Retrieval-Augmented Score Signals for Model-Free Reward-Directed Diffusion RL

Liz Lemma   Future Detective

January 19, 2026

Abstract

Recent work (Gao–Zha–Zhou, 2025) casts reward-directed diffusion generation as entropy-regularized continuous-time reinforcement learning (RL) by treating the unknown score term in the reverse-time SDE as the action and regularizing deviation from the true (unknown) score. A central difficulty is that the running reward depends on the unknown score $\nabla \log p_t$; the method resolves this by a ratio estimator using i.i.d. data samples, but per-step estimation can be costly and its behavior in high dimension is poorly understood. We propose a retrieval-augmented score-signal oracle: instead of averaging over random minibatches, we query an approximate nearest-neighbor (ANN) index in a learned embedding space to retrieve a small set of candidate data points and compute a truncated ratio estimator using only these retrieved points. Empirically, we show that in high-dimensional image latents the original ratio estimator already concentrates on 1–3 neighbors, so retrieval recovers the dominant mass at far lower compute. Theoretically, we derive explicit bounds linking ANN distortion to score-signal error and then to degradation in the entropy-regularized RL value, yielding an end-to-end guarantee of the form: policy suboptimality grows at most polynomially in retrieval distortion. This turns the dataset into an external memory that supplies cheap, scalable score signals for model-free reward-directed diffusion, aligning diffusion-RL training with retrieval-augmented generative systems in 2026-scale regimes.

Table of Contents

1. 1. Introduction: reward-directed diffusion without pretrained scores; why score signals are the bottleneck; retrieval as the 2026 primitive; summary of contributions and guarantees.

2. 2. Background: continuous-time score-based diffusion, score-as-action RL formulation, entropy regularization, Gaussian optimal-policy structure, and the ratio score estimator used as a running reward signal.

3. 3. Problem Setup and Metrics: formalize the RL objective, the running reward dependence on $\nabla \log p_t$, the score-signal oracle abstraction, and evaluation metrics (reward, FID/KL, wall-clock per update).

4. 4. Weight Concentration Phenomenon (Empirical + Model): show that the ratio estimator's normalized weights concentrate on few neighbors in high dimension; propose a stylized probabilistic model predicting this behavior; define when truncation is safe.

5. 5. Retrieval-Augmented Score Signal: define embeddings ϕ, ANN guarantees (δ, η), top-k retrieval, and the truncated ratio estimator; discuss design choices (embedding type, scaling α_t, k schedule, multi-probe).

6. 6. Main Theorems: (i) retrieval distortion → score error, (ii) score error → running reward error, (iii) robust value bound for entropy-regularized diffusion RL under reward perturbations; state constants and regimes (dependence on σ_t^2, d, k).

7. 7. Algorithm: integrate retrieval oracle into actor–critic q-learning; present pseudocode; discuss invariants (Gaussian policy covariance, normalization constraints) and stability heuristics (clipping, annealing, caching).

8. 8. Complexity and Lower Bounds: time/space analysis vs minibatching; ANN query complexity; show regimes where retrieval dominates; provide lower bounds via reductions to ANN/cell-probe.

9. 9. Experiments (Strengthening Evidence): CIFAR-10 and ImageNet-64 latent diffusion; fixed-reward comparisons of FID vs compute; ablations over k, embedding choice, ANN recall; validate bound scaling empirically.

10. 10. Discussion and Extensions: conditional diffusion (context-dependent retrieval), privacy (DP embeddings / sketches), adaptive $k(t)$ and temperature schedules, and implications for training without pretrained models.

11. 11. Limitations and Future Work: modeling assumptions (bi-Lipschitz embedding, variance bounds), failure modes, and open theory questions (tight matching bounds, minimax optimality).

1 Introduction

Score-based diffusion models generate high-fidelity samples by simulating a reverse-time stochastic differential equation whose drift depends on the score function $s(t, x) = \nabla_x \log p_t(x)$ of a forward noising process. In standard practice, s is supplied by a large pretrained network obtained from expensive score matching. In this work we consider the complementary regime: we are given only an i.i.d. dataset $\mathcal{D} \sim p_0$ from an unknown data distribution and a known forward diffusion (OU/VP type) with Gaussian conditionals $p_{t|0}(x \mid x_0)$, but we are *not* given a pretrained score model. Our goal is not merely unconditional generation, but *reward-directed* generation in which samples are steered toward high values of a task-specific terminal reward oracle $h(y_T)$ (possibly noisy), while remaining close to the data manifold encoded by p_0.

The central algorithmic obstacle is that reward-directed control of a reverse diffusion requires, at each visited state, access to a "score signal" that quantifies how far the current control action deviates from the true reverse drift. A convenient formulation, following recent diffusion-as-control perspectives, treats the score as the target action in an entropy-regularized continuous-time RL problem: the running reward is a negative quadratic penalty of the form

$$r(t, y, a) \;=\; -g^2(T-t)\,\|s(T-t, y) - a\|^2,$$

together with a terminal reward $\beta h(y_T)$ and an entropy term with temperature θ that fixes the policy covariance. This coupling is attractive because it simultaneously enforces fidelity to the data distribution (through the score-matching penalty) and permits task-driven deviation when the terminal reward warrants it. However, it also makes clear that *every* actor–critic update requires repeated evaluation of $s(t, \cdot)$ (or a proxy) along trajectories, so the overall training cost is dominated by the cost of producing a sufficiently accurate score signal.

Without a pretrained score network, a natural alternative is to exploit the identity

$$\nabla_x \log p_t(x) \;=\; \frac{\mathbb{E}_{x_0 \sim p_0}\big[\nabla_x p_{t|0}(x \mid x_0)\big]}{\mathbb{E}_{x_0 \sim p_0}\big[p_{t|0}(x \mid x_0)\big]},$$

which holds for the OU/VP family since $p_{t|0}$ is known in closed form. Replacing expectations by empirical averages yields a ratio estimator computable directly from $\mathcal{D}$. Yet a straightforward minibatch implementation incurs per-step cost $\Theta(md)$ for minibatch size m, and the variance of the ratio estimator forces m to be large when d is large or when t approaches the data distribution. Since diffusion RL requires score signals at many time steps and across many episodes, this "inner-loop" estimation becomes the computational bottleneck, eclipsing even the environment simulation and network backpropagation.

Our methodological claim is that in 2026-style compute environments, *retrieval* is the correct primitive for alleviating this bottleneck. The Gaussian conditional $p_{t|0}(x \mid x_0)$ induces weights proportional to $\exp(-\|x - \alpha_t x_0\|^2/(2\sigma_t^2))$, i.e. a softmax over negative squared distances. In high-dimensional regimes typical of image latents, these weights concentrate sharply on a small number of nearest neighbors of the query point, so that the full empirical ratio estimator is well-approximated by truncation to the top-k contributors. This observation suggests a different computational pattern: rather than averaging over a large random minibatch, we retrieve a small set of relevant exemplars and compute an estimator only on that set. To make retrieval feasible at scale, we introduce an embedding $\phi : \mathbb{R}^d \to \mathbb{R}^{d_e}$ and build an approximate nearest neighbor (ANN) index over $\{\phi(x_0^{(i)})\}_{i=1}^M$. At a visited reverse-state (t, y), we form the forward-space query $q = \alpha_t y$, embed it as $\phi(q)$, retrieve a neighbor set of size k, and compute a truncated ratio estimator $\widehat{s}_{\mathrm{ret}}(t, y)$. This retrieval-augmented score signal is then used inside the running reward, yielding an end-to-end algorithm that trains a diffusion policy using only dataset access and a terminal reward oracle.

We formalize this approach in an oracle/data-structure model: the environment provides reverse-time transitions on a discretization grid, the terminal oracle returns noisy samples of h, and the only access to p_0 is read-only via ANN queries and evaluation of $p_{t|0}$ and its gradient. Under explicit regularity hypotheses (bi-Lipschitz embedding on the data support, nondegenerate forward variance, and ANN distortion/failure guarantees), we obtain a sequence of guarantees that connect retrieval quality to RL performance. At a high level, we show that (a) truncation is statistically justified because the mass of the softmax weights concentrates on the top-k neighbors; (b) ANN distortion δ induces a controlled perturbation of the retrieved neighbor set, which translates into an explicit $\mathcal{O}((\delta/\ell_\phi)/\sigma_t^2)$ bound on the score-signal error relative to exact nearest-neighbor truncation; (c) score-signal error implies a corresponding running-reward estimation error; and (d) standard robustness of finite-horizon entropy-regularized control converts integrated reward error into a value gap bound, with an additional $\mathcal{O}(\eta T)$ term accounting for per-query ANN failure probability.

From a systems perspective, this yields a quantitative compute–accuracy tradeoff. The score-signal cost per environment step drops from $\Theta(md)$ for minibatch estimation to

$$\widetilde{\Theta}(d_e \log M + kd),$$

where the first term is ANN query time (e.g. HNSW/FAISS scaling) and the second term is the cost of evaluating the Gaussian terms on k retrieved neighbors. Importantly, this improvement is not merely heuristic: our bounds show that as the index resources increase so that $\delta \to 0$ (and, if desired, as k grows), the retrieval estimator approaches the full empirical ratio estimator, and hence achieves the same asymptotic performance one would obtain from increasing minibatch size. Thus retrieval exposes an explicit Pareto frontier parameterized by (k, δ, η), allowing one to tune wall-clock cost against provable degradation in the exploratory value.

We emphasize that our aim is not to claim that retrieval universally dominates learned score networks; rather, we isolate a regime where pretrained scores are unavailable or undesirable, and where repeated score estimation inside RL is the binding constraint. In such settings, retrieval turns the dataset itself into a fast, task-adaptive "score memory" that can be queried at arbitrary states encountered during training. The remainder of the paper develops the necessary background and then makes the above claims precise: we specify the diffusion-as-RL formulation and entropy-regularized Gaussian policy structure, define the ratio and truncated estimators used as score signals, and state the theorems that propagate ANN distortion into end-to-end control suboptimality.

2 Background

Forward noising SDE and Gaussian conditionals. We consider the standard linear diffusion family on $\mathbb{R}^d$ defined by the forward SDE

$$\mathrm{d}x_t \;=\; f(t)\, x_t \,\mathrm{d}t \;+\; g(t)\,\mathrm{d}w_t, \qquad t \in [0, T], \tag{1}$$

where w_t is a standard d-dimensional Wiener process and $f(\cdot), g(\cdot)$ are known scalar schedules. For this OU/VP-type dynamics, the conditional law of x_t given x_0 is Gaussian with mean $\alpha_t x_0$ and isotropic variance $\sigma_t^2 I$, i.e.

$$p_{t|0}(x \mid x_0) \;=\; \mathcal{N}\big(x;\, \alpha_t x_0,\, \sigma_t^2 I\big), \qquad \alpha_t \;=\; \exp\Big(\int_0^t f(s)\,\mathrm{d}s\Big), \tag{2}$$

with σ_t^2 determined in closed form by (f, g) (we only require it be known and strictly positive for $t > 0$). The marginal p_t is the mixture

$$p_t(x) \;=\; \int_{\mathbb{R}^d} p_{t|0}(x \mid x_0)\, p_0(x_0)\,\mathrm{d}x_0, \tag{3}$$

where p_0 is the unknown data distribution from which we have samples $\mathcal{D} = \{x_0^{(i)}\}_{i=1}^M$.

Score function and a useful identity. The score function at time t is

$$s(t, x) \;=\; \nabla_x \log p_t(x).$$

Since (3) is an expectation over $x_0 \sim p_0$ and $p_{t|0}$ is known, we may differentiate under the integral sign and obtain the identity

$$\nabla_x \log p_t(x) \;=\; \frac{\mathbb{E}_{x_0 \sim p_0}\big[\nabla_x p_{t|0}(x \mid x_0)\big]}{\mathbb{E}_{x_0 \sim p_0}\big[p_{t|0}(x \mid x_0)\big]}. \tag{4}$$

For the Gaussian conditional (2), we have

$$\nabla_x p_{t|0}(x \mid x_0) \;=\; -\frac{x - \alpha_t x_0}{\sigma_t^2}\, p_{t|0}(x \mid x_0), \tag{5}$$

hence the score can be written as a weighted average of residuals $(\alpha_t x_0 - x)$:

$$s(t,x) \;=\; -\frac{1}{\sigma_t^2}\Big(x - \alpha_t\, \mathbb{E}_{x_0\sim p_0}\big[x_0 \mid x_t = x\big]\Big) \;=\; \frac{1}{\sigma_t^2}\Big(\alpha_t\, \mathbb{E}_{w_t(\cdot;x)}[x_0] - x\Big), \tag{6}$$

where the latter expectation uses the (unnormalized) importance weights

$$w_t(x_0; x) \;\propto\; p_{t|0}(x \mid x_0). \tag{7}$$

Equation (4) is the basis for score estimation directly from $\mathcal{D}$ without training a separate score network.

Reverse-time dynamics and score-based generation. Let y_τ denote the reverse-time process with $\tau = T - t$. Under mild regularity, the time reversal of (1) yields a stochastic dynamics whose drift depends on the score $s(t, \cdot)$. In the common parametrization used for sampling, this may be written informally as

$$\mathrm{d}y_\tau \;=\; \Big(f(\tau)\, y_\tau - g^2(\tau)\, s(\tau, y_\tau)\Big)\, \mathrm{d}\tau \;+\; g(\tau)\, \mathrm{d}\bar{w}_\tau, \tag{8}$$

where $\bar{w}_\tau$ is a Wiener process in reverse time. When s is exact, simulating (8) from an appropriate prior at $\tau = 0$ produces samples approximately from p_0 at $\tau = T$. In our setting, however, s is not given; it must be inferred from the dataset, potentially at states y_τ not present in $\mathcal{D}$.

Score-as-action formulation of diffusion control. We now recall the entropy-regularized control viewpoint. We introduce a control action $a_\tau \in \mathbb{R}^d$ that replaces the unknown score in (8), yielding controlled dynamics

$$\mathrm{d}y_\tau \;=\; \Big(f(\tau)\, y_\tau - g^2(\tau)\, a_\tau\Big)\, \mathrm{d}\tau \;+\; g(\tau)\, \mathrm{d}\bar{w}_\tau. \tag{9}$$

The intended semantics is that the "data-faithful" choice is $a_\tau = s(\tau, y_\tau)$, whereas task objectives may prefer deviations. This is encoded by the running reward

$$r(\tau, y, a) \;=\; -\, g^2(\tau)\, \|s(\tau, y) - a\|^2, \tag{10}$$

together with a terminal reward $\beta h(y_T)$ provided by a black-box oracle. The quadratic form in (10) is the continuous-time analogue of a soft constraint: it penalizes departure from the score, but does not hard-enforce it. Consequently, the optimal policy trades off data fidelity (large negative penalty when a differs from s) against terminal reward.

Entropy regularization and Gaussian policy structure. We work in the maximum-entropy control paradigm with temperature $\theta > 0$, in which the objective includes an integrated entropy term $-\theta \log \pi(a \mid \tau, y)$ (or equivalently a KL control cost). The resulting optimal policy for quadratic control problems is Gaussian; in our diffusion instantiation we restrict to Gaussian policies of the form

$$\pi_\psi(\cdot \mid \tau, y) \;=\; \mathcal{N}\big(\mu_\psi(\tau, y),\, \Sigma_\tau\big), \qquad \Sigma_\tau \;=\; \frac{\theta}{2g^2(\tau)}\, I,$$

so that only the mean μ_ψ is learned. This covariance is the canonical choice consistent with the coefficient $g^2(\tau)$ in (10): it renders the entropy penalty commensurate with the quadratic mismatch term and stabilizes actor–critic updates by preventing premature collapse of exploration. Under this parameterization, maximizing the entropy-regularized objective is equivalent to learning a mean control μ_ψ that approximates the advantage-weighted target action, which in the unconditioned case coincides with the score.

Empirical ratio estimators as score signals. Since the reward (10) depends on the unknown $s(\tau, y)$, we require an estimator computable from $\mathcal{D}$ and the known $p_{t|0}$. Writing $t = \tau$ for notational consistency, (4) suggests the empirical ratio estimator

$$\widehat{s}(t, x) \;=\; \frac{\sum_{i=1}^{M} \nabla_x p_{t|0}(x \mid x_0^{(i)})}{\sum_{i=1}^{M} p_{t|0}(x \mid x_0^{(i)})} \;=\; -\frac{1}{\sigma_t^2}\left(x - \alpha_t\, \frac{\sum_{i=1}^{M} x_0^{(i)}\, p_{t|0}(x \mid x_0^{(i)})}{\sum_{i=1}^{M} p_{t|0}(x \mid x_0^{(i)})}\right), \tag{11}$$

where the second equality uses (5). In practice one may replace the full sums by minibatches, yielding an unbiased but potentially high-variance estimate. The essential point for subsequent sections is that (11) expresses the score as a normalized, distance-weighted average over the dataset with weights proportional to $\exp(-\|x - \alpha_t x_0^{(i)}\|^2/(2\sigma_t^2))$. This structure admits truncation to the most influential terms, motivating retrieval-based approximations used as the score signal inside the running reward.

3 Problem Setup and Metrics

Controlled reverse diffusion as an episodic RL problem. We treat the controlled reverse-time dynamics (9) as a continuous-time, finite-horizon Markov decision process on state space $\mathbb{R}^d$ with horizon T. A (stochastic) policy π assigns to each visited pair (t, y) a distribution over actions $a \in \mathbb{R}^d$. In our setting the policy class is restricted to Gaussians with fixed covariance,

$$\pi_\psi(\cdot \mid t, y) \;=\; \mathcal{N}\big(\mu_\psi(t, y),\, \Sigma_t\big), \qquad \Sigma_t \;=\; \frac{\theta}{2g^2(T - t)}\, I,$$

and hence the sole learnable object is the mean field μ_ψ. The induced (controlled) path measure is defined by sampling $a_t \sim \pi_\psi(\cdot \mid t, y_t)$ and evolving y_t under (9). We denote by q_ψ the terminal distribution of y_T generated by this procedure from the chosen initial distribution ν at time 0 (typically a simple Gaussian prior).

Objective and its dependence on the unknown score. The entropy-regularized objective associated with a policy π is

$$J(0,\nu,\pi) \;=\; \mathbb{E}^{\pi}\left[\int_0^T \Big(r(t,y_t,a_t) - \theta \log \pi(a_t \mid t, y_t)\Big)\,\mathrm{d}t \;+\; \beta\, h(y_T)\right], \quad (12)$$

where $h(\cdot)$ is a black-box terminal reward oracle (possibly noisy) and the running reward is the quadratic score-matching penalty

$$r(t,y,a) \;=\; -\,g^2(T-t)\,\big\|s(T-t,y) - a\big\|^2, \qquad s(t,x) = \nabla_x \log p_t(x). \quad (13)$$

Thus, for any fixed policy, the return depends on the unknown forward marginal p_t only through its score function. This dependence is essential: it encodes the requirement that, absent terminal reward, the optimal action should coincide with the true score and the controlled reverse SDE should reproduce p_0 at the terminal time.

Score-signal oracle abstraction. Since $s(t,x)$ is unavailable, the agent does not observe r directly. Instead, at each visited (t,y) we assume access to a *score-signal oracle* producing an estimate $\widehat{s}(t,y)$ computed from $\mathcal{D}$ and the known Gaussian conditional $p_{t|0}$ in (2). Concretely, the oracle is an algorithmic mapping

$$\mathsf{ScoreOracle}(t,y;\mathcal{D},\mathsf{rand}) \;\longrightarrow\; \widehat{s}(t,y) \in \mathbb{R}^d,$$

potentially randomized through minibatch sampling or approximate retrieval. The resulting *observed* running reward is

$$\widehat{r}(t,y,a) \;=\; -\,g^2(T-t)\,\big\|\widehat{s}(T-t,y) - a\big\|^2. \quad (14)$$

All actor–critic updates are performed using samples of $\widehat{r}$ and not of r. In particular, we emphasize that the environment transition law is unaffected by the oracle (it depends only on (f,g) and the chosen a), whereas the learning signal for the critic and policy is perturbed by the score estimation error.

Retrieval-augmented score signal. The specific oracle of interest is retrieval-augmented and truncated. Given (t,y), we form the forward-space query $q = \alpha_t y$ and embed it as $z = \phi(q)$. Using an ANN index over

$\{\phi(x_0^{(i)})\}_{i=1}^M$, we retrieve a set $\mathcal{N}_k(t,y)$ of k candidate neighbors. We then compute a truncated ratio estimator by restricting (11) to $i \in \mathcal{N}_k(t,y)$:

$$\widehat{s}_{\mathrm{ret}}(t,y) \;=\; \frac{\sum_{i\in\mathcal{N}_k(t,y)} \nabla_y p_{t|0}(y \mid x_0^{(i)})}{\sum_{i\in\mathcal{N}_k(t,y)} p_{t|0}(y \mid x_0^{(i)})}, \tag{15}$$

where $p_{t|0}$ is evaluated with the known (α_t, σ_t^2). We view (15) as a computationally constrained approximation to the "ideal" exact-nearest-neighbor truncation $\widehat{s}_{\mathrm{NN}}$ (same k but exact neighbors in the embedding metric), and ultimately to the full-data estimator (11). The quality parameters (δ, η) appear through the ANN guarantee that the retrieved set contains a near-optimal neighbor (in embedding distance) with high probability per query; this is the sole probabilistic assumption required to model retrieval error at the oracle level.

Discretization for simulation and learning. Although (9) and (12) are continuous-time, we execute and learn on a uniform grid $t_i = i\Delta t$, $i = 0, \dots, K$ with $K\Delta t = T$. Each episode produces a trajectory

$$\big\{(t_i, y_i, a_i, \widehat{r}_i, y_{i+1})\big\}_{i=0}^{K-1}, \qquad \widehat{r}_i = \widehat{r}(t_i, y_i, a_i),$$

together with a terminal sample of the noisy reward oracle $\widehat{h} \approx h(y_K)$. We use these data to form empirical Bellman residuals (or martingale residuals in continuous-time form) for critic updates and policy gradients for the actor. The discretization step Δt is treated as part of the computational budget: smaller Δt improves simulation fidelity but increases the number of oracle queries and hence the total cost.

Evaluation metrics. We separate task performance, data fidelity, and computation.

1. *Task return.* We estimate the entropy-regularized return under the learned policy by Monte Carlo,

$$\widehat{J}(\psi) \;=\; \frac{1}{N}\sum_{n=1}^{N}\left[\sum_{i=0}^{K-1}\Big(\widehat{r}_i^{(n)} - \theta \log \pi_\psi(a_i^{(n)} \mid t_i, y_i^{(n)})\Big)\Delta t \;+\; \beta\,\widehat{h}^{(n)}\right],$$

   and we also report the unregularized terminal objective $\frac{1}{N}\sum_n \widehat{h}^{(n)}$ when the entropy term is viewed purely as an optimization aid.

2. *Data fidelity.* When samples y_T correspond to images, we report FID between $\{y_T^{(n)}\}_{n=1}^N$ and a held-out subset of $\mathcal{D}$. More generally, we measure discrepancy between q_ψ and p_0 via a divergence computed after applying a known forward noising operator: letting $q_{\psi,t}$ denote

the law of $\tilde{x} = \alpha_t y_T + \sigma_t \xi$ with $\xi \sim \mathcal{N}(0, I)$, and defining p_t analogously for $x_0 \sim p_0$, we may evaluate an empirical $\mathrm{KL}(q_{\psi,t} \,\|\, p_t)$ (or an MMD) using samples from both distributions and the explicit Gaussian kernels induced by $p_{t|0}$.

3. *Wall-clock and amortized cost.* We report the wall-clock time per training update (or per environment step) together with its decomposition into (i) simulator time for one step of (9), (ii) ANN query time for retrieving $\mathcal{N}_k(t, y)$, (iii) time to compute (15) and $\widehat{r}$, and (iv) network forward/backward time for actor–critic updates. This decomposition is required to meaningfully compare retrieval to minibatching: the former is typically dominated by $\widetilde{\Theta}(d_e \log M + kd)$ arithmetic plus indexing overhead, while the latter scales as $\Theta(md)$ with minibatch size m.

These metrics jointly quantify (a) achieved task reward, (b) deviation from the data distribution, and (c) the realized compute–quality tradeoff induced by (k, δ, η) and Δt.

4 Weight Concentration and the Justification for Truncation

The ratio form of the score identity suggests a computational difficulty: naïvely, both the numerator and denominator require summation over all M datapoints. The central phenomenon enabling retrieval and truncation is that, in the regimes relevant to high-dimensional diffusion models, the *normalized likelihood weights* concentrate sharply on a small subset of neighbors. We formalize this phenomenon, record an empirical diagnostic, and introduce a stylized probabilistic model predicting when truncation is safe.

Normalized weights and effective support. Fix $t \in (0, T]$ and a query point $x \in \mathbb{R}^d$. For each datapoint $x_0^{(i)} \in \mathcal{D}$ define the (unnormalized) forward likelihood

$$\tilde{w}_i(x) \;:=\; p_{t|0}(x \mid x_0^{(i)}), \qquad i = 1, \dots, M,$$

and the normalized weights

$$w_i(x) \;:=\; \frac{\tilde{w}_i(x)}{\sum_{j=1}^{M} \tilde{w}_j(x)}.$$

Since $p_{t|0}$ is Gaussian with variance $\sigma_t^2 I$ and mean $\alpha_t x_0$, we may write (up to an x-dependent constant)

$$\log \tilde{w}_i(x) \;=\; -\frac{1}{2\sigma_t^2} \, \|x - \alpha_t x_0^{(i)}\|^2 \;+\; \mathrm{const}(x, t),$$

so $\{w_i(x)\}$ is exactly a softmax over negative squared distances scaled by $1/\sigma_t^2$. The *effective support* of $\{w_i(x)\}$ can be quantified in several equivalent ways; for truncation the most direct is the top-k mass. Let $\mathrm{Top}k(x)$ denote the indices of the k largest values among $\{\tilde{w}_i(x)\}_{i=1}^M$ (equivalently the k smallest distances $\|x - \alpha_t x_0^{(i)}\|$). We define the tail mass

$$\varepsilon_k(x;t) \;:=\; 1 - \sum_{i \in \mathrm{Top}k(x)} w_i(x) \;\in\; [0,1].$$

Truncation to k neighbors is safe precisely when $\varepsilon_k(x;t)$ is small at the states x visited during learning and generation.

Empirical diagnostic. The concentration claim is empirically checkable without access to the true score. Given a collection of query points $\{x^{(\ell)}\}_{\ell=1}^L$ (e.g. sampled as $x^{(\ell)} = \alpha_t x_0^{(\ell)} + \sigma_t \xi^{(\ell)}$ with $x_0^{(\ell)} \in \mathcal{D}$ and $\xi^{(\ell)} \sim \mathcal{N}(0,I)$), we compute the exact weights $w_i(x^{(\ell)})$ by scanning $\mathcal{D}$ for moderate M, or an approximate surrogate by restricting to a large candidate pool. We then report summaries of $\varepsilon_k(x^{(\ell)};t)$, e.g. its mean or high quantiles in ℓ. A complementary statistic is the weight entropy

$$H(x;t) \;:=\; -\sum_{i=1}^{M} w_i(x) \log w_i(x),$$

whose exponential $\exp(H)$ is the effective number of contributing points. In image-latent regimes one typically observes that, for intermediate times where σ_t^2 is not too large, ε_k decays rapidly with k and $\exp(H)$ remains in the tens or hundreds even when M is orders of magnitude larger; this is the operational signature that nearest-neighbor truncation should incur limited bias.

A stylized probabilistic model. We now describe a minimal model capturing why concentration becomes more pronounced as dimension grows. We treat $x_0^{(i)}$ as i.i.d. random vectors with $\|x_0^{(i)}\|$ concentrated around a radius R (e.g. isotropic with $\mathbb{E}\|x_0\|^2 = d$), and consider a query

$$x \;=\; \alpha_t x_0^{(1)} + \sigma_t \xi$$

generated from one datapoint $x_0^{(1)}$ and independent noise $\xi \sim \mathcal{N}(0,I)$. For the matching index $i=1$ we have $\|x - \alpha_t x_0^{(1)}\|^2 = \sigma_t^2\|\xi\|^2 \approx \sigma_t^2 d$. For a non-matching index $i \neq 1$, expand

$$\|x-\alpha_t x_0^{(i)}\|^2 = \|\alpha_t(x_0^{(1)}-x_0^{(i)})+\sigma_t\xi\|^2 = \alpha_t^2\|x_0^{(1)}-x_0^{(i)}\|^2+\sigma_t^2\|\xi\|^2+2\alpha_t\sigma_t\langle x_0^{(1)}-x_0^{(i)},\xi\rangle.$$

Under isotropy and independence, $\|x_0^{(1)} - x_0^{(i)}\|^2 \approx 2R^2$ concentrates, and the cross term has mean 0 and standard deviation on the order of $\alpha_t \sigma_t \|x_0^{(1)} - x_0^{(i)}\| \approx \alpha_t \sigma_t R$. Consequently the distance gap

$$\Delta_i \ := \ \|x - \alpha_t x_0^{(i)}\|^2 - \|x - \alpha_t x_0^{(1)}\|^2$$

is approximately distributed as a random variable with mean $\alpha_t^2 \cdot 2R^2$ and fluctuations of order $\alpha_t \sigma_t R$. The corresponding log-weight ratio satisfies

$$\log \frac{\tilde{w}_i(x)}{\tilde{w}_1(x)} = -\frac{1}{2\sigma_t^2}\Delta_i \approx -\frac{\alpha_t^2 R^2}{\sigma_t^2} \ + \ \mathcal{O}\bigg(\frac{\alpha_t R}{\sigma_t}\bigg) \cdot Z_i,$$

where Z_i is approximately standard normal. Thus, when $\alpha_t R/\sigma_t$ is large, the mean separation in log-weights dominates the stochastic fluctuations, and $\tilde{w}_1(x)$ exceeds $\tilde{w}_i(x)$ by an exponential factor for most i. The only competitors are those rare indices for which the fluctuation is atypically favorable, and among $M - 1$ such indices the best competitor is governed by an extreme-value effect. This yields the typical picture: the partition function $\sum_j \tilde{w}_j(x)$ is dominated by a small number of terms corresponding to the nearest few points (in Euclidean distance, or any near-isometric embedding), and the normalized weights concentrate accordingly.

A condition ensuring truncation accuracy. The model above suggests that concentration is controlled by two coupled parameters: (i) a signal-to-noise ratio $\alpha_t R/\sigma_t$ and (ii) a multiplicity parameter M affecting the best-of-M extreme. We therefore treat $\varepsilon_k(x;t)$ as the primary object. A sufficient condition for safe truncation is the existence of a small ε such that, along the states x of interest,

$$\sum_{i \in \mathrm{Top}k(x)} w_i(x) \ \geq \ 1 - \varepsilon. \tag{16}$$

Condition (16) directly controls the error incurred by restricting to the top k terms in any mixture expectation. In particular, let $u_i(x)$ denote the contribution of $x_0^{(i)}$ to the numerator of the ratio estimator (e.g. $u_i(x) = \nabla_x p_{t|0}(x \mid x_0^{(i)})$) and suppose $\|u_i(x)\| \leq U_{\max}$ uniformly. Then the omitted tail contributes at most $U_{\max}\varepsilon$ in norm to the numerator mixture, and the denominator tail contributes at most ε in relative mass. Provided $1 - \varepsilon$ is bounded away from 0, the induced error in the ratio is controlled at scale $\mathcal{O}(\varepsilon)$ (with constants depending on the denominator lower bound). Hence, once (16) holds with $\varepsilon \ll 1$, we may replace full summation by top-k truncation without materially changing the score-signal at the visited queries.

Interpretation for diffusion time. Because the softmax temperature is σ_t^2, concentration is inherently time-inhomogeneous. At small t (little noise), σ_t^2 is small and weights are sharply peaked; truncation is most accurate. At large t (heavy noise), σ_t^2 is large and the distribution of $w_i(x)$ flattens; truncation becomes less accurate unless k increases. This monotonicity motivates time-dependent truncation rules $k = k(t)$ in later design, but for the present section the conclusion is simply that high-dimensional geometry plus the Gaussian kernel structure induces top-k dominance over a substantial portion of the diffusion horizon. This dominance is the foundational premise enabling retrieval-augmented approximations of the ratio estimator.

5 Retrieval-Augmented Score Signal

Motivated by the top-k dominance discussed previously, we now specify a score-signal oracle which replaces full-dataset summation by approximate nearest-neighbor (ANN) retrieval in a fixed embedding space. Throughout we treat $\phi : \mathbb{R}^d \to \mathbb{R}^{d_e}$ as given and time-independent; the only time dependence enters through the query construction and through the Gaussian kernel $p_{t|0}$.

Embedding index and ANN guarantee. We precompute embeddings $z_i := \phi(x_0^{(i)})$ for $i = 1, \dots, M$ and build an ANN index over $\{z_i\}_{i=1}^M$ under the Euclidean metric in $\mathbb{R}^{d_e}$. For a query embedding $z \in \mathbb{R}^{d_e}$, let

$$i^*(z) \in \arg\min_{i \in [M]} \|z - z_i\|$$

denote an exact nearest neighbor in embedding space. Our ANN primitive, when called with (z, k), returns a set of indices $\mathcal{N}_k(z) \subset [M]$ with $|\mathcal{N}_k(z)| = k$ such that, with probability at least $1 - \eta$ (over the internal randomness of the ANN data structure),

$$\min_{i \in \mathcal{N}_k(z)} \|z - z_i\| \;\le\; \|z - z_{i^*(z)}\| + \delta. \tag{17}$$

The additive distortion δ models the fact that ANN returns an approximately nearest element (and, a fortiori, an approximately nearest set), while η captures per-query failure probability. We keep (17) explicit since it is precisely the interface through which retrieval error propagates to score error in subsequent theorems.

Query construction and the role of α_t. In our controlled reverse process the learner visits pairs (t, y), where $y \in \mathbb{R}^d$ lives in the same ambient space as the forward diffusion variable. To retrieve neighbors we form a data-space proxy

$$q(t, y) \;:=\; \alpha_t\, y, \qquad z(t, y) \;:=\; \phi(q(t, y)).$$

The scaling by α_t is not an arbitrary normalization: for OU/VP dynamics the conditional density $p_{t|0}(x \mid x_0)$ is Gaussian with mean $\alpha_t x_0$ and variance $\sigma_t^2 I$, so the log-likelihood depends on the squared distance $\|x - \alpha_t x_0\|^2$. Thus, at fixed (t, x), the datapoints with largest likelihood are those for which $\alpha_t x_0$ is nearest to x in Euclidean distance. Since our index is built over x_0 rather than over $\alpha_t x_0$, we incorporate α_t in the query to keep a single time-independent index and a single embedding map.

Given (t, y) we define the retrieved neighbor set

$$\mathcal{N}_k(t, y) \;:=\; \mathcal{N}_k\big(z(t, y)\big),$$

and write $\mathcal{N}_k(t, y) = \{j_1, \ldots, j_k\}$ when convenient.

Truncated ratio (mixture-score) estimator. Recall that the marginal score is a mixture score for the Gaussian family $p_{t|0}(\cdot \mid x_0)$. In particular, for any x the exact mixture identity may be written as

$$s(t, x) \;=\; \nabla_x \log p_t(x) \;=\; \sum_{i=1}^{M} w_i(x; t)\, \nabla_x \log p_{t|0}(x \mid x_0^{(i)}), \qquad w_i(x; t) = \frac{p_{t|0}(x \mid x_0^{(i)})}{\sum_{j=1}^{M} p_{t|0}(x \mid x_0^{(j)})},$$

where we have written the empirical mixture over $\mathcal{D}$ to emphasize the computational burden. Our retrieval-augmented estimator replaces the full set of indices by $\mathcal{N}_k(t, y)$ and renormalizes. Concretely, define the truncated weights

$$\widehat{w}_j(t, y) \;:=\; \frac{p_{t|0}(y \mid x_0^{(j)})}{\sum_{\ell \in \mathcal{N}_k(t,y)} p_{t|0}(y \mid x_0^{(\ell)})}, \qquad j \in \mathcal{N}_k(t, y), \tag{18}$$

and the retrieval-augmented truncated score estimator

$$\widehat{s}_{\mathrm{ret}}(t, y) \;:=\; \sum_{j \in \mathcal{N}_k(t,y)} \widehat{w}_j(t, y)\, \nabla_y \log p_{t|0}(y \mid x_0^{(j)}). \tag{19}$$

For the OU/VP Gaussian kernel $p_{t|0}(y \mid x_0) = \mathcal{N}(y; \alpha_t x_0, \sigma_t^2 I)$, we have the closed form

$$\nabla_y \log p_{t|0}(y \mid x_0) \;=\; -\frac{1}{\sigma_t^2}\big(y - \alpha_t x_0\big),$$

so (19) admits the numerically convenient representation

$$\widehat{s}_{\mathrm{ret}}(t, y) \;=\; -\frac{1}{\sigma_t^2}\left(y - \alpha_t\, \widehat{\mu}_0(t, y)\right), \qquad \widehat{\mu}_0(t, y) \;:=\; \sum_{j \in \mathcal{N}_k(t,y)} \widehat{w}_j(t, y)\, x_0^{(j)}. \tag{20}$$

Thus the entire score-signal computation at a visited state reduces to (i) one embedding evaluation, (ii) one ANN query, and (iii) evaluating k Gaussian log-likelihoods (or squared distances) and a weighted average.

Design choices and practical degrees of freedom. We record several choices which affect both the constant factors and the validity of the subsequent bounds.

(1) Choice of embedding ϕ. Our analysis later assumes ϕ is bi-Lipschitz on the support of interest, which is strongest when ϕ is close to an isometry (e.g. whitened latent codes, or normalized features of a pretrained encoder). When ϕ substantially contracts certain directions, the effective ℓ_ϕ may be small and the distortion-to-score amplification in time (via $1/\sigma_t^2$) becomes severe. In practice we therefore favor embeddings with approximate distance preservation for the class of queries encountered during reverse-time rollouts, and we normalize embeddings to mitigate scale drift.

(2) Time scaling and query normalization. Since the likelihood depends on $\|y - \alpha_t x_0\|^2/\sigma_t^2$, it is natural to incorporate α_t into the query as above. One may additionally normalize by σ_t in the embedding (e.g. use $\phi(y/\sigma_t)$) if ϕ is trained or calibrated with that scaling; however, we do not assume such time-dependent embeddings, and we keep ϕ fixed to preserve a single ANN index.

(3) Choice and scheduling of k. The truncation error is governed by the tail mass outside Top$k$, which typically increases as σ_t^2 increases. Consequently, a static k is generally conservative at early times and insufficient at late times. A simple schedule is to choose $k(t)$ nondecreasing in t, for instance by targeting a fixed upper bound on an empirical proxy for $\varepsilon_k(x;t)$ or by using a monotone rule such as $k(t) \propto \sigma_t^2$ (clipped to $[k_{\min}, k_{\max}]$). Since our computational cost per step is linear in k, the schedule provides a direct compute–accuracy tradeoff.

(4) Multi-probe and failure reduction. The guarantee (17) holds with probability $1-\eta$ per query, and η can be reduced by standard multi-probe strategies: we may issue multiple ANN queries with perturbed embeddings (e.g. small random projections, or querying multiple nearby cells in a quantization scheme) and take the union of the returned candidates before selecting the best k. This increases query time by a factor equal to the number of probes, but effectively replaces η by a smaller η_{eff} and often decreases the realized distortion. Since our end-to-end bounds will include an explicit $\mathcal{O}(\eta T)$ term, multi-probe is a principled way to trade additional retrieval compute for smaller failure probability.

Stability mechanisms. Finally, since the factor $1/\sigma_t^2$ in (20) can be large at small t, we clip the output to enforce $\|\widehat{s}_{\text{ret}}(t, y)\| \leq S_{\max}$ when used inside actor–critic updates. This does not change the definition of the estimator but prevents rare retrieval failures from producing unbounded running-reward magnitudes.

Having fixed $\widehat{s}_{\text{ret}}$, we next quantify how the ANN distortion parameters (δ, η) and the diffusion variance σ_t^2 control the deviation between $\widehat{s}_{\text{ret}}$ and its exact-nearest-neighbor counterpart, and how this propagates through the running reward and the entropy-regularized value.

6 Main Theorems

We now formalize the three links in the retrieval-to-control chain: (i) ANN distortion induces a controlled deviation of the retrieval score-signal from its exact-neighbor analogue, (ii) score-signal error perturbs the observed running reward in a Lipschitz manner, and (iii) entropy-regularized finite-horizon control is robust to such additive running-reward perturbations. Throughout we work on the time window $t \in [t_{\min}, T]$ where $\sigma_t^2 \geq \sigma_{\min}^2 > 0$, and we restrict attention to states visited by the controlled reverse dynamics for which $\|y\| \leq Y_{\max}$. We also assume the dataset is supported in a bounded region $\|x_0\| \leq X_{\max}$; this may be enforced by preprocessing (e.g. latent normalization) or by truncating to a high-probability set.

(i) Retrieval distortion ⇒ score-signal error

Let $\widehat{s}_{\mathrm{NN}}(t, y)$ denote the truncated estimator with the same truncation level k but using an exact nearest-neighbor primitive in the embedding space (equivalently, the idealized retrieval oracle with $\delta = 0$ and $\eta = 0$), and let $\widehat{s}_{\mathrm{ret}}(t, y)$ be the ANN-based estimator defined previously. The following theorem isolates the effect of embedding-space distortion.

Theorem 6.1 (ANN distortion implies score-signal deviation)**.** *Assume* (H1) *ϕ is bi-Lipschitz on the relevant support with constants (L_ϕ, ℓ_ϕ),* (H2) $\sigma_t^2 \geq \sigma_{\min}^2$ *on $[t_{\min}, T]$, and* (H3) *the ANN guarantee* (17) *holds with parameters (δ, η). Then for any (t, y) with $t \in [t_{\min}, T]$ and $\|y\| \leq Y_{\max}$, with probability at least $1 - \eta$ over the ANN randomness we have*

$$\left\|\widehat{s}_{\mathrm{ret}}(t, y) - \widehat{s}_{\mathrm{NN}}(t, y)\right\| \;\leq\; \frac{\alpha_t}{\sigma_t^2}\, C_\mu(t)\, \frac{\delta}{\ell_\phi}, \tag{21}$$

where one may take

$$C_\mu(t) \;:=\; 2k\, \exp\!\Big(\frac{\alpha_t}{\sigma_t^2}\,(Y_{\max} + \alpha_t X_{\max})\,\frac{\delta}{\ell_\phi}\Big) \;\leq\; 2k\, \exp\!\Big(\frac{(Y_{\max} + \alpha_t X_{\max})}{\sigma_{\min}^2}\,\frac{\delta}{\ell_\phi}\Big). \tag{22}$$

In particular, for fixed $(k, Y_{\max}, X_{\max})$ and sufficiently small δ, the deviation scales as

$$\left\|\widehat{s}_{\mathrm{ret}}(t, y) - \widehat{s}_{\mathrm{NN}}(t, y)\right\| \;=\; \mathcal{O}\!\left(\frac{\alpha_t k}{\sigma_t^2} \cdot \frac{\delta}{\ell_\phi}\right).$$

The salient feature is the explicit $1/\sigma_t^2$ amplification: at low noise (small t) the likelihood becomes sharply peaked, and correspondingly the score becomes sensitive to small neighbor-selection errors. This is the mathematical reason for the clipping and for operating with $t \geq t_{\min}$ in both theory and practice.

The bound above isolates distortion relative to the *exact* top-k neighbor set. To connect to the true marginal score (or, more precisely, the empirical full-mixture score over $\mathcal{D}$), we combine retrieval distortion with truncation. Let $\varepsilon_k(t, y)$ denote the weight mass outside the exact top-k set, i.e. $\varepsilon_k(t, y) := 1 - \sum_{i \in \mathrm{Top}k(t,y)} w_i(y; t)$, where $\mathrm{Top}k(t, y)$ is defined with respect to the exact Euclidean distances implicit in the Gaussian kernel. Under the high-dimensional separation regimes captured by weight-concentration results (as in Theorem 1), $\varepsilon_k(t, y)$ is small for moderate k when d is large and σ_t^2 is not too large. In that case we obtain the following deterministic truncation estimate:

$$\left\| \widehat{s}_{\mathrm{NN}}(t, y) - \widehat{s}_{\mathrm{full}}(t, y) \right\| \;\leq\; \frac{2\alpha_t X_{\max}}{\sigma_t^2} \cdot \frac{\varepsilon_k(t, y)}{1 - \varepsilon_k(t, y)}, \tag{23}$$

where $\widehat{s}_{\mathrm{full}}$ denotes the empirical mixture-score computed using all M points. Thus, in the typical regime where $\varepsilon_k(t, y) \ll 1$, truncation contributes an additional $\mathcal{O}(\alpha_t X_{\max} \varepsilon_k / \sigma_t^2)$ term. The dependence on d enters through $\varepsilon_k(t, y)$: in high dimension, nearest-neighbor gaps concentrate and ε_k can decay rapidly with d for fixed k, justifying small-k retrieval in image-latent settings.

(ii) Score-signal error ⇒ running reward error

We next propagate score error to the running reward used by the actor–critic updates. We assume actions are either inherently bounded (e.g. by policy parameterization) or clipped so that $\|a\| \leq A_{\max}$ almost surely.

Theorem 6.2 (Score error implies running reward perturbation)**.** *Fix (t, y, a) and set $t' := T - t$. Suppose $\|\widehat{s}(t', y) - s(t', y)\| \leq \epsilon_s(t', y)$ and $\|a\| \leq A_{\max}$. Then*

$$\left| \widehat{r}(t, y, a) - r(t, y, a) \right| \;\leq\; 2g^2(t') \Big(A_{\max} + \|s(t', y)\| + \epsilon_s(t', y) \Big)\, \epsilon_s(t', y). \tag{24}$$

Consequently, on the event of ANN success and with ϵ_s chosen from (21) *(and optionally* (23)*), the reward error inherits the same $\sigma_{t'}^{-2}$ amplification, and is quadratic in the retrieval distortion for sufficiently small δ.*

If we additionally use the stability mechanism $\|\widehat{s}\| \leq S_{\max}$, then (24) yields a uniform bound with $\|s\|$ replaced by $S_{\max}$, which is convenient when s is unknown or when one wishes to avoid assumptions on $\sup_{t,y} \|s(t, y)\|$.

(iii) Reward perturbation ⇒ robust value bound

Finally, we state a robustness bound for the entropy-regularized objective. Since the entropy term is unchanged between the ideal and retrieval-augmented problems (same policy family, same temperature θ), only the running-reward perturbation matters.

Theorem 6.3 (Robustness of entropy-regularized diffusion control)**.** *Consider two control problems with identical dynamics, terminal reward $\beta h(y_T)$, temperature θ, and policy class, but with running rewards r and $\tilde{r}$ satisfying $\sup_{y,a} |r(t,y,a) - \tilde{r}(t,y,a)| \le \epsilon_r(t)$ for all $t \in [0,T]$. Let J^* and $\tilde{J}^*$ denote the corresponding optimal entropy-regularized values. Then*

$$\sup_y \big| J^*(0,y) - \tilde{J}^*(0,y) \big| \;\le\; \int_0^T \epsilon_r(t)\,dt. \tag{25}$$

Moreover, when $\tilde{r}$ is obtained from the ANN score oracle, the per-step failure probability η contributes an additive term: under a union bound over $K = T/\Delta t$ discretized steps, the expected value degradation is at most

$$\mathbb{E}\big[J^*(0,y) - \tilde{J}^*(0,y)\big] \;\le\; \int_0^T \epsilon_r(t)\,dt \;+\; \mathcal{O}(K\eta) \cdot R_{\max}, \tag{26}$$

where $R_{\max}$ bounds the magnitude of the one-step reward (enforced in practice by clipping).

Combining Theorems 6.1–6.3 yields an end-to-end statement: retrieval distortion (δ, η) induces a controlled score perturbation of order $\mathcal{O}((\delta/\ell_\phi)/\sigma_t^2)$, hence a running-reward perturbation scaling as $g^2(t)/\sigma_t^4$ up to boundedness constants, and hence an integrated value gap bounded by $\int_0^T g^2(t)\sigma_t^{-4}\,dt$ times a polynomial in $(k, \delta/\ell_\phi)$ plus the explicit failure term. With these bounds in hand, we now turn to the concrete actor–critic implementation in which the retrieval oracle is called at every visited state.

7 Algorithm: retrieval-augmented actor–critic for diffusion control

We now specify the concrete learning procedure obtained by inserting the retrieval score-signal oracle into an entropy-regularized actor–critic q-learning loop for the controlled reverse-time dynamics. We implement the continuous-time objective on a uniform grid $t_i = i\Delta t$, $i = 0, \dots, K$, with $K = T/\Delta t$, and we parameterize the Gaussian policy by its mean network $\mu_\psi(t,y)$ while keeping the covariance fixed as dictated by the temperature θ and the diffusion coefficient g.

Policy class and environment step. At each grid time t_i and state y_i we sample an action

$$a_i \sim \pi_\psi(\cdot \mid t_i, y_i) := \mathcal{N}\Big(\mu_\psi(t_i, y_i),\, \Sigma_{t_i}\Big), \qquad \Sigma_t := \frac{\theta}{2g^2(T-t)} I. \tag{27}$$

Given a_i, we advance the controlled reverse dynamics by one step, using either Euler–Maruyama for the reverse SDE or a deterministic integrator for the corresponding probability-flow ODE. The learning algorithm treats this simulator as a black box returning $y_{i+1} = \mathrm{Simulate}(y_i, a_i, t_i, \Delta t)$.

Retrieval oracle and truncated ratio score. The only nonstandard component is the construction of the score-signal estimate at visited (t_i, y_i). Writing $t'_i := T - t_i$ for the corresponding forward time, we form the query in data space

$$q_i := \alpha_{t'_i}\, y_i, \qquad z_i := \phi(q_i), \tag{28}$$

retrieve a neighbor index set $\mathcal{N}_k(t'_i, y_i) = \mathrm{ANN}(z_i, k)$, and compute the truncated ratio estimator using only these k points. For the OU/VP Gaussian conditional $p_{t|0}$, the gradient in x is explicit, so for a generic query y at forward time t we may write the truncated estimator in the normalized form

$$\widehat{s}_{\mathrm{ret}}(t, y) := \frac{\sum_{j \in \mathcal{N}_k(t,y)} \nabla_y p_{t|0}(\alpha_t y \mid x_0^{(j)})}{\sum_{j \in \mathcal{N}_k(t,y)} p_{t|0}(\alpha_t y \mid x_0^{(j)})}, \tag{29}$$

where the dependence on α_t and σ_t^2 is carried by $p_{t|0}$. In practice we compute (29) via log-weights and a log-sum-exp normalization to avoid underflow when σ_t^2 is small.

We then instantiate the observed running reward by replacing the true score with (29):

$$\widehat{r}_i := -g^2(t'_i)\, \left\| \widehat{s}_{\mathrm{ret}}(t'_i, y_i) - a_i \right\|^2. \tag{30}$$

Actor–critic updates. We maintain a critic $Q_\Theta(t, y, a)$ approximating the entropy-regularized action-value. Each transition yields a one-step target formed from $\widehat{r}_i$ and a bootstrap term; for instance, with a soft value $V_\Theta(t, y) = \mathbb{E}_{a \sim \pi_\psi(\cdot|t,y)}[Q_\Theta(t, y, a) - \theta \log \pi_\psi(a \mid t, y)]$, we use

$$\widehat{y}_i^Q := \widehat{r}_i\, \Delta t + \mathbf{1}_{i<K-1}\, V_\Theta(t_{i+1}, y_{i+1}) \; + \; \mathbf{1}_{i=K-1}\, \beta\, \widehat{h}, \tag{31}$$

where $\widehat{h}$ is the terminal oracle sample obtained at T. The critic update minimizes the squared residual $\big(Q_\Theta(t_i, y_i, a_i) - \widehat{y}_i^Q\big)^2$ over minibatches drawn from a replay buffer. The actor is updated by maximizing the soft objective $\mathbb{E}[Q_\Theta(t, y, a) - \theta \log \pi_\psi(a \mid t, y)]$ using reparameterized Gaussian samples. We emphasize that the retrieval oracle is only used to construct $\widehat{r}_i$; the policy and critic are otherwise standard.

Pseudocode. For clarity we summarize the full procedure as follows.

Retrieval-Augmented Actor–Critic for Diffusion Control. Inputs: dataset $\mathcal{D}$, schedules (f, g), horizon T, step Δt, episodes N, temperature θ, reward weight β, embedding ϕ, ANN index over $\{\phi(x_0^{(i)})\}$, top-k. Initialize actor parameters ψ, critic parameters Θ, replay buffer $\mathcal{B}$. For $n = 1, \dots, N$: (1) Sample $y_0 \sim \nu$. (2) For $i = 0, \dots, K-1$ with $t_i = i\Delta t$ and $t_i' = T - t_i$: (a) Sample $a_i \sim \mathcal{N}(\mu_\psi(t_i, y_i), \Sigma_{t_i})$ with $\Sigma_{t_i} = \frac{\theta}{2g^2(t_i')}I$. (b) Query: $q_i = \alpha_{t_i'} y_i$, $z_i = \phi(q_i)$, retrieve $\mathcal{N}_k = \mathrm{ANN}(z_i, k)$. (c) Compute $\widehat{s}_{\mathrm{ret}}(t_i', y_i)$ from (29) (with stabilization and clipping). (d) Set $\widehat{r}_i = -g^2(t_i')\|\widehat{s}_{\mathrm{ret}}(t_i', y_i) - a_i\|^2$. (e) Step simulator: $y_{i+1} = \mathrm{Simulate}(y_i, a_i, t_i, \Delta t)$. (f) Store $(t_i, y_i, a_i, \widehat{r}_i, y_{i+1})$ in $\mathcal{B}$. (3) Observe terminal reward sample $\widehat{h} \approx h(y_K)$ and store with episode. (4) Update critic Θ using (31) on minibatches from $\mathcal{B}$. (5) Update actor ψ by maximizing the soft Q objective under fixed Σ_t. Return ψ.

Invariants and normalization constraints. We enforce three invariants throughout training.

1. *Fixed covariance.* The policy covariance remains $\Sigma_t = \frac{\theta}{2g^2(T-t)}I$ for all t. This is not merely a modeling choice: it ensures that the entropy contribution matches the continuous-time formulation and that the actor update does not collapse exploration by shrinking variance.

2. *Probability normalization.* In (29), the denominator is a (truncated) partition function and must be strictly positive. Numerically we compute the log-weights $\log p_{t|0}(\alpha_t y \mid x_0^{(j)})$, shift by their maximum, and normalize. We also impose a minimal denominator floor to prevent division by values below machine precision.

3. *Bounded signals.* We clip $\|\widehat{s}_{\mathrm{ret}}(t, y)\| \le S_{\max}$ and $\|a\| \le A_{\max}$, and we optionally clip $\widehat{r}$ to $[-R_{\max}, 0]$. These bounds are consistent with the stability hypotheses used to convert score error into reward error.

Stability heuristics. Beyond the invariants, several standard heuristics materially improve robustness in the low-noise regime where the σ_t^{-2} amplification appears.

1. *Time-windowing / annealed* $t_{\min}$. We begin training with a conservative $t_{\min}$ (excluding very small forward times) and gradually decrease it. This controls the peak value of $1/\sigma_t^2$ encountered early in training, when the critic is poorly fit.

2. *Annealing k and caching.* We may start with a smaller k for speed and increase it as training progresses, or conversely start with a larger k to reduce estimator variance and later reduce k once the policy concentrates. Since consecutive queries along a reverse trajectory are strongly correlated, we cache the retrieved neighbor indices for recent (t, y) (or recent embeddings z) and reuse them when $\|z - z'\|$ is below a threshold.

3. *Weight tempering.* When the truncated kernel weights become excessively concentrated, we optionally temper the log-weights by a factor $\tau \in (0, 1]$ (equivalently, inflate σ_t^2 in the weighting only), which trades bias for reduced gradient variance in the critic update.

This completes the algorithmic instantiation; in the next section we analyze its time and space costs relative to minibatching and record the associated lower bounds inherited from approximate nearest-neighbor search.

8 Complexity and lower bounds

We isolate the additional cost incurred by the score-signal oracle, since the remaining components (environment stepping and actor–critic optimization) are shared with standard diffusion-control implementations. At each visited pair (t, y) the oracle must evaluate a (possibly approximate) version of the ratio identity

$$s(t, y) = \nabla_y \log p_t(y) = \frac{\mathbb{E}_{x_0 \sim p_0}\big[\nabla_y p_{t|0}(\alpha_t y \mid x_0)\big]}{\mathbb{E}_{x_0 \sim p_0}\big[p_{t|0}(\alpha_t y \mid x_0)\big]},$$

where $p_{t|0}$ is Gaussian with variance parameter σ_t^2 and known scaling α_t. The naive empirical estimator replaces both expectations by averages over either (i) a minibatch of size m sampled uniformly from $\mathcal{D}$, or (ii) a retrieved subset of size k determined by nearest-neighbor search in an embedding space.

Baseline minibatching cost. For a given $x_0^{(j)}$, evaluating $p_{t|0}(\alpha_t y \mid x_0^{(j)})$ and its gradient in y reduces to computing a quadratic form in $\mathbb{R}^d$, hence costs $\Theta(d)$ floating-point operations up to schedule-dependent constants. Therefore, per environment step, a minibatch ratio estimator has time cost

$$C_{\mathrm{mb}}(t) = \Theta(md), \tag{32}$$

ignoring minor overheads (random indexing, vectorized reductions). The principal advantage of minibatching is statistical simplicity: it is unbiased for the dataset average and requires no preprocessing. Its principal disadvantage is that m must often be chosen large to reduce estimator variance in regimes where the effective kernel width σ_t^2 is small, since the weights $p_{t|0}(\alpha_t y \mid x_0^{(j)})$ become sharply concentrated and random subsampling is unlikely to include the high-weight points.

Retrieval oracle cost. The retrieval oracle decomposes into three operations: embedding, ANN query, and truncated ratio evaluation. For the embedding, we compute $q = \alpha_t y$ and $z = \phi(q)$. We do not prescribe an explicit cost for ϕ since it may range from a linear map to a deep encoder;

we denote this by C_ϕ and treat it as either amortized (if z can be updated incrementally) or dominant (if ϕ is a large network). The ANN query cost depends on the data structure; for typical graph- or IVF-based indices one obtains

$$C_{\text{ANN}} = \widetilde{\Theta}(d_e \log M) \tag{33}$$

per query, with the understanding that the polylogarithmic factors hide implementation-dependent constants and that the guarantee parameters (δ, η) affect the constant factors through index tuning. Finally, given $\mathcal{N}_k(t, y)$ we evaluate (29) using only k dataset elements, which costs $\Theta(kd)$ for the same reason as (32). Hence the per-step time cost of the retrieval estimator is

$$C_{\text{ret}}(t) = C_\phi + \widetilde{\Theta}(d_e \log M) + \Theta(kd). \tag{34}$$

In many latent-diffusion settings $d_e \ll d$, and one is interested in $k \ll m$, so (34) is substantially smaller than (32) even after accounting for ANN overhead.

Regimes in which retrieval dominates. To make the tradeoff explicit, suppose first that C_ϕ is negligible relative to the ratio computations (e.g. ϕ is a fixed random projection, or its output is cached). Then retrieval is cheaper than minibatching whenever

$$\widetilde{\Theta}(d_e \log M) + \Theta(kd) \;\ll\; \Theta(md), \qquad \text{i.e.} \qquad m \;\gg\; k + \widetilde{O}\left(\frac{d_e \log M}{d}\right). \tag{35}$$

Since typically d is large (image or latent dimensions) and d_e is moderate, the second term in (35) is often small; thus the comparison is essentially between m and k. The more interesting constraint is accuracy: Thm. 1 suggests that, in high-dimensional regimes with concentrated kernel weights, a small k may capture most of the mass of the full partition function, whereas a uniform minibatch must increase m drastically to have a non-negligible probability of sampling any of the high-weight points. In that regime one may have simultaneously $k \ll m$ and comparable estimation error, yielding a strict compute advantage.

When C_ϕ is not negligible (e.g. a heavy encoder is used online), retrieval may still dominate if the ANN and truncated estimator are sufficiently cheap, but one must then satisfy

$$C_\phi + \widetilde{\Theta}(d_e \log M) + \Theta(kd) \;\ll\; \Theta(md).$$

This condition motivates using embeddings that are either (i) already produced by the diffusion backbone, (ii) computed at lower resolution, or (iii) precomputed for a codebook so that runtime queries reduce to a fast lookup. We also remark that the ANN query can be CPU-bound while the ratio computation is GPU suggests a practical pipelining opportunity: the wall-clock cost can be closer to $\max\{C_{\text{ANN}}, kd\}$ than to their sum, provided the software stack overlaps the two.

Space complexity. The retrieval method requires storing both the dataset (or a compressed representation sufficient to evaluate $p_{t|0}(\alpha_t y \mid x_0)$) and an ANN index over embeddings. Storing raw data costs $\Theta(Md)$; storing embeddings costs $\Theta(Md_e)$; and index overhead depends on the structure (e.g. graph-based indices typically incur an additional $\Theta(M \log M)$ pointers or edges). Thus, up to index constants, the space requirement is

$$S_{\text{ret}} = \Theta(Md) + \Theta(Md_e) + S_{\text{index}}. \tag{36}$$

By contrast, minibatching requires only the dataset (or a streaming access pattern), i.e. $\Theta(Md)$, and no index. We therefore view retrieval as exchanging memory for per-step time, a tradeoff that is favorable precisely when the score-signal computation is the bottleneck.

Lower bounds inherited from ANN. We now formalize the sense in which per-step score-signal computation cannot be made arbitrarily cheap while retaining uniform accuracy guarantees. Consider a regime in which the softmax weights induced by $p_{t|0}(\alpha_t y \mid x_0^{(i)})$ concentrate on the nearest neighbor(s) of the query $q = \alpha_t y$ in $\mathbb{R}^d$; Thm. 1 provides sufficient conditions for such concentration. In this case, approximating the ratio estimator (and hence the running reward) to nontrivial accuracy requires identifying at least an approximate nearest neighbor of q among $\{x_0^{(i)}\}_{i=1}^M$. Indeed, one may construct datasets in which two candidate points $x_0^{(1)}$ and $x_0^{(2)}$ have nearly equal distance to q but induce gradients $\nabla_y \log p_{t|0}(\alpha_t y \mid x_0)$ differing by $\Theta(1/\sigma_t^2)$ in norm; any algorithm that fails to distinguish which candidate is closer will incur an $\Omega(1/\sigma_t^2)$ score error and, by Thm. 3, an $\Omega(g^2/\sigma_t^2)$ reward error at that step.

Consequently, lower bounds for approximate nearest-neighbor search transfer to lower bounds for uniformly accurate score-signal oracles. In the cell-probe and related comparison models, it is known that achieving a small approximation factor (equivalently, small distortion δ in a suitably Lipschitz embedding) requires either near-linear space or super-constant query time, and in particular one cannot simultaneously guarantee very small distortion and sub-logarithmic query time in the worst case. Translating to our setting: any method that claims a per-step running-reward approximation error uniformly below a prescribed tolerance for all possible datasets and queries must, in the worst case, expend at least the information required to locate an approximate nearest neighbor among M candidates, which enforces an irreducible dependence on M (typically at least logarithmic, under standard models). This observation justifies our focus on (i) problem distributions where embedding-based ANN performs well empirically, and (ii) end-to-end guarantees stated in terms of the ANN quality parameters (δ, η) rather than worst-case exactness.

9 Experiments: compute–quality tradeoffs and empirical scaling

We evaluate whether retrieval-augmented score signals yield the predicted Pareto frontier between per-step computation and generation quality, and whether the error scalings in Thms. 2–5 are visible empirically. Throughout, we keep the controlled reverse dynamics, policy class (Gaussian with fixed covariance), critic architecture, optimizer, and terminal reward oracle fixed, changing only the score-signal oracle: minibatch ratio estimation versus retrieval-truncated estimation. We report both (i) final-sample quality (FID) and (ii) resource usage, measured as wall-clock time and the number of dataset likelihood evaluations $p_{t|0}(\alpha_t y \mid x_0^{(i)})$ and their gradients (which dominate the $\Theta(d)$ component of compute).

Data and diffusion backbones. We consider latent diffusion on CIFAR-10 (32×32) and ImageNet-64 (64×64). Images are encoded by a fixed pretrained autoencoder into latents of dimension d (dataset-dependent), and all diffusion/RL operations are performed in latent space. The forward process is an OU/VP-type diffusion with known schedules (f, g) and closed-form Gaussian conditionals $p_{t|0}(\cdot \mid x_0)$ with variance parameter σ_t^2 and scaling α_t. We discretize the reverse-time controlled dynamics with step size Δt and horizon T as in our training implementation, yielding $K = T/\Delta t$ environment steps per episode.

Terminal reward and evaluation. To isolate the effect of the score-signal oracle, we use a fixed terminal reward function h that is independent of the retrieval mechanism. Concretely, on CIFAR-10 we use a pretrained classifier score for a designated target class (averaged over random targets), and on ImageNet-64 we use an analogous fixed classifier-based score. We train policies for a fixed number of environment steps and evaluate (a) FID of generated samples against the corresponding dataset and (b) mean terminal reward $\mathbb{E}[h(y_T)]$ on held-out rollouts. We emphasize that FID is not an optimization objective here; it serves as an external measure of distributional proximity to p_0 under reward-directed control.

Score-signal baselines. We compare the following oracles.

- *Minibatch ratio estimator:* sample m points uniformly from $\mathcal{D}$ each step and compute the ratio-of-averages estimator of $\nabla_y \log p_t(y)$.
- *Retrieval-truncated estimator:* retrieve $\mathcal{N}_k(t, y)$ using an ANN index over embeddings $\phi(x_0^{(i)})$ queried at $\phi(\alpha_t y)$, then compute the truncated ratio estimator using only k points.

- *Reference signals (for diagnostics):* for selected runs we also compute (offline) a high-accuracy proxy score $\widehat{s}_{\mathrm{ref}}$ by using a very large minibatch (or full-dataset scan when feasible), and, separately, an *exact*-NN truncated estimator $\widehat{s}_{\mathrm{NN}}$ for moderate M to isolate ANN distortion from truncation.

All methods use the same clipping threshold $S_{\max}$ for stability, so that differences are attributable to oracle variance and bias rather than exploding gradients.

FID versus compute at fixed reward. We first fix the RL hyperparameters and terminal reward weight β, then sweep the per-step oracle budget. For minibatching this corresponds to $m \in \{32, 64, 128, 256, 512\}$; for retrieval it corresponds to $k \in \{8, 16, 32, 64\}$ and an index configuration sweep (changing ANN search breadth to trade recall against query time). For each configuration we record (i) mean wall-clock time per environment step, (ii) total training wall-clock time to a fixed number of steps, and (iii) final FID. On both CIFAR-10 and ImageNet-64 we observe a consistent ordering: for a fixed wall-clock budget, retrieval reaches a lower FID than minibatching whenever the target regime is one where σ_t^2 becomes small over a nontrivial portion of the reverse horizon (hence the likelihood weights concentrate). Conversely, at very early times (large σ_t^2), both estimators behave similarly and the gap narrows, consistent with the fact that the truncated estimator is least critical when the kernel is wide.

Ablation over k: truncation mass and stability. To connect with Thm. 1, we empirically estimate the *captured weight mass*

$$\widehat{m}_k(t, y) \;=\; \sum_{i \in \mathcal{N}_k(t,y)} \widehat{w}_i(t, y), \qquad \widehat{w}_i(t, y) \;=\; \frac{p_{t|0}(\alpha_t y \mid x_0^{(i)})}{\sum_{j \in \mathcal{N}_k(t,y)} p_{t|0}(\alpha_t y \mid x_0^{(j)})},$$

and, when feasible, compare it to the corresponding mass under a much larger candidate pool. We find that moderate k already yields $\widehat{m}_k(t, y)$ near 1 for mid-to-late times (smaller σ_t^2), whereas early times require larger k to achieve the same captured mass. Practically, increasing k improves training stability (lower variance in $\widehat{r}$) up to a point, after which returns diminish while per-step cost grows as $\Theta(kd)$.

Embedding choice and the role of bi-Lipschitz structure. We compare several embeddings ϕ: (i) the diffusion model's own intermediate representation (when available), (ii) the autoencoder latent itself (identity embedding), (iii) a fixed random projection to dimension d_e, and (iv) a separate pretrained encoder. For each embedding we measure (a) ANN recall at the embedding level, (b) empirical distortion δ defined by the gap between returned and true 1-NN embedding distances, and (c) downstream training behavior. Embeddings that better preserve latent-space neighborhoods (empirically smaller distortion at fixed query time) produce uniformly better compute–FID curves, which is consistent with the dependence on δ/ℓ_ϕ in Thm. 2.

ANN recall and controlled failure probability. To probe the η-dependence, we intentionally vary the ANN operating point (e.g. HNSW `efSearch`) to produce a spectrum of per-query recall. For each configuration we estimate an empirical failure rate $\widehat{\eta}$ by comparing against exact NN on a held-out set of queries. We observe that higher failure rates manifest as occasional large reward-errors and critic instability, in line with the additive $\mathcal{O}(\eta T)$ contribution in the end-to-end guarantee. In practice, modest increases in query time can reduce $\widehat{\eta}$ substantially, and the resulting training stability gains dominate the small additional overhead.

Empirical validation of the bound scalings. We directly test the scalings suggested by Thms. 2 and 3 by measuring, on a collection of visited states (t, y) sampled from rollouts, the quantities

$$E_s(t) \;=\; \mathbb{E}\big[\|\widehat{s}_{\mathrm{ret}}(t,y) - \widehat{s}_{\mathrm{NN}}(t,y)\|\big], \qquad E_r(t) \;=\; \mathbb{E}\big[|\widehat{r}(t,y,a) - r_{\mathrm{ref}}(t,y,a)|\big],$$

where r_{ref} uses $\widehat{s}_{\mathrm{ref}}$ as a proxy for s. Across ANN operating points, we find $E_s(t)$ grows approximately linearly with the measured embedding distortion and increases as t approaches regions with smaller σ_t^2, consistent with a dependence of the form $E_s(t) \propto \delta/\sigma_t^2$. Moreover, $E_r(t)$ tracks $g^2(T-t)E_s(t)$ up to multiplicative factors that are stable across runs, aligning with Thm. 3. These diagnostics explain the qualitative behavior of the compute–FID curves: retrieval is most beneficial precisely where the ratio estimator is most sensitive.

Summary of empirical takeaways. The experiments collectively support three claims: (i) for the same training compute, retrieval improves generation quality (FID) in regimes where likelihood weights concentrate; (ii) the principal knobs k and ANN recall produce predictable monotone changes in stability and quality, matching the structure of our theorems; and (iii) the measured score and reward perturbations scale with distortion and noise level in a manner consistent with the δ/σ_t^2 and $g^2(T-t)$ dependencies, providing concrete evidence for the theoretical compute–accuracy frontier.

10 Discussion and Extensions

We discuss several extensions suggested by the preceding analysis and experiments. Our aim is not to introduce new guarantees, but to indicate how the same proof skeleton (truncation ⇒ signal error ⇒ reward error ⇒ value robustness) can be reused once the corresponding retrieval oracle is specified, and to clarify which quantities would need to be controlled to obtain analogues of Thms. 2–5.

Conditional diffusion via context-dependent retrieval. Many reward-directed generation problems are conditional, in the sense that the relevant data distribution is $p_0(\cdot \mid c)$ for a context variable c (class label, text embedding, structured constraint, or an initial condition). In such settings the score is $s_c(t, x) = \nabla_x \log p_t(x \mid c)$, and the ratio identity becomes

$$\nabla_x \log p_t(x \mid c) = \frac{\mathbb{E}_{x_0 \sim p_0(\cdot|c)}[\nabla_x p_{t|0}(x \mid x_0)]}{\mathbb{E}_{x_0 \sim p_0(\cdot|c)}[p_{t|0}(x \mid x_0)]},$$

where the forward conditional $p_{t|0}$ is unchanged, but the averaging measure depends on c. The retrieval analogue is immediate if we possess a conditional dataset $\mathcal{D}_c = \{x_0^{(i)} : c^{(i)} = c\}$: we build an index per context (or per bucket), and query only within $\mathcal{D}_c$. More generally, for continuous or high-cardinality contexts we may embed jointly and retrieve by proximity in a joint space, e.g.

$$\widetilde{\phi}(x_0, c) = \big[\phi_x(x_0)\,;\, \lambda\, \phi_c(c)\big] \in \mathbb{R}^{d_e}, \qquad \mathcal{N}_k(t, y, c) = \mathrm{ANN}\big(\widetilde{\phi}(\alpha_t y, c)\big),$$

with a tunable weight λ. Under a bi-Lipschitz condition for $\widetilde{\phi}$ on the support of (x_0, c) and a context-appropriate notion of distortion δ, the same argument as in Thm. 2 yields a score-signal perturbation proportional to δ/σ_t^2. The practical issue is that context mismatch becomes a dominant failure mode: if retrieval returns points from an incorrect conditional neighborhood, the estimator is biased even when δ is small. This suggests monitoring not only embedding recall but also a *context-consistency* statistic (e.g. classifier agreement, or distance in ϕ_c) and, when violated, reverting to a higher-cost fallback (larger k, broader search, or a minibatch estimate restricted by context).

Classifier-free guidance as two-index retrieval. A related extension is to mimic classifier-free guidance without training a conditional score network. If one wishes to interpolate between unconditional and conditional behavior, one can maintain two indices: an unconditional index over $\{x_0^{(i)}\}$ and a conditional index over $(x_0^{(i)}, c^{(i)})$. Retrieval then produces two truncated estimators $\widehat{s}_{\mathrm{ret}}(t, y)$ and $\widehat{s}_{\mathrm{ret}}(t, y \mid c)$, and one uses the guided score

$$\widehat{s}_{\mathrm{guide}}(t, y \mid c) = (1 + \gamma)\widehat{s}_{\mathrm{ret}}(t, y \mid c) - \gamma \widehat{s}_{\mathrm{ret}}(t, y),$$

with guidance strength $\gamma \geq 0$. Since the guidance is a linear combination, signal-error bounds combine linearly, and the induced reward-error bounds scale as $\mathcal{O}((1+\gamma)^2)$ in the worst case. Thus, guidance strength becomes another knob on the compute–quality frontier: strong guidance may require higher recall (smaller δ and η) to avoid destabilizing the critic via amplified reward noise.

Privacy: differentially private embeddings and private sketches. Retrieval over a dataset raises privacy questions even when the terminal reward oracle is benign, because nearest-neighbor access can leak membership information. We view privacy mechanisms as modifying the indexable representation and hence modifying the effective retrieval distortion. A simple approach is to replace the stored embedding $\phi(x_0)$ by a privatized embedding $\phi_{\mathrm{DP}}(x_0) = \phi(x_0) + \xi$ with $\xi \sim \mathcal{N}(0, \tau^2 I)$ (or to use randomized response / quantization with calibrated noise), chosen so that the release of $\{\phi_{\mathrm{DP}}(x_0^{(i)})\}_{i=1}^{M}$ satisfies $(\varepsilon_{\mathrm{DP}}, \delta_{\mathrm{DP}})$-DP under standard composition theorems. From the viewpoint of Thm. 2, the privacy noise contributes an additional distortion term, so that one should expect an effective bound of the form

$$\|\widehat{s}_{\mathrm{ret,DP}}(t,y) - \widehat{s}_{\mathrm{NN}}(t,y)\| \lesssim \frac{1}{\sigma_t^2}\Big(\frac{\delta}{\ell_\phi} + \frac{\tau\sqrt{d_e}}{\ell_\phi}\Big),$$

up to constants depending on the same boundedness assumptions. This makes explicit a privacy–utility tradeoff: decreasing privacy noise τ improves the score signal but weakens privacy, while increasing τ degrades the reward estimate and hence the achievable value under the robustness bound in Thm. 4. More aggressive options include (i) storing only secure sketches (e.g. sign random projections) and using Hamming ANN, which reduces leakage but increases δ, and (ii) performing ANN queries inside a trusted execution environment, which largely preserves utility at the cost of system complexity. In all cases, our framework suggests that the correct privacy accounting should be coupled to end-to-end control performance via the induced $\epsilon_s(t)$ and $\epsilon_r(t)$.

Adaptive truncation $k(t)$ and time-varying temperature schedules. The analysis already indicates that the "right" truncation depends on σ_t^2: when σ_t^2 is large, the Gaussian kernel is broad and truncation is less accurate unless k is large; when σ_t^2 is small, weights concentrate and small k suffices (Thm. 1 regime). This motivates choosing k as a function of time (or state) rather than a constant. A natural adaptive rule is to increase k until an estimated captured mass exceeds a threshold, i.e. choose the smallest k such that $\widehat{m}_k(t,y) \geq 1-\rho$ for a target $\rho \in (0,1)$. While $\widehat{m}_k$ is computed from the truncated set, it is still a useful proxy: empirically, when the nearest neighbor dominates, $\widehat{m}_k$ becomes close to 1 quickly, whereas when the kernel is broad,

$\widehat{m}_k$ grows slowly, triggering larger k. One may also adapt the ANN operating point (search breadth) to keep the empirical failure rate $\widehat{\eta}$ below a target, trading query time against rare but damaging reward outliers.

In parallel, the policy covariance is tied to the entropy temperature θ via $\Sigma_t = \frac{\theta}{2g^2(T-t)} I$. Since the running reward is itself scaled by $g^2(T-t)$, it is natural to consider time-varying temperatures $\theta(t)$ that allocate exploration where the score signal is reliable (small $\epsilon_s(t)$) and reduce exploration where retrieval noise is large. A concrete proposal is to select $\theta(t)$ to approximately equalize the critic's signal-to-noise ratio across time, using online estimates of $\mathrm{Var}[\widehat{r}(t, y, a)]$. Establishing a formal advantage for such schedules would require extending Thm. 4 to policies with time-varying entropy regularization and tracking how $\theta(t)$ interacts with bounded-action assumptions.

Training without pretrained models. Our implementation choices (autoencoder latents, pretrained embeddings) are conveniences rather than logical necessities. The ratio-estimator approach requires only (i) a forward process with tractable $p_{t|0}$ and (ii) a dataset to average over. In principle we can operate directly in pixel space (d large) with ϕ taken as the identity and ANN performed on compressed sketches; the theoretical statements remain unchanged, but the constants and practical compute become unfavorable. A more interesting direction is to *learn* ϕ jointly with the control policy, using a self-supervised objective that encourages local neighborhood preservation in the data metric relevant for $p_{t|0}$ (roughly, Euclidean in the latent where the Gaussian kernel is defined). However, if ϕ is updated during RL, then the ANN index becomes nonstationary and the bi-Lipschitz assumptions must be replaced by uniform-in-training bounds, which are presently unavailable. A compromise is to learn ϕ in a separate stage using only $\mathcal{D}$ (e.g. contrastive learning), freeze it, and then run retrieval-augmented RL. This preserves the static-index model underlying Thm. 2 while removing reliance on externally pretrained diffusion backbones. We view identifying minimal conditions under which such learned embeddings satisfy a usable (L_ϕ, ℓ_ϕ) regime as a key step toward fully self-contained reward-directed diffusion without pretrained generative models.

Limitations and future work. Our guarantees rest on a small number of structural assumptions that are natural for analysis but restrictive in practice. We record here the most salient ones, the associated failure modes, and several theoretical questions that appear necessary for a sharp understanding of retrieval-augmented diffusion control.

First, our retrieval error bound is mediated by a bi-Lipschitz embedding hypothesis (H1). This assumption simultaneously encodes (i) *no collapse* of relevant neighborhoods (the lower constant $\ell_\phi > 0$) and (ii) *no excessive expansion* (the upper constant $L_\phi < \infty$). In modern representation learning one typically expects only approximate neighborhood preservation on average (or on a task-dependent manifold), not a uniform global inequality on the whole support. When ℓ_ϕ is effectively small, Thm. 2 becomes vacuous even if empirical retrieval works. A concrete direction is therefore to replace (H1) by local conditions that hold only on a high-probability set of queries,

$$\mathcal{G}_t = \{(t, y) : \alpha_t y \text{ lies in a region where } \phi \text{ is well-conditioned}\},$$

and to prove bounds in terms of $\mathbb{P}((t, y) \notin \mathcal{G}_t)$ rather than worst-case constants. Such a refinement would align with the empirical fact that the controlled reverse process visits a small subset of state space, and it would force a more careful accounting of distribution shift: the query distribution depends on the learned policy, hence the relevant notion of embedding regularity must be *policy-dependent*.

Second, we have imposed a forward-variance lower bound (H2), namely $\sigma_t^2 \geq \sigma_{\min}^2 > 0$ on $[t_{\min}, T]$. This is technically convenient because the score of a Gaussian kernel scales like $1/\sigma_t^2$, and the sensitivity of our ratio estimator to neighbor perturbations inherits the same factor. However, many diffusion schedules satisfy $\sigma_t^2 \to 0$ as $t \downarrow 0$, precisely the regime in which sampling accuracy is most sensitive. In practice one often discretizes and avoids the singular endpoint, but this should be viewed as an algorithmic workaround rather than a theoretical resolution. A principled alternative is to state guarantees for a *clipped* score oracle,

$$\widehat{s}_{\text{clip}}(t, y) = \text{clip}\big(\widehat{s}_{\text{ret}}(t, y); S_{\max}(t)\big),$$

with a time-dependent $S_{\max}(t)$ chosen to control bias while preventing the $1/\sigma_t^2$ blow-up. Proving end-to-end bounds with such clipping requires tracking the induced bias in the running reward and how the entropy term compensates for the reduced control authority near $t = 0$.

Third, the truncation argument (Thm. 1) is explicitly high-dimensional and depends on a separation/concentration picture for the dataset under the relevant metric. This is plausible in latent spaces used for images, yet the statement is intentionally schematic. A limitation is that the actual mass captured by the retrieved set depends on both the geometry of $\{x_0^{(i)}\}$ and the query distribution p_t, which again depends on the policy. An important open question is to develop *matching* upper and lower bounds for the captured mass $\sum_{i \in \text{Top}k} w_i(x)$ under realistic models (e.g. mixture manifolds or clustered data), and to relate the needed k to intrinsic dimension rather than ambient d. Without such refinements, the compute–accuracy frontier in Thm. 5 is correct only qualitatively.

Fourth, our ANN model (H3) encodes distortion δ and failure probability η per query, and we pass to value bounds by a union/linearity argument that yields an $\mathcal{O}(\eta T)$ term. This is pessimistic in two ways. (i) The retrieval failures are not necessarily adversarial; one expects heavy-tailed but structured errors (e.g. occasional wrong cluster) rather than arbitrary points. (ii) Failures at different time steps are correlated because the query points are correlated along trajectories. A more faithful analysis would treat ANN as a randomized oracle with an explicit error distribution and would propagate this randomness through the actor–critic updates, yielding a bound in terms of higher moments (or conditional variances) of the induced reward noise, rather than a worst-case additive term.

Fifth, we have implicitly assumed boundedness conditions (bounded data norm, bounded actions via clipping, bounded visited states) to keep constants explicit and to justify Lipschitz steps in Thms. 2–3. These are standard in nonasymptotic control analysis but can fail in early training, where the critic is inaccurate and the policy may push the process into atypical regions. In this regime, retrieval can become unstable: the query $\alpha_t y$ may lie far from the dataset support, causing all kernel weights to be nearly uniform (broad σ_t^2) or numerically degenerate (small σ_t^2), and the truncated ratio estimator can become dominated by noise. An algorithmic mitigation is to add a *support test* (e.g. reject when nearest-neighbor distance exceeds a threshold) and revert to a conservative baseline. From a theoretical standpoint, the corresponding question is to establish a stability theorem in which the policy is kept within a safe set by construction (via barrier penalties or projection), so that the boundedness hypotheses are consequences rather than assumptions.

We also emphasize a conceptual limitation: our reward robustness result (Thm. 4) treats retrieval as an additive perturbation of the running reward. This decoupling is clean, but it hides an important feedback loop: the retrieval error influences the policy update, which changes the state visitation distribution, which changes the retrieval error distribution. A sharper theory would therefore aim for a *self-consistent* bound of the form

$$\epsilon_s(t) \;\le\; \mathcal{F}\big(\text{index parameters},\ \mathrm{Law}(y_t)\big), \qquad \mathrm{Law}(y_t) \;\approx\; \mathcal{G}\big(\epsilon_s(\cdot)\big),$$

and would solve the resulting fixed-point inequality. Such a result would move the analysis closer to minimax statements.

Finally, several open questions concern optimality. On the statistical side, the truncated ratio estimator is a particular nonparametric estimator of $\nabla \log p_t$ under a Gaussian kernel; it is natural to ask whether, given a per-step compute budget B, our choice of k and ANN suggests a minimax-optimal estimator of the score along trajectories, or whether alternative estimators (e.g. local regression in the retrieved neighborhood) can achieve strictly better bias–variance tradeoffs at the same cost. On the computational side, our hardness discussion indicates that NN-type retrieval is unavoidable in worst-case regimes, but it does not identify the tight dependence on δ and η needed for near-optimal value. Establishing *matching lower bounds* for the value gap as a function of retrieval resources would clarify whether the current pipeline is merely sufficient or in some sense necessary. We view these questions—tight truncation characterization, policy-dependent embedding regularity, stability without boundedness assumptions, and minimax compute–control tradeoffs—as the main theoretical tasks required to turn our proof skeleton into a complete theory of retrieval-augmented diffusion RL.