D}),$$

then $\mathcal{A}$ cannot distinguish which instance generated the data. Consequently, $\mathcal{A}$ must output the same (distribution over) policy under either instance, and an adversary may choose the instance on which that policy performs poorly. To turn this indistinguishability into a quantitative regret lower bound, we force $\mathcal{M}_0$ and $\mathcal{M}_1$ to coincide on the data-support region and disagree only on a far-OOD region that is visited with nontrivial probability under the evaluation rollout.

A concrete lower bound (bandit form). We first record the one-step version, which isolates the statistical phenomenon without dynamical complications.

Theorem 6.1 (Unavoidable $\Omega(p_{\text{OOD}})$ loss in offline contextual bandits)**.** *There exist two contextual bandit instances $\mathcal{M}_0, \mathcal{M}_1$ with rewards in $[-R_{\max}, R_{\max}]$, a data distribution $d_{\mathcal{D}}$ over contexts, and an evaluation distribution d_{test} such that:*

1. *The induced offline data laws are identical:* $\mathsf{Law}_{\mathcal{M}_0}(\mathcal{D}) = \mathsf{Law}_{\mathcal{M}_1}(\mathcal{D})$ *for any sample size.*
2. *There is an OOD region $\mathcal{S}_{\text{out}}$ with $p_{\text{OOD}} := \mathbb{P}_{s\sim d_{\text{test}}}[s \in \mathcal{S}_{\text{out}}] > 0$ and $\mathbb{P}_{s\sim d_{\mathcal{D}}}[s \in \mathcal{S}_{\text{out}}] = 0$.*
3. *For any (possibly randomized) offline algorithm $\mathcal{A}$ outputting $\pi = \mathcal{A}(\mathcal{D})$, we have*

$$\max\Big\{\mathbb{E}_{s\sim d_{\text{test}}}\big[V^*_{\mathcal{M}_0}(s) - V^\pi_{\mathcal{M}_0}(s)\big],\ \mathbb{E}_{s\sim d_{\text{test}}}\big[V^*_{\mathcal{M}_1}(s) - V^\pi_{\mathcal{M}_1}(s)\big]\Big\} \ \geq\ R_{\max}\, p_{\text{OOD}}.$$

Proof sketch. Let $\mathcal{S}$ contain two disjoint subsets $\mathcal{S}_{\text{in}}$ and $\mathcal{S}_{\text{out}}$, and take the dataset contexts to be supported only on $\mathcal{S}_{\text{in}}$. Define two actions a_+, a_-. On $\mathcal{S}_{\text{in}}$, set $r(s, a_+) = r(s, a_-) = 0$ in both instances. On $\mathcal{S}_{\text{out}}$, swap the rewards:

$$r_0(s, a_+) = +R_{\max}, \quad r_0(s, a_-) = -R_{\max}, \qquad r_1(s, a_+) = -R_{\max}, \quad r_1(s, a_-) = +R_{\max}.$$

Because $\mathcal{S}_{\text{out}}$ is never observed in $\mathcal{D}$, the joint law of (s, a, r) under the behavior policy is identical in $\mathcal{M}_0$ and $\mathcal{M}_1$. Hence π (as a function of $\mathcal{D}$) has the same distribution under both instances. For any fixed context $s \in \mathcal{S}_{\text{out}}$, the policy places some probability on a_+; whichever action it favors is optimal in one instance and suboptimal in the other, incurring instantaneous regret at least $R_{\max}$ at that s. Averaging over $s \sim d_{\text{test}}$ yields the stated $R_{\max} p_{\text{OOD}}$ lower bound.

Extension to discounted MDPs. The same construction yields a discounted lower bound by embedding the bandit ambiguity behind a transition that is never seen in the dataset but is reached under evaluation with probability comparable to an OOD occupancy mass. Concretely, we may take an initial region on which the dataset is collected and create a transition into an absorbing OOD state s_{out} that occurs with probability p_{OOD} under the evaluation rollout. In s_{out}, the two instances again swap which action yields $+R_{\max}$ versus $-R_{\max}$ at every step, so that the per-step value gap is amplified by the geometric sum. One obtains, for a universal constant $c > 0$,

$$\max\{J^*_{\mathcal{M}_0} - J^\pi_{\mathcal{M}_0},\ J^*_{\mathcal{M}_1} - J^\pi_{\mathcal{M}_1}\} \ \geq \ c\, \frac{R_{\max}}{1-\gamma}\, p_{\text{OOD}}.$$

The key point is that the dataset law is unchanged by construction, since the OOD transition and the OOD region have zero probability under the behavior distribution.

Interpretation: what gating can and cannot accomplish. The lower bound isolates the best-possible guarantee one may hope for in the absence of further structure: *only the probability of encountering far-OOD states can appear*, multiplied by the worst-case return scale $R_{\max}/(1-\gamma)$. In particular, any method that claims uniform improvement on far-OOD regions without assumptions contradicts Theorem 6.1 (and its discounted analogue). From this perspective, the role of uncertainty gating is not to “solve” OOD decision-making, but to ensure that the deployed procedure does not incur *additional* unidentifiable loss through extrapolative critic guidance. Once guidance is disabled on $\{u_\psi \geq \tau\}$, the remaining worst-case degradation is precisely of the unavoidable form above, up to constants and the realized OOD occupancy under deployment.

This also clarifies what remains to be controlled algorithmically: not the values on the far-OOD region (which are unidentifiable), but rather (i) the false-negative rate of the gate (entering far-OOD while still applying guidance), and (ii) the induced OOD occupancy of the deployed policy. The next section addresses (i) via finite-sample calibration of the threshold τ under $d_{\mathcal{D}}$, thereby quantifying how often the gate activates on in-distribution states and providing a principled means of trading off abstention against risk.

7 Calibration Guarantees: Split Conformal Thresholds for Reconstruction-Based OOD Scores

We now formalize how we set the gating threshold τ from data, and what (limited, but distribution-free) guarantees this calibration provides. Throughout, we treat $u_\psi : \mathcal{S} \to \mathbb{R}$ as an arbitrary measurable score produced by the reconstruction head (possibly using an ensemble), with the convention that larger $u_\psi(s)$ indicates "more OOD." The gate activates (disables critic guidance, and optionally switches to a BC sampler) when $u_\psi(s) \geq \tau$.

Split calibration protocol. We split the dataset $\mathcal{D}$ into a training part $\mathcal{D}_{\mathrm{tr}}$ and a calibration part $\mathcal{D}_{\mathrm{cal}} = \{s_i\}_{i=1}^m$, where the s_i are states (or observations) treated as i.i.d. samples from the data state distribution $d_\mathcal{D}$.[1] We train $(f_\phi, f_\theta, f_\psi)$ on $\mathcal{D}_{\mathrm{tr}}$, then compute scores $u_i := u_\psi(s_i)$ on $\mathcal{D}_{\mathrm{cal}}$. For a miscoverage level $\alpha \in (0, 1)$, we define

$$k := \lceil (m+1)(1-\alpha) \rceil, \qquad \tau := u_{(k)},$$

where $u_{(1)} \leq \cdots \leq u_{(m)}$ are the order statistics of the calibration scores. This is precisely the split conformal quantile (with the standard $(m + 1)$ correction) and requires no parametric assumptions on u_ψ.

Theorem 7.1 (Distribution-free control of false-positive gating under $d_\mathcal{D}$). *Assume $(s_1, \ldots, s_m, s)$ are exchangeable draws from $d_\mathcal{D}$, where s is an independent fresh draw and τ is computed as above from $\{u_\psi(s_i)\}_{i=1}^m$. Then*

$$\mathbb{P}_{s\sim d_\mathcal{D}}\big[u_\psi(s) \leq \tau\big] \;\geq\; 1-\alpha, \qquad \textit{equivalently} \qquad \mathbb{P}_{s\sim d_\mathcal{D}}\big[u_\psi(s) \geq \tau\big] \;\leq\; \alpha.$$

In particular, under $d_\mathcal{D}$ the hard gate $g(s) = \mathbf{1}[u_\psi(s) \geq \tau]$ has a marginal false-positive (abstention) rate at most α.

Interpretation as an abstention bound. Theorem 7.1 should be read as a statement about how often we will disable guidance on states distributed like the offline data. Since guidance is the mechanism that can amplify critic error, this calibration gives a concrete knob: increasing α raises guidance usage (fewer abstentions), while decreasing α yields a more conservative deployment rule. Notably, the guarantee is *finite-sample* and depends only on exchangeability, not on the correctness of the reconstruction model.

[1] In sequential datasets, exact i.i.d. sampling is not literal; in practice we subsample widely separated time indices or treat the empirical distribution over logged states as approximately exchangeable for calibration purposes. The formal statement below assumes exchangeability of the calibration scores.

What calibration does *not* guarantee. First, the bound is marginal rather than conditional: it does not assert that $\mathbb{P}[u_\psi(s) \geq \tau \mid s \in \text{some subgroup}]$ is controlled for every subgroup (indeed this is impossible without further assumptions). Second, the guarantee is with respect to $d_\mathcal{D}$, not an arbitrary shifted evaluation distribution d_{test}. Under shift, the abstention probability

$$p_{\text{gate}}(d_{\text{test}}) := \mathbb{P}_{s \sim d_{\text{test}}}\big[u_\psi(s) \geq \tau\big]$$

may be larger (sometimes substantially), and this is precisely the quantity that appears in the performance bounds through the OOD occupancy term. Thus, calibration controls false positives on-distribution, while deployment-time abstention reflects the actual shift.

Beyond deterministic residuals: the autoencoder–OOD paradox. If we take $u_\psi(s) = \|s - \hat{s}\|$ for a deterministic decoder $\hat{s} = f_\psi(f_\phi(s))$, then a well-known pathology is that low reconstruction error need not imply "in-distribution." Highly expressive autoencoders may reconstruct OOD inputs almost as well as in-distribution ones (because the decoder learns a near-identity map on a large region of input space), while bottlenecked autoencoders may distort rare but legitimate in-distribution states, producing spuriously large residuals. From the viewpoint of gating, these effects translate into (i) false negatives, where far-OOD states are not gated (dangerous, since critic guidance remains active), and (ii) excessive false positives on infrequent in-distribution modes (undesirable, since we unnecessarily disable guidance).

This motivates using scores that expose *uncertainty* rather than mere reconstruction error. Two practical refinements are particularly compatible with SRDP.

Probabilistic reconstruction and calibrated likelihood scores. We may let the reconstruction head output a conditional density $p_\psi(s \mid z)$ with $z = f_\phi(s)$, e.g., a diagonal Gaussian with mean $\mu_\psi(z)$ and variance $\sigma^2_\psi(z)$, and define

$$u_\psi(s) := -\log p_\psi(s \mid f_\phi(s)).$$

In this case, the score penalizes not only large residuals but also confident misreconstructions: for Gaussian decoders, $u_\psi(s)$ contains a variance-normalized squared error term. This mitigates the tendency of a deterministic decoder to "explain away" atypical inputs by projecting them onto a nearby high-density region without reflecting epistemic uncertainty. Split conformal calibration applies verbatim to u_ψ as a real-valued score, irrespective of whether the density model is well-specified.

Ensemble and epistemic scores. Alternatively, we train an ensemble $\{f_{\psi^{(e)}}\}_{e=1}^{E}$ sharing the SRDP trunk f_ϕ but using independent decoder heads, and define $u_\psi(s)$ via predictive dispersion, e.g.,

$$u_\psi(s) := \frac{1}{d}\sum_{j=1}^{d} \mathrm{Var}_{e\in[E]}\Big(\mu_{\psi^{(e)},j}(f_\phi(s))\Big),$$

or via disagreement in negative log-likelihood. Such ensemble-based scores empirically correlate better with support mismatch because they target epistemic uncertainty induced by finite data. Again, conformal calibration treats the resulting $u_\psi(s)$ as an opaque score and provides a threshold with controlled in-data abstention.

Consequence for UG-SRDP deployment. Combining Theorem 7.1 with the invariants of the gating rule (in particular, $\eta(s) = 0$ whenever $u_\psi(s) \geq \tau$ and $\eta(\cdot)$ non-increasing in u_ψ), we obtain a deployable procedure whose rate of disabling guidance on in-distribution states is controlled at level α without modeling assumptions. The remaining question is computational: once the gate is calibrated, we still require that the resulting state-dependent sampler and guidance schedule be feasible under real-time constraints. We address this systems aspect next.

8 Complexity & Systems Considerations: Compute–Safety Trade-offs Under Gated Sampling

We now make explicit the computational consequences of deploying a diffusion policy with state-dependent guidance. Since deployment may occur under tight latency constraints (e.g., closed-loop robotics), our objective is not merely asymptotic complexity but rather *predictable* per-decision wall-clock time while retaining the safety benefit of disabling critic guidance when the OOD score is large.

Decomposing test-time cost. At a given state s, the inference pipeline consists of: (i) computing a representation $z = f_\phi(s)$, (ii) computing the OOD score $u_\psi(s)$ (possibly requiring a reconstruction pass or an ensemble), (iii) selecting a sampler/guidance schedule, and (iv) executing $T(s)$ denoising steps to sample an action a_0. We therefore write a simple additive model for expected per-decision cost,

$$\mathrm{Cost}(s) \;\approx\; C_{\mathrm{score}}(s) \;+\; \sum_{t=1}^{T(s)} C_{\mathrm{step}}(s,t),$$

where $C_{\mathrm{score}}(s)$ includes the SRDP trunk forward pass plus any decoder/ensemble computation, and $C_{\mathrm{step}}(s,t)$ accounts for the diffusion head evaluation at timestep t and (when enabled) critic guidance evaluation.

Guidance is computationally expensive in diffusion policies. In common implementations of guided diffusion for control, each denoising step requires (a) a forward pass through the diffusion head f_θ to predict $\hat{\epsilon}$, and (b) one or more critic evaluations to compute a guidance direction, sometimes including a gradient $\nabla_a Q_\phi(s, a_t)$ with respect to the current noisy action a_t. Denoting by C_{DP} the cost of one diffusion-head evaluation and by C_Q (resp. $C_{\nabla Q}$) the cost of evaluating Q_ϕ (resp. its action-gradient), a crude but operational upper bound is

$$C_{\mathrm{step}}(s,t) \;\le\; C_{\mathrm{DP}} \;+\; \mathbf{1}[\eta(s) > 0]\,(C_Q + C_{\nabla Q}),$$

up to constant factors from multiple critics or classifier-free style guidance. Thus, even if T is fixed, *turning off* guidance via the gate yields a substantial multiplicative reduction in cost per denoising step. This is one of the central systems motivations for the hard gate $\eta(s) = 0$ when $u_\psi(s) \ge \tau$: it eliminates both the risk channel (critic error amplification) and a major compute channel (critic-gradient guidance).

Sampler switching and expected denoising steps. Beyond disabling guidance, we may also change the sampler itself when $u_\psi(s) \ge \tau$, e.g., by switching to a behavior-cloning diffusion sampler with fewer steps, or to a deterministic distilled sampler. Let T_{guided} be the step budget for Q-guided sampling and T_{BC} the step budget for the fallback sampler (typically $T_{\mathrm{BC}} < T_{\mathrm{guided}}$). Under the hard gate with a statewise switch,

$$T(s) \;=\; T_{\mathrm{guided}}\mathbf{1}[u_\psi(s) < \tau] \;+\; T_{\mathrm{BC}}\mathbf{1}[u_\psi(s) \ge \tau],$$

and hence, for any evaluation distribution over states,

$$\mathbb{E}[T(s)] \;=\; T_{\mathrm{guided}}\mathbb{P}[u_\psi(s) < \tau] \;+\; T_{\mathrm{BC}}\mathbb{P}[u_\psi(s) \ge \tau].$$

This expression makes the compute–safety coupling explicit: increasing conservatism (larger gating probability) simultaneously reduces critic usage *and* reduces the expected number of denoising steps if the fallback is cheaper. In particular, under covariate shift where $\mathbb{P}_{d_{\mathrm{test}}}[u_\psi(s) \ge \tau]$ may be large, the method tends to *self-throttle* compute in precisely those regimes where guidance is least trustworthy.

Worst-case latency and real-time control. For real-time robotics, expected cost is insufficient: we must also control worst-case latency to avoid missing control deadlines. The hard-gated sampler switch provides a deterministic cap

$$T(s) \le \max\{T_{\mathrm{guided}}, T_{\mathrm{BC}}\},$$

and one may further enforce a strict per-step compute budget by (i) fixing the diffusion network width, (ii) avoiding backpropagation through Q_ϕ by using guidance forms that do not require $\nabla_a Q_\phi$ (at some loss of performance), or (iii) precomputing guidance-relevant quantities when the state remains constant across the denoising loop. When critic gradients are required, implementation should use fused kernels and automatic mixed precision, and (in GPU settings) keep the entire denoising loop on-device to avoid host-device synchronization.

Cost of computing the OOD score. The OOD computation itself must be cheap relative to the sampling loop. If u_ψ is a deterministic residual produced by a decoder that shares the trunk f_ϕ, then $C_{\text{score}}(s)$ is approximately one additional decoder head forward pass. If u_ψ is ensemble-based with E decoder heads, then $C_{\text{score}}(s)$ scales roughly linearly in E, but can be parallelized because all heads consume the same trunk representation. In practice, we prefer architectural choices in which f_ϕ dominates cost and decoder heads are shallow, so that ensembles improve OOD sensitivity without imposing prohibitive overhead. Moreover, $u_\psi(s)$ is computed once per environment step, whereas diffusion sampling incurs $T(s)$ repeated evaluations; hence, for moderate $T(s)$, even a small reduction in $T(s)$ dominates the cost of richer scoring.

Practical deployment rule and monotonic compute schedules. To ensure predictability, we recommend a *one-shot* gating decision: compute $u_\psi(s)$ at the start of action selection, then commit to either the guided sampler or the fallback for the entire denoising trajectory. This avoids intra-trajectory branching and ensures that $T(s)$ and the presence/absence of critic calls are known before sampling begins. If one desires a smoother trade-off than a hard switch, a monotone map $u \mapsto T(u)$ can be used (fewer denoising steps as u grows), preserving the invariant that compute and critic reliance do not increase with OOD score.

Training-time overhead. Training remains dominated by diffusion-policy optimization and critic learning. The reconstruction head introduces an additional loss term and backward pass through the shared trunk; with shared features, this overhead is typically a constant factor. Ensembles multiply only the decoder-head parameters (and their gradients) and can be trained with shared-trunk minibatches to amortize representation learning cost. From a systems perspective, this is attractive: we accept a modest training-time constant factor to obtain a deployment-time rule that can *reduce* compute by disabling guidance and possibly reducing denoising steps under shift.

Summary of the compute–safety mechanism. The gate simultaneously (i) limits sensitivity to critic error by setting $\eta(s) = 0$ when $u_\psi(s) \geq \tau$, and (ii) can reduce deployment cost via sampler switching and step-budget reduction. The next section specifies an experimental plan designed to quantify both effects: return improvements when guidance is reliable, and controlled degradation (with reduced compute) when guidance is likely to be unsafe.

9 Experimental Plan: Stress-Testing UG-SRDP Under Support Mismatch and Critic Failure Modes

We outline an experimental program aimed at isolating the two claims implicit in the gated design: (i) when guidance is reliable, UG-SRDP retains the benefits of Q-guided diffusion relative to pure imitation; (ii) when guidance is unreliable due to covariate shift or critic miscalibration, the OOD gate prevents catastrophic degradation and yields a predictable compute–safety trade-off.

Common protocol and reporting. Across all domains, we train the SRDP diffusion policy and the critic(s) strictly offline on the training portion of $\mathcal{D}$. We reserve a held-out calibration split (disjoint from training) to fit the conformal threshold τ at target miscoverage α and to select any gating hyperparameters not determined by conformal calibration (e.g., slope parameters for a soft gate). We report (a) average discounted return, (b) estimated OOD encounter rate $\widehat{\mathbb{P}}_{s\sim d_\pi}[u_\psi(s) \geq \tau]$ along rollouts, (c) an application-specific safety metric (catastrophe probability or cumulative cost when available), and (d) inference compute proxies (average denoising steps and number of critic-gradient calls per environment step). For statistical stability, we evaluate over multiple random seeds and report confidence intervals.

Baselines. We recommend comparing against: (i) π_{BC} (pure behavior-cloning diffusion, no guidance), (ii) ungated guided diffusion (fixed $\eta \equiv \eta_0$, i.e., always-on critic guidance), (iii) a conservative offline RL baseline (e.g., IQL/CQL) where applicable, and (iv) a naive OOD heuristic (e.g., density model score without calibration) to separate the effect of *calibrated* gating from generic uncertainty estimates. Where sampler switching is used, we additionally compare to a fixed-step budget baseline with the same average denoising steps to control for compute.

(1) Missing-goal Maze2D: goal shift and corridor-induced OOD. Maze2D provides a clean mechanism for inducing covariate shift by changing the goal specification while retaining similar local dynamics. We propose training on standard Maze2D datasets where goals occupy a subset of the maze (or a subset of rooms), and evaluating on (a) unseen goal locations in held-out rooms (far-OOD) and (b) goals near but not identical to training goals (near-OOD). Since diffusion policies often exploit critic gradients to "snap" actions toward high-value regions, this setting is well-suited to revealing critic overgeneralization beyond the demonstrated manifold. We measure: success rate (reaching the goal), wall-collision rate (as a proxy for catastrophic actions), and the alignment between the gate events $\{u_\psi(s) \geq \tau\}$ and failure events (precision/recall or AUROC of u_ψ for failure prediction).

(2) Synthetic region removal in continuous-control D4RL. To systematically study support mismatch while keeping the environment fixed, we recommend constructing modified offline datasets by removing transitions whose states fall in a specified region $\mathcal{R} \subset \mathcal{S}$ (or whose actions fall in $\mathcal{R} \subset \mathcal{A}$). Concretely, for locomotion tasks (HalfCheetah/Hopper/Walker2d), one can define $\mathcal{R}$ by joint-angle ranges, torso height, or velocity thresholds, thereby creating "holes" in the data manifold. Evaluation then forces policies to traverse $\mathcal{R}$ (e.g., via modified initial-state distributions or by adding external perturbations that push the agent into the removed region). This permits controlled sweeps over the OOD mass by varying the size/placement of $\mathcal{R}$. The key outcome is whether UG-SRDP degrades gracefully as $\mathbb{P}_{d_\pi}[s \in \mathcal{R}]$ increases, while ungated guidance exhibits a sharper collapse due to extrapolation error.

(3) Critic-miscalibration stress tests: injected extrapolation error. Because UG-SRDP is explicitly designed to mitigate critic error amplification, we recommend experiments that *deliberately* corrupt the critic. We consider three mechanisms: (i) *Data thinning*: train the critic on a strict subset of $\mathcal{D}$ while keeping the diffusion model fixed, increasing critic variance out of support; (ii) *Label corruption*: add structured noise to Bellman targets or rewards in a localized region of state space, mimicking reward misspecification; (iii) *Adversarial critic heads*: train an auxiliary critic that agrees with the true critic on in-support states but deviates elsewhere, and use it for guidance. We then compare ungated guidance, UG-SRDP, and π_{BC}. The intended diagnostic is a phase transition: ungated guidance may improve in-distribution performance yet catastrophically fail under miscalibration, whereas UG-SRDP should recover a bounded-loss behavior by disabling guidance when u_ψ is large.

(4) Real-robot "forbidden zones": explicit safety constraints under offline shift. We propose a real-robot evaluation where costs are physically meaningful and can be monitored, such as a planar end-effector reaching/pushing task with forbidden workspace regions (e.g., near joint limits, a fragile object, or a no-go boundary). The offline dataset $\mathcal{D}$ is collected under a conservative teleoperation or impedance controller that avoids the forbidden region. At test time, we introduce distribution shift via new target placements or obstacles that increase the likelihood of encountering the boundary. We define $c(s,a) \in [0,1]$ as an indicator (or smooth function) of forbidden-zone violation and report both return and discounted cost. The central question is whether calibrated gating reduces violation probability without requiring online intervention.

Ablations: calibration, gate shape, and decoder choice. To attribute performance to specific components, we recommend: (i) *Calibration*: compare conformal τ at various α to uncalibrated fixed thresholds and to thresholds tuned on test performance (oracle); (ii) *Gate shape*: hard gate $\eta(s) = \eta_0 \mathbf{1}[u < \tau]$ versus soft gates $\eta(s) = \eta_0 h(u)$ (e.g., logistic or piecewise linear), and optionally a monotone $u \mapsto T(u)$ schedule; (iii) *OOD score construction*: deterministic reconstruction residual, probabilistic negative log-likelihood, and ensemble variance (holding SRDP trunk capacity fixed). We additionally record calibration diagnostics (empirical coverage on the held-out split) and the correlation between $u_\psi(s)$ and downstream failure, as the latter mediates whether gating events occur at semantically meaningful times.

Expected outcomes. The experimental objective is not only to improve mean return, but to demonstrate a consistent pattern: UG-SRDP matches ungated guided diffusion when u_ψ indicates in-support states, and reverts toward π_{BC}-like behavior (with reduced violations and reduced compute) when encountering far-OOD states or critic miscalibration.

10 Discussion & Limitations

The central role of uncertainty gating in UG-SRDP is to prevent *critic-error amplification* by ensuring that the guidance coefficient $\eta(s)$ vanishes on states deemed far from the data support. This design is most beneficial precisely in the regimes where offline RL is brittle: covariate shift, "holes" in the dataset support, and critics that extrapolate with high error. In such regimes, turning off guidance converts a potentially unbounded degradation (since $\sup_{s \notin \mathcal{S}_{\text{in}}(\tau)} |Q^*(s,a) - Q_\phi(s,a)|$ is uncontrolled) into an additive term proportional to the gate-activation probability $\mathbb{P}[u_\psi(s) \geq \tau]$ along the learned policy's occupancy. From an algorithmic standpoint, this is not merely a safety heuristic: it is a structural restriction that enforces a decomposition between an "improve when reliable" mode and a "revert when uncertain" mode.

At the same time, gating can hurt performance in regimes where leaving the behavior manifold is *necessary* for high return and the critic is in fact accurate there. In such cases, a hard gate $\eta(s) = 0$ for $u_\psi(s) \geq \tau$ may over-regularize and effectively constrain the policy toward π_{BC}, producing avoidable suboptimality. This limitation is intrinsic: in the strict offline setting, if improved behavior requires visiting states with little or no support in $\mathcal{D}$, then any guarantee must trade off improvement against unidentifiability. The best we can hope for is to expose this trade-off explicitly through τ (or through a soft gate $h(u)$), rather than implicitly through unstable optimization of a miscalibrated critic.

A second limitation is that the calibration guarantee for τ is *distribution-specific.* Split conformal calibration yields marginal coverage $\mathbb{P}_{s \sim d_{\mathcal{D}}}[u_\psi(s) \leq \tau] \geq 1 - \alpha$ under exchangeability with the calibration split. This ensures that, on data drawn like the dataset, the gate does not trigger more often than intended. However, this statement does not extend to the evaluation distribution d_{test} (which is precisely where covariate shift occurs), nor does it provide conditional coverage (e.g., coverage conditioned on task-relevant events). Consequently, the calibrated α should be interpreted as controlling false-positive gating *on-support*, not as providing a universal bound on false negatives (i.e., failing to gate when guidance is unsafe) under arbitrary shift. In particular, if u_ψ is only weakly correlated with critic error, then conformal calibration controls the frequency of gating but does not guarantee that gating triggers on the “right” states.

Relatedly, our choice of u_ψ as a reconstruction-based score is only a proxy for the relevant quantity: the critic’s out-of-support error. Reconstruction residuals can be low in states that are visually or geometrically familiar yet semantically dangerous, and can be high for benign nuisance shifts (lighting, textures, camera intrinsics) that do not meaningfully change the dynamics or rewards. In such settings, gating can be either overly conservative (unnecessary reversion to π_{BC}) or insufficiently protective (failure to detect critic unreliability). Probabilistic decoders and ensembles partially address this by capturing epistemic uncertainty, but they do not resolve the fundamental mismatch between “state likelihood” and “value uncertainty.” An important direction is to couple the gate to quantities more directly linked to value estimation error, for example disagreement across critic ensembles, Bellman residual-based scores, or hybrid scores that combine reconstruction uncertainty with critic uncertainty in a calibrated manner.

Partial observability exacerbates these issues. In many practical domains (notably vision-based control), the learner observes o_t rather than the Markov state s_t, and the diffusion policy is conditioned on an encoder representation. Then both Q_ϕ and u_ψ are functions of observations (or latent features), and reconstruction may be ill-posed: many distinct latent states can produce similar observations, and conversely small observation shifts can correspond to large changes in the underlying state. Under such aliasing, a gate calibrated on observation-level reconstruction error may provide weak protection against entering latent regions where the critic extrapolates. Addressing this likely requires temporal information (trajectory-level diffusion or recurrent encoders), state-estimation modules, or gate definitions that operate on belief states rather than instantaneous observations. From the theory side, one would need to replace the MDP assumptions with POMDP-appropriate notions of occupancy and support, and to state explicitly what is being calibrated (coverage in observation space versus coverage in latent belief space).

It is also useful to situate UG-SRDP relative to pessimism and conservative offline RL. Conservative methods such as CQL or lower-confidence-bound style critics seek to make Q_ϕ pessimistic on OOD actions, thereby discouraging the policy from exploiting extrapolation. Our gating mechanism is complementary: it controls *whether* the policy is allowed to follow critic gradients at all in a given state, rather than modifying the critic globally. In particular, even a pessimistic critic can be wrong far OOD, and a diffusion policy that follows its gradients may still be destabilized; conversely, gating can be combined with conservative critic training to reduce both the magnitude of extrapolation error and the sensitivity of the policy to whatever error remains. Conceptually, UG-SRDP can be viewed as implementing a state-dependent interpolation between improvement (guided diffusion) and stability (imitation), whereas pessimism modifies the improvement signal itself.

Finally, the approach inherits the limitations of the fallback policy: when the dataset is low-quality or highly suboptimal, reverting to π_{BC} may be safe yet ineffective. Moreover, in tasks where safety requires actions outside the behavior support (e.g., evasive maneuvers not present in $\mathcal{D}$), gating cannot synthesize such competence without additional structure. Thus, UG-SRDP should be understood as a method for *graceful degradation* under support mismatch, not as a mechanism that removes the need for coverage assumptions. We view the main open problems as (i) calibration under explicit covariate shift (beyond i.i.d. conformal guarantees), (ii) principled gate scores tied to value uncertainty rather than reconstruction alone, and (iii) extensions to vision and trajectory-level diffusion where both support and safety are intrinsically temporal.

11 Conclusion

We introduced UG-SRDP, an uncertainty-gated variant of critic-guided diffusion policies for strict offline reinforcement learning. The method is motivated by a simple structural observation: when a learned critic is used as a guidance signal in a generative policy update, extrapolation error in regions of state space with little or no dataset support can be *amplified* by the policy optimization itself. UG-SRDP intervenes precisely at this amplification pathway by making the guidance coefficient state-dependent, $\eta(s) = \eta(u_\psi(s))$, where u_ψ is an OOD score derived from an SRDP reconstruction head and $\eta(\cdot)$ is monotone non-increasing with a hard cutoff

$\eta(s) = 0$ when $u_\psi(s) \geq \tau$. This produces an explicit separation between (i) a *guided improvement* regime in which we allow the policy to exploit the critic, and (ii) an *imitation fallback* regime in which we revert to a behavior-regularized sampler (e.g. pure diffusion behavior cloning).

Our theoretical contribution is to make this decomposition quantitative in both bandit and discounted MDP settings. In the contextual bandit case, we bound test-time suboptimality by a sum of two interpretable terms: a critic-error term scaled by the maximum guidance strength on in-support states, and an additive term proportional to the probability mass of far-OOD contexts under the test distribution. Concretely, the gate replaces an otherwise uncontrolled dependence on $\sup_{s \notin \mathcal{S}_{\text{in}}(\tau)} \sup_a |Q^*(s,a) - Q_\phi(s,a)|$ with a bounded loss term of order $R_{\max}\, p_{\text{OOD}}$. In the discounted MDP case, we extend the argument via an occupancy decomposition: we express performance differences as expectations over $d_{\pi_{\text{UG}}}$ and isolate the contribution of OOD encounters as an additive penalty of size at most $\frac{2R_{\max}}{1-\gamma}\, \mathbb{P}_{s \sim d_{\pi_{\text{UG}}}}[u_\psi(s) \geq \tau]$, while the on-support guidance term scales as $\varepsilon\, \mathbb{E}_{s \sim d_{\pi_{\text{UG}}}}[\eta(s)]$ up to an explicit factor depending on γ and the stability/Lipschitz parameters of the diffusion update. These statements formalize the intuition that gating does not eliminate the inherent difficulty of support mismatch, but converts potentially unbounded degradation into a controlled trade-off between (a) the frequency with which the learned policy visits OOD regions and (b) the extent to which we rely on a possibly imperfect critic within the calibrated in-support set.

A second theoretical contribution is the use of split conformal calibration to select the gating threshold τ with a finite-sample guarantee under the dataset state distribution. Given a calibration split of size m, we choose τ as an appropriate order statistic of $\{u_\psi(s_i)\}_{i=1}^m$ and obtain marginal coverage $\mathbb{P}_{s \sim d_{\mathcal{D}}}[u_\psi(s) \leq \tau] \geq 1 - \alpha$ without parametric assumptions on the score distribution. While this does not control coverage under arbitrary evaluation shift, it provides a principled mechanism to tune conservatism in deployment while certifying that the gate does not spuriously over-trigger on in-distribution states beyond the user-chosen miscoverage level α.

We also clarify the limits of what strict offline learning can guarantee in the presence of far-OOD test mass. By an unidentifiability construction, one can exhibit pairs of environments that agree on the support of the offline data distribution yet disagree on rewards (or optimal actions) in unobserved regions. Any offline algorithm must then incur an additive loss proportional to the probability of encountering those regions. This lower bound matches the qualitative form of the p_{OOD} (or occupancy-level) term in our gated bounds and supports the interpretation of UG-SRDP as *graceful degradation*: when the test distribution forces the policy into unobserved regimes, the best achievable guarantee necessarily pays an additive price that depends on how often this occurs.

On the practical side, UG-SRDP is a minimal modification of existing

SRDP/Diffusion-QL style pipelines: we retain diffusion-based action generation, double-critic learning, and behavior cloning objectives, and add (i) a reconstruction head used only to compute $u_\psi(s)$ and (ii) a deployment-time rule mapping $u_\psi(s)$ to guidance and (optionally) to a sampler switch or reduced denoising budget. This yields a favorable compute–robustness trade-off: the expected inference cost becomes $\mathbb{E}[T(s)] = T_{\text{guided}}\mathbb{P}[u_\psi < \tau] + T_{\text{BC}}\mathbb{P}[u_\psi \geq \tau]$, which can be substantially smaller than always running a fully guided sampler, while simultaneously reducing the risk of catastrophic critic-driven errors in OOD states.

Several directions appear technically natural and practically important. First, vision-based control requires that both the critic and the gate operate on representations induced by high-dimensional observations; here we expect that reconstruction-based scores should be augmented with representation uncertainty and temporal consistency, potentially via sequence models or latent-state estimators. Second, trajectory-level diffusion policies offer a direct way to incorporate temporal constraints (including safety constraints) into the generative process; extending uncertainty gating to operate on trajectory prefixes or belief states may better align the gate with the onset of compounding error. Third, calibration beyond the i.i.d. split conformal setting remains open: we would like guarantees that remain meaningful under explicit covariate shift, for example by using weighted conformal schemes, conditional calibration on task-relevant strata, or hybrid gates that incorporate critic disagreement/Bellman residuals. Finally, the iterative denoising cost of diffusion remains a barrier in real-time control; integrating UG-SRDP with accelerated samplers, and in particular with consistency distillation to obtain few-step or one-step policies while preserving the gated guidance semantics, is an immediate avenue for making the method more deployable.

In summary, UG-SRDP provides a principled mechanism for controlling critic reliance in offline diffusion policies via a calibrated, state-dependent gate. The resulting theory makes explicit the unavoidable role of OOD occupancy and the bounded nature of any strict offline improvement guarantee, while the algorithm offers a practical path toward robust deployment under distribution shift with predictable failure modes and tunable conservatism.