Calibrated LRT Guidance for Offline Diffusion Policies

Ximan Sun
Duke University
Durham, United States
ximan.sun@duke.edu

Xiang Cheng
Duke University
Durham, United States
xiang.cheng@duke.edu

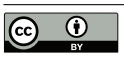
ABSTRACT

Diffusion policies are competitive for offline Reinforcement Learning but are typically guided at sampling time by heuristics that lack a statistical notion of risk. We introduce LRT-Diffusion, a risk-aware sampling rule that performs evidence accumulation between two inference-time heads: an unconditional background head and a state-conditional good head. Concretely, we accumulate a log-likelihood ratio and gate the conditional mean with a logistic controller whose threshold τ is calibrated once per task and per sampler under H_0 to meet a user-specified Type-I level α. This turns guidance from a fixed push into an *evidence-driven* adjustment with a user-interpretable risk budget. Importantly, we deliberately leave training vanilla (two heads with standard ϵ-prediction) under the structure of DDPM. LRT guidance composes naturally with Q-gradients: critic-gradient updates can be taken at the unconditional mean, at the LRT-gated mean, or a blend, exposing a continuum from exploitation to conservatism. We standardize states/actions consistently at train and test time and report a state-conditional OOD metric alongside return. On D4RL MuJoCo tasks, LRT-Diffusion yields a calibrated return–risk frontier: LRT often reduces state-conditional OOD, and combining with a small Q-step increases return along the frontier. Theoretically, we establish level-α calibration, stability bounds, and a return comparison showing when evidence-gated guidance is preferable to pure Q-guidance. Overall, LRT-Diffusion is a drop-in, inference-time method that adds principled, calibrated risk control to diffusion policies for offline RL.

KEYWORDS

Offline Reinforcement learning; Diffusion Policies; Likelihood-ratio Test; Inference; Risk-aware Guidance; Calibration; Distribution Shift; Q-guidance; Out-of-distribution Detection; D4RL.

ACM Reference Format:
Ximan Sun and Xiang Cheng. 2026. Calibrated LRT Guidance for Offline Diffusion Policies. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 9 pages. https://doi.org/10.65109/VBFF4869



Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/VBFF4869

1 INTRODUCTION

Offline reinforcement learning (RL) aims to learn high-performing policies from fixed datasets without further environment interaction.[1] A central difficulty is *distributional shift*: actions proposed by the learned policy can drift away from the behavioral support where value estimates are reliable. Diffusion policies have recently emerged as strong generative decision-makers for offline RL [11]: by learning a conditional diffusion model over actions given state, they produce smooth, high-fidelity samples that respect support better than direct regression. However, *how* these policies are guided at sampling time remains largely heuristic. Common practices—injecting Q-gradients with hand-tuned schedules and ad-hoc clipping—lack a statistical notion of risk and offer limited control of the return–shift trade-off. Unlike prior works, our approach keeps training strictly vanilla (no critic-guided losses) and moves all risk control to inference via a calibrated likelihood-ratio gate (LRT), yielding an interpretable, reproducible risk knob.

Intuition. Rather than always pulling samples toward the conditional head, we ask at every denoising step: *is there enough evidence to move toward the "good-action" direction for this state?* Concretely, we split offline actions into a *good* subset and a *background* subset, train a two-head diffusion model (an *unconditional* head on all data and a *conditional* head on good data) with class-balancing and optional advantage-based soft weights, and then gate the conditional pull at inference by a *calibrated* likelihood ratio. The gate is motivated by the Neyman–Pearson test [17]: a single user knob Type-I rate α controls the tolerated false activations under H_0 ("background is correct"); while the hard LRT is UMP at level α under equal covariances, we use a smooth gate in practice for numerical stability and keep α interpretable via calibration on held-out states.

Method overview. We introduce *LRT-Diffusion*, a risk-aware, inference-only sampling scheme for diffusion policies (see Fig 1). At each denoising step we make a binary decision between a background prediction and a data-conditioned prediction. We accumulate evidence during sampling and open the gate only when the evidence is strong, interpolating between the two predictions with a data-dependent weight. A single threshold is calibrated once per task so that the empirical false-activation rate does not exceed a user-chosen α, turning guidance from a fixed push into an *evidence-driven* adjustment: weak evidence $\Rightarrow$ stay near the background prior; strong evidence $\Rightarrow$ move decisively toward the conditional policy. Training is unchanged. A detailed, step-by-step illustration of the inference-time pipeline is provided in the extended version at https://arxiv.org/abs/2510.24983.

Composition with value guidance. We optionally combine LRT gating with a small critic step: at each denoising update, take a

[1] Due to space constraints, all theoretical proofs, supplementary ablations, and detailed experimental setups are available in the extended version at https://arxiv.org/abs/2510.24983.

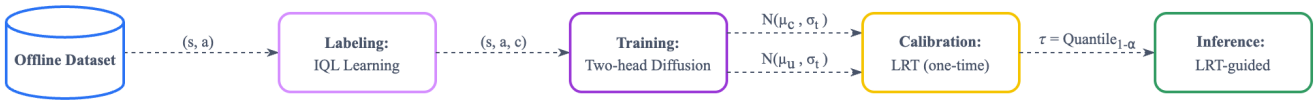


Figure 1: Overall pipeline of LRT-Diffusion. Training is vanilla; all risk control is applied at inference via a calibrated gate.

capped step that increases the learned critic at the current proposal (standard decreasing schedule). The gradient can be evaluated near the background proposal, the LRT-gated proposal, or an interpolation; the threshold is calibrated with the same choice, so the Type-I semantics are preserved.

Theoretical guarantees at a glance. Under equal-covariance heads, the hard gate is the uniformly most powerful test at level α; our soft gate (a numerically stable surrogate) retains the threshold's semantics and improves stability. Beyond testing optimality, we analyze offline-RL error propagation and show that reducing the policy's *state-conditional* OOD rate—which LRT lowers in many regimes—tightens a lower bound on true return relative to pure Q-guidance when the critic is unreliable off-support.

Practical picture. The gate exposes monotone, interpretable knobs: α (risk), $\beta_{\max}$ (max pull), and δ (sharpness). LRT can be combined with a small critic step evaluated at the background mean, the LRT-gated mean, or a blend; the threshold is calibrated with that same choice so the Type-I semantics are preserved. Empirically on D4RL MuJoCo, LRT tracks the target α and yields a calibrated return-risk frontier: LRT serves as a low-risk anchor, while LRT+Q increases return along the frontier.

1.1 Contributions

Calibrated LRT guidance for diffusion policies. Unlike prior research integrating value or energy guidance into the diffusion model, we cast each denoising step as a likelihood-ratio test between a background head and a "good" head, and calibrate a single threshold τ on held-out states to bound the trajectory-level Type-I rate at a user-chosen α. Training remains vanilla (two heads with class balancing and optional advantage weights).

Labeling-and-weighting recipe for the conditional head. Rather than fitting a single conditional diffusion model to all (state, action) pairs, as is common in return-dominated diffusion or Q-guided training, we split offline actions into a *good* subset and a *background* subset using IQL advantages (top-p quantile), and train a two-head diffusion model with class-balancing and advantage-aware soft weights. This improves the quality of the conditional head without extra supervision. [2]

Actionable theory with finite-sample and stability guarantees. Under equal covariances the hard LRT is UMP at level α (Prop. 5.1); a one-shot Monte-Carlo calibration yields $\mathbb{P}_{H_0}(\ell_{\text{cum}} \geq \hat{\tau}) \leq \alpha + \varepsilon_n$ (Thm. 5.4). We bound the per-step deterministic drift and show the cumulative LLR is sub-Gaussian with an explicit variance proxy (Sec. 5.2), explaining why calibration is numerically stable.

[2] E.g., Diffuser/Decision-Diffuser use return conditioning and sampling-time guidance; Diffusion-QL injects value signals into the training loss, all without a binary good/background head.

Return bounds that connect risk control to distribution shift. With a standard offline error split, we prove a comparison bound for true returns between LRT-guided and Q-guided sampling (Prop. 5.6). A mild monotonicity assumption turns the calibrated level α into a conservative OOD upper bound (Prop. 5.7), clarifying when and why LRT is preferable to pure Q-guidance.

Compatibility with value gradients via matched calibration. LRT controls *risk* while the Q-step pursues *return*, and these roles are orthogonal. We show that composing LRT with a small Q-step preserves level-α semantics *so long as* calibration uses the exact deployed sampler (Prop. 5.3). Thus, risk control (through α and its calibrated τ) is decoupled from return-seeking (through $\nabla_a \hat{Q}$), making the Q-update a plug-and-play module. The knobs $(\alpha, \beta_{\max}, \delta)$ remain monotone and interpretable.

Empirical validation. On D4RL MuJoCo, LRT-Diffusion honors the target Type-I rate and yields a calibrated return-risk frontier, with ablations over α, $\beta_{\max}$, δ, and gradient anchoring.

2 RELATED WORK

Distribution shift in offline RL and conservative learning. A central obstacle in offline RL is that policy actions may fall outside the behavior support, where critic estimates are brittle. Conservative algorithms explicitly counter this by penalizing or constraining OOD actions, e.g. [7, 8, 13, 14]. IQL [12] learns values via expectile regression and performs advantage-weighted improvement without explicit behavior cloning. Our work is complementary: we keep training unchanged (e.g., use IQL advantages only to label/weight "good" actions) and control OOD at *sampling time* via a calibrated gate.

Diffusion policies for decision-making and their guidance. Diffusion models have been adopted for control and planning [11]. To improve returns, many systems apply *Q-guidance*—adding an action-space step along $\nabla_a \hat{Q}(s, a)$ with hand-tuned schedules and clips—conceptually analogous to guidance in image diffusion [4, 10]. However, such heuristics lack a statistical notion of risk: there is no global control on the probability of "falsely" pulling away from the background prior.

Risk-aware control and calibration. Risk-sensitive and conservative offline RL methods (e.g., behavior-regularized or pessimistic objectives) encode risk *in the training loss* by penalizing value estimates or constraining policy deviation from the dataset [3, 8, 12, 14, 19, 24]. A complementary line uses *distribution-free calibration* to turn data-driven thresholds into finite-sample guarantees (e.g., conformal prediction) [2, 22]. In diffusion-based decision making, sampling is typically guided heuristically (e.g., Q-gradient pushes with hand-tuned schedules/clipping) without an explicit statistical notion of risk. Our approach bridges these threads: we cast

each reverse step as a simple-vs-simple test between an unconditional (background) head and a conditional (good) head, accumulate a log-likelihood ratio (LLR), and *calibrate one threshold* on held-out states so that the empirical Type-I rate under H_0 does not exceed a user-chosen α. Under equal covariances—satisfied by our two-head design—the hard likelihood-ratio test is uniformly most powerful [17], and we use a smooth gate for numerical stability while preserving the same α-semantics via calibration. This gives a statistically grounded alternative to heuristic mixing: guidance becomes evidence-driven with an explicit level-α risk budget at inference time, without changing training.

Hypothesis testing and likelihood ratios in RL Testing in RL is typically used for deployment-time decisions (high-confidence OPE [21], safe improvement such as SPIBB [15]), rather than within a generative sampler. Safety-constrained methods (e.g., CPO [1]) control violations but do not apply likelihood-ratio gating during action generation. Although sequential tests like SPRT [23] inspire sequential criteria, we are not aware of calibrated likelihood-ratio gates used to steer diffusion denoising at inference. Our approach turns guidance into an evidence-controlled, level-α procedure.

Advantage-based labeling and weighting. For higher return actions, advantage-weighted and critic-regularized schemes [3, 19, 20] are proposed. We use IQL advantages to (i) define a top-p "good" subset for the conditional head and (ii) optionally apply a temperature-controlled soft weight on positives, while keeping a background head trained on all data. This two-head setup strengthens the conditional signal without altering the base training pipeline, and remains compatible with our calibrated LRT at inference. Our two-head design enforces equal covariances by construction, which simplifies the LLR and aligns with the NP test.

3 BACKGROUND

Offline RL is given a fixed dataset $\mathcal{D} = \{(s^{(i)}, a^{(i)}, r^{(i)}, s^{(i)\prime})\}_{i=1}^{N}$, collected by an unknown behavior policy π, and aims to learn a high-return policy while controlling out-of-distribution risk. In the dataset, $s \in \mathbb{R}^{d_s}$ is the current state, $a \in \mathbb{R}^{d_a}$ is the action, $r \in \mathbb{R}$ is the reward, s' is the next state. We treat the process as a discounted Markov Decision Process with continuous state/action spaces.

3.1 IQL Advantages and "Good vs. Background" Labels

For each pair of state and action $(s, a) \in \mathcal{D}$, we train an IQL-style critic [12] $(\hat{Q}(s,a), \hat{V}(s))$ on standardized inputs and define an *advantage* $A(s,a)$ as $A(s,a) = \hat{Q}(s,a) - \hat{V}(s)$, which quantifies the value contributed by action a at state s relative to the state's baseline. Let κ be the top (p)-quantile of $\{A(s_i, a_i)\}_{i=1}^{N}$. We label each pair of state-action data by $c = \mathbf{1}\{A(s,a) \geq \kappa\} \in \{0, 1\}$.

Pairs with c=1 are considered *good*, as their advantages fall within the top (p) fraction of the dataset, while those with c=0 form the non-advantage subset. In practice, we set $p = 0.2$, allocating approximately 20% of the data as *good* examples.

3.2 Diffusion Policies with Two Heads

We train a diffusion policy with two heads: an *unconditional* head trained on all state–action pairs $\mathcal{D}$, and a *conditional* head trained only on the high-advantage subset we labeled beforehand. Formally, let $\mathcal{D}_{\text{all}} := \mathcal{D}$ and $\mathcal{D}_{\text{good}} := \{(s, a, r, s') \in \mathcal{D} : c = 1\}$. The unconditional head learns broad dataset coverage from $\mathcal{D}_{\text{all}}$, while the conditional head specializes to advantage behavior using $\mathcal{D}_{\text{good}}$.

Two heads and notation. We use a shared backbone $\phi_\theta(s, a_t, t)$ with two output branches ("heads"): an *unconditional/background* head trained on $\mathcal{D}_{\text{all}}$, and a *conditional/good* head trained on $\mathcal{D}_{\text{good}}$. Each head maps features to a DDPM-style noise prediction,

$$\hat{\varepsilon}_u = h_{u,\theta}\big(\phi_\theta(s, a_t, t)\big), \qquad \hat{\varepsilon}_c = h_{c,\theta}\big(\phi_\theta(s, a_t, t)\big), \tag{1}$$

where the subscripts u and c will consistently denote *unconditional* and *conditional* quantities, respectively.

DDPM parameterization. We adopt the predict-ε parameterization [9, 18] with schedule $\{\alpha_t, \bar{\alpha}_t\}_{t=1}^{T}$, where $\bar{\alpha}_t = \prod_{\tau=1}^{t} \alpha_\tau$. Given the current latent action a_t at step t, a noise prediction $\hat{\varepsilon}$ induces a Gaussian reverse kernel with shared covariance $\sigma_t^2 I$ and mean

$$\mu(\hat{\varepsilon}; t, s, a_t) = \frac{1}{\sqrt{\alpha_t}}\left(a_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\,\hat{\varepsilon}\right).$$

Throughout this subsection, we assume that the two diffusion heads share the same reverse-time variance σ_t^2, as given by the diffusion schedule, which enables a closed-form expression for the step-wise log-likelihood ratio. We discuss this equal-variance choice and examine robustness when it is deliberately violated at inference time in Sec. 6.5. We define the head-specific means

$$\mu_u(t, s, a_t) := \mu(\hat{\varepsilon}_u; t, s, a_t), \qquad \mu_c(t, s, a_t) := \mu(\hat{\varepsilon}_c; t, s, a_t),$$

and use the same $\sigma_t > 0$ from the diffusion schedule for both heads. This yields two reverse kernels for the next-step action:

$$a_{t-1} \mid a_t, s \sim \mathcal{N}(\mu_u(t, s, a_t),\, \sigma_t^2 I), \quad a_{t-1} \mid a_t, s \sim \mathcal{N}(\mu_c(t, s, a_t),\, \sigma_t^2 I),$$

where $\mathcal{N}(\mu, \Sigma)$ is Gaussian with mean μ and variance Σ.

After training, the parameters θ are frozen, so $\mu_u(t, s, a_t)$ and $\mu_c(t, s, a_t)$ are deterministic functions of (t, s, a_t) (with the same σ_t). We will interpret a single reverse step through these two proposals in the next subsection.

Notation clarification. We write dataset indices as parenthesized superscripts, e.g., $a^{(i)}$ is the $i-th$ action in the dataset $\mathcal{D}$. We reserve subscripts t for diffusion steps, e.g., a_t is the latent at step t and a_0 is the final action.

3.3 Hypotheses for a Reverse Step

With the trained parameters frozen, each reverse step t and conditioning state–latent pair (s, a_t) yields two Gaussian proposals for the next latent action a_{t-1}:

$$H_0 \text{ (background)}: \; a_{t-1} \mid a_t, s \sim \mathcal{N}\big(\mu_u(t, s, a_t),\, \sigma_t^2 I\big),$$
$$H_1 \text{ (good)}: \; a_{t-1} \mid a_t, s \sim \mathcal{N}\big(\mu_c(t, s, a_t),\, \sigma_t^2 I\big),$$

where μ_u and μ_c are the head-specific means defined in the previous subsection 3.2. Intuitively, H_0 favors broad, background behavior supported by the entire dataset, while H_1 emphasizes high-advantage behavior learned from the c=1 subset.

For compactness, denote the head-induced one-step densities by

$$p_c(a_{t-1} \mid a_t, s, t) := \mathcal{N}\big(a_{t-1};\, \mu_c(t, s, a_t),\, \sigma_t^2 I\big), \tag{2}$$
$$p_u(a_{t-1} \mid a_t, s, t) := \mathcal{N}\big(a_{t-1};\, \mu_u(t, s, a_t),\, \sigma_t^2 I\big). \tag{3}$$

3.4 Likelihood-Ratio View of a Reverse Step

Unknown labels ⇒ a trajectory-level test. At inference the label c is unknown. Given two trajectory models $p_u(a_{T:1} \mid s)$ (background) and $p_c(a_{T:1} \mid s)$ (good), the Neyman–Pearson (NP) lemma implies that the level-α test that rejects H_0 when the cumulative log–likelihood ratio (LLR) exceeds a threshold is uniformly most powerful (UMP). We thus base our gate on the trajectory statistic

$$\ell_{\text{cum}}(a_{T:1}) \;:=\; \log \frac{p_c(a_{T:1} \mid s)}{p_u(a_{T:1} \mid s)}.$$

How to compute ℓ_{cum}: step-wise LLRs. With the reverse chain factorization in DDPM, for $i \in \{u, c\}$

$$p_i(a_{T:1} \mid s) = q(a_T) \prod_{t=1}^{T} p_i(a_{t-1} \mid a_t, s, t),$$

so the trajectory LLR decomposes into a sum of step-wise terms:

$$\ell_{\text{cum}}(a_{T:1}) \;=\; \sum_{t=1}^{T} \ell_t, \qquad \ell_t \;:=\; \log \frac{p_c(a_{t-1} \mid a_t, s, t)}{p_u(a_{t-1} \mid a_t, s, t)}. \tag{4}$$

Head-induced one-step densities and quadratic form. From Sec. 3.3, each head induces a Gaussian kernel with shared covariance. Hence the step-wise LLR admits the shared-variance simplification

$$\ell_t \;=\; \frac{1}{2\sigma_t^2}\Big(\|a_{t-1} - \mu_u(t, s, a_t)\|_2^2 - \|a_{t-1} - \mu_c(t, s, a_t)\|_2^2\Big), \tag{5}$$

i.e., a linear discriminant in a_{t-1} and cheap to evaluate.

Decision rule for assigning labels. The NP test uses ℓ_{cum} with a calibrated threshold τ to control the level-α Type-I rate. We implement a smooth gate that monotonically approximates the NP decision and calibrate τ under H_0; details are given in Sec. 4.

4 METHOD: LRT-DIFFUSION

We split offline actions by an advantage threshold into a *good* subset and a *background* subset. We then train a *vanilla* two-head diffusion policy (DDPM ϵ-prediction [9]): an unconditional head on all data and a conditional head on the good subset, with class-balancing and optional advantage-based soft weights; no value/energy-guided losses are added. At inference, each reverse step is treated as a binary test between heads: we accumulate a cumulative log-likelihood ratio (LLR; Sec. 3.4) and apply a calibrated logistic gate to interpolate the mean,

$$\mu_{\text{LRT},t} \;=\; \mu_u(t, s, a_t) \;+\; \beta_t\,\big(\mu_c(t, s, a_t) - \mu_u(t, s, a_t)\big), \tag{6}$$

then sample $a_{t-1} \sim \mathcal{N}(\mu_{\text{LRT},t}, \sigma_t^2 I)$. The threshold τ is calibrated under H_0 (background) to meet a user-chosen Type-I rate α. The end-to-end pipeline is: (1) Train an IQL critic $(\hat{Q}, \hat{V})$ on standardized inputs; (2) Compute advantages and label top-p pairs (§4.1); (3) Train a two-head diffusion model on ϵ-prediction (§4.2); (4) Calibrate a single threshold τ under H_0 (§4.4); (5) At inference, accumulate LLR and gate the conditional pull (§4.5).

4.1 Good-vs-background Labeling

We rank state–action pairs by the IQL advantage $A(s, a) = \hat{Q}(s, a) - \hat{V}(s)$ computed on standardized inputs, and mark the top-p global quantile as "good." Although one could define state-wise thresholds, the advantage already subtracts a state-dependent baseline $\hat{V}(s)$, making A comparable across states in practice. Moreover, global top-p is a non-parametric surrogate of advantage-weighted learning (e.g., AWR/AWAC), corresponding to a small-temperature limit without tuning an extra temperature. Practically, offline datasets have highly uneven coverage—many states appear once or with only a handful of actions—so per-state ranking is statistically brittle; a single global threshold is more stable and reproducible. Crucially, our risk control via LRT is orthogonal to this choice: once the two-head model is trained, Type-I error is calibrated at inference regardless of how the conditional head's subset was selected.

4.2 Two-head Training

Under the two-head diffusion policy in Sec. 3.2, we optimize the ϵ-prediction loss (refer to Eq. 1) with per-sample weights

$$\mathcal{L} = \tfrac{1}{B}\sum_{i=1}^{B} \tilde{w}_i \left\|\hat{\epsilon}_\phi(a_{t,i}, t, \tilde{s}_i) - \epsilon_i\right\|_2^2, \quad \tilde{w}_i = \frac{w^{\text{cb}}(c_i; \hat{\rho}) \cdot u_i}{\frac{1}{B}\sum_{j=1}^{B} w^{\text{cb}}(c_j; \hat{\rho}) \cdot u_j},$$

where $w^{\text{cb}}(c; \hat{\rho})$ balances positive/negative contributions per batch via an EMA estimate, and u_i optionally emphasizes stronger positives.

$$w^{\text{cb}}(c; \hat{\rho}) = \frac{\mathbf{1}\{c = 1\}}{2\hat{\rho} + \varepsilon} + \frac{\mathbf{1}\{c = 0\}}{2(1 - \hat{\rho}) + \varepsilon}, \tag{7}$$

$$u_i = 1 + \mathbf{1}\{c_i = 1\}\,\min\Big\{\max\big(0, \tfrac{A_i - \kappa}{\tau_A}\big),\, u_{\max} - 1\Big\}. \tag{8}$$

Multiplying w^{cb} and u_i separates roles: the former fixes class imbalance between conditional and unconditional datasets, while the latter reallocates mass *within* c=1 without changing the effective class ratio; the batch normalization of $\tilde{w}_i$ stabilizes the step size.

4.3 Evidence-gated Sampler

4.3.1 Motivation.

1D evidence direction. From Eq. 5, under equal covariances the one-step LLR has constant gradient $\nabla_{a_{t-1}} \ell_t = \Sigma_t^{-1}(\mu_c - \mu_u)$; among unit directions u the directional derivative $\langle u, \nabla \ell_t \rangle$ is maximized by $u \parallel \Sigma_t^{-1}(\mu_c - \mu_u)$ (isotropic: $u \parallel \mu_c - \mu_u$). We therefore restrict the reverse-step mean to the 1D ray $\mu_u + \beta(\mu_c - \mu_u)$, and let the scalar gate β depend monotonically on the cumulative evidence ℓ_{cum}.

Why not always use μ_c? Always setting a_{t-1}=μ_c+$\sigma_t z_t$ is equivalent to *always accepting* H_1, which (i) removes any control on false activations of the conditional head and thus increases off-support mass in offline RL, and (ii) couples performance to critic/label errors where the model is most brittle. Our gate opens *only when there is sufficient evidence* (large ℓ_{cum}), yielding a calibrated budget on false activations (Prop. 5.3 and Thm. 5.4). Section 5.3 further shows that, when off-support errors dominate ($\varepsilon_{\text{out}} \gg \varepsilon_{\text{in}}$), reducing the state-conditional OOD rate via a smaller α tightens a lower bound on the true return (Prop. 5.6).

4.3.2 Gate Selection.

Hard gate (UMP at level-α). Let τ be a threshold. The *hard* likelihood-ratio test accepts H_1 when $\ell_{\text{cum}} \geq \tau$ and rejects otherwise. Equivalently, with a cap $\beta_{\max} \in [0, 1]$, $\beta_t^{\text{h}} = \beta_{\max}\,\mathbf{1}\{\ell_{\text{cum}} \geq \tau\}$, and $a_{t-1} = \mu_u + \beta_t^{\text{h}}(\mu_c - \mu_u) + \sigma_t z_t$. Under equal covariances and the joint factorization of the reverse chain, this hard test is uniformly most powerful (UMP) among all level-α tests (Prop. 5.1).

Algorithm 1 Calibration under H_0 to obtain τ (pre-inference sampling)

Require: Frozen policy θ, steps T, risk α, gate $(\beta_{\max}, \delta)$, noises $\{\sigma_t\}$, states $\mathcal{S}_{\text{cal}}$, budget n
Ensure: Threshold τ

$\tau \leftarrow +\infty$
for $k = 1..K_{\max}$ **do**
  $\mathcal{L} \leftarrow \emptyset$
  while $|\mathcal{L}| < n$ **do**
    Sample $s \sim \text{Unif}(\mathcal{S}_{\text{cal}})$, $a_T \sim \mathcal{N}(0, I)$; $\ell \leftarrow 0$
    for $t = T..1$ **do**
      $(\mu_u, \mu_c) \leftarrow \mu_\theta(t, s, a_t)$
      $\beta_t \leftarrow \beta_{\max}\text{Sigmoid}((\ell - \tau)/\delta)$(soft-gate)
      $a_{t-1} \sim \mathcal{N}(\mu_u + \beta_t(\mu_c - \mu_u), \sigma_t^2 I)$
      $\ell \leftarrow \ell - \frac{\|a_{t-1}-\mu_c\|^2 - \|a_{t-1}-\mu_u\|^2}{2\sigma_t^2}$
    end for
    $\mathcal{L} \leftarrow \mathcal{L} \cup \{\ell\}$
  end while
  $\tau \leftarrow \text{Quantile}_{1-\alpha}(\mathcal{L})$
end for
return τ

Soft gate (stable surrogate) and its hard-limit. For numerical stability we use a logistic surrogate

$$\beta_t \;=\; \beta_{\max}\,\text{Sigmoid}\Big(\tfrac{\ell_{\text{cum}}-\tau}{\delta}\Big), \quad a_{t-1} \;=\; \mu_u \;+\; \beta_t(\mu_c - \mu_u) \;+\; \sigma_t z_t,$$

where $z_t \sim \mathcal{N}(0, I)$ and $\delta > 0$ controls sharpness (smaller $\delta \to$ sharper switch). As $\delta \downarrow 0$, the soft rule converges to the hard gate both pointwise and in trajectory law (Lemma 5.2). Thus, the α, τ, β_t can be seen as monotone, interpretable risk knobs. The resulting sampler is summarized in Alg. 2.

4.4 Calibration

Fix $(\beta_{\max}, \delta)$ and freeze the sampler in Alg.2. On a held-out state set matched to deployment, simulate reverse chains *under* H_0 using the same sampler and collect the realized ℓ_{cum}; set $\hat{\tau} = \text{Quantile}_{1-\alpha}\{\ell_{\text{cum}}^{(i)}\}$. Calibrating with the *exact* deployment sampler preserves the level-α semantics (Prop. 5.3). By the Dvoretzky–Kiefer–Wolfowitz bound [5, 16], with probability $\geq 1 - \eta$ over n calibration draws,

$$\mathbb{P}_{H_0}^{(\text{sampler})}\big(\ell_{\text{cum}} \geq \hat{\tau}\big) \;\leq\; \alpha + \sqrt{\tfrac{1}{2n}\log\tfrac{2}{\eta}},$$

giving a finite-sample guarantee (Thm. 5.4); see Alg. 1.

4.5 Inference

We calibrate τ once per *(task, model, gate hyperparameters)* on a held-out state set, using the same frozen sampler as in Alg. 2 (and Alg. 1 for the Monte-Carlo procedure). Thereafter, all rollouts simply call Alg. 2 with this fixed $\hat{\tau}$; no re-calibration is needed unless hyperparameters change. The final a_0 is un-standardized and clipped before deployment (Alg. 2).

4.6 Composition with Value Gradients

To cleanly separate *risk control* from *return seeking*, we optionally apply a value-guidance step. Specifically, we ascend the *action-space* gradient of a learned critic with a simple schedule and clipping:

$$a_{t-1} \;\leftarrow\; a_{t-1} \;+\; \lambda_t\,\sigma_t^2\,\nabla_a \hat{Q}_\theta(s,a)\big|_{a=a_c}, \tag{9}$$

where $a_c \in \{\mu_u,\ \mu_{\text{LRT},t},\ (1-\rho)\mu_u + \rho\,\mu_{\text{LRT},t}\}$, and $\mu_{\text{LRT},t} = \mu_u + \beta_t(\mu_c - \mu_u)$ is the LRT-gated mean. We use a light, hand-tuned schedule (e.g., $\lambda_t \propto \sigma_t$) and gradient clipping to keep updates stable.

We use the evaluation center $a_c = \mu_{\text{LRT},t}$ as evidence opens the gate, suggesting the gradient is closer to the good direction. For more evaluation center choice discussion, refer to the extended version.

Algorithm 2 LRT-Guided Inference

Require: State s, steps T, τ, $(\beta_{\max}, \delta)$, noises $\{\sigma_t\}$
Ensure: Action a_0

$a_T \sim \mathcal{N}(0, I)$; $\ell \leftarrow 0$
for $t = T..1$ **do**
  $(\mu_u, \mu_c) \leftarrow \mu_\theta(t, s, a_t)$
  $\beta_t \leftarrow \beta_{\max}\text{Sigmoid}((\ell - \tau)/\delta)$
  $a_{t-1} \sim \mathcal{N}(\mu_u + \beta_t(\mu_c - \mu_u), \sigma_t^2 I)$
  $\ell \leftarrow \ell - \frac{\|a_{t-1}-\mu_c\|^2 - \|a_{t-1}-\mu_u\|^2}{2\sigma_t^2}$
  $a_t \leftarrow a_{t-1}$ (optionally use Eq. 9)
end for
return a_0

5 THEORY AND PROPERTIES

5.1 Gate Selection: UMP vs. Stability

In this part, we discuss the properties of hard and soft gates. For the full mathematical proof, see the extended version.

Proposition 5.1 (Neyman–Pearson optimality). *For the reverse chain conditioned on s with two simple hypotheses follows the two-head reverse model of Sec. 3.4. Then the Neyman-Pearson test that rejects H_0 when $\ell_{\text{cum}} \geq \tau$ is uniformly most powerful among all level-α tests.*

Lemma 5.2 (Soft→hard limit under logistic gate). *Assume the two-head reverse model of Sec. 3.4 and a* soft *gate denoted in Sec. 4.3.2. At step t, the proposal mean is $\mu_t^{\text{soft}} = \mu_u + \beta_t(\mu_c - \mu_u)$ and the reverse variance is $\sigma_t^2 I$. Fix $(\tau, \beta_{\max})$. As $\delta \to 0$, we have pointwise*

$$\mu_t^{\text{soft}} \;\to\; \mu_t^{\text{hard}} := \mu_u + \beta_{\max}\,\mathbf{1}\{\ell_{\text{cum}} \geq \tau\}\,(\mu_c - \mu_u),$$

and the trajectory law *induced by the soft-gated sampler converges weakly to that of the hard-gated sampler.*

Proposition 5.3 (Calibrated semantics under the deployment sampler). *Fix sampler hyperparameters $(\beta_{\max}, \delta)$ (and any deterministic Q-composition). Let $\hat{\tau}$ be the empirical $(1-\alpha)$ quantile of ℓ_{cum} computed from i.i.d. rollouts under H_0* with the same frozen sampler. *Then the realized false-activation rate at deployment satisfies*

$$\mathbb{P}_{H_0}^{(\text{sampler})}\big(\ell_{\text{cum}} \geq \hat{\tau}\big) \approx \alpha,$$

up to finite-sample fluctuations (see Thm. 5.4 for a DKW bound).

5.2 Finite-sample Calibration and Stability

Let F_0 be the CDF of the cumulative LLR ℓ_{cum} under H_0 for the *frozen* sampler (soft gate and, if present, the fixed Q-step). Given i.i.d.

calibration draws $\ell^{(1)}, \ldots, \ell^{(n)}$ and the empirical CDF $\widehat{F}_n$, define the plug-in quantile $\hat{\tau} = \inf\{x : \widehat{F}_n(x) \geq 1 - \alpha\}$.

5.2.1 Finite-sample Guarantee.

Theorem 5.4 (Calibration accuracy via DKW [5, 16]). *For any $\zeta \in (0, 1)$, with probability at least $1 - \zeta$ over the calibration sample,*

$$\mathbb{P}_{H_0}^{(sampler)}\big(\ell_{\text{cum}} \geq \hat{\tau}\big) \;\leq\; \alpha + \varepsilon_n, \qquad \varepsilon_n = \sqrt{\tfrac{1}{2n} \log \tfrac{2}{\zeta}}.$$

Rearranging the bound gives a one-line rule: to guarantee $\mathbb{P}_{H_0}^{(\text{sampler})}(\ell_{\text{cum}} \geq \hat{\tau}) \leq \alpha + \varepsilon$ with confidence at least $1 - \zeta$, it suffices to set sample size $n \;\geq\; \frac{1}{2\varepsilon^2} \log \frac{2}{\zeta}$. See the extended version for proof.

5.2.2 Stability.

Lemma 5.5 (Deterministic displacement bound). *At reverse step t, write $\Delta\mu_t := \mu_{c,t} - \mu_{u,t}$ and let g_t be the* clipped *critic gradient with $\|g_t\| \leq G$. Our update has deterministic mean*

$$m_t \;=\; \mu_{u,t} \,+\, \beta_t\, \Delta\mu_t \,+\, \lambda_t\, \sigma_t^2\, g_t, \qquad 0 \leq \beta_t \leq \beta_{\max},\; 0 \leq \lambda_t \leq \lambda_{\max}.$$

Hence the per-step deterministic displacement *from the background anchor is bounded by*

$$\|m_t - \mu_{u,t}\| \;\leq\; \beta_{\max}\, \|\Delta\mu_t\| \;+\; \lambda_{\max}\, \sigma_t^2\, G \;=:\; B_t. \tag{10}$$

If, additionally, $\|\Delta\mu_t\| \leq D$ (e.g. clamp on $\Delta\mu_t$) and $\sigma_t^2 \leq S^2$, then

$$\|m_t - \mu_{u,t}\| \leq \beta_{\max} D + \lambda_{\max} S^2 G \;=:\; B_{step}$$

for all t, and the cumulative deterministic *deviation from the background chain across T steps is at most $\sum_{t=1}^{T} B_t \leq T\, B_{step}$.*

Thus, the deterministic component of the sampler admits a uniform, hyperparameter-controlled drift bound both per step and over the full reverse trajectory; see the extended version for the detailed proof.

Implication for one-step LLR fluctuations. Under equal covariances, the one-step LLR increment admits the identity

$$\Delta\ell_t \;=\; \underbrace{\frac{\Delta\mu_t^\top \big(m_t - \frac{\mu_{c,t} + \mu_{u,t}}{2}\big)}{\sigma_t^2}}_{\text{deterministic part}} \;+\; \underbrace{\frac{\Delta\mu_t^\top z_t}{\sigma_t}}_{\text{zero-mean Gaussian}},$$

where $z_t \sim \mathcal{N}(0, I)$. Therefore, conditionally on $(m_t, \mu_{u,t}, \mu_{c,t})$,

$$\big|\mathbb{E}[\Delta\ell_t]\big| \;\leq\; \frac{\|\Delta\mu_t\|}{\sigma_t^2}\Big(B_t + \tfrac{1}{2}\|\Delta\mu_t\|\Big),$$
$$\mathrm{Var}(\Delta\ell_t) \;=\; \frac{\|\Delta\mu_t\|^2}{\sigma_t^2}.$$

In particular, if $\|\Delta\mu_t\| \leq D$ and is the fixed schedule $\sigma_t \geq \sigma_{\min} > 0$, then $\mathrm{Var}(\Delta\ell_t) \;\leq\; D^2/\sigma_{\min}^2$ and $\big|\mathbb{E}[\Delta\ell_t]\big| \;\leq\; \frac{D}{\sigma_{\min}^2}\big(B_{\text{step}} + \frac{D}{2}\big)$, so $\ell_{\text{cum}} = \sum_t \Delta\ell_t$ is sub-Gaussian with variance proxy $\sum_t \|\Delta\mu_t\|^2/\sigma_t^2 \leq T\, D^2/\sigma_{\min}^2$. This quantifies that our gate and gradient clipping keep both the *magnitude* of mean shifts and the *variance* of the accumulated evidence controlled. See proofs in the extended version.

5.3 Distributional Control and Return Bounds

Let $\mathcal{S}(s)$ denote the dataset action support at state s, and define the state-conditional OOD rate of a policy π by

$$\eta(\pi \,|\, s) = \Pr_{a \sim \pi(\cdot|s)}[a \notin \mathcal{S}(s)], \quad \eta(\pi) = \mathbb{E}_{s \sim d_{\text{eval}}}[\eta(\pi \,|\, s)].$$

Let $\hat{Q}$ be a learned critic and Q^{true} the environment action-value.[3] Assume the standard offline error split

$$\varepsilon_{\text{in}} := \sup_{a \in \mathcal{S}(s)} |\hat{Q}(s,a) - Q^{\text{true}}(s,a)|, \;\; \varepsilon_{\text{out}} := \sup_{a \notin \mathcal{S}(s)} |\hat{Q}(s,a) - Q^{\text{true}}(s,a)|,$$

and that Q^{true} is L-Lipschitz in a. Intuitively, $\nu := \varepsilon_{\text{out}} - \varepsilon_{\text{in}} \geq 0$ in offline RL due to extrapolation error. Additionally, write π_{LRT} for the LRT-gated policy and π_Q for pure Q-guided sampling, with $a_{\text{LRT}} \sim \pi_{\text{LRT}}(\cdot|s)$ and $a_Q \sim \pi_Q(\cdot|s)$.

Proposition 5.6 (Return comparison under offline errors). *Under the assumption above, let*

$$\Delta_{Q^{true}} := \mathbb{E}_s[Q^{true}(s, a_{LRT}) - Q^{true}(s, a_Q)],$$
$$\Delta_{\hat{Q}} := \mathbb{E}_s[\hat{Q}(s, a_{LRT}) - \hat{Q}(s, a_Q)],$$

the following inequality holds:

$$\Delta_{Q^{true}} \geq \Delta_{\hat{Q}} - 2\varepsilon_{\text{in}} - \nu\, \big(\eta(\pi_Q) + \eta(\pi_{\text{LRT}})\big).$$

Sketch. Decompose expectations into on-support and OOD parts. On-support deviations are bounded by ε_{in}, off-support by ε_{out}. Refer to the extended version for complete proof. □

Assumption (Monotone support w.r.t. gating). To relate $\eta(\pi_{\text{LRT}})$ to the calibrated level α, we use a mild monotonicity assumption. With the background head fixed, opening an LRT gate at any step does not decrease the probability that the final a_0 is outside $\mathcal{S}(s)$; if all T gates reject, the final a_0 lies in $\mathcal{S}(s)$ with high probability.

Proposition 5.7 (Trajectory-level OOD bound via single LLR threshold). *Assuming monotone support, let π_{LRT} be the LRT-gated policy with cumulative LLR threshold τ calibrated under H_0 at level α. Then the expected state-conditional OOD rate satisfies*

$$\eta(\pi_{\text{LRT}}) \;=\; \mathbb{E}_s\Big[\Pr_{a_0 \sim \pi_{\text{LRT}}(\cdot|s)}[a_0 \notin \mathcal{S}(s)]\Big] \;\lesssim\; \Pr_{H_0}(\ell_{\text{cum}} \geq \tau) \;\approx\; \alpha.$$

Here, ℓ_{cum} is the cumulative log-likelihood ratio over the reverse trajectory, and the approximation holds up to finite-sample calibration error (Thm. 5.4). Proof see the extended version.

Thus, smaller α directly yields fewer activations and lower expected OOD.

5.4 A sufficient condition for LRT to dominate Q

Combining Prop. 5.6 and Prop. 5.7 yields

$$\Delta_{Q^{\text{true}}} \;\geq\; \Delta_{\hat{Q}} \;-\; 2\varepsilon_{\text{in}} \;-\; \nu\big(\alpha + \eta(\pi_Q)\big).$$

Hence, if

$$\alpha \;\leq\; \alpha_{\max} \;:=\; \frac{\Delta_{\hat{Q}} \;-\; 2\varepsilon_{\text{in}} \;-\; \nu\, \eta(\pi_Q)}{\nu},$$

then $\mathbb{E}_s[Q^{\text{true}}(s, a_{\text{LRT}})] \geq \mathbb{E}_s[Q^{\text{true}}(s, a_Q)]$. Moreover, $\alpha_{\max} > 0$ if and only if $\Delta_{\hat{Q}} > 2\varepsilon_{\text{in}} + \nu\, \eta(\pi_Q)$.

[3] Any fixed evaluation state distribution d_{eval} may be used; in our experiments it is the dataset state marginal.

The bound is sufficient (and conservative): it uses the calibrated bound $\eta(\pi_{\mathrm{LRT}}) \lesssim \alpha$ and the critic gap $\Delta_{\hat{Q}}$, which may be biased off–support. Nevertheless it yields actionable levers: (i) decreasing $\beta_{\max}$ reduces ν and enlarges the feasible range; (ii) making $\eta(\pi_Q)$ small (a conservative anchor for the Q–step) increases $\alpha_{\max}$; (iii) evidence–tied guidance near μ_{LRT} can raise $\Delta_{\hat{Q}}$. In the favorable regime where $\eta(\pi_Q)$ is small, $\alpha_{\max}$ grows as $\alpha_{\max} = (\Delta_{\hat{Q}} - 2\varepsilon_{\mathrm{in}} - \nu\,\eta(\pi_Q))/\nu$, so moderate α can still be certified. In practice we treat $\widehat{\alpha}_{\max}$ as a diagnostic and sweep α on a log grid; the selected α is the knee of the return–risk curve (Fig. 2), while the certificate provides a sanity upper bound. For further discussion, see the extended version.

6 EXPERIMENTS

We empirically study whether LRT guidance delivers a calibrated, *interpretable* risk knob at inference time and yields a calibrated return–risk frontier relative to standard Q-guided sampling. Unless otherwise noted, diffusion training is vanilla (Sec. 4.2); all risk control is applied *only* at inference via the calibrated LRT gate.

6.1 Tasks and Datasets

We evaluate on continuous-control D4RL MuJoCo tasks [6]. Throughout, we standardize states and actions using dataset statistics. When we interact with the environment we map actions back to the original scale. For each task, we standardize states/actions using dataset means/stds and adopt the D4RL raw return (higher is better).

We report three metrics. *(i) Return:* per seed we evaluate N_{roll} episodes and compute the mean return, and tables report mean±std across seeds; *(ii) Realized Type-I:* under H_0 with the deployed sampler, the frequency $\Pr[\ell_{\mathrm{cum}} \geq \hat{\tau}]$ (target α), shown in Figure 2; *(iii) State-conditional OOD:* a k-NN proxy flagging actions as OOD if they exceed the q-th percentile distance of in-dataset k state-neighbors (default: $k = 50$, $q = 95\%$).

Our goal is to isolate the effect of *inference-time guidance* rather than to push absolute SOTA returns. Accordingly, we compare a standard Q-guided sampler (QG) against our LRT-guided sampler (LRT) and their simple composition (LRT+Q) under the same vanilla training pipeline, on representative D4RL MuJoCo tasks. This design controls for training confounders and highlights the *structure* of the sampler: does replacing a heuristic push by a calibrated, level-α gate improve the return–OOD trade-off? We therefore report raw return alongside a state-conditional OOD metric and realized Type-I, and interpret results through risk–performance curves and Pareto fronts rather than absolute leaderboards.

6.2 Baselines

- **LRT:** evidence-gated sampler (no Q step).
- **QG:** standard action-space Q update with schedules/clipping.
- **LRT+Q:** LRT mean + small Q step.

6.3 Implementation & Reproducibility

Setup. IQL critic (2×256 MLP, γ=0.99, expectile 0.7); advantages on standardized (s, a); labels: global top-p (p=0.2). Diffusion: T=50, DDPM linear noise $(1-\alpha_t)$ with endpoints $10^{-4} \to 2\times10^{-2}$, MLP backbone (SiLU), two ϵ heads; AdamW 2×10^{-4}, batch 1024, 150 epochs. Class balancing via EMA positive rate; optional within-positive soft weights (τ_A, $u_{\max}$). *Full configs in the extended version.*

Inference & calibration. Use reverse/posterior variance $\tilde{\sigma}_t^2$ for both heads and LLR; gate defaults $\beta_{\max}$=1, $\delta\in[1, 2]$. Calibrate a single threshold $\hat{\tau}$ on a held-out state set by the fixed-point update $\tau \leftarrow \mathrm{Quantile}_{1-\alpha}\{\ell_{\mathrm{cum}}(\tau)\}$, then *reuse* $\hat{\tau}$ for all deployments with the same $(\beta_{\max}, \delta)$ and Q-composition.

Protocol & compute. Each configuration: 10 seeds × 10 episodes over the entire training and evaluation time; report mean ± std; actions mapped back to env scale and clipped; using D4RL-raw returns. Calibration uses $n \in [3000, 5000]$ reverse trajectories.

6.4 Main Results

We perform structured hyperparameter sweeps to characterize the return–risk frontier. For **LRT**, we sweep the risk level $\alpha \in \{0.20, 0.10, 0.01, 0.005, 0.001\}$ with fixed $(\beta_{\max}, \delta)$. For **QG**, we sweep the maximum guidance step $\lambda_{\max} \in \{0.2, 0.1, 0.05, 0.02, 0.005\}$ (same schedule and clipping). For **LRT+Q**, we sweep the Cartesian product $\alpha \times \lambda_{\max}$ (25 settings per seed).

Table 1 reports a single operating point per method by selecting the setting that maximizes the *mean return across seeds* among the swept candidates, and then reporting mean±std across seeds; Fig. 2 shows the full sweeps as risk–performance curves and Pareto fronts. To mitigate seed-to-seed variability, we also report paired relative gaps vs. Q, see the extended version. Two notable cases are: (i) `halfcheetah-medium-replay-v2` exhibits substantially higher seed-to-seed variability, especially for value-guided samplers; and (ii) `walker2d-medium-v2` is an exception where LRT achieves lower return and higher OOD than value-guided baselines under the selected hyperparameters. We discuss both phenomena and report relative-gap statistics in the extended version.

Results at a glance. Across tasks, **LRT** often yields the lowest state-conditional OOD, acting as a conservative, low-risk anchor. Adding a small critic step (**LRT+Q**) typically increases return but can increase OOD, tracing a return–risk frontier. Notably, on `medium-replay` datasets LRT's OOD advantage is most pronounced, while on `medium` datasets value guidance often delivers larger return gains. See the extended version for per-task interpretation.

Reading the curves and fronts. In the left panels of Fig. 2, the realized Type-I (middle row) closely tracks the target α within DKW bands, validating calibration. As α decreases, return (top) drops modestly while state-conditional OOD (bottom) decreases consistently, exhibiting a smooth, monotone risk knob. The right panels (Pareto fronts) make the trade-off explicit: on Hopper and Walker2d, LRT shifts the frontier up-and-left relative to QG—higher return at lower OOD for a fixed α—whereas on HalfCheetah, adding a small Q-step can push return further at the cost of OOD, matching the intended "anchor-plus-exploitation" behavior.

6.5 Ablations

Effect of α: Smaller $\alpha \Rightarrow$ larger $\hat{\tau}$, fewer gate activations, *lower* state-conditional OOD with *modest* return drop (consistent with the left panels in Fig. 2). Cap $\beta_{\max}$ & temperature δ: Decreasing $\beta_{\max}$ contracts updates toward μ_u (uniform OOD reduction even at large α); decreasing δ sharpens the switch and approaches hard LRT (shifting Pareto up–left). Labeling p: Moderate $p \in [0.1, 0.3]$ is robust; calibration preserves Type-I semantics regardless of p, so Pareto trends

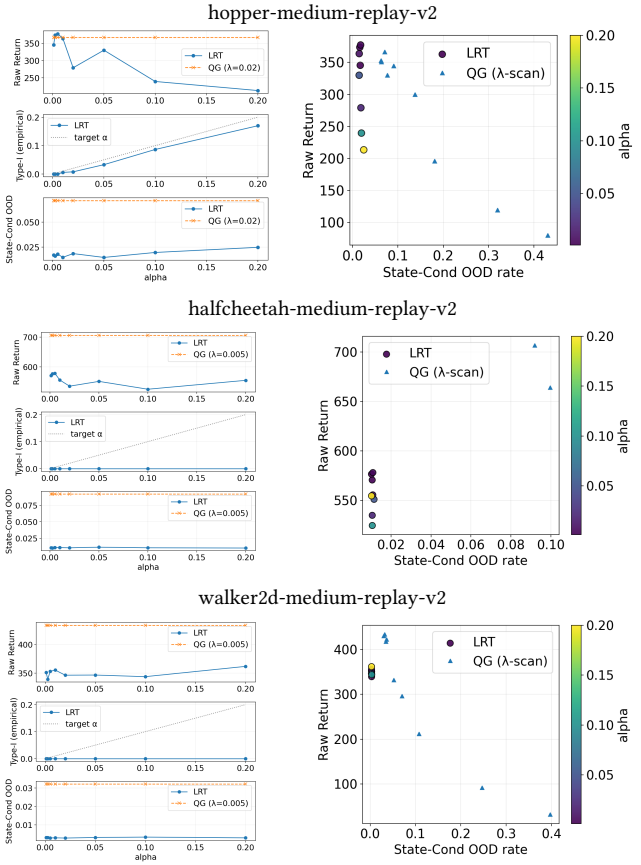


Figure 2: Risk–performance and Pareto fronts across tasks. Left of each row: risk–performance curves versus target α. We report return (top), realized Type-I (middle), and state-conditional OOD (bottom) for LRT (solid) and QG (dashed); the realized Type-I tracks the target (gray) within finite-sample DKW bands.Right of each row: Pareto fronts (OOD vs. return; color encodes α). LRT shifts the frontier up-and-left relative to QG on tasks where off-support critic error dominates, yielding higher return at lower OOD for the same α. Error bars denote standard errors over evaluation rollouts.

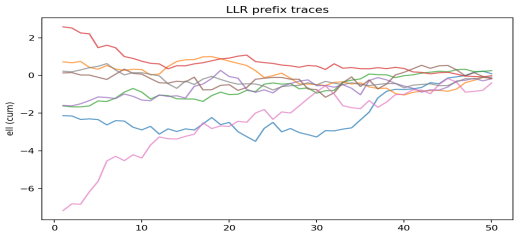


Figure 3: Prefix LLR across denoising steps on random states; lines are trajectories. Data: `hopper-medium-replay-v2`.

Table 1: Main results on D4RL MuJoCo. For each task/dataset and method, we report return (D4RL raw return) and state-conditional OOD (reported in units of $\times 10^{-2}$; lower is better). Results correspond to the operating point that maximizes mean return across seeds within each method's sweep.

Data Name	Mode	Return±std	OOD(10^{-2})±std
hopper-medium-replay-v2	lrt	329±22	**1.84±0.15**
hopper-medium-replay-v2	q	363±34	6.32±0.53
hopper-medium-replay-v2	lrt+q	**366±28**	6.67±0.77
halfcheetah-medium-replay-v2	lrt	558±111	**1.14±0.25**
halfcheetah-medium-replay-v2	q	598±216	13.13±2.72
halfcheetah-medium-replay-v2	lrt+q	**615±252**	13.93±3.20
walker2d-medium-replay-v2	lrt	315±21	**0.41±0.07**
walker2d-medium-replay-v2	q	373±44	3.76±0.43
walker2d-medium-replay-v2	lrt+q	**375±46**	3.86±0.87
hopper-medium-v2	lrt	744±71	**9.21±0.46**
hopper-medium-v2	q	1176±60	9.30±0.55
hopper-medium-v2	lrt+q	**1197±47**	11.32±0.83
halfcheetah-medium-v2	lrt	3526±38	**3.61±0.97**
halfcheetah-medium-v2	q	4404±31	5.11±0.80
halfcheetah-medium-v2	lrt+q	**4452±54**	5.02±0.87
walker2d-medium-v2	lrt	568±49	10.79±2.51
walker2d-medium-v2	q	2282±178	5.28±0.96
walker2d-medium-v2	lrt+q	**2448±196**	**4.94±0.95**

remain unchanged (see the extended version). Evidence dynamics: Prefix LLRs typically hover near zero at early high-noise steps and exceed $\hat{\tau}$ only later, if at all; the gate $\beta_t/\beta_{\max}$ remains near zero before crossing and then rises smoothly (Fig. 3). This matches the monotone risk knob in Fig. 2. Stress test under variance mismatch: We conduct a stress test under deliberate variance mis-specification of the background head at inference. Increasing mismatch inflates Type-I error and OOD, causing return degradation; matched variance ($s = 1$) preserves risk control, while severe mismatch ($s \geq 2$) degrades guarantees. See the extended version for details.

7 CONCLUSION

We presented **LRT-Diffusion**, a calibrated, inference-only guidance rule that turns diffusion-policy denoising into an evidence-gated process governed by a single, interpretable risk knob α. Under equal covariances, the hard likelihood-ratio test (LRT) is uniformly most powerful at level α; our smooth gate retains the same statistical semantics through a simple Monte-Carlo calibration matched to the deployed sampler. Empirically, LRT delivers calibrated control of Type-I risk and often improves the return–OOD trade-off over standard Q-guided sampling; moreover, in the experiments in Table 1, the composed sampler (LRT+Q) consistently achieves the highest return, highlighting that evidence-gated guidance can stabilize and amplify the benefits of value-gradient updates. Together, these demonstrate that *risk-aware diffusion guidance* can be achieved entirely at inference without modifying training objectives.

REFERENCES

[1] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. 2017. Constrained Policy Optimization. In *International Conference on Machine Learning*. PMLR, 22–31.

[2] Anastasios N Angelopoulos and Stephen Bates. 2021. A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification. *arXiv:2107.07511* (2021).

[3] Nair Ashvin, Dalal Murtaza, Gupta Abhishek, and L Sergey. 2020. Accelerating Online Reinforcement Learning with Offline Datasets. *CoRR, vol. abs/2006.09359* (2020).

[4] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems*, Vol. 34. 8780–8794.

[5] Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. 1956. Asymptotic Minimax Character of the Sample Distribution Function and of the Classical Multinomial Estimator. *Annals of Mathematical Statistics* 27, 3 (1956), 642–669.

[6] Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. 2020. D4RL: Datasets for Deep Data-Driven Reinforcement Learning. *arXiv preprint arXiv:2004.07219* (2020).

[7] Scott Fujimoto and Shixiang Shane Gu. 2021. A Minimalist Approach to Offline Reinforcement Learning. In *Advances in Neural Information Processing Systems*, Vol. 34. 20132–20145.

[8] Scott Fujimoto, David Meger, and Doina Precup. 2019. Off-Policy Deep Reinforcement Learning without Exploration. In *Proceedings of the 36th International Conference on Machine Learning*, Vol. 97. 2052–2062.

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, Vol. 33. 6840–6851.

[10] Jonathan Ho and Tim Salimans. 2022. Classifier-Free Diffusion Guidance. *arXiv preprint arXiv:2207.12598* (2022).

[11] Michael Janner, Yilun Du, Joshua Tenenbaum, and Sergey Levine. 2022. Planning with Diffusion for Flexible Behavior Synthesis. In *Proceedings of the 39th International Conference on Machine Learning*, Vol. 162. 9902–9915.

[12] Ilya Kostrikov, Ashvin Nair, and Sergey Levine. 2022. Offline Reinforcement Learning with Implicit Q-Learning. In *International Conference on Learning Representations*.

[13] Aviral Kumar, Justin Fu, George Tucker, and Sergey Levine. 2019. Stabilizing Off-policy Q-learning via Bootstrapping Error Reduction. In *Advances in Neural Information Processing Systems*, Vol. 32.

[14] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. 2020. Conservative Q-Learning for Offline Reinforcement Learning. In *Advances in Neural Information Processing Systems*, Vol. 33. 1179–1191.

[15] Romain Laroche, Paul Trichelair, and Remi Tachet Des Combes. 2019. Safe Policy Improvement with Baseline Bootstrapping. In *International Conference on Machine Learning*. PMLR, 3652–3661.

[16] Pascal Massart. 1990. The Tight Constant in the Dvoretzky–Kiefer–Wolfowitz Inequality. *Annals of Probability* 18, 3 (1990), 1269–1283.

[17] Jerzy Neyman and Egon S. Pearson. 1933. On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 231, 694-706 (1933), 289–337.

[18] Alexander Quinn Nichol and Prafulla Dhariwal. 2021. Improved Denoising Diffusion Probabilistic Models. In *International Conference on Machine Learning*. 8162–8171.

[19] Xue Bin Peng, Aviral Kumar, et al. 2019. Advantage-Weighted Regression: Simple and Scalable Off-Policy Reinforcement Learning. *arXiv preprint arXiv:1910.00177* (2019).

[20] Jan Peters and Stefan Schaal. 2007. Reinforcement Learning by Reward-Weighted Regression. In *Proceedings of the 24th International Conference on Machine Learning*. 745–750.

[21] Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. 2015. High-Confidence Off-Policy Evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 29.

[22] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. 2005. *Algorithmic Learning in a Random World.* Springer.

[23] Abraham Wald. 1992. Sequential Tests of Statistical Hypotheses. (1992), 256–298.

[24] Yifan Wu, George Tucker, and Ofir Nachum. 2019. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*.