TEMPORAL-DIFFERENCE GUIDANCE FOR DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Learned diffusion guidance aims to sample from a tilted target distribution using gradient-based corrections derived from a noise-dependent value function. Standard regression-based guidance, however, trains this function through value prediction alone: small value error need not imply accurate gradients, creating a mismatch between the training objective and the quantity that drives sampling. We introduce temporal-difference (TD) guidance, a simple, easily implementable modification of existing guidance training that bootstraps across adjacent noise levels. By exploiting the temporal structure of the diffusion process, TD guidance addresses this value–gradient mismatch. We establish a finite-sample oracle inequality in a Sobolev norm that jointly controls value and spatial-gradient errors, balancing approximation, discretization, and statistical errors. Across image and molecular generation benchmarks, TD guidance achieves better quality–control and reward–validity trade-offs than direct regression at comparable conditioning strength.

1 INTRODUCTION

Diffusion models generate data by learning to reverse a gradual noising process: starting from noise, they apply a sequence of learned denoising steps to recover samples from the data distribution (Ho et al., 2020; Song et al., 2021). Many applications require more than reproducing this distribution; generated samples should belong to a specified class or exhibit properties favored by a reward. Guidance incorporates such target information into the sampling dynamics, allowing a pretrained model to be adapted to a downstream task (Dhariwal & Nichol, 2021; Uehara et al., 2025). When the base model is held fixed, the quality of this adaptation depends on how accurately the guidance correction is estimated from noisy states.

To guide generation toward a desired terminal target, the canonical approach is to learn a time-dependent value function. Let $\overline{X}_t$ denote the pretrained reverse diffusion, with $t = 0$ at the noise endpoint and $t = T$ at the data endpoint. For a terminal weight q, such as a class likelihood or an exponentiated reward, the associated value function is $f_t^*(x) = \mathbb{E}[q(\overline{X}_T) \mid \overline{X}_t = x]$, where the expectation is taken under the pretrained reverse process.

Existing methods learn an estimate f_t through supervised prediction: direct regression predicts terminal targets from noisy inputs (Uehara et al., 2025), while classifier guidance fits class probabilities at each noise level (Dhariwal & Nichol, 2021). These losses assess function values. At deployment, however, a learned function f_t enters the sampler through its guidance score $\nabla_x \log f_t$, with drift correction $a_t \nabla_x \log f_t$, where a_t is the diffusion coefficient. The training and sampling criteria are therefore mismatched: small error in f_t relative to f_t^* does not, by itself, guarantee that $\nabla_x \log f_t$ approximates $\nabla_x \log f_t^*$.

This mismatch can cause a nonvanishing error in the guided sampler's terminal distribution even as value estimates converge uniformly to the true function (Proposition 1). Our experiments also show the same distinction in learned models: lower value error can accompany higher guidance-score error (Section 4.1). Some existing approaches address this gap through derivative regularization or auxiliary gradient estimation (Chang et al., 2026; Tang & Xu, 2024; Guo et al., 2026; Mou, 2025), but these methods introduce additional derivative computations or estimation stages.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

In this paper, we take an alternative perspective, and address this mismatch through temporal-difference (TD) guidance. TD guidance is a lightweight modification of regression-based guidance training. Training uses only value evaluations and ordinary parameter gradients, with no explicit spatial derivatives. Nonetheless, the learned value function is guaranteed to have a small guidance-score error, and the resulting sampler is guaranteed to be close to the target distribution.

The method is based on the martingale structure of $(f_t^*(\overline{X}_t))_{t\in[0,T]}$: the current value equals the conditional expectation of the value at a later, less noisy state. This is a Bellman consistency relation for the fixed pretrained diffusion. TD guidance learns this relation by bootstrapping value predictions between neighboring noise levels, with terminal supervision anchoring the recursion. With function approximation, we formulate the population TD target as a projected fixed point, while the finite-sample implementation uses sample averaging or stochastic approximation. We undertake a comprehensive theoretical analysis of TD guidance, accompanied by empirical evaluation. Our main contributions are as follows.

- **Variational and statistical theory.** We characterize the TD variational solution and establish population and finite-sample oracle inequalities in an energy norm that jointly controls value and spatial-gradient errors; the bounds account for function approximation, time discretization, and statistical error (Theorems 1 and 2; see also Theorem 4), and energy-norm error in turn controls guidance error and terminal KL divergence (Theorem 3).
- **Algorithms and convergence.** We provide a unified training procedure for classifier and reward guidance (Algorithm 1) and establish mean-square convergence for a TD stochastic-approximation scheme with fixed linear features (Theorem 5 in Appendix C).
- **Empirical evaluation.** On four synthetic benchmarks with exact reference scores, TD achieves lower mean guidance error than regression-based methods; we also study CIFAR-10 classifier guidance and COMPAS reward guidance, where TD guidance improves the trade-off between sample quality and target validity (Sections 4.1 to 4.3).

1.1 Related work

Guidance methods differ in what they estimate and in how gradient information enters; Appendix F places this work in that broader landscape, and we record here only the comparisons closest to ours. Tang & Xu (2024) and Guo et al. (2026) learn the conditioning function by martingale regression and estimate its gradient by a separate quadratic-covariation regression, so their gradient bound retains a covariation error inherited from the fitted conditioning function, at a rate not supplied by the first-stage value bound (Guo et al., 2026, Theorems 4.9 and 4.11). Chang et al. (2026) obtain joint value–gradient error bounds from gradient-regularized Doob's matching, at the cost of mixed input–parameter derivatives during optimization. For classifier guidance, Sahu et al. (2026) show that population excess cross-entropy controls guidance error for bounded data under quantitative derivative bounds on the learned classifier, with counterexamples absent such control. Closest in setting, Ying et al. (2026) impose TD consistency on the base score model, whereas we estimate a guidance value on top of a frozen base model, and Mou (2025) casts diffusion fine-tuning with general value approximation as a Hamilton–Jacobi–Bellman variational inequality, whereas we analyse the projected policy-evaluation equation induced by terminal-value semi-gradient TD.

2 Problem setup

Let $s \in [0,T]$ denote forward noising time, and let $t = T - s$ denote reverse sampling time. The forward process satisfies

$$dX_s = h_s(X_s)\,ds + G_s^{1/2}\,dB_s, \qquad X_0 \sim p_{\mathrm{data}},$$

where G_s is a positive-definite covariance matrix and p_s is the law of X_s. The corresponding reverse process is

$$d\overline{X}_t = b_t(\overline{X}_t)\,dt + \sigma_t\,d\overline{B}_t, \qquad \overline{X}_0 \sim p_T,$$
$$b_t(x) := -h_{T-t}(x) + G_{T-t}\nabla \log p_{T-t}(x), \qquad a_t := G_{T-t}, \quad \sigma_t := G_{T-t}^{1/2}.$$

Let $\nu_t = \mathrm{Law}(\overline{X}_t)$; in particular, $\nu_T = p_{\mathrm{data}}$. The formal regularity and quantitative assumptions are specified in Assumption 1

The target is determined by a nonnegative terminal observation $Q \in L^2$: reward guidance takes $Q = e^{r(\overline{X}_T)}$, classifier guidance for a fixed class y takes $Q = \mathbf{1}\{Y = y\}$. Define $q(x) := \mathbb{E}[Q \mid \overline{X}_T = x]$ and $Z_Q := \mathbb{E}[Q] > 0$, the target distribution is

$$\mu^q(dx) = \frac{q(x)}{Z_Q}\, p_{\mathrm{data}}(dx). \tag{1}$$

We assume terminal sufficiency:

$$\mathbb{E}[Q \mid \mathcal{F}_T^{\overline{X}}] = q(\overline{X}_T), \qquad \mathcal{F}_t^{\overline{X}} := \sigma(\overline{X}_u : 0 \le u \le t), \tag{2}$$

so the path carries no information about Q beyond the clean endpoint. The value function is

$$f_t^*(x) := \mathbb{E}[q(\overline{X}_T) \mid \overline{X}_t = x] = \mathbb{E}[Q \mid \overline{X}_t = x], \qquad f_T^* = q. \tag{3}$$

Since $Q \in L^2$, the process $f_t^*(\overline{X}_t)$ is a martingale under the base reverse dynamics. Under standard smoothness conditions, f^* solves the backward Kolmogorov equation

$$(\partial_t + \mathcal{L}_t) f_t^* = 0, \qquad \mathcal{L}_t \varphi := b_t \cdot \nabla \varphi + \tfrac{1}{2} \operatorname{Tr}(a_t \nabla^2 \varphi). \tag{4}$$

By the Doob h-transform, the conditioned reverse process solves

$$d\overline{X}_t = \big(b_t(\overline{X}_t) + a_t \nabla_x \log f_t^*(\overline{X}_t)\big)dt + \sigma_t\, d\overline{B}_t, \tag{5}$$

started from $\mu_0^q(dx) = \frac{f_0^*(x)}{Z_Q}\, p_T(dx)$. Approximating the guided drift requires controlling the error in the gradient of the log-value, not just the value itself (Proposition 1).

3 Temporal difference guidance

In this section, we introduce the TD guidance methods and provide theoretical guarantees. We first discuss the limitations of standard regression-based guidance, and then introduce the fixed-point condition that TD guidance aims to solve.

Direct regression. Direct regression methods independently fit the function f_t at each noise level (Dhariwal & Nichol, 2021; Denker et al., 2024; Zhang et al., 2025; Li et al., 2025). For a loss ℓ whose population minimizer is the conditional mean (such as squared error or cross-entropy), the standard regression objective $\mathcal{J}(f) := \int_0^T \mathbb{E}[\ell(f_t(\overline{X}_t), Q)]\, dt$ is minimized exactly at $f_t = f_t^*$ for all t. The objective learns a value function rather than directly matching a score function. This creates a gap: achieving low error in the value estimate does not mathematically guarantee the accuracy of its spatial gradient $\nabla_x \log f_t$, which is the actual quantity required for sampling (5). A secondary limitation is that this independent regression struggles at high noise levels, where the state carries extremely weak signal about the terminal target.

The following proposition shows a concrete instance where small value error does not imply small sampling error. We use the notation $d_{\mathrm{TV}}(\mu, \nu) := \sup_A |\mu(A) - \nu(A)|$.

Proposition 1 (Value accuracy alone does not ensure tilted sampling)**.** *There exist a one-dimensional reverse diffusion on $[0, T]$, a fixed positive terminal target q, and smooth value estimates $(\widehat{f}^{(k)})_{k \ge 1}$ with common bounds $0 < m \le f_t^*(x), \widehat{f}_t^{(k)}(x) \le M < \infty$ for all t, x, k and $\widehat{f}_T^{(k)} = q$, such that, for all sufficiently large k,*

$$\sup_{0 \le t \le T} \|\widehat{f}_t^{(k)} - f_t^*\|^2_{L^2(\nu_t)} = O(k^{-1}), \qquad d_{\mathrm{TV}}(\widehat{\mu}_T^{(k)}, \mu_T^q) > 0.15. \tag{6}$$

Here $\widehat{\mu}_T^{(k)}$ is the terminal law of the guided SDE with drift $b_t + a_t \nabla \log \widehat{f}_t^{(k)}$, started from the exact tilted initial law μ_0^q.

The proof is in Appendix D.2. This counterexample demonstrates that uniform value convergence alone—absent spatial gradient control—cannot guarantee score or sampling convergence, underscoring the necessity of the energy norm introduced below.

Temporal-difference method. Since $(f_t^*(\overline{X}_t))_{t\in[0,T]}$ is a martingale, it satisfies an orthogonality condition for any $0 \le t < t' \le T$ and any square-integrable test function $g_t(\overline{X}_t)$: $\mathbb{E}[(f_t^*(\overline{X}_t) - f_{t'}^*(\overline{X}_{t'}))\, g_t(\overline{X}_t)] = 0$. This identity provides the mathematical foundation for bootstrapping values between adjacent noise levels. We use it to formulate a population variational problem and its empirical counterpart. The population solution controls value and spatial-gradient errors through the approximation bound in Theorem 1; Theorem 2 adds time-discretization and statistical errors.

Variational formulation. For $t \in [0,T]$, let $\nu_t = \mathrm{Law}(\overline{X}_t)$, equipped with the $L^2(\nu_t)$ inner product $\langle r, s\rangle_t = \mathbb{E}[r(\overline{X}_t)s(\overline{X}_t)]$ and norm $\|r\|_t^2 = \langle r, r\rangle_t$.

To ensure the variational problem is well-posed, we introduce a discount factor $\lambda > 0$ and define the discounted target $u_t^* := e^{-\lambda(T-t)} f_t^*$. This discounted factor leaves the guidance field invariant ($\nabla_x \log u_t^* = \nabla_x \log f_t^*$). By definition, u^* satisfies the discounted backward Kolmogorov equation:

$$(\partial_t + \mathcal{L}_t - \lambda)u_t^* = 0, \qquad u_T^* = q. \tag{7}$$

At $t = T$, the discount is exactly one, allowing the observable Q to serve as an unbiased squared-loss target for q.

To quantify approximation errors, we define the energy norm $\|\cdot\|_{\mathscr{E}}$ and the stronger parabolic Sobolev norm $\|\cdot\|_{\mathscr{V}^{(4)}}$ for any regular function h:

$$\|h\|_{\mathscr{E}}^2 := \|h_0\|_0^2 + \|h_T\|_T^2 + \int_0^T \left(\|h_t\|_t^2 + \|\nabla h_t\|_t^2\right) dt, \tag{8}$$

$$\|h\|_{\mathscr{V}^{(4)}}^2 := \|h\|_{\mathscr{E}}^2 + \int_0^T \left(\|\partial_t h_t\|_t^2 + \|h_t\|_{L^4(\nu_t)}^2 + \|\nabla h_t\|_{L^4(\nu_t)}^2\right) dt. \tag{9}$$

The $\|\cdot\|_{\mathscr{E}}$ norm fundamentally controls the value and spatial-gradient errors required for guidance. The $\|\cdot\|_{\mathscr{V}^{(4)}}$ norm additionally bounds time derivatives and L^4 moments. Our variational identities are first established on a smooth function class and then extended to these full spaces via standard approximation methods.

The continuous variational problem requires regularity of the reverse diffusion, regularity of the target, and a weakly compact approximation class. We group these conditions by their roles.

Assumption 1 (Regularity)**.** *The following conditions hold on* $[0,T]$.

(i) Diffusion regularity. *The reverse process is a nonexplosive Itô diffusion and admits a positive regular density ρ_t. In particular, its terminal law $\nu_T = p_{\mathrm{data}}$ has a strictly positive C^2 density on $\mathbb{R}^d$. There are constants $0 < \kappa \le \Lambda < \infty$ such that $\kappa I \preceq a_t \preceq \Lambda I$ for every $0 \le t \le T$. The probability current velocity $v_t^{\mathrm{cur}} := b_t - \frac{1}{2}a_t \nabla \log \rho_t$ satisfies*

$$M_4 := \sup_{0\le t\le T} \|v_t^{\mathrm{cur}}\|_{L^4(\nu_t)} < \infty. \tag{10}$$

(ii) Trial space, target, and approximation class. *Let $\mathscr{C}$ be a dense core of smooth functions with compact spatial support, and define $\mathscr{V}^{(4)}$ as its closure under $\|\cdot\|_{\mathscr{V}^{(4)}}$. The discounted target u^* belongs to $\mathscr{V}^{(4)}$. The class $\mathcal{F} \subset \mathscr{V}^{(4)}$ is nonempty, closed, bounded, and convex.*

In our variational framework, uniform ellipticity guarantees coercivity in the energy norm, while the $L^4(\nu_t)$ probability current bound ensures continuity of the TD formulation. Furthermore, because $\mathscr{V}^{(4)}$ is a reflexive Banach space, the closed, bounded, and convex class $\mathcal{F}$ is weakly compact, ensuring the existence of a projection solution.

Fix a terminal weight $\beta > 1/2$. We define the continuous TD bilinear form and the matching linear functional by

$$\mathfrak{B}_\beta(h, g) := \beta\langle h_T, g_T\rangle_T - \int_0^T \left\langle (\partial_t + \mathcal{L}_t - \lambda)h_t, g_t \right\rangle_t dt, \tag{11}$$

$$\mathfrak{l}_\beta(g) := \beta\mathbb{E}\left[Q g_T(\overline{X}_T)\right]. \tag{12}$$

By continuity, $\mathfrak{B}_\beta$ uniquely extends to $\mathscr{V}^{(4)} \times \mathscr{E}$ (Appendix A). We define its coercivity and continuity constants as:

$$\alpha_{B,\beta} := \min\left\{\tfrac{1}{2},\, \beta - \tfrac{1}{2},\, \lambda,\, \tfrac{\kappa}{2}\right\}, \qquad C_{D,\beta} := 1 + \beta + \lambda + \tfrac{\Lambda}{2} + M_4. \tag{13}$$

The following theorem characterizes the population TD solution and bounds its approximation error.

Theorem 1 (Population approximation error)**.** *Suppose Assumption 1 holds. Then*

$$\mathfrak{B}_\beta(u^*, g) = \mathfrak{l}_\beta(g) \qquad \textit{for every } g \in \mathscr{V}^{(4)}.$$

The variational inequality

$$\mathfrak{B}_\beta(\bar{u}, g - \bar{u}) \geq \mathfrak{l}_\beta(g - \bar{u}) \qquad \textit{for every } g \in \mathcal{F} \tag{14}$$

has a unique solution $\bar{u} \in \mathcal{F}$, and this solution satisfies

$$\|\bar{u} - u^*\|_{\mathscr{E}} \leq \left(1 + \frac{C_{D,\beta}}{\alpha_{B,\beta}}\right) \inf_{v \in \mathcal{F}} \|v - u^*\|_{\mathscr{V}^{(4)}}. \tag{15}$$

See Appendix A for the proof. The bound controls both value and spatial-gradient errors in the energy norm by the best approximation in the stronger $\mathscr{V}^{(4)}$ norm. If $u^* \in \mathcal{F}$, it gives exact recovery $\bar{u} = u^*$. When a nonsmooth terminal target prevents $u^* \in \mathscr{V}^{(4)}$, Remark 2 describes how to apply the theory on a shortened interval and account for the additional stopping bias.

3.1 Statistical error over a convex class

We now add time-discretization and sampling errors to the population approximation bound in Theorem 1. We retain the fixed convex class $\mathcal{F}$ from Assumption 1 and its continuous population solution $\bar{u}$. The statistical bound depends on the local geometry of this class. Define its approximation error by

$$a_{\mathrm{app}} := \inf_{v \in \mathcal{F}} \|v - u^*\|_{\mathscr{V}^{(4)}}.$$

For a deterministic uniform grid $t_k = k\Delta$, $\Delta = T/K$, and $\gamma = e^{-\lambda\Delta}$, observe i.i.d. full trajectory blocks $W_i = (Z_{i,0}, \ldots, Z_{i,K}, Q_i)$, $i = 1, \ldots, n$, from the exact joint diffusion and terminal-target law in Section 2, where $Z_{i,k} = \overline{X}^{(i)}_{t_k}$. For deterministic functions h, g, set

$$\begin{aligned} b_{W_i}(h, g) &:= \sum_{k=0}^{K-1} \{h_{t_k}(Z_{i,k}) - \gamma h_{t_{k+1}}(Z_{i,k+1})\} g_{t_k}(Z_{i,k}) + \beta h_T(Z_{i,K}) g_T(Z_{i,K}), \\ \ell_{W_i}(g) &:= \beta Q_i g_T(Z_{i,K}). \end{aligned} \tag{16}$$

Write P for expectation over one block and $P_n = n^{-1} \sum_{i=1}^n \delta_{W_i}$. The independent sampling units are the n trajectories; observations within each trajectory may be dependent. The empirical estimator is any $\widehat{u}_{n,\Delta} \in \mathcal{F}$ satisfying

$$P_n b_W(\widehat{u}_{n,\Delta}, v - \widehat{u}_{n,\Delta}) \geq P_n \ell_W(v - \widehat{u}_{n,\Delta}), \qquad v \in \mathcal{F}. \tag{17}$$

Two discounted generators control the error caused by observing a grid. Using the forward drift term h_r and diffusion matrix G_r, we define

$$\begin{aligned} &\mathcal{L}_r^{\mathrm{fwd}} \varphi := h_r \cdot \nabla\varphi + \tfrac{1}{2} \operatorname{Tr}(G_r \nabla^2 \varphi), \\ &R_\lambda^+ h_t := (\partial_t + \mathcal{L}_t - \lambda) h_t, \qquad\qquad R_\lambda^- h_t := (-\partial_t + \mathcal{L}_{T-t}^{\mathrm{fwd}} - \lambda) h_t. \end{aligned} \tag{18}$$

Both residuals are written in the reverse sampling coordinate t. All suprema below range over $[0, T] \times \mathbb{R}^d$. For smooth h, use the single regularity norm

$$J(h) := \beta \|h\|_\infty + \sqrt{T} \|a^{1/2} \nabla h\|_\infty + T \|R_\lambda^+ h\|_\infty + \|R_\lambda^- h\|_\infty. \tag{19}$$

This norm controls both trajectory fluctuations and grid error. All covering numbers and diameters below use the metric $J(h - g)$. We make the following assumption on the target and the convex class.

Assumption 2 (Bounded target and regularity of the convex class)**.**

(i) $|Q| \le Q_{\max} < \infty$ *almost surely.*

(ii) The elements of $\mathcal{F}$ *have fixed* $C^{1,2}$ *representatives. The class is compact in the uniform norm, separable in* J*, and* $\sup_{u\in\mathcal{F}} J(u) < \infty$*.*

Uniform compactness guarantees the existence of an empirical VI solution, while separability ensures the measurability of local suprema.

For $r > 0$, define the local difference class

$$\mathcal{D}(r) := \{v - \bar{u} : v \in \mathcal{F}, \|v - \bar{u}\|_{\mathscr{E}} \le r\}. \tag{20}$$

Write $N(D, J, \epsilon)$ for the covering number of D by J-balls of radius ϵ centered in D.

We now state the main statistical bound, which is a corollary of the more general Theorem 4 in Appendix B. For simplicity, we focus on the parametric case where the local covering number grows polynomially in $1/\epsilon$, while the general theorem allows for nonparametric classes with slower growth.

Theorem 2 (Statistical error for a parametric class)**.** *Suppose Assumptions 1 and 2 hold and, for some* $A \ge 1$ *and* $d \ge 1$*,*

$$N\big(r^{-1}\mathcal{D}(r), J, \epsilon\big) \le \max\big\{1, (A/\epsilon)^d\big\} \qquad \textit{for every } r, \epsilon > 0. \tag{21}$$

There is a constant $C \ge 1$*, depending only on* $A, \beta, Q_{\max}, \sup_{u\in\mathcal{F}} J(u), \alpha_{B,\beta}, C_{D,\beta}$*, such that, for every* $n \ge 1$*,* $0 < \delta < 1$ *and deterministic uniform grid, with probability at least* $1 - \delta$*, every empirical VI solution satisfies*

$$\|\widehat{u}_{n,\Delta} - u^*\|_{\mathscr{E}}^2 \le C\left\{a_{\mathrm{app}}^2 + \frac{d + \log(4/\delta)}{n} + \Delta^2\right\}. \tag{22}$$

See Appendix B.5 for the proof. Theorem 2 provides an oracle inequality for the empirical TD solution over a general convex class, with explicit dependence on the approximation error, sample size, and time discretization. The error is given in the energy norm instead of the standard $\mathbb{L}^2$ norm, which is crucial for controlling the guidance error and terminal KL divergence, as shown in Theorem 3 below. The statistical error term has the same order as the standard parametric rate d/n for a d-dimensional class.

The estimator $\widehat{u}_{n,\Delta}$ is defined by a variational inequality, which can be computed by solving a sequence of convex optimization problems (see Mou (2025) for a similar approach). In practice, one can also use stochastic approximation methods to solve the equation using online samples. Focused on finite-dimensional linear subspaces, Appendix C develops a stochastic approximation scheme for TD guidance and proves a mean-square convergence rate (Theorem 5). As in classical TD learning, $O(1/t)$ step sizes give an $O(1/n)$ rate in the number n of independent trajectories.

3.2 From value error to guidance and tilted sampling

An energy-norm bound controls the guidance error and hence terminal KL divergence through Girsanov's theorem (Mou, 2025, Proposition 5).

Let $\mathbb{P}$ be the path law of the exact reverse diffusion in Section 2, and let $\mathbb{P}^q$ be its terminal tilt $d\mathbb{P}^q/d\mathbb{P} = q(X_T)/Z_Q$, with marginals μ_t^q. Conditional on an independent training sample, fix a positive estimate $\widehat{u}$. Let $\widehat{\mathbb{P}}$ denote the path law of the SDE with drift $b_t + a_t\nabla \log \widehat{u}_t$ and diffusion coefficient σ_t, started from the same initial law μ_0^q; write $\widehat{\mu}_t$ for its marginals.

Theorem 3 (Guidance error and terminal KL)**.** *Suppose* $0 < q_{\min} \le q(x) \le q_{\max} < \infty$ *for every* $x \in \mathbb{R}^d$ *and* $a_t \preceq \Lambda I$ *for every* $t \in [0,T]$*, where* $\Lambda < \infty$*. Assume* $u^*, \widehat{u} \in C^{1,2}([0,T] \times \mathbb{R}^d)$*,* $\widehat{u} \ge m > 0$*, and*

$$\sup_{t,x} \|\nabla \log u_t^*(x)\|_2 \le G < \infty, \qquad \sup_{t,x} \|\sigma_t^\top \nabla \log \widehat{u}_t(x)\|_2 < \infty,$$

where both suprema range over $[0,T] \times \mathbb{R}^d$*. If the plug-in SDE is nonexplosive and unique in law, then*

$$\begin{aligned} D_{\mathrm{KL}}(\mu_T^q \| \widehat{\mu}_T) &\le \frac{1}{2} \int_0^T \mathbb{E}_{\mu_t^q} \left\| \sigma_t^\top (\nabla \log \widehat{u}_t - \nabla \log u_t^*) \right\|_2^2 dt \\ &\le \frac{\Lambda q_{\max}}{m^2 Z_Q} \max\{1, G^2\} \|\widehat{u} - u^*\|_{\mathscr{E}}^2. \end{aligned} \tag{23}$$

See Appendix D.1 for the proof. The energy bounds in Theorems 1 and 2 therefore give population and finite-sample terminal KL bounds, respectively, provided the corresponding estimate satisfies the hypotheses of Theorem 3. For the population solution, the bound is proportional to a_{app}^2; for the empirical solution, it has the same squared order as (22) on the high-probability event in Theorem 2. The factor m^{-2} requires a quantitative positive floor. The result concerns the exact base SDE with unit, unclipped guidance; discretization, scaling, and clipping introduce separate errors.

Remark 1 (Different initial laws). If the plug-in sampler instead starts from $\widehat{\mu}_0$, the same argument adds $D_{\mathrm{KL}}(\mu_0^q \| \widehat{\mu}_0)$ to both upper bounds in (23), provided this divergence is finite and the SDE is also nonexplosive and unique in law from $\widehat{\mu}_0$.

3.3 Practical guidance algorithm

The preceding variational inequalities characterize the temporal-difference solution over convex function classes, where the fixed point can be reached via stochastic approximation using sampled trajectories. To adapt this continuous-time formulation for non-linear neural networks—to which these formal guarantees do not directly apply—we discretize the diffusion trajectory into a time grid $0 = t_0 < \cdots < t_K = T$. This temporal-difference approach enforces local consistency between value predictions at adjacent noise levels. We parameterize the value function via a time-conditioned neural network $f_t(x, y; \theta)$, where x is the diffusion state and y the conditioning label, defining the function space $\mathcal{F} = \{(f_t(\cdot, \cdot; \theta))_{t \in [0,T]} : \theta \in \Theta\}$. For reward guidance, we omit y and write $f_t(x; \theta)$.

Since stochastic approximation with non-linear function approximators is prone to numerical divergence during bootstrapping, we decouple the online network θ from the evaluation target by introducing an Exponential Moving Average (EMA) target network parametrized by $\widetilde{\theta}$ (Lillicrap et al., 2016). This algorithmic device ensures the bootstrap target changes more slowly than the online network, stabilizing the empirical update.

This discrete formulation unifies classifier and reward guidance into a single procedure: they share identical temporal-difference updates, differing only in the task-specific boundary loss $\mathcal{L}_{\mathrm{bdry}}(\theta)$ at $t = T$.

Given the empirical temporal-difference error $\delta_k = f_{t_{k+1}}(\overline{X}_{t_{k+1}}, Y; \widetilde{\theta}) - f_{t_k}(\overline{X}_{t_k}, Y; \theta)$, the complete training objective combines the dynamic TD consistency with this terminal constraint:

$$\mathcal{L}(\theta) = \frac{1}{2} \sum_{k=0}^{K-1} \mathbb{E}\big[\delta_k^2\big] + \alpha\, \mathbb{E}\big[\mathcal{L}_{\mathrm{bdry}}(\theta)\big].$$

This boundary loss $\mathcal{L}_{\mathrm{bdry}}(\theta)$ explicitly dictates the target distribution. For classifier guidance, the output $f_t(x, \cdot; \theta)$ is simplex-valued in the label, and the terminal condition is enforced by cross-entropy: $\mathcal{L}_{\mathrm{bdry}}^{(\mathrm{cls})}(\theta) = -\log f_T(\overline{X}_T, y; \theta)$. For reward guidance, the network instead enforces $f_T = e^r$ in squared error: $\mathcal{L}_{\mathrm{bdry}}^{(\mathrm{rew})}(\theta) = \frac{1}{2}\big(f_T(\overline{X}_T; \theta) - e^{r(\overline{X}_T)}\big)^2$.

Treating the target parameters $\widetilde{\theta}$ as constant and replacing the expectation with a sampled discrete trajectory yields the semi-gradient update rule:

$$\theta \leftarrow \theta + \eta \left(\sum_{k=0}^{K-1} \delta_k \nabla_\theta f_{t_k}(\overline{X}_{t_k}, Y; \theta) - \alpha \nabla_\theta \mathcal{L}_{\mathrm{bdry}}(\theta) \right). \tag{24}$$

Algorithm 1: Temporal difference guidance

Input: Dataset $\mathcal{D}$, learning rate η, EMA rate τ, boundary weight α, time grid $0 = t_0 < \cdots < t_K = T$.

Initialize θ and set $\widetilde{\theta} \leftarrow \theta$
while *not converged* **do**
  Sample a mini-batch $\{(x_i, y_i)\}_{i=1}^{B}$ from $\mathcal{D}$ (the label argument is absent for reward guidance)
  for $i = 1, \ldots, B$ **do**
    Noise x_i forward and reverse-index to obtain $\overline{X}_{i,t_k}$ for $k = 0, \ldots, K$
    Compute $\delta_{i,k} = f_{t_{k+1}}(\overline{X}_{i,t_{k+1}}, y_i; \widetilde{\theta}) - f_{t_k}(\overline{X}_{i,t_k}, y_i; \theta)$ for $k = 0, \ldots, K-1$
  end
  $\widehat{g}_\theta = \frac{1}{B} \sum_{i=1}^{B} \left[\sum_{k=0}^{K-1} \delta_{i,k} \nabla_\theta f_{t_k}(\overline{X}_{i,t_k}, y_i; \theta) - \alpha \nabla_\theta \mathcal{L}_{\mathrm{bdry}}(\theta) \right]$
  $\theta \leftarrow \theta + \eta \widehat{g}_\theta; \quad \widetilde{\theta} \leftarrow \tau\theta + (1-\tau)\widetilde{\theta}$
end

4 Experiments

We compare TD learning with direct regression in three settings. Four two-dimensional benchmarks with closed-form references let us measure both value and guidance errors. CIFAR-10 tests classifier guidance using generated-image quality at matched control levels, while COMPAS tests reward guidance through the tradeoff between the molecular objective and corpus match rate.

4.1 Two-dimensional guidance benchmarks

We evaluate on four two-dimensional problems where the exact value f^* and the guidance field $\nabla_x \log f^*$ are available in closed form, permitting exact computation of the guidance error. As shown in Table 1, TD achieves lower mean guidance-field error on all four problems, despite yielding *higher* mean value error on three of them. This observation demonstrates that value accuracy does not guarantee guidance fidelity.

	Value MSE ($\times 10^{-4}$)		Log-guidance MSE ($\times 10^{-2}$)		
Task	Direct	TD	Direct	TD	TD/Direct
Eight-Gaussian ring	7.54	5.87	4.982	3.612	0.725
Rare bistable	5.17	11.83	4.290	3.367	0.785
Two moons	25.61	29.70	18.41	13.04	0.709
Spiral reward	4.29	6.38	14.03	11.47	0.818

Table 1: Value and guidance-field errors on four two-dimensional problems (means over five paired training seeds per task; lower is better).

4.2 CIFAR-10 classifier guidance

For CIFAR-10 classifier guidance, we freeze the unconditional DDPM generator `google/ddpm-cifar10-32`. Direct regression and TD learning train separate time-conditioned, 10-class ADM-style guidance classifiers with identical architectures and paired initializations; the frozen ResNet-56 and VGG-16-BN networks serve as external evaluators. Across five paired training seeds, we compare the methods at seven matched external-control targets, sharing base noise and conditioning labels within each pair. Each evaluation cell contains 2,000 images from a 250-step DDPM sampler. The two arms share a near-clean Brier anchor but differ in the noisy-time objective: observed-label cross-entropy for direct regression, a categorical forward-KL semi-gradient for TD. Appendix E.2 gives architectures, objectives, and control matching.

Figure 1 reports the absolute quality metrics across all control quantiles. Across the five independent training seeds, TD learning strictly improves upon direct regression at every evaluated quantile for Inception Score, precision, recall, and conditional FID/KID. To provide qualitative validation at

matched control levels, Figure 3 details the corresponding generated samples from the base generator, direct regression, and TD learning.

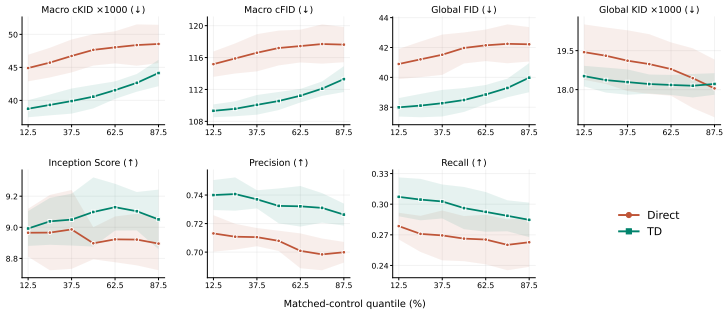

Figure 1: Selected exploratory CIFAR-10 generation results, not a fresh-seed confirmation. Solid lines and shaded bands are the mean and nominal pointwise 95% Student-t intervals over five paired training seeds, unadjusted for method selection. Top panels: lower is better (FID and KID variants); bottom panels: higher is better (Inception Score, precision, recall).

4.3 Molecular design on the COMPAS corpus

We evaluate on the COMPAS corpus (Wahab et al., 2022), a complete enumeration of cata-condensed polybenzenoid hydrocarbons with up to eleven rings. The minimization objective $J = 3y_{\mathrm{HLG}} + y_{\mathrm{IP}} - y_{\mathrm{EA}}$, combining the HOMO–LUMO gap, ionization potential, and electron affinity, is obtained by table lookup; we report its normalized value $\widetilde{J}$. Structures absent from the corpus carry no label, so the mean objective is computed over the matched subset. Both arms use the same frozen E(3)-equivariant diffusion model (Hoogeboom et al., 2022) on graph-of-rings geometries, with a 9-layer EGNN denoiser and a $1{,}000$-step polynomial schedule, and are trained over ten seeds generating 512 samples per seed and scale.

Figure 2 reports the tradeoff between the mean objective and the fraction of corpus-matched samples. As the guidance scale increases, both methods attain lower objective values at the cost of reduced match rate. TD learning exhibits Pareto dominance in the high-guidance regime. Furthermore, TD learning attains a lower objective minimum ($\widetilde{J} \approx 0.19$ at $s = 3.5$) than direct regression (≈ 0.21).

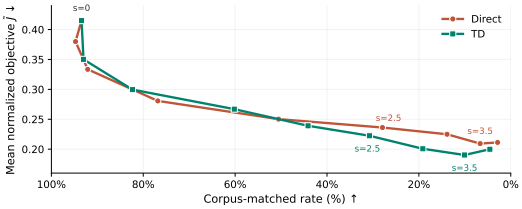

Figure 2: Trade-off between the corpus-matched fraction and the normalized objective $\widetilde{J}$, averaged over the matched subset only (lower is better). Annotated values denote the guidance scale s.

5 Conclusion

Diffusion guidance relies on the spatial gradient of the value function, not the value itself; uniformly accurate value estimates can still yield samplers that fail to converge. We therefore formulate the temporal-difference fixed point in an energy norm that controls value and gradient together. The projected solution is then well posed, and the empirical optimization error bounds the terminal Kullback–Leibler divergence of the tilted sampler. Two-dimensional benchmarks, CIFAR-10, and the COMPAS corpus support this account.

Several limitations remain. First, the formulation assumes continuous-time SDEs and does not cover other generative families such as flow matching or discrete masked diffusion. Second, practical samplers introduce time discretization, EMA targets, and nonlinear function approximators, none of which the analysis covers; the guarantees also describe the exact reverse process under unit, unclipped guidance, whereas the experiments sweep a guidance scale and clip the guidance field. Closing that distance and extending the objective to other generative families are the natural next steps.

Reproducibility Statement

The assumptions behind every theoretical claim are collected in Assumptions 1 and 2, and complete proofs of all results stated in the main text, alongside their required auxiliary lemmas, are provided in Appendices A to D. Appendix E provides details for the three empirical studies: the closed-form targets and error measurements for the two-dimensional benchmarks; the base generator, classifier, sampler settings, matched-control construction, and seed budgets for CIFAR-10; and the corpus details, guidance scales, attempt budgets, and in-corpus definitions for COMPAS. An anonymized zip file containing the training, sampling, and evaluation code, the frozen summary data for all figures and tables, and the exact scripts required to reproduce them is submitted as supplementary material.

AI Use Statement

In this work, the authors used generative AI tools to assist with drafting and polishing the manuscript, checking theoretical proofs, implementing experimental code, and providing critical feedback. AI-generated outputs were treated as suggestions and checked against mathematical derivations, executable tests, experimental records, or cited sources, as appropriate. The authors reviewed the AI-assisted material included in the manuscript and take responsibility for the final content, including all claims, proofs, code, experiments, citations, and prose.

References

Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Nearly d-linear convergence bounds for diffusion models via stochastic localization. In *International Conference on Learning Representations*, pp. 36916–36936, 2024. URL https://openreview.net/forum?id=r5njV3BsuD.

Jinyuan Chang, Chenguang Duan, Yuling Jiao, Yi Xu, and Jerry Zhijian Yang. Inference-time alignment for diffusion models via variationally stable Doob's matching, 2026. URL https://arxiv.org/abs/2601.06514.

Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 4735–4763, 2023a. URL https://proceedings.mlr.press/v202/chen23q.html.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R. Zhang. Sampling is as easy as learning the score: Theory for diffusion models with minimal data assumptions. In *The Eleventh International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=zyLVMgsZ0U_.

Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=MhK5aXo3gB.

Alexander Denker, Francisco Vargas, Shreyas Padhy, Kieran Didi, Simon Mathis, Vincent Dutordoir, Riccardo Barbano, Emile Mathieu, Urszula Julia Komorowska, and Pietro Lio. DEFT: Efficient fine-tuning of diffusion models by learning the generalised h-transform. In *Advances in Neural Information Processing Systems*, volume 37, pp. 19636–19682. Curran Associates, Inc., 2024. doi: 10.52202/079017-0620. URL https://doi.org/10.52202/079017-0620.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/49ad23d1ec9fa4bd8d77d02681df5cfa-Abstract.html.

Sjoerd Dirksen. Tail bounds via generic chaining. *Electronic Journal of Probability*, 20(53):1–29, 2015. doi: 10.1214/EJP.v20-3760. URL https://doi.org/10.1214/EJP.v20-3760.

Hengyu Fu, Zhuoran Yang, Mengdi Wang, and Minshuo Chen. Unveil conditional diffusion models with classifier-free guidance: A sharp statistical theory, 2024. URL https://arxiv.org/abs/2403.11968.

Zhengyi Guo, Wenpin Tang, and Renyuan Xu. Conditional diffusion guidance under hard constraint: A stochastic analysis approach, 2026. URL https://arxiv.org/abs/2602.05533.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. URL https://arxiv.org/abs/2207.12598.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Emiel Hoogeboom, Víctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3D. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8867–8887. PMLR, 2022. URL https://proceedings.mlr.press/v162/hoogeboom22a.html.

Vineet Jain, Kusha Sareen, Mohammad Pedramfar, and Siamak Ravanbakhsh. Diffusion tree sampling: Scalable inference-time alignment of diffusion models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 145576–145615. Curran Associates, Inc., 2025. doi: 10.52202/085713-4871. URL https://doi.org/10.52202/085713-4871.

Sunwoo Kim, Minkyu Kim, and Dongmin Park. Test-time alignment of diffusion models without reward over-optimization. In *International Conference on Learning Representations*, pp. 87238–87272, 2025. URL https://openreview.net/forum?id=vi3DjUhFVm.

David Kinderlehrer and Guido Stampacchia. *An Introduction to Variational Inequalities and Their Applications*, volume 31 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics, 2000. doi: 10.1137/1.9780898719451. URL https://epubs.siam.org/doi/book/10.1137/1.9780898719451.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence of score-based generative modeling for general data distributions. In *Proceedings of the 34th International Conference on Algorithmic Learning Theory*, volume 201 of *Proceedings of Machine Learning Research*, pp. 946–985, 2023. URL https://proceedings.mlr.press/v201/lee23a.html.

Xiner Li, Yulai Zhao, Chenyu Wang, Gabriele Scalia, Gokcen Eraslan, Surag Nair, Tommaso Biancalani, Shuiwang Ji, Aviv Regev, Sergey Levine, and Masatoshi Uehara. Derivative-free guidance in continuous and discrete diffusion models with soft value-based decoding. In *Advances in Neural Information Processing Systems*, volume 38, pp. 95507–95545. Curran Associates, Inc., 2025. doi: 10.52202/085713-3194. URL https://doi.org/10.52202/085713-3194.

Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1509.02971.

Wenlong Mou. Is RL fine-tuning harder than regression? a PDE learning approach for diffusion models, 2025. URL https://arxiv.org/abs/2509.02528.

Sharan Sahu, Arisina Banerjee, and Yuchen Wu. Provably reliable classifier guidance via cross-entropy control, 2026. URL https://arxiv.org/abs/2601.21200.

Bruno Scherrer. Should one compute the temporal difference fix point or minimize the Bellman residual? the unified oblique projection view. In *Proceedings of the 27th International Conference on Machine Learning*, pp. 959–966. Omnipress, 2010. URL https://icml.cc/Conferences/2010/papers/654.pdf.

Raghav Singhal, Zachary Horvitz, Ryan Teehan, Mengye Ren, Zhou Yu, Kathleen McKeown, and Rajesh Ranganath. A general framework for inference-time scaling and steering of diffusion models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 55810–55827. PMLR, 2025. URL https://proceedings.mlr.press/v267/singhal25b.html.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Wenpin Tang and Renyuan Xu. A stochastic analysis approach to conditional diffusion guidance, 2024. URL https://www.columbia.edu/~wt2319/CDG.pdf. Manuscript dated December 29, 2024.

John N. Tsitsiklis and Benjamin Van Roy. Analysis of temporal-difference learning with function approximation. In *Advances in Neural Information Processing Systems*, volume 9, pp. 1075–1081. MIT Press, 1997. URL https://proceedings.neurips.cc/paper/1996/hash/e00406144c1e7e35240afed70f34166a-Abstract.html.

Masatoshi Uehara, Yulai Zhao, Tommaso Biancalani, and Sergey Levine. Understanding reinforcement learning-based fine-tuning of diffusion models: A tutorial and review, 2024. URL https://arxiv.org/abs/2407.13734.

Masatoshi Uehara, Yulai Zhao, Chenyu Wang, Xiner Li, Aviv Regev, Sergey Levine, and Tommaso Biancalani. Inference-time alignment in diffusion models with reward-guided generation: Tutorial and review, 2025. URL https://arxiv.org/abs/2501.09685.

Alexandra Wahab, Lara Pfuderer, Eno Paenurk, and Renana Gershoni-Poranne. The COMPAS project: A computational database of polycyclic aromatic systems. phase 1: cata-condensed polybenzenoid hydrocarbons. *Journal of Chemical Information and Modeling*, 62(16):3704–3713, 2022. doi: 10.1021/acs.jcim.2c00503. URL https://pubs.acs.org/doi/10.1021/acs.jcim.2c00503.

Weixin Wang, Yu Yang, Wei Deng, and Pan Xu. Inference-time alignment of diffusion models via trust-region iterative twisted sequential monte carlo, 2026. URL https://arxiv.org/abs/2605.25123.

Luhuan Wu, Brian L. Trippe, Christian A. Naesseth, David M. Blei, and John P. Cunningham. Practical and asymptotically exact conditional sampling in diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 31372–31403. Curran Associates, Inc., 2023. doi: 10.52202/075280-1363. URL https://doi.org/10.52202/075280-1363.

Qizhen Ying, Yangchen Pan, Victor Adrian Prisacariu, and Junfeng Wen. Temporal difference learning for diffusion models. In *International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/62087.

Tao Zhang, Cheng Da, Kun Ding, Huan Yang, Kun Jin, Yan Li, Tingting Gao, Di Zhang, Shiming Xiang, and Chunhong Pan. Diffusion model as a noise-aware latent reward model for step-level preference optimization. In *Advances in Neural Information Processing Systems*, volume 38, pp. 106202–106235. Curran Associates, Inc., 2025. doi: 10.52202/085713-3545. URL https://doi.org/10.52202/085713-3545.

A THE POPULATION VARIATIONAL PROBLEM

This appendix proves Theorem 1, the population approximation bound stated in Section 3, and discusses early stopping when the terminal target lacks the required regularity.

Remark 2 (Early stopping when endpoint regularity fails). A nonsmooth terminal target q can prevent $u^* \in \mathscr{V}^{(4)}$. If endpoint regularity fails, fix $\varepsilon \in (0, T)$ and stop the sampler at $S = T - \varepsilon$. Apply the results on $[0, S]$, viewing $X_{\varepsilon+s}$ as the forward process and verifying the assumptions on this shortened interval. Replace the terminal regression by $f_S^*(x) = \mathbb{E}[Q \mid \overline{X}_S = x]$ and define the discount relative to S. The terminal observation remains Q, paired with $\overline{X}_S = X_\varepsilon$; thus the guidance field is unchanged.

To quantify the additional stopping bias, let $\mu^*(dx) = q(x)p_{\mathrm{data}}(dx)/Z_Q$ and let μ_s^* be its law after forward noising for time s. The exact tilted reverse process has law μ_ε^* at time $T - \varepsilon$. Suppose μ^* has a finite second moment and, for some $\varepsilon_0 \in (0, T)$, $H_*^2 := \sup_{0\le s\le\varepsilon_0} \mathbb{E}_{\mu^*}\|h_s(X_s)\|^2 < \infty$ and $\Gamma_* := \sup_{0\le s\le\varepsilon_0} \operatorname{Tr} G_s < \infty$, where $\mathbb{E}_{\mu^*}$ starts the forward process from μ^*. The forward coupling, Cauchy–Schwarz, and Itô isometry give

$$W_2^2(\mu_\varepsilon^*, \mu^*) \le \mathbb{E}_{\mu^*}\|X_\varepsilon - X_0\|^2 \le 2H_*^2\varepsilon^2 + 2\Gamma_*\varepsilon, \qquad 0 < \varepsilon \le \varepsilon_0. \tag{25}$$

Hence the additional Wasserstein bias can be made arbitrarily small, as in standard early-stopping analyses (Lee et al., 2023).

To ensure coercivity in the energy norm implies uniqueness in the stronger trial space, we first establish its definiteness.

Remark 3 (Definiteness of the energy norm). Because each ν_t has a strictly positive density, $\|h\|_{\mathscr{E}} = 0$ implies $h = 0$ almost everywhere. Consequently, the remaining $L^4(\nu_t)$ and weak time-derivative seminorms in $\mathscr{V}^{(4)}$ trivially vanish. Thus, $h = 0$ in $\mathscr{V}^{(4)}$, which implies that $\|\cdot\|_{\mathscr{E}}$ is a definite norm on $\mathscr{V}^{(4)}$ and its subspaces.

To complete the proof, we firstly establish three lemmas:

Lemma 1. *Suppose Assumption 1(i) holds and $u^* \in \mathscr{V}^{(4)}$. Then*

$$u_T^* = q, \qquad \mathfrak{B}_\beta(u^*, g) = \mathfrak{l}_\beta(g) \quad \textit{for every } g \in \mathscr{V}^{(4)}. \tag{26}$$

See Appendix A.1 for the proof.

Continuity ensures that the VI operator is well defined and bounds its response to an approximation error in the trial space. Recall the continuity constant from (13):

$$C_{D,\beta} := 1 + \beta + \lambda + \frac{\Lambda}{2} + M_4. \tag{27}$$

Lemma 2. *For every $h \in \mathscr{V}^{(4)}$ and $g \in \mathscr{E}$, the continuous extension of the TD bilinear form satisfies*

$$|\mathfrak{B}_\beta(h, g)| \le C_{D,\beta}\|h\|_{\mathscr{V}^{(4)}}\|g\|_{\mathscr{E}}. \tag{28}$$

See Appendix A.2 for the proof.

The energy identity gives monotonicity of the VI operator and controls the distance between two candidate solutions. Recall the positive coercivity constant from (13):

$$\alpha_{B,\beta} := \min\left\{\frac{1}{2},\, \beta - \frac{1}{2},\, \lambda,\, \frac{\kappa}{2}\right\} > 0. \tag{29}$$

Lemma 3. *For every $h \in \mathscr{V}^{(4)}$,*

$$\begin{aligned}\mathfrak{B}_\beta(h, h) = \frac{1}{2}\|h_0\|_0^2 + \left(\beta - \frac{1}{2}\right)\|h_T\|_T^2 + \lambda\int_0^T \|h_t\|_t^2\, dt \\ + \frac{1}{2}\int_0^T \langle a_t\nabla h_t, \nabla h_t\rangle_t\, dt \ge \alpha_{B,\beta}\|h\|_{\mathscr{E}}^2.\end{aligned} \tag{30}$$

See Appendix A.3 for the proof.

Taking Lemmas 1 to 3 as given, we now prove Theorem 1. Lemma 1 gives the claimed population moment. For existence of the VI solution, set $X = \mathscr{V}^{(4)}$ and $K = \mathcal{F}$. The space X is reflexive, and the set K is nonempty, closed, bounded, and convex, so K is weakly compact. Cauchy–Schwarz and the boundary trace embedding give, for every $g \in X$,

$$\begin{aligned} |\mathfrak{l}_\beta(g)| &\le \beta \|Q\|_{L^2} \|g_T\|_T \\ &\le \beta \|Q\|_{L^2} \|g\|_{\mathscr{V}^{(4)}}. \end{aligned}$$

Thus $\mathfrak{l}_\beta \in X^*$. Define $T : K \to X^*$ by

$$\langle Tu, g \rangle := \mathfrak{B}_\beta(u, g) - \mathfrak{l}_\beta(g).$$

Lemma 2 and the embedding $\|g\|_{\mathscr{E}} \le \|g\|_{\mathscr{V}^{(4)}}$ imply that T is bounded and norm-continuous, hence continuous on finite-dimensional subspaces. For $u, v \in K$, Lemma 3 gives

$$\begin{aligned} \langle Tu - Tv, u - v \rangle &= \mathfrak{B}_\beta(u - v, u - v) \\ &\ge \alpha_{B,\beta} \|u - v\|_{\mathscr{E}}^2 \ge 0, \end{aligned}$$

so T is monotone. The monotone-operator VI existence theorem (Kinderlehrer & Stampacchia, 2000, Chapter 3) therefore gives a solution $\bar{u} \in K$.

For uniqueness, let $\bar{u}_1, \bar{u}_2 \in \mathcal{F}$ be two VI solutions. Testing each VI against the other solution and adding the inequalities gives

$$\mathfrak{B}_\beta(\bar{u}_1 - \bar{u}_2, \bar{u}_1 - \bar{u}_2) \le 0.$$

Lemma 3 then implies $\alpha_{B,\beta} \|\bar{u}_1 - \bar{u}_2\|_{\mathscr{E}}^2 \le 0$. Since $\alpha_{B,\beta} > 0$, the energy norm of the difference vanishes. Both solutions lie in $\mathscr{V}^{(4)}$, so Remark 3 gives $\bar{u}_1 = \bar{u}_2$.

It remains to prove the oracle inequality. Fix $v \in \mathcal{F}$ and set $d = \bar{u} - v$. Subtracting the exact moment equation from the VI (14) gives

$$\mathfrak{B}_\beta(\bar{u} - u^*, g - \bar{u}) \ge 0, \qquad g \in \mathcal{F}.$$

Taking $g = v$ yields $\mathfrak{B}_\beta(\bar{u} - u^*, d) \le 0$. The decomposition $d = (\bar{u} - u^*) + (u^* - v)$, followed by Lemmas 2 and 3, gives

$$\begin{aligned} \alpha_{B,\beta} \|d\|_{\mathscr{E}}^2 &\le \mathfrak{B}_\beta(d, d) \\ &= \mathfrak{B}_\beta(\bar{u} - u^*, d) + \mathfrak{B}_\beta(u^* - v, d) \\ &\le C_{D,\beta} \|u^* - v\|_{\mathscr{V}^{(4)}} \|d\|_{\mathscr{E}}. \end{aligned}$$

If $\|d\|_{\mathscr{E}} > 0$, division by $\alpha_{B,\beta} \|d\|_{\mathscr{E}}$ gives

$$\|\bar{u} - v\|_{\mathscr{E}} \le \frac{C_{D,\beta}}{\alpha_{B,\beta}} \|u^* - v\|_{\mathscr{V}^{(4)}}.$$

This inequality also holds when $\|d\|_{\mathscr{E}} = 0$. The triangle inequality and the embedding of $\mathscr{V}^{(4)}$ into $\mathscr{E}$ now give

$$\begin{aligned} \|\bar{u} - u^*\|_{\mathscr{E}} &\le \|\bar{u} - v\|_{\mathscr{E}} + \|v - u^*\|_{\mathscr{E}} \\ &\le \left(1 + \frac{C_{D,\beta}}{\alpha_{B,\beta}}\right) \|v - u^*\|_{\mathscr{V}^{(4)}}. \end{aligned}$$

Taking the infimum over $v \in \mathcal{F}$ proves (15) and completes the proof of Theorem 1.

A.1 Proof of Lemma 1

The conditional expectation (3) is the backward transition semigroup applied to q. Choose smooth, compactly supported $q_m \to q$ in $L^2(\nu_T)$ and let $f_t^{(m)}(x) = \mathbb{E}[q_m(\overline{X}_T) \mid \overline{X}_t = x]$. Under the diffusion regularity in Section 2, each $f^{(m)}$ satisfies the classical backward equation on $(0, T)$. Conditional Jensen gives $\sup_t \|f_t^{(m)} - f_t^*\|_t \le \|q_m - q\|_T \to 0$. Since ρ_t is positive and regular, this convergence is also locally in space–time L^2, so the backward equation passes to f^* in distributions

on $(0,T)\times\mathbb{R}^d$. Multiplication by $e^{-\lambda(T-t)}$ gives (7) in the same sense. Testing against $\rho_t g_t$ for a smooth, compactly supported space–time test function g and integrating by parts in space identifies the dynamic pairing in (32). The bounds in Lemma 2 make its integrand absolutely integrable because $u^*\in\mathscr{V}^{(4)}$. Dominated convergence with time cutoffs therefore extends the identity to every $g\in\mathscr{C}$, including tests with nonzero endpoint traces. Finally, density of $\mathscr{C}$ and continuity of the bilinear form and terminal pairing give, for every $g\in\mathscr{V}^{(4)}$,

$$\int_0^T \big\langle(\partial_t+\mathcal{L}_t-\lambda)u_t^*, g_t\big\rangle_t\,dt = 0. \tag{31}$$

The integral denotes the continuous extension of the smooth pairing. Thus the dynamic term in $\mathfrak{B}_\beta(u^*,g)$ vanishes.

The definition of u^* in Section 3 and that of f^* in (3) imply

$$u_T^*(\overline{X}_T) = \mathbb{E}[Q\mid\sigma(\overline{X}_T)] = q(\overline{X}_T).$$

This is the claimed terminal condition $u_T^*=q$. Substituting into the terminal pairing and using the tower property gives

$$\begin{aligned}\beta\langle u_T^*, g_T\rangle_T &= \beta\mathbb{E}\big[\mathbb{E}[Q\mid\sigma(\overline{X}_T)]g_T(\overline{X}_T)\big]\\ &= \beta\mathbb{E}\big[Qg_T(\overline{X}_T)\big]\\ &= \mathfrak{l}_\beta(g).\end{aligned}$$

All expectations are under the joint data-and-forward-noising measure. Combining the terminal identity with the vanishing dynamic integral proves (26) and completes the proof.

A.2 Proof of Lemma 2

Fix $g\in\mathscr{E}$ and $h\in\mathscr{C}$. Since a_t is state independent, $\mathrm{Tr}(a_t\nabla^2 h)=\nabla\cdot(a_t\nabla h)$. Spatial integration by parts has no boundary term because h has compact spatial support. Using $\nabla\rho_t=\rho_t\nabla\log\rho_t$, we obtain

$$\begin{aligned}\langle\mathcal{L}_t h, g\rangle_t &= \langle b_t\cdot\nabla h, g\rangle_t - \frac{1}{2}\int a_t\nabla h\cdot\nabla(g\rho_t)\,dx\\ &= \left\langle\left(b_t-\frac{1}{2}a_t\nabla\log\rho_t\right)\cdot\nabla h, g\right\rangle_t - \frac{1}{2}\langle a_t\nabla h,\nabla g\rangle_t\\ &= \langle v_t^{\mathrm{cur}}\cdot\nabla h, g\rangle_t - \frac{1}{2}\langle a_t\nabla h,\nabla g\rangle_t.\end{aligned}$$

Substituting this identity into the TD form gives

$$\begin{aligned}\mathfrak{B}_\beta(h,g) = \beta\langle h_T, g_T\rangle_T &- \int_0^T\langle\partial_t h_t, g_t\rangle_t\,dt\\ &+\frac{1}{2}\int_0^T\langle a_t\nabla h_t,\nabla g_t\rangle_t\,dt - \int_0^T\langle v_t^{\mathrm{cur}}\cdot\nabla h_t, g_t\rangle_t\,dt\\ &+\lambda\int_0^T\langle h_t, g_t\rangle_t\,dt.\end{aligned}\tag{32}$$

Cauchy–Schwarz and $a_t\preceq\Lambda I$ bound the endpoint, time-derivative, diffusion, and zeroth-order terms by β, 1, $\Lambda/2$, and λ, respectively, times $\|h\|_{\mathscr{V}^{(4)}}\|g\|_{\mathscr{E}}$.

For the current term, Hölder's inequality in space with $1/4+1/4+1/2=1$, followed by Cauchy–Schwarz in time, gives

$$\begin{aligned}\int_0^T|\langle v_t^{\mathrm{cur}}\cdot\nabla h_t, g_t\rangle_t|\,dt &\le M_4\int_0^T\|\nabla h_t\|_{L^4(\nu_t)}\|g_t\|_t\,dt\\ &\le M_4\|h\|_{\mathscr{V}^{(4)}}\|g\|_{\mathscr{E}}.\end{aligned}$$

Adding the five bounds proves (28) on the smooth core. The density of $\mathscr{C}$ in $\mathscr{V}^{(4)}$ extends this bound to every $h\in\mathscr{V}^{(4)}$, which completes the proof.

A.3 Proof of Lemma 3

Fix $h \in \mathscr{C}$. Applying Itô's formula to $h_t(\overline{X}_t)^2$ and taking expectations gives

$$\frac{d}{dt}\|h_t\|_t^2 = 2\langle(\partial_t + \mathcal{L}_t)h_t, h_t\rangle_t + \langle a_t \nabla h_t, \nabla h_t\rangle_t.$$

Integrating from 0 to T and rearranging yields

$$-\int_0^T \langle(\partial_t + \mathcal{L}_t)h_t, h_t\rangle_t \, dt = \frac{1}{2}\|h_0\|_0^2 - \frac{1}{2}\|h_T\|_T^2 + \frac{1}{2}\int_0^T \langle a_t \nabla h_t, \nabla h_t\rangle_t \, dt.$$

By the definition of the TD form,

$$\begin{aligned}\mathfrak{B}_\beta(h,h) &= \beta\|h_T\|_T^2 + \lambda\int_0^T \|h_t\|_t^2 \, dt \\ &\quad - \int_0^T \langle(\partial_t + \mathcal{L}_t)h_t, h_t\rangle_t \, dt \\ &= \frac{1}{2}\|h_0\|_0^2 + \left(\beta - \frac{1}{2}\right)\|h_T\|_T^2 + \lambda\int_0^T \|h_t\|_t^2 \, dt \\ &\quad + \frac{1}{2}\int_0^T \langle a_t \nabla h_t, \nabla h_t\rangle_t \, dt.\end{aligned}$$

This proves the energy identity on $\mathscr{C}$. Lemma 2 makes the bilinear form continuous in the $\mathscr{V}^{(4)}$. The boundary traces and the remaining energy terms are continuous in the same space by construction. The density of $\mathscr{C}$ therefore extends the identity to every $h \in \mathscr{V}^{(4)}$.

Uniform ellipticity gives $a_t \succeq \kappa I$. Lower-bounding the four energy coefficients by $\alpha_{B,\beta} = \min\{1/2, \beta - 1/2, \lambda, \kappa/2\} > 0$ proves the inequality in (30) ,which completes the proof.

A.4 A bound on the population projection

The next consequence of coercivity replaces the projected norm by a bound on the terminal target. The same argument controls the full-space Galerkin solution in the stochastic approximation analysis.

Corollary 1 (Removing the unknown projected norm)**.** *Let $\bar{u}$ be the population VI solution from Theorem 1. If $0 \in \mathcal{F}$, then*

$$\|\bar{u}\|_{\mathscr{E}} \le \frac{\beta}{\alpha_{B,\beta}}\|Q\|_{L^2}. \tag{33}$$

Proof. Taking 0 as the competitor in (14) gives $\mathfrak{B}_\beta(\bar{u}, \bar{u}) \le \mathfrak{l}_\beta(\bar{u})$. Lemma 3, Cauchy–Schwarz, and the endpoint bound in the energy norm imply

$$\begin{aligned}\alpha_{B,\beta}\|\bar{u}\|_{\mathscr{E}}^2 &\le \beta\|Q\|_{L^2}\|\bar{u}_T\|_T \\ &\le \beta\|Q\|_{L^2}\|\bar{u}\|_{\mathscr{E}}.\end{aligned}$$

If $\|\bar{u}\|_{\mathscr{E}} = 0$, the claimed inequality holds. Otherwise, dividing by $\alpha_{B,\beta}\|\bar{u}\|_{\mathscr{E}}$ proves (33), which completes the proof.

B Proofs for the convex-class statistical bound

This appendix states the general localized bound behind Theorem 2 and proves the results of Section 3.1.

For $n \ge 1$ and $0 < \delta < 1$, define the local complexity

$$\Psi(r) := \frac{1}{\sqrt{n}}\left\{\int_0^{\operatorname{diam}_J \mathcal{D}(r)} \sqrt{\log N(\mathcal{D}(r), J, \epsilon)} \, d\epsilon + \operatorname{diam}_J \mathcal{D}(r)\sqrt{\log(4/\delta)}\right\}. \tag{34}$$

Set $\Psi(r) = 0$ if the diameter is zero.

Theorem 4 (Localized statistical bound for a convex class)**.** *Suppose Assumptions 1 and 2 hold. There is a constant $c > 0$, depending only on $\beta, Q_{\max}$ and $\sup_{u\in\mathcal{F}} J(u)$, such that the following holds. Fix $n \geq 1$, $0 < \delta < 1$, a deterministic uniform grid, and a deterministic $r > 0$ satisfying*

$$\Psi(r) + \Delta \operatorname{diam}_J \mathcal{D}(r) \leq c\alpha_{B,\beta} r^2. \tag{35}$$

An empirical VI solution exists for every sample realization and, with probability at least $1-\delta$, every such solution satisfies

$$\|\widehat{u}_{n,\Delta} - u^*\|_{\mathscr{E}} \leq \left(1 + \frac{C_{D,\beta}}{\alpha_{B,\beta}}\right) a_{\mathrm{app}} + r. \tag{36}$$

The error is the sum of approximation error and a radius at which coercivity dominates sampling fluctuations and grid bias. The same J metric controls both terms in (35); all constants are independent of n, Δ, δ and r. The proof uses empirical-process chaining (Dirksen, 2015).

Corollary 2 (Realizable target)**.** *Under the assumptions of Theorem 4, if $u^* \in \mathcal{F}$, its conclusion holds with $a_{\mathrm{app}} = 0$ and with $\operatorname{diam}_J \mathcal{D}(r)$ in the grid term of (35) replaced by $\{\operatorname{diam}_J \mathcal{D}(r)\}^2$.*

Grid observations alone do not control the continuous energy norm: $h(t,x) = \sin(\pi t/\Delta)\psi(x)$ vanishes at every grid point but can have positive energy norm. The derivative bounds in J control these inter-grid fluctuations. The local J-diameter quantifies this grid error at the scale r.

Let $\bar{u}$ be the continuous population VI solution. We adopt the shorthand $B := \mathfrak{B}_\beta$, $B_\Delta := Pb_W$, and $\widehat{B} := P_n b_W$. By the triangle inequality, the error decomposes as:

$$\|\widehat{u}_{n,\Delta} - u^*\|_{\mathscr{E}} \leq \|\bar{u} - u^*\|_{\mathscr{E}} + \|\widehat{u}_{n,\Delta} - \bar{u}\|_{\mathscr{E}}. \tag{37}$$

The population theorem controls the first term. For the second term, define $e = \widehat{u}_{n,\Delta} - \bar{u}$, the two VI inequalities and continuous coercivity give,

$$\begin{aligned}\alpha_{B,\beta}\|e\|_{\mathscr{E}}^2 \leq (P-P_n)b_W(e,e) + (P-P_n)\{b_W(\bar{u},e) - \ell_W(e)\} \\ + (B - B_\Delta)(e,e) + (B - B_\Delta)(\bar{u},e).\end{aligned} \tag{38}$$

The following lemmas establish existence and bound the four terms on a deterministic local class. In the following lemmas, $C \geq 1$ denotes a constant depending only on $\beta, Q_{\max}$ and $\sup_{u\in\mathcal{F}} J(u)$, whose value may change between displays. The unified norm controls the energy norm, since

$$\|h\|_{\mathscr{E}}^2 \leq \left\{\frac{2+T}{\beta^2} + \frac{1}{\kappa}\right\} J(h)^2.$$

Lemma 4 (Grid identities and empirical existence)**.** *Let $t_k = k\Delta$, $\gamma = e^{-\lambda\Delta}$, and $Z_k = \overline{X}_{t_k}$. For every function h with square-integrable grid evaluations,*

$$\begin{aligned}b_W(h,h) = \frac{1}{2}h_0(Z_0)^2 + \frac{1}{2}\sum_{k=0}^{K-1}\{h_{t_k}(Z_k) - \gamma h_{t_{k+1}}(Z_{k+1})\}^2 \\ + \frac{1-\gamma^2}{2}\sum_{j=1}^{K-1} h_{t_j}(Z_j)^2 + \left(\beta - \frac{\gamma^2}{2}\right) h_T(Z_K)^2 \geq 0.\end{aligned} \tag{39}$$

Taking expectations yields $B_\Delta(h,h) \geq 0$ and the target satisfies

$$B_\Delta(u^*, g) = \mathfrak{l}_\beta(g). \tag{40}$$

If $\mathcal{F}$ is nonempty and convex and Assumption 2 holds, an empirical VI solution exists for every sample realization.

See Appendix B.1 for the proof.

The next lemma compares the continuous and discrete forms using derivatives of the particular functions being tested.

Lemma 5 (Finite-grid consistency). *Suppose Assumption 1(i) holds and the trajectories follow the joint forward and reverse path law defined in Section 2. Let h, g be deterministic $C^{1,2}$ functions in $\mathscr{V}^{(4)}$ with bounded values and bounded diffusion-weighted spatial gradients. Suppose $\|R_\lambda^+ h\|_\infty$ and $\|R_\lambda^- g\|_\infty$ are finite. Then*

$$(B_\Delta - B)(h,g) = \sum_{k=0}^{K-1} \int_{t_k}^{t_{k+1}} \mathbb{E}\Big[(R_\lambda^+ h_s)(\overline{X}_s)\{g_s(\overline{X}_s) - e^{-\lambda(s-t_k)} g_{t_k}(Z_k)\}\Big]\, ds, \tag{41}$$

$$|(B_\Delta - B)(h,g)| \le \frac{T\Delta}{2} \|R_\lambda^+ h\|_\infty \|R_\lambda^- g\|_\infty. \tag{42}$$

In particular, under Assumption 2,

$$\sup_{d \in \mathcal{D}(r)} |(B_\Delta - B)(d,d)| \le C\Delta \operatorname{diam}_J \mathcal{D}(r),$$
$$\sup_{d \in \mathcal{D}(r)} |(B_\Delta - B)(\bar{u},d)| \le C\Delta \operatorname{diam}_J \mathcal{D}(r). \tag{43}$$

See Appendix B.2 for the proof.

To bound the statistical fluctuations, we define the local empirical processes over the function differences $\mathcal{D}(r)$:

$$\mathcal{Q}_n(r) := \sup_{d \in \mathcal{D}(r)} |(P_n - P) b_W(d,d)|,$$
$$\mathcal{S}_n(r) := \sup_{d \in \mathcal{D}(r)} |(P_n - P)\{b_W(\bar{u},d) - \ell_W(d)\}|. \tag{44}$$

The following lemma bounds these processes using the Dudley entropy integrals.

Lemma 6 (Localized trajectory concentration). *Suppose Assumption 2 holds. For every $n \ge 1$, $0 < \delta < 1$ and deterministic $r > 0$ with $\Psi(r) < \infty$, with probability at least $1 - \delta$,*

$$\mathcal{Q}_n(r) \le C\Psi(r), \qquad \mathcal{S}_n(r) \le C\Psi(r). \tag{45}$$

See Appendix B.3 for the proof.

The following deterministic lemma leverages convexity to translate local empirical process bounds at a fixed radius r into a global guarantee for the empirical solution.

Lemma 7 (Localized VI comparison). *Let $\mathcal{F}$ be convex, let $\bar{u} \in \mathcal{F}$ solve the continuous population VI, and let $\widehat{u} \in \mathcal{F}$ solve the empirical VI. Suppose $B(d,d) \ge \alpha \|d\|^2$ for every secant of $\mathcal{F}$, with $\alpha > 0$. Define $\mathcal{D}(r)$, $\mathcal{Q}_n(r)$, and $\mathcal{S}_n(r)$ using this norm and center. If deterministic $r > 0$ and $\eta_Q, \eta_S \ge 0$ satisfy $\eta_Q + \eta_S < 1$ and*

$$\mathcal{Q}_n(r) + \sup_{d \in \mathcal{D}(r)} |(B - B_\Delta)(d,d)| \le \eta_Q \alpha r^2,$$
$$\mathcal{S}_n(r) + \sup_{d \in \mathcal{D}(r)} |(B - B_\Delta)(\bar{u},d)| \le \eta_S \alpha r^2, \tag{46}$$

then $\|\widehat{u} - \bar{u}\| \le r$.

See Appendix B.4 for the proof.

First, Theorem 1 and lemma 4 guarantee existence of the population and empirical solutions. Choose C to dominate the constants in Lemmas 5 and 6 and take $c \le 1/(4C)$ in Theorem 4. On the event (45), each left-hand side of (46) is at most

$$C\{\Psi(r) + \Delta \operatorname{diam}_J \mathcal{D}(r)\} \le \tfrac{1}{4}\alpha_{B,\beta} r^2.$$

Thus Lemma 7 applies with $\alpha = \alpha_{B,\beta}$ and $\eta_Q = \eta_S = 1/4$. It bounds every empirical solution on the same event by $\|\widehat{u}_{n,\Delta} - \bar{u}\|_{\mathscr{E}} \le r$. Combining this with the population oracle inequality yields

$$\|\widehat{u}_{n,\Delta} - u^*\|_{\mathscr{E}} \le \left(1 + \frac{C_{D,\beta}}{\alpha_{B,\beta}}\right) a_{\mathrm{app}} + r.$$

This proves Theorem 4. If $u^* \in \mathcal{F}$, then $\bar{u} = u^*$ and $a_{\mathrm{app}} = 0$. The exact grid identity in Lemma 4 gives $(B - B_\Delta)(\bar{u}, d) = 0$ for every $d \in \mathcal{D}(r)$. Since $0 \in \mathcal{D}(r)$, Lemma 5 also gives

$$\sup_{d \in \mathcal{D}(r)} |(B - B_\Delta)(d,d)| \le \tfrac{1}{2}\Delta\{\operatorname{diam}_J \mathcal{D}(r)\}^2.$$

The same choice of c, with $C \ge 1$, therefore proves Corollary 2.

B.1 Proof of Lemma 4

Apply the scalar identity $2(a-\gamma b)a = a^2 - \gamma^2 b^2 + (a-\gamma b)^2$ to each adjacent pair of evaluations of h. Summing over the transitions and adding the terminal term gives (39). Every coefficient in that identity is nonnegative because $0 < \gamma < 1$ and $\beta > 1/2$. Taking expectations gives the population grid energy identity.

The martingale property of $f_t^*(\overline{X}_t)$ under the reverse diffusion law directly implies

$$\mathbb{E}[u_{t_k}^*(Z_k) - \gamma u_{t_{k+1}}^*(Z_{k+1}) \mid Z_k] = 0.$$

Multiplying by $g_{t_k}(Z_k)$ and taking expectations eliminates all dynamic terms in $B_\Delta(u^*, g)$. Using the fact that $u_T^* = q$ and $\mathbb{E}[Q \mid Z_K] = q(Z_K)$ yields (40).

For empirical existence, fix a sample and consider the map

$$E_n : \mathcal{F} \longrightarrow \mathbb{R}^{n(K+1)}, \qquad E_n u = (u_{t_k}(Z_{i,k}))_{1\le i\le n,\, 0\le k\le K}.$$

Its image $\mathcal{K}_n = E_n\mathcal{F}$ is nonempty, compact, and convex: the evaluation map is linear and continuous in the uniform norm, and the class is compact in that norm by Assumption 2. For a block vector $y = (y_{i,k})$, define the continuous affine map F_n by

$$(F_n(y))_{i,k} := \begin{cases} y_{i,k} - \gamma y_{i,k+1}, & k < K, \\ \beta(y_{i,K} - Q_i), & k = K. \end{cases}$$

The map $y \mapsto \Pi_{\mathcal{K}_n}(y - F_n(y))$ is continuous from the compact convex set $\mathcal{K}_n$ into itself. Brouwer's fixed point theorem therefore gives $\widehat{y} \in \mathcal{K}_n$. The characterization of Euclidean projection implies

$$\langle F_n(\widehat{y}), z - \widehat{y}\rangle \ge 0, \qquad z \in \mathcal{K}_n.$$

Choose any $\widehat{u} \in \mathcal{F}$ with $E_n\widehat{u} = \widehat{y}$. The last inequality with $z = E_n v$, divided by n, is exactly the empirical VI for every $v \in \mathcal{F}$. This proves existence and completes the proof of Lemma 4.

B.2 Proof of Lemma 5

Applying Itô's formula to the discounted process $u \mapsto e^{-\lambda(u-t)}h_u(\overline{X}_u)$ over the reverse-time interval $[t, s]$ yields

$$\begin{aligned} &e^{-\lambda(s-t)}h_s(\overline{X}_s) - h_t(\overline{X}_t) \\ &\quad = \int_t^s e^{-\lambda(u-t)}(R_\lambda^+ h_u)(\overline{X}_u)\, du + \int_t^s e^{-\lambda(u-t)}(a_u^{1/2}\nabla h_u)(\overline{X}_u)^\top d\overline{B}_u. \end{aligned}$$

By the uniform boundedness of the diffusion-weighted gradient $a_u^{1/2}\nabla h_u$, the stochastic integral is square-integrable and thus is a martingale. Multiplying by the $\mathcal{F}_{t_k}$-measurable variable $g_{t_k}(Z_k)$ and taking expectations over $[t_k, t_{k+1}]$ eliminates this term, leaving the kth dynamic term of $B_\Delta(h, g)$ as

$$-\int_{t_k}^{t_{k+1}} e^{-\lambda(s-t_k)}\mathbb{E}[(R_\lambda^+ h_s)(\overline{X}_s)g_{t_k}(Z_k)]\, ds.$$

Applying spatial integration by parts with a smooth cutoff χ_R, the boundary term scales as $\mathcal{O}(1/R)$ and vanishes as $R \to \infty$. Under the $\mathscr{V}^{(4)}$ hypotheses, this establishes that the continuous extension $B(h, g)$ coincides with the classical generator pairing. Subtracting this continuous pairing from the grid expansion precisely cancels the terminal terms, yielding (41).

To bound its integrand, we apply the forward Itô formula under the same joint path law to $u \mapsto e^{-\lambda(u-(T-s))}g_{T-u}(X_u)$ on $T - s \le u \le T - t$. The time derivative contributes $-\partial_t g$ and the diffusion covariance is $G_u = a_{T-u}$, so the drift is exactly $R_\lambda^- g$; the forward stochastic integral is again a martingale by the gradient bound on g. Conditioning on X_{T-s} and returning to reverse time gives

$$\begin{aligned} &e^{-\lambda(s-t)}\mathbb{E}[g_t(\overline{X}_t) \mid \overline{X}_s] - g_s(\overline{X}_s) \\ &\qquad = \int_t^s e^{-\lambda(s-u)}\mathbb{E}[(R_\lambda^- g)_u(\overline{X}_u) \mid \overline{X}_s]\, du. \end{aligned} \tag{47}$$

Multiply (47) by $(R_\lambda^+ h_s)(\overline{X}_s)$ and take expectations. The tower property and the two residual bounds yield

$$\Big|\mathbb{E}\Big[(R_\lambda^+ h_s)(\overline{X}_s)\{g_s(\overline{X}_s) - e^{-\lambda(s-t)} g_t(\overline{X}_t)\}\Big]\Big| \le (s-t)\|R_\lambda^+ h\|_\infty \|R_\lambda^- g\|_\infty.$$

Integrating over $[t_k, t_{k+1}]$ and summing over the $K = T/\Delta$ intervals proves (42). By the definition of J, this bound implies

$$|(B_\Delta - B)(h, g)| \le \tfrac{1}{2}\Delta J(h) J(g).$$

For $d \in \mathcal{D}(r)$, we have $J(d) \le \operatorname{diam}_J \mathcal{D}(r)$ and $J(d) \le 2\sup_{u\in\mathcal{F}} J(u)$. Substituting $(h, g) = (d, d)$ and $(h, g) = (\bar{u}, d)$ and absorbing one factor into the class bound gives (43), completing the proof.

B.3 Proof of Lemma 6

Write $Z_t = \overline{X}_t$ and $\tau(s) = t_k$ for $s \in (t_k, t_{k+1}]$. For deterministic smooth h, g, applying Itô's formula to $e^{-\lambda(s-\tau(s))} h_s(Z_s)$ and multiplying by $g_{\tau(s)}(Z_{\tau(s)})$ yields the decomposition

$$\begin{aligned} b_W(h, g) &= D_W(h, g) + M_W(h, g), \\ D_W(h, g) &= -\int_0^T e^{-\lambda(s-\tau(s))} g_{\tau(s)}(Z_{\tau(s)})(R_\lambda^+ h_s)(Z_s)\, ds + \beta h_T(Z_T) g_T(Z_T), \\ M_W(h, g) &= -\int_0^T e^{-\lambda(s-\tau(s))} g_{\tau(s)}(Z_{\tau(s)})(a_s^{1/2}\nabla h_s)(Z_s)^\top d\overline{B}_s. \end{aligned} \tag{48}$$

The assumed bounds imply

$$\begin{aligned} |D_W(h, g)| &\le \|g\|_\infty\{T\|R_\lambda^+ h\|_\infty + \beta\|h\|_\infty\}, \\ \langle M(h, g)\rangle_T &\le T\|g\|_\infty^2 \|a^{1/2}\nabla h\|_\infty^2. \end{aligned}$$

Combining the exponential martingale inequality for M_W and Hoeffding's lemma for $D_W - PD_W$ via Cauchy–Schwarz yields, for every $\theta \in \mathbb{R}$,

$$\mathbb{E}\exp(\theta\{b_W(h, g) - Pb_W(h, g)\}) \le \exp\big(\theta^2\|g\|_\infty^2 J(h)^2\big), \tag{49}$$

Subtracting $\ell_W(g)$ from the bounded drift term, the same argument gives

$$\mathbb{E}\exp\big(\theta\big[b_W(\bar{u}, g) - \ell_W(g) - P\{b_W(\bar{u}, g) - \ell_W(g)\}\big]\big) \le \exp\big(\theta^2[J(\bar{u}) + \beta Q_{\max}]^2\|g\|_\infty^2\big). \tag{50}$$

For $d, e \in \mathcal{D}(r)$, bilinearity gives

$$b_W(d, d) - b_W(e, e) = b_W(d - e, d) + b_W(e, d - e).$$

Let $L := \sup_{u\in\mathcal{F}} J(u)$ within this proof. For $d, e \in \mathcal{D}(r)$, the inequalities $J(d), J(e) \le 2L$ and $\beta\|h\|_\infty \le J(h)$ give

$$\begin{aligned} \|d\|_\infty J(d - e) + J(e)\|d - e\|_\infty &\le \frac{4L}{\beta} J(d - e), \\ [J(\bar{u}) + \beta Q_{\max}]\|d - e\|_\infty &\le \left(\frac{L}{\beta} + Q_{\max}\right) J(d - e). \end{aligned}$$

Applying (49) to the bilinear expansion and using Cauchy–Schwarz therefore bounds the centered quadratic increments by a sub-Gaussian scale $CJ(d - e)$. Equation (50) gives the same scale for the residual increments. Independence across the n trajectories reduces both scales to $CJ(d - e)/\sqrt{n}$.

Both processes vanish at $d = 0$. Their finite-grid evaluations and expectations are continuous in J, so separability allows their suprema on $\mathcal{D}(r)$ to be taken over a countable dense subset. Applying the sub-Gaussian chaining bound (Dirksen, 2015, Theorem 3.2) in the J metric gives

$$\Pr\{\mathcal{Q}_n(r) > C\Psi(r)\} \le \delta/2, \qquad \Pr\{\mathcal{S}_n(r) > C\Psi(r)\} \le \delta/2.$$

Here the entropy integral bounds the chaining functional, and $\sqrt{\log(4/\delta)}$ controls the diameter term in each tail. If the diameter is zero, both processes are identically zero. A union bound proves (45), completing the proof of Lemma 6.

B.4 Proof of Lemma 7

Let $e = \widehat{u} - \bar{u}$. Evaluating the population VI at the test function $v = \widehat{u}$ and the empirical VI at $v = \bar{u}$ yields

$$B(\bar{u}, e) \geq \mathfrak{l}_\beta(e), \qquad \widehat{B}(\widehat{u}, e) \leq P_n \ell_W(e).$$

Rearranging the difference $\widehat{B}(\widehat{u}, e) - B(\bar{u}, e) \leq P_n \ell_W(e) - \mathfrak{l}_\beta(e)$ via the bilinear expansion $\widehat{u} = \bar{u} + e$ and adding $B(e, e)$ to both sides yields

$$B(e, e) \leq (B - \widehat{B})(e, e) + (B - \widehat{B})(\bar{u}, e) + (P_n - P)\ell_W(e).$$

Since $P\ell_W = \mathfrak{l}_\beta$, splitting $B - \widehat{B} = (B - B_\Delta) + (P - P_n)b_W$ proves the basic inequality (38), with α and the chosen norm in place of $\alpha_{B,\beta}$ and $\|\cdot\|_{\mathscr{E}}$.

Assume for contradiction that $r_e := \|e\| > r$. By convexity, $d = (r/r_e)e \in \mathcal{D}(r)$. Substituting $e = (r_e/r)d$ into (38) and applying the local assumptions to these scaled terms yields

$$\alpha r_e^2 \leq \left(\frac{r_e}{r}\right)^2 \eta_Q \alpha r^2 + \frac{r_e}{r} \eta_S \alpha r^2.$$

Dividing by $\alpha r_e > 0$ yields $(1 - \eta_Q) r_e \leq \eta_S r$. Since $\eta_Q + \eta_S < 1$, this strict inequality requires $r_e < r$, which contradicts the assumption $r_e > r$. This completes the proof of Lemma 7.

B.5 Proof of Theorem 2

Put $\eta = \sqrt{\{d + \log(4/\delta)\}/n}$. By (21), one J-ball of radius A covers $r^{-1}\mathcal{D}(r)$. Homogeneity of J gives

$$\operatorname{diam}_J \mathcal{D}(r) \leq 2Ar,$$
$$\int_0^{\operatorname{diam}_J \mathcal{D}(r)} \sqrt{\log N(\mathcal{D}(r), J, \epsilon)}\, d\epsilon \leq \int_0^{Ar} \sqrt{d \log(Ar/\epsilon)}\, d\epsilon = \frac{\sqrt{\pi}}{2} Ar\sqrt{d}.$$

The integrand is zero above Ar because one ball suffices there. Consequently, for every $r > 0$,

$$\Psi(r) + \Delta \operatorname{diam}_J \mathcal{D}(r) \leq C_1 r(\eta + \Delta),$$

where $C_1 \geq 1$ depends only on A. Let c be the constant in Theorem 4 and choose the deterministic radius

$$r = \frac{C_1}{c\alpha_{B,\beta}}(\eta + \Delta).$$

This radius satisfies (35). The theorem therefore gives, on one event of probability at least $1 - \delta$, for every empirical VI solution,

$$\|\widehat{u}_{n,\Delta} - u^*\|_{\mathscr{E}}^2 \leq 2\left(1 + \frac{C_{D,\beta}}{\alpha_{B,\beta}}\right)^2 a_{\text{app}}^2 + \frac{4C_1^2}{c^2 \alpha_{B,\beta}^2}(\eta^2 + \Delta^2).$$

Choosing the constant C in the corollary to dominate these coefficients proves (22), completing the proof.

C Stochastic approximation on a linear space

Over a fixed linear space, the TD moment equation can be solved via an unconstrained stochastic approximation algorithm. We maintain the reverse sampling framework, the discounted target u^* with terminal condition $u_T^* = q$, and the use of independent full trajectories from the diffusion law as in (16).

Assumption 3 (Linear features and fresh trajectories)**.** *The diffusion satisfies Assumption 1(i), and $u^* \in \mathscr{V}^{(4)}$, with $\mathscr{V}^{(4)}$ defined as the closure of the smooth core under* (9)*. The deterministic space*

$\mathcal{X} \subset \mathscr{V}^{(4)} \cap C^{1,2}$ *has dimension* $d_0 \geq 1$. *There are finite constants* H, L, D *such that every* $h \in \mathcal{X}$ *satisfies*

$$\|h\|_\infty \leq H\|h\|_{\mathscr{E}}, \qquad \|a^{1/2}\nabla h\|_\infty \leq L\|h\|_{\mathscr{E}}, \qquad \max_{\pm} \|R_\lambda^{\pm} h\|_\infty \leq D\|h\|_{\mathscr{E}}. \tag{51}$$

Here the suprema range over $[0,T] \times \mathbb{R}^d$, *with the Euclidean norm for vector-valued quantities, and* $R_\lambda^{\pm}$ *are defined in* (18). *The blocks* $W_1, W_2, \ldots$ *are i.i.d. from the exact joint law of the diffusion and terminal target, and* $\mathbb{E}Q^2 < \infty$. *The constants* H, L, D *are independent of the number of trajectories and the grid spacing, but may depend on the fixed space and diffusion.*

Choose a basis $\phi = (\phi_1, \ldots, \phi_{d_0})$ of $\mathcal{X}$ and write $u^\theta = \theta^\top \phi$. Its energy Gram matrix $M_E \succ 0$ is defined by $\|u^\theta\|_{\mathscr{E}}^2 = \theta^\top M_E \theta$. For a uniform grid $t_k = k\Delta$, $K\Delta = T$, and $\gamma = e^{-\lambda\Delta}$, put $\phi_{W,k} = \phi(t_k, Z_k)$ and define

$$A_W := \sum_{k=0}^{K-1} \phi_{W,k}(\phi_{W,k} - \gamma\phi_{W,k+1})^\top + \beta\phi_{W,K}\phi_{W,K}^\top, \qquad b_W^{\mathrm{vec}} := \beta Q \phi_{W,K}. \tag{52}$$

Thus $\eta^\top A_W \theta = b_W(u^\theta, u^\eta)$ and $\eta^\top b_W^{\mathrm{vec}} = \ell_W(u^\eta)$, where the scalar forms are those in (16). Let $A_\Delta = \mathbb{E}A_W$ and $b = \mathbb{E}b_W^{\mathrm{vec}}$. The iteration uses an ordinary coefficient update:

$$\theta_{j+1} = \theta_j + \eta_j(b_{W_{j+1}}^{\mathrm{vec}} - A_{W_{j+1}}\theta_j). \tag{53}$$

For this linear space, define the grid margin and moment bound

$$\begin{aligned} &\alpha_{\mathrm{lin}} := \alpha_{B,\beta} - \tfrac{1}{2}TD^2\Delta, && \mu := \alpha_{\mathrm{lin}}\lambda_{\min}(M_E), \\ &L_2^2 := 2\lambda_{\max}(M_E)\,\mathrm{Tr}(M_E)\{TH^2L^2 + (THD + \beta H^2)^2\}. \end{aligned} \tag{54}$$

When $A_\Delta\theta_\Delta = b$ has a unique solution, define its trajectory noise level by

$$\sigma_\Delta^2 := \mathbb{E}\|b_W^{\mathrm{vec}} - A_W\theta_\Delta\|_2^2. \tag{55}$$

The existence of this solution and an upper bound for σ_Δ^2, uniform over grids with α_{lin} bounded away from zero, are established in Lemmas 8 and 9.

Theorem 5 (Linear stochastic approximation)**.** *Suppose Assumption 3 holds and* $\alpha_{\mathrm{lin}} > 0$. *Then* $A_\Delta\theta_\Delta = b$ *has a unique solution. Fix a deterministic* $\theta_0 \in \mathbb{R}^{d_0}$, *take*

$$j_0 \geq \max\{2, 4L_2^2/\mu^2\}, \qquad \eta_j = \frac{2}{\mu(j + j_0)}, \qquad j \geq 0,$$

and run (53). *For every integer* $n \geq 0$,

$$\mathbb{E}\|\theta_n - \theta_\Delta\|_2^2 \leq e_n^2 := \frac{\max\{j_0\|\theta_0 - \theta_\Delta\|_2^2, 8\sigma_\Delta^2/\mu^2\}}{n + j_0}, \tag{56}$$

$$\left(\mathbb{E}\|u^{\theta_n} - u^*\|_{\mathscr{E}}^2\right)^{1/2} \leq \left(1 + \frac{C_{D,\beta}}{\alpha_{B,\beta}}\right) \inf_{v \in \mathcal{X}} \|v - u^*\|_{\mathscr{V}^{(4)}} + \frac{\beta TD^2\|Q\|_{L^2}}{2\alpha_{B,\beta}\alpha_{\mathrm{lin}}}\Delta + \sqrt{\lambda_{\max}(M_E)}\, e_n. \tag{57}$$

If $u^* \in \mathcal{X}$, *the approximation and grid terms in* (57) *can both be omitted.*

See Appendix C for the proof.

The oracle inequality (57) explicitly decouples the spatial approximation and time-discretization biases from the stochastic optimization error. Notably, in an energy-orthonormal basis with a grid margin bounded away from zero, the trajectory variance scales as $\sigma_\Delta^2 = \mathcal{O}(d_0)$. This full-trajectory formulation achieves an $\mathcal{O}(d_0/n)$ statistical error rate, uniformly over grids whose margin stays bounded away from zero.

We separate the error of the population Galerkin approximation from the stochastic iteration:

$$\left(\mathbb{E}\|u^{\theta_n} - u^*\|_{\mathscr{E}}^2\right)^{1/2} \leq \left(\mathbb{E}\|u^{\theta_n} - u^{\theta_\Delta}\|_{\mathscr{E}}^2\right)^{1/2} + \|u^{\theta_\Delta} - \bar{u}_{\mathcal{X}}\|_{\mathscr{E}} + \|\bar{u}_{\mathcal{X}} - u^*\|_{\mathscr{E}}, \tag{58}$$

where $\bar{u}_{\mathcal{X}}$ is the continuous Galerkin solution defined below. Throughout this section, Assumption 3 holds and $\alpha_{\mathrm{lin}} > 0$. The continuous moment, continuity, and coercivity identities in Lemmas 1 to 3 use the diffusion and target conditions and the definition of $\mathscr{V}^{(4)}$.

The first lemma proves existence directly on the entire linear space and controls the two deterministic terms in (58).

Lemma 8 (Population Galerkin solutions). *Under Assumption 3, if $\alpha_{\mathrm{lin}} > 0$, the equations*

$$\mathfrak{B}_\beta(\bar{u}_{\mathcal{X}}, g) = \mathfrak{l}_\beta(g), \qquad Pb_W(u^{\theta_\Delta}, g) = \mathfrak{l}_\beta(g), \qquad g \in \mathcal{X}, \tag{59}$$

have unique solutions in $\mathcal{X}$, and the second equation is $A_\Delta \theta_\Delta = b$. They satisfy

$$\|\bar{u}_{\mathcal{X}} - u^*\|_{\mathscr{E}} \le \left(1 + \frac{C_{D,\beta}}{\alpha_{B,\beta}}\right) \inf_{v \in \mathcal{X}} \|v - u^*\|_{\mathscr{V}^{(4)}}, \tag{60}$$

$$\|u^{\theta_\Delta} - \bar{u}_{\mathcal{X}}\|_{\mathscr{E}} \le \frac{TD^2\Delta}{2\alpha_{\mathrm{lin}}} \|\bar{u}_{\mathcal{X}}\|_{\mathscr{E}}, \tag{61}$$

$$\|\bar{u}_{\mathcal{X}}\|_{\mathscr{E}} \le \frac{\beta \|Q\|_{L^2}}{\alpha_{B,\beta}}, \qquad \|u^{\theta_\Delta}\|_{\mathscr{E}} \le \frac{\beta \|Q\|_{L^2}}{\alpha_{\mathrm{lin}}}. \tag{62}$$

Moreover, $v^\top A_\Delta v \ge \mu \|v\|_2^2$ for every $v \in \mathbb{R}^{d_0}$. If $u^* \in \mathcal{X}$, both solutions equal u^*.

See Appendix C.1 for the proof.

The second lemma derives the stochastic moments from the trajectory structure, so the iteration does not need a bounded coefficient set.

Lemma 9 (Trajectory moments for linear stochastic approximation). *Under Assumption 3, if $\alpha_{\mathrm{lin}} > 0$, then for every deterministic $v \in \mathbb{R}^{d_0}$,*

$$\mathbb{E}\|A_W v\|_2^2 \le L_2^2 \|v\|_2^2, \tag{63}$$

$$\sigma_\Delta^2 \le 2 \operatorname{Tr}(M_E) \left\{ TH^2 L^2 R_*^2 + \left[(THD + \beta H^2) R_* + \beta H \|Q\|_{L^2} \right]^2 \right\}, \tag{64}$$

where $R_* := \beta \|Q\|_{L^2} / \alpha_{\mathrm{lin}}$.

See Appendix C.2 for the proof.

Taking Lemmas 8 and 9 as given, we prove Theorem 5. Let $\mathcal{G}_j = \sigma(W_1, \dots, W_j)$, augmented by the deterministic initialization, and put $e_j^{\mathrm{vec}} = \theta_j - \theta_\Delta$ and $\xi_{j+1} = b_{W_{j+1}}^{\mathrm{vec}} - A_{W_{j+1}} \theta_\Delta$. The update makes θ_j measurable with respect to $\mathcal{G}_j$. Independence of the fresh block gives $\mathbb{E}[\xi_{j+1} \mid \mathcal{G}_j] = 0$ and $\mathbb{E}[A_{W_{j+1}} \mid \mathcal{G}_j] = A_\Delta$. The moment bounds and induction give finite second moments for every iterate. Expanding the update conditionally on $\mathcal{G}_j$ yields

$$\begin{aligned} \mathbb{E}[\|e_{j+1}^{\mathrm{vec}}\|_2^2 \mid \mathcal{G}_j] &\le (1 - 2\mu\eta_j) \|e_j^{\mathrm{vec}}\|_2^2 + \eta_j^2 \mathbb{E}[\|A_{W_{j+1}} e_j^{\mathrm{vec}} - \xi_{j+1}\|_2^2 \mid \mathcal{G}_j] \\ &\le (1 - 2\mu\eta_j + 2L_2^2 \eta_j^2) \|e_j^{\mathrm{vec}}\|_2^2 + 2\sigma_\Delta^2 \eta_j^2. \end{aligned} \tag{65}$$

The first inequality uses the population coercivity in Lemma 8; the second applies $\|x - y\|_2^2 \le 2\|x\|_2^2 + 2\|y\|_2^2$ and Lemma 9.

The condition on j_0 implies $\eta_j \le \mu / (2L_2^2)$. Taking expectations in (65), and writing $s_j = \mathbb{E}\|e_j^{\mathrm{vec}}\|_2^2$, therefore gives

$$s_{j+1} \le \left(1 - \frac{2}{j + j_0}\right) s_j + \frac{8\sigma_\Delta^2}{\mu^2 (j + j_0)^2}. \tag{66}$$

Set $c_* := \max\{j_0 \|\theta_0 - \theta_\Delta\|_2^2, 8\sigma_\Delta^2 / \mu^2\}$. The bound $s_0 \le c_* / j_0$ holds by definition. Since $j + j_0 \ge 2$, an induction using (66) gives

$$s_{j+1} \le \frac{c_*}{j + j_0} - \frac{c_*}{(j + j_0)^2} \le \frac{c_*}{j + j_0 + 1}.$$

This proves (56). The identity defining M_E then implies

$$\left(\mathbb{E}\|u^{\theta_n} - u^{\theta_\Delta}\|_{\mathscr{E}}^2\right)^{1/2} \le \sqrt{\lambda_{\max}(M_E)}\, e_n.$$

Substituting this estimate and Lemma 8 into (58) proves (57). When $u^* \in \mathcal{X}$, the same lemma makes both deterministic terms zero. These observations complete the proof of Theorem 5.

C.1 Proof of Lemma 8

For $h, g \in \mathfrak{X}$, apply Lemma 5 with the feature bounds (51) to obtain

$$|(Pb_W - \mathfrak{B}_\beta)(h, g)| \le \tfrac{1}{2} T D^2 \Delta \|h\|_{\mathscr{E}} \|g\|_{\mathscr{E}}. \tag{67}$$

The smoothness and the two bounded discounted drifts in Assumption 3 supply the hypotheses of that lemma. Combining (67) with the continuous energy identity gives

$$Pb_W(h, h) \ge \alpha_{\mathrm{lin}} \|h\|_{\mathscr{E}}^2, \qquad h \in \mathfrak{X}.$$

In coordinates, this is $v^\top A_\Delta v \ge \alpha_{\mathrm{lin}} v^\top M_E v \ge \mu \|v\|_2^2$. Consequently, A_Δ is invertible in finite dimension, proving existence and uniqueness of θ_Δ. The same argument with the continuous coercivity constant $\alpha_{B,\beta}$ proves existence and uniqueness of $\bar{u}_{\mathfrak{X}}$.

The exact moment equation and the first Galerkin equation give $\mathfrak{B}_\beta(\bar{u}_{\mathfrak{X}} - u^*, g) = 0$ for every $g \in \mathfrak{X}$. For $v \in \mathfrak{X}$, set $d = \bar{u}_{\mathfrak{X}} - v$. Coercivity and continuity give

$$\alpha_{B,\beta} \|d\|_{\mathscr{E}}^2 \le \mathfrak{B}_\beta(d, d) = \mathfrak{B}_\beta(u^* - v, d) \le C_{D,\beta} \|u^* - v\|_{\mathscr{V}^{(4)}} \|d\|_{\mathscr{E}}.$$

Dividing both sides by $\|d\|_{\mathscr{E}}$, applying the triangle inequality, and taking the infimum over v proves (60).

Subtracting the Galerkin equations gives, for $d_\Delta = u^{\theta_\Delta} - \bar{u}_{\mathfrak{X}}$,

$$Pb_W(d_\Delta, d_\Delta) = (\mathfrak{B}_\beta - Pb_W)(\bar{u}_{\mathfrak{X}}, d_\Delta).$$

Grid coercivity and (67) imply (61). Evaluating each Galerkin equation at its solution and using $|\mathfrak{l}_\beta(h)| \le \beta \|Q\|_{L^2} \|h\|_{\mathscr{E}}$ proves both bounds in (62).

If $u^* \in \mathfrak{X}$, the continuous exact moment equation gives $\bar{u}_{\mathfrak{X}} = u^*$ by uniqueness. The martingale property of $f_t^*(\overline{X}_t)$ implies

$$\mathbb{E}[\gamma u_{t_{k+1}}^*(\overline{X}_{t_{k+1}}) \mid \overline{X}_{t_k}] = u_{t_k}^*(\overline{X}_{t_k}).$$

Together with the terminal condition $u_T^* = q$ and $\mathbb{E}[Q \mid \overline{X}_T] = q(\overline{X}_T)$, this makes u^* a solution of the grid equation. Grid uniqueness therefore gives $u^{\theta_\Delta} = u^*$, which completes the proof.

C.2 Proof of Lemma 9

Let $h, g \in \mathfrak{X}$, applying Itô's formula to the discounted process $e^{-\lambda(r-t_k)} h_r(\overline{X}_r)$ on each interval $[t_k, t_{k+1}]$ and summing across the grid yields

$$b_W(h, g) = N_W(h, g) - M_W(h, g), \tag{68}$$

$$N_W(h, g) := \beta h_T(Z_K) g_T(Z_K) - \sum_{k=0}^{K-1} \int_{t_k}^{t_{k+1}} e^{-\lambda(r-t_k)} g_{t_k}(Z_k) R_\lambda^+ h_r(\overline{X}_r)\, dr,$$

$$M_W(h, g) := \sum_{k=0}^{K-1} \int_{t_k}^{t_{k+1}} e^{-\lambda(r-t_k)} g_{t_k}(Z_k) \nabla h_r(\overline{X}_r)^\top a_r^{1/2}\, d\overline{B}_r.$$

Combining the pointwise bounds on h, ∇h, and $R_\lambda^+ h$ from Assumption 3 with the Itô isometry yields

$$|N_W(h, g)| \le (THD + \beta H^2) \|h\|_{\mathscr{E}} \|g\|_{\mathscr{E}},$$
$$\mathbb{E}|M_W(h, g)|^2 \le T H^2 L^2 \|h\|_{\mathscr{E}}^2 \|g\|_{\mathscr{E}}^2. \tag{69}$$

Using $(x - y)^2 \le 2x^2 + 2y^2$ in (68) gives

$$\mathbb{E}|b_W(h, g)|^2 \le 2\{T H^2 L^2 + (THD + \beta H^2)^2\} \|h\|_{\mathscr{E}}^2 \|g\|_{\mathscr{E}}^2.$$

For $h = u^v$ and $g = \phi_i$, the left side is the second moment of the ith coordinate of $A_W v$. Summing over i, using $\sum_i \|\phi_i\|_{\mathscr{E}}^2 = \mathrm{Tr}(M_E)$ and $\|u^v\|_{\mathscr{E}}^2 \le \lambda_{\max}(M_E) \|v\|_2^2$, we obtain(63).

For the trajectory noise, put $h = u^{\theta_\Delta}$ and subtract $\ell_W(g) = \beta Q g_T(Z_K)$ from (68). Minkowski's inequality and (69) imply

$$\|N_W(h, g) - \ell_W(g)\|_{L^2(P)} \le \{(THD + \beta H^2) \|h\|_{\mathscr{E}} + \beta H \|Q\|_{L^2}\} \|g\|_{\mathscr{E}}.$$

Consequently,

$$\mathbb{E}|b_W(h,g) - \ell_W(g)|^2 \le 2\big\{TH^2L^2\|h\|_{\mathscr{E}}^2 + [(THD + \beta H^2)\|h\|_{\mathscr{E}} + \beta H\|Q\|_{L^2}]^2\big\}\|g\|_{\mathscr{E}}^2.$$

Apply this bound to $g = \phi_i$, sum over i, and use $\|u^{\theta_\Delta}\|_{\mathscr{E}} \le R_*$ from Lemma 8. The resulting inequality is (64). These bounds require only $\mathbb{E}Q^2 < \infty$ and allow dependence of Q on the trajectory, which completes the proof.

D Proofs for guidance transfer and the Gaussian counterexample

D.1 Proof of Theorem 3

Condition on the training sample and use the canonical natural filtration, completed separately under each path law. The terminal tilt has drift $b_t + a_t\nabla \log u_t^*$ by (5), since $\nabla \log u^* = \nabla \log f^*$. The drift difference in diffusion coordinates is $\sigma_t^\top(\nabla \log \widehat{u}_t - \nabla \log u_t^*)$. It is bounded by the hypotheses, so Novikov's condition holds. Girsanov's theorem and uniqueness in law identify the changed law with $\widehat{\mathbb{P}}$. Since the initial laws agree, the relative-entropy identity and data processing at time T give

$$\begin{aligned} D_{\mathrm{KL}}(\mu_T^q\|\widehat{\mu}_T) &\le D_{\mathrm{KL}}(\mathbb{P}^q\|\widehat{\mathbb{P}}) \\ &= \frac{1}{2}\int_0^T \mathbb{E}_{\mu_t^q}\big\|\sigma_t^\top(\nabla \log \widehat{u}_t - \nabla \log u_t^*)\big\|_2^2\, dt. \end{aligned} \tag{70}$$

To bound the guidance error, set $e = \widehat{u} - u^*$. The positive floor and the bounds on a_t and $\nabla \log u^*$ yield the pointwise estimates

$$\nabla \log \widehat{u}_t - \nabla \log u_t^* = \frac{\nabla e_t - e_t\nabla \log u_t^*}{\widehat{u}_t},$$

$$\big\|\sigma_t^\top(\nabla \log \widehat{u}_t - \nabla \log u_t^*)\big\|_2^2 \le \frac{2\Lambda}{m^2}\big(\|\nabla e_t\|_2^2 + G^2|e_t|^2\big).$$

The tower property and (3) give

$$\frac{d\mu_t^q}{d\nu_t}(x) = \frac{f_t^*(x)}{Z_Q} \le \frac{q_{\max}}{Z_Q}. \tag{71}$$

After integrating the pointwise estimate and applying (71), we bound the right-hand side of (70) by

$$\frac{\Lambda q_{\max}}{m^2 Z_Q}\int_0^T \big(\|\nabla e_t\|_t^2 + G^2\|e_t\|_t^2\big)\, dt \le \frac{\Lambda q_{\max}}{m^2 Z_Q}\max\{1, G^2\}\|e\|_{\mathscr{E}}^2.$$

Combining this estimate with (70) proves (23), completing the proof.

D.2 Proof of Proposition 1

Let $T = 1$, $\overline{X}_0 \sim N(0, 10^{-2})$, and $d\overline{X}_t = d\overline{B}_t$. This is a Gaussian reverse diffusion in the formal setup, with $\nu_t = N(0, 10^{-2} + t)$; its forward counterpart is $dX_s = -X_s/(1.01 - s)\, ds + dB_s$ from $X_0 \sim N(0, 1.01)$. Take $q \equiv 1$, so $f_t^* \equiv 1$ and the desired tilted terminal law is $\mu_1^q = \nu_1 = N(0, 1.01)$. Fix a smooth $a : [0,1] \to [0,1]$ with $a = 1$ on $[0, 0.99]$ and $a(1) = 0$. For integers $k \ge 1$, put $\delta_k = k^{-1/2}$ and

$$\widehat{f}_t^{(k)}(x) := \exp\big\{\delta_k a(t)\sin(kx - k^2 t)\big\}. \tag{72}$$

These estimates are smooth, satisfy $\widehat{f}_1^{(k)} = q = 1$, and obey the common bounds $e^{-1} \le \widehat{f}_t^{(k)} \le e$. The plug-in SDE starts from $\mu_0^q = \nu_0$ and has drift $\partial_x \log \widehat{f}_t^{(k)}$. Write $\delta = \delta_k$ and $v_t = 10^{-2} + t$. Since $f_t^* \equiv 1$ and $\sup_{t,x}|\widehat{f}_t^{(k)}(x) - 1| \le e^\delta - 1 = O(k^{-1/2})$, the value estimates converge uniformly; in particular,

$$\sup_{0\le t\le 1}\|\widehat{f}_t^{(k)} - 1\|_{L^2(\nu_t)}^2 = O(k^{-1}).$$

On $[0, 0.99]$ the plug-in score equals $\sqrt{k}\cos(kx - k^2t)$. For $X \sim \nu_t = N(0, v_t)$,

$$\mathbb{E}\cos^2(kX - k^2t) = \tfrac{1}{2}\{1 + e^{-2k^2 v_t}\cos(2k^2t)\} \geq \tfrac{1}{2}(1 - e^{-0.02k^2}).$$

Thus, for all sufficiently large k,

$$\int_0^{0.99} \|\partial_x \log \widehat{f}_t^{(k)}\|^2_{L^2(\nu_t)}\, dt \gtrsim k.$$

Because $\widehat{f}_t^{(k)} \geq e^{-1}$, the spatial-gradient component of $\|\widehat{f}^{(k)} - f^*\|^2_{\mathscr{E}}$ is likewise bounded below by a constant multiple of k.

Let Y solve the plug-in SDE with $Y_0 \sim \nu_0$, independent of its driving Brownian motion B. Its drift is smooth and bounded for each k, so the SDE is nonexplosive. Set

$$\theta_t := kY_t - k^2t, \qquad A_t := \sqrt{k}\int_0^t a(s)\cos\theta_s\, ds.$$

Then $Y_t = Y_0 + B_t + A_t$ and $d\theta_t = k^2[-1 + \delta a(t)\cos\theta_t]dt + k\, dB_t$. Define the periodic differential operator $L_0 = -\partial_\theta + \frac{1}{2}\partial_{\theta\theta}$ and

$$\phi(\theta) := \tfrac{4}{5}\sin\theta + \tfrac{2}{5}\cos\theta, \qquad L_0\phi = -\cos\theta,$$
$$h(\theta) := \cos\theta\,\phi'(\theta) = \tfrac{2}{5} + \tfrac{2}{5}\cos(2\theta) - \tfrac{1}{5}\sin(2\theta).$$

Applying Itô's formulato $\delta a(t)\phi(\theta_t)/k$ gives the exact identity

$$\begin{aligned} A_t = &-\frac{\delta}{k}[a(t)\phi(\theta_t) - a(0)\phi(\theta_0)] + \frac{\delta}{k}\int_0^t a'(s)\phi(\theta_s)\, ds \\ &+ \int_0^t a(s)^2 h(\theta_s)\, ds + \delta\int_0^t a(s)\phi'(\theta_s)\, dB_s. \end{aligned} \tag{73}$$

To eliminate the oscillations of h, let

$$\psi(\theta) := \tfrac{1}{20}\sin(2\theta) + \tfrac{3}{20}\cos(2\theta), \qquad L_0\psi = -(h - \tfrac{2}{5}).$$

Applying the same formula to $a(t)^2\psi(\theta_t)/k^2$ gives

$$\begin{aligned} \int_0^t a(s)^2(h(\theta_s) - \tfrac{2}{5})\, ds = &-\frac{a(t)^2\psi(\theta_t) - a(0)^2\psi(\theta_0)}{k^2} + \frac{1}{k^2}\int_0^t 2a(s)a'(s)\psi(\theta_s)\, ds \\ &+ \delta\int_0^t a(s)^3\cos\theta_s\,\psi'(\theta_s)\, ds + \frac{1}{k}\int_0^t a(s)^2\psi'(\theta_s)\, dB_s. \end{aligned} \tag{74}$$

All periodic functions here and the a, a' are bounded. The deterministic remainder terms in (73)–(74) are $O(k^{-1/2})$ or smaller; the two stochastic terms have maximal second moments $O(k^{-1})$ by the Burkholder–Davis–Gundy inequality. Consequently,

$$\mathbb{E}\sup_{0\leq t\leq 1}\left|A_t - \frac{2}{5}\int_0^t a(s)^2\, ds\right|^2 = O(k^{-1}). \tag{75}$$

Letting $c_a := \frac{2}{5}\int_0^1 a(s)^2 ds \geq 0.396$. Since $Y_0 + B_1 \sim N(0, 1.01)$, (75) implies the weak convergence $\widehat{\mu}_1^{(k)} \Rightarrow N(c_a, 1.01)$, whereas the target is exactly $\mu_1^q = N(0, 1.01)$. Denoting the standard normal cdf by Φ, evaluating the total variation via the test set $(c_a/2, \infty)$ yields

$$\liminf_{k\to\infty} d_{\mathrm{TV}}(\widehat{\mu}_1^{(k)}, \mu_1^q) \geq 2\Phi\left(\frac{c_a}{2\sqrt{1.01}}\right) - 1 \geq 2\Phi\left(\frac{0.396}{2\sqrt{1.01}}\right) - 1 > 0.15.$$

Hence $d_{\mathrm{TV}}(\widehat{\mu}_1^{(k)}, \mu_1^q) > 0.15$ for all sufficiently large k, completing the proof of Proposition 1.

E EXPERIMENTAL FORMULATIONS AND EVALUATION

This section specifies the exact target, baseline methods, and evaluation metrics for each empirical study. Complete implementation configurations, including exact hyperparameters, are documented in the supplementary code archive.

E.1 TWO-DIMENSIONAL GUIDANCE BENCHMARKS

Common diffusion and TD comparison. All four tasks employ the variance-preserving forward diffusion

$$dX_s = -\tfrac{3}{2}X_s\,ds + \sqrt{3}\,dB_s, \qquad 0 \le s \le 1,$$

where $s = 0$ denotes the clean data state and $s = 1$ the terminal noise state; the reverse sampling time is $t = 1-s$. For a fixed class or reward observation Q, we denote $h_s^*(x) = \mathbb{E}[Q \mid X_s = x] = f_{1-s}^*(x)$. The actual quantity required for guidance is the spatial gradient $\nabla_x \log h_s^*$, rather than the value h_s^* itself. In these two-dimensional benchmarks, the TD prediction at the noisier state X_s is trained against an exponential-moving-average target network evaluated at the Markov-coupled cleaner state X_c, with $c = \max\{0, s - 0.1\}$. Within each task, both the Direct and TD objectives include an additional clean-endpoint loss at forward time $s = 0$, with coefficient 0.5. This loss is squared error against the clean reward for scalar reward tasks and cross-entropy against the true label for two moons. Scalar outputs use a sigmoid activation, while class outputs use a softmax; the log-guidance field is differentiated prior to any test-time clipping.

Targets and reference fields. The eight-Gaussian ring and the bistable task possess finite Gaussian-mixture clean distributions. Let J denote the latent component with base weights w_j, component rewards q_j, and noised component densities $\phi_{j,s}$. The terminal reward is the observable clean-state posterior mean $Q(X_0) = \mathbb{E}[q_J \mid X_0]$. Consequently, their reference value function is the finite sum

$$h_s^*(x) = \frac{\sum_j w_j q_j \phi_{j,s}(x)}{\sum_j w_j \phi_{j,s}(x)},$$

yielding an analytically tractable log-gradient. The ring task comprises eight equal-weight modes with heterogeneous rewards; in the bistable task, the high-reward mode possesses a base mass of 0.05 but a target-tilted mass of 0.5. For the two-moons task, $Q_y = \mathbf{1}\{Y = y\}$, making the reference the noisy class posterior $h_s^{*,y}(x) = \mathbb{P}(Y = y \mid X_s = x)$. For the spiral task, the reward is a strictly positive function of its clean latent coordinate U, $Q = \exp\{-0.15 - 1.4[1 - \cos(2\pi(U - 0.72))]\}$, and $h_s^*(x) = \mathbb{E}[Q \mid X_s = x]$. These continuous-latent integrals are computed via $2{,}048$-node midpoint quadrature, with approximation accuracy verified at $4{,}096$ nodes for both values and log-gradients.

Training setup and error measures. Within each task, direct regression and TD learning share paired data and initializations across five training seeds. Table 1 compares these two methods across the ring, bistable, spiral, and two-moons benchmarks.

For the scalar reward tasks, the two reported metrics under the base measure are the time-averaged value MSE $\mathbb{E}[(\widehat{h}_s(X_s) - h_s^*(X_s))^2]$ and the log-guidance MSE $\mathbb{E}[\|\nabla \log \widehat{h}_s(X_s) - \nabla \log h_s^*(X_s)\|^2]$. For the two-moons task, the value error evaluates the two-class posterior vector, and the guidance error computes the class-specific log-posterior gradient averaged over both classes. The evaluation states are drawn from a fixed base-sampling bank, divided into equal-width forward-time bins (ten for reward mixtures, eight for two-moons/spiral), with 384 independent, uniformly sampled states per bin. Errors are averaged across time bins and then over the five seeds. The reported TD/Direct column represents the ratio of these overall seed-mean errors, not the mean of seed-wise ratios.

The numerical parameters defining the four tasks—mixture locations, weights and component rewards for the ring and bistable targets, the two-moons noise level, and the latent curve of the spiral reward—are given by the task configuration files in the supplementary archive, which also contains the code that constructs each task and computes its reference value and guidance fields.

E.2 CIFAR-10 CLASSIFIER GUIDANCE

Generator and guidance networks. The frozen base generator is the unconditional Diffusers `google/ddpm-cifar10-32 UNet2DModel` for 32×32 RGB images, with block widths

$(128, 256, 256, 256)$. Its DDPM schedule has $1{,}000$ training steps and linearly spaced noise variances from 10^{-4} to 0.02; generation uses 250 reverse steps. The separately trained guidance network in each arm is a randomly initialized, time-conditioned, 10-class ADM `EncoderUNetModel` with base width 64, channel multipliers $(1, 2, 3, 4)$, two residual blocks per level, and $10{,}206{,}922$ parameters. The direct and TD methods use paired initializations and data but have distinct learned weights.

Classifier objectives. Here $n = 0$ denotes clean data and $n \in \{1, \dots, 1000\}$ is the physical forward-noising index (the network receives timestep $n - 1$). Writing $\bar{\alpha}_n = \prod_{j=1}^{n}(1 - \beta_j)$, the marginal noising law and the learned class probability are

$$X_n \mid X_0 \sim \mathcal{N}\big(\sqrt{\bar{\alpha}_n} X_0, (1 - \bar{\alpha}_n) I\big)\,,$$
$$P_\theta(c \mid x, n) := \mathrm{softmax}\big(z_\theta(x, n-1)\big)_c.$$

Temporal-difference learning requires a terminal boundary condition to ground its recursive updates. Let R be a forward timestep drawn uniformly from the near-clean regime $\{1, \dots, 10\}$. Denoting the predicted class probability vector at step n as $p_n^\theta := (P_\theta(c \mid X_n, n))_{c=1}^{10}$, and the one-hot encoding of the true label Y as e_Y, we define the root grounding loss via the mean squared error:

$$\mathcal{L}_{\mathrm{root}}(\theta) := \mathbb{E}\big[\tfrac{1}{2}\|p_R^\theta - e_Y\|_2^2\big].$$

It is evaluated at a *noisy* near-clean state, not at the exact endpoint X_0. Let N_{all} be uniform on $\{1, \dots, 1000\}$ and N_{int} uniform on $\{11, \dots, 1000\}$. The per-update training losses used in the plotted comparison are

$$\mathcal{L}_{\mathrm{Direct}}(\theta) = \mathbb{E}[-\log p_{N_{\mathrm{all}}}^\theta(Y)] + \mathcal{L}_{\mathrm{root}}(\theta),$$
$$\mathcal{L}_{\mathrm{TD}}(\theta) = \mathcal{L}_{\mathrm{root}}(\theta) + \lambda_{\mathrm{TD}}\mathbb{E}\big[D_{\mathrm{KL}}\big(\mathrm{sg}\, p_{N_{\mathrm{int}}-10}^\theta \,\big\|\, p_{N_{\mathrm{int}}}^\theta\big)\big]\,,$$

where $(X_{N_{\mathrm{int}}-10}, X_{N_{\mathrm{int}}})$ is a coupled forward-Markov pair, sg holds the cleaner prediction from the same pre-update network fixed while differentiating the noisier prediction, and $\lambda_{\mathrm{TD}} = 0.3735$.

Sampling and matched-control evaluation. The target class probability is $f_t^*(x, y) = \mathbb{P}(Y = y \mid \overline{X}_t = x)$, with reverse time $t = T - s$ for forward time s. In the discrete DDPM sampler, a class y and scale w modify the base noise prediction according to

$$\widehat{\epsilon}_{\theta,w}(x, n, y) = \epsilon_{\mathrm{base}}(x, n) - w\sqrt{1 - \bar{\alpha}_n}\ \mathrm{clip}_{[-1,1]}(\nabla_x \log P_\theta(y \mid x, n))\,,$$

where $\epsilon_{\mathrm{base}}(x, n)$ is the frozen U-Net prediction at model timestep $n - 1$, and clipping acts coordinatewise. The same frozen generator and sampler are used in both methods; only the guidance classifier and calibrated scale change.

For a class-balanced generated bank $\{(x_i, y_i)\}_{i=1}^M$, external control is the average top-1 hit fraction of two frozen evaluators, ResNet-56 (R) and VGG-16-BN (V):

$$C := \frac{1}{2M}\sum_{i=1}^{M}\Big(\mathbf{1}\{\arg\max R(x_i) = y_i\} + \mathbf{1}\{\arg\max V(x_i) = y_i\}\Big).$$

Thus C is not the guidance classifier's softmax confidence. For each model seed j, an independent calibration bank estimates the control-versus-guidance-scale curve for each method. The common reachable interval $[L_j, U_j]$ is the intersection of attainable controls for Direct, Pure TD. At quantiles $q \in \{0.125, 0.25, 0.375, 0.5, 0.625, 0.75, 0.875\}$, the seed-specific nominal target is $L_j + q(U_j - L_j)$; each method obtains its scale by monotone calibration and interpolation, without extrapolation. Achieved controls on a disjoint quality bank need not match exactly: a Direct–TD pair is declared matched when their achieved controls differ by at most 0.03, and every plotted pair passes this check. Each quality cell contains $2{,}000$ class-balanced images from the 250-step DDPM sampler.

Figure 1 reports seven quality metrics. Each curve displays the arithmetic mean across five model seeds. Shaded bands are method-wise, nominal pointwise 95% Student-t intervals.

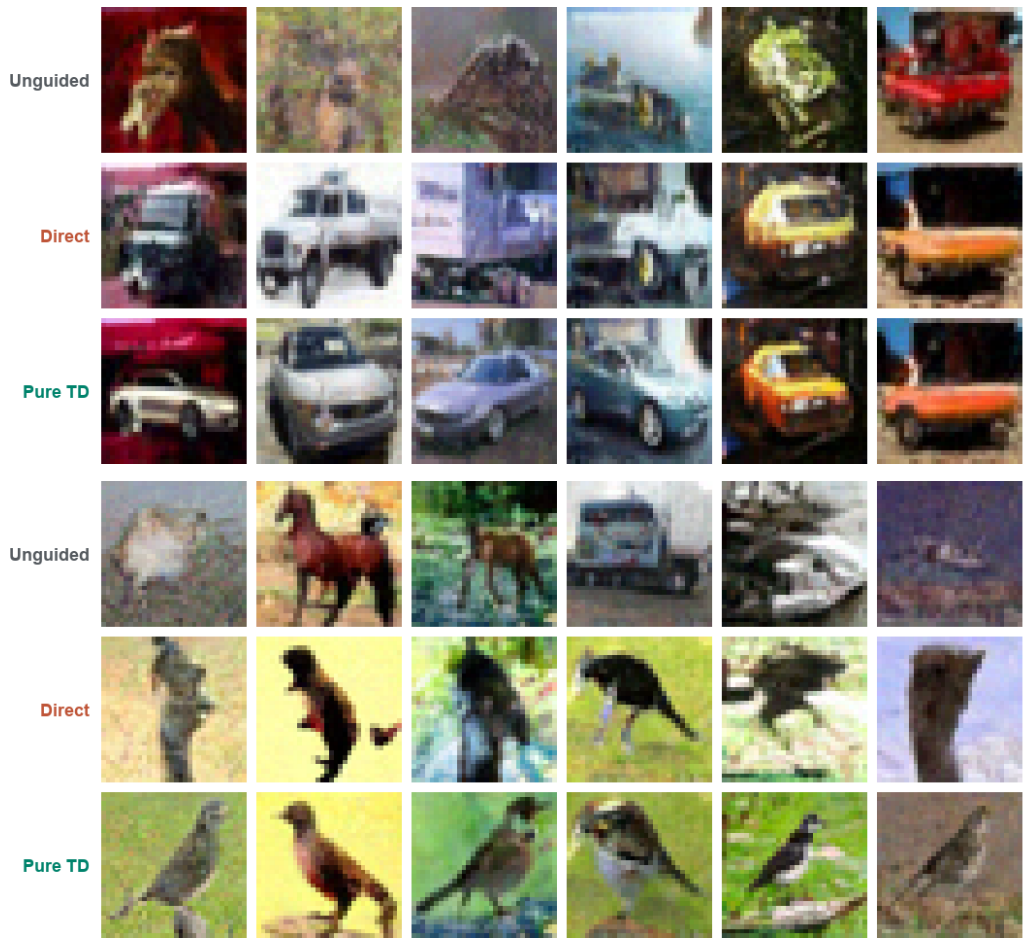

Figure 3: Car (top) and bird (bottom) examples at $q = 0.5$ calibration. Each row shows the same five seeds for Direct (left) and TD (right). The top row of each panel is the base generator output, the middle row is Direct, and the bottom row is TD. The Direct–TD pairs are matched by control C within 0.03.

E.3 Molecular design on the COMPAS corpus

Corpus split and objective formulation. The COMPAS corpus contains 34,071 enumerated catacondensed polybenzenoid hydrocarbons. For any generated structure successfully matched to the corpus, its electronic properties are retrieved via exact table lookup to compute the minimization objective $J = 3y_{\mathrm{HLG}} + y_{\mathrm{IP}} - y_{\mathrm{EA}}$. The reported evaluation uses the normalized objective $\widetilde{J} = (J - J^{\mathrm{all}}_{\mathrm{min}})/(J^{\mathrm{all}}_{\mathrm{max}} - J^{\mathrm{all}}_{\mathrm{min}})$, defined via the full-table extrema $J^{\mathrm{all}}_{\mathrm{min}} = 20.105$ eV and $J^{\mathrm{all}}_{\mathrm{max}} = 27.412$ eV. During guidance training, the objective is normalized using the subset extrema of 11-ring structures (20.105 and 24.994 eV); both methods share this identically scaled training target. The 25,394 eleven-ring structures are partitioned into independent identically distributed (IID) splits of 16,973 (training), 3,883 (validation), and 4,538 (test) structures, leaving 8,677 smaller-ring entries unassigned but available for evaluation matching.

Evaluation protocol and trade-off metrics. The experimental comparison employs a shared base generator and ten training seeds per method. At each guidance scale $s \in \{0, 0.5, \ldots, 4\}$, each seed generates 512 samples, yielding 5,120 total attempts per method and scale. The evaluated models are direct regression and TD learning.

The two axes of Figure 2 decouple the generation quality from the matching rate. Let M_i be the binary indicator that generation attempt i is both chemically valid and exactly matches an entry in the COMPAS database. The horizontal coordinate reports the corpus-matched fraction, $x = N^{-1} \sum_{i=1}^{N} M_i$. The vertical coordinate reports the conditional mean objective $\widetilde{J}$ computed strictly over this matched subset, $y = (\sum_i M_i \widetilde{J}_i)/(\sum_i M_i)$, which is mathematically well-defined provided at least one attempt matches the corpus. Consequently, the pooled quality metric y naturally weights individual seeds by their respective match counts, reflecting the expected objective value of a successfully matched sample.

F Extended related work

Section 1.1 records the comparisons closest to our setting. This appendix places them in the broader landscape.

Guidance for diffusion models. Beyond the supervised value and classifier guidance of Section 3, other methods estimate different objects. Classifier-free guidance combines conditional and unconditional score estimates (Ho & Salimans, 2022), while DEFT learns an auxiliary score correction through denoising score matching with the base model frozen (Denker et al., 2024). Noise-aware reward models provide another source of step-level target information (Zhang et al., 2025). Learned values can also be used without differentiating them, as in soft value-based decoding (Li et al., 2025).

A related line of work uses particle reweighting, resampling, or tree search to allocate additional computation at inference time (Wu et al., 2023; Jain et al., 2025; Kim et al., 2025; Singhal et al., 2025). Such methods can also incorporate learned value information: Wang et al. (2026) learn twisting functions for sequential Monte Carlo using trust-region updates. We study a complementary approach in which a trained value network and its gradient are reused across samples and denoising steps.

Statistical guarantees for learned guidance. Conditional-score learning also admits statistical guarantees in the classifier-free setting (Fu et al., 2024). These learning guarantees complement diffusion convergence analyses that propagate given score errors into sampling errors (De Bortoli, 2022; Chen et al., 2023b;a; Benton et al., 2024).

Temporal-difference methods. Adjacent-step regression has been proposed for diffusion guidance (Uehara et al., 2024; Li et al., 2025; Uehara et al., 2025). This connects diffusion guidance with classical TD policy evaluation and its projected fixed-point interpretation (Tsitsiklis & Van Roy, 1997; Scherrer, 2010).