Dynamical Consistency Is All You Need

Few-Step, High-Fidelity Sampling via a Variance-Minimizing Blend of Score Estimators

Aloïs Duston de Villerglan
The Oden Institute for Computational Engineering, Sciences, & Mathematics
University of Texas at Austin

Tan Bui-Thanh
Department of Aerospace Engineering & Engineering Mechanics,
The Oden Institute for Computational Science, Engineering, & Mathematics
University of Texas at Austin

July 2025

Abstract

We study *statistical estimation of score fields* for diffusion/flow models and show that a simple, nonparametric combination of two complementary estimators yields strictly lower risk at any time–location (y, t). The approach is built on the **Conditional Score Expectation Matching (CSEM)** principle, which links scores across the forward semigroup. For the Ornstein–Uhlenbeck (OU) case this specializes to

$$\boxed{s(y, t + \Delta t) = e^{\Delta t}\, \mathbb{E}\big[s(X_t, t) \mid X_{t+\Delta t} = y\big]}$$

and we estimate the right–hand side with a simple importance–weighted Monte Carlo procedure using prior samples and forward noise. We pair this *CSEM* estimator with the classical *Tweedie* estimator and show that, under broad conditions, their Monte Carlo errors are negatively aligned, yielding a *variance–minimizing convex blend* with an explicit local weight $\lambda^*(y, t)$.

Our goal is to *demonstrate the utility of this estimator–centric, variance–minimized blended score* over the purely Tweedie-based mode used in standard DSM/flow pipelines—both as a plug–in score for few–step samplers and as a superior training signal for neural distillation. Empirically, the blended estimator reduces variance and improves few–step sampling. Beyond OU, we formulate the CSEM viewpoint for general *affine diffusions* via gradient–semigroup commutation; OU is the canonical worked example. We also give a one–line extension to Bayesian inverse problems by likelihood tilting of the importance weights, turning prior estimators into posterior ones without retraining.

Notation

1 Introduction

Diffusion and flow models set the state of the art across modalities by learning the *score* $\nabla \log p_t$ along a decreasing-noise trajectory and integrating a reverse dynamics to synthesize samples [1–6]. Despite rapid progress, a central bottleneck remains: **sample efficiency**. High fidelity typically requires many denoising steps (large NFE), driving compute and energy costs[1–3, 7, 8].

A prominent response is to compress trajectories via *consistency* or *short-flow* objectives [4–6], but most such formulations tie predictions across time heuristically or via teacher signals, offering

Table 1: Notation. d dimension; N reference set size; K step budget.

Symbol	Meaning
$x \in \mathbb{R}^d$	latent variable with prior density p_0
$y \in \mathbb{R}^d$	forward state at time t
K_t	forward Markov kernel (OU in examples; affine in general)
$s_0(x)$	initial score $\nabla_x \log p_0(x)$
$s(y,t)$	time–t score of the forward marginal p_t
$\hat{s}_{\mathrm{TWD}}(y,t)$	Tweedie estimator at (y,t)
$\hat{s}_{\mathrm{CSEM}}(y,t)$	CSEM estimator at (y,t)
$\hat{s}_{\mathrm{BLEND}}(y,t)$	variance-optimal convex blend of the two
N	number of reference samples / function evaluations
$\pi_t(x_0 \mid y)$	OU posterior over X_0 conditional on $X_t{=}y$

limited control of estimator variance and bias. Our perspective is different: *improve the statistical estimator of the score field itself*, pointwise in (y,t), so that any downstream sampler or distillation scheme inherits lower variance and needs fewer steps.

This paper. We develop a nonparametric, plug-and-play estimator based on a new *Conditional Score Expectation Matching (CSEM)* principle that relates scores at adjacent times through the forward semigroup. In the OU case this yields an exact finite-time identity, which we estimate with self-normalized importance sampling (SNIS). We pair this CSEM estimator with the classical Tweedie estimator and show that their Monte Carlo errors are oppositely aligned—*exactly* for linear–Gaussian priors and *in expectation* under a single-basin posterior regime. This anti-alignment gives a closed-form, *variance-minimizing convex blend* with weight $\lambda^*(y,t)$ determined from the same SNIS batch. The result is a *superior local score estimator* that drops into standard reverse SDE/ODE integrators or consistency distillation, improving few-step sampling without architectural changes.

Beyond OU, we formulate the CSEM viewpoint for general *affine diffusions* via gradient–semigroup commutation; OU is used as the canonical worked example because it exposes all calculations in closed form. We also show that Bayesian inverse problems are handled by a likelihood tilt of the SNIS weights, converting prior estimators to posterior ones with no change to the integrator.

Contributions.

- **CSEM framework.** A conditional-expectation relation for score evolution (exact for OU; formulated for affine diffusions) that supplies a direct, data-driven estimator.
- **Variance-optimal blending.** Closed-form $\lambda^*(y,t)$ from local (co)variances; *pathwise* inner-product negativity for linear–Gaussian priors, and *expected* negativity under OU single-basin dominance.
- **Plug-in estimator for DSM.** A sampler-agnostic score estimator that reduces variance and improves few-step sampling when used with standard reverse integrators or consistency distillation.
- **Posterior sampling by tilting.** A one-line likelihood tilt of SNIS weights yields posterior versions of the same estimators for Bayesian inverse problems.

2 Relation to Prior Work

Diffusion and score-based generative modeling. Modern diffusion and score-based models achieve state-of-the-art synthesis by learning noise-conditional scores and reversing a corruption process. Denoising diffusion probabilistic models (DDPM) introduced the modern denoising formulation [9], while the SDE view unified score-based diffusion with reverse-time dynamics and predictor–corrector samplers [10]. Subsequent architectural and training improvements further cemented performance and scalability [3, 11].

Few-step generation and temporal consistency. A major thrust to reduce the number of function evaluations (NFE) either accelerates integration or imposes temporal ties. Training-free or post-hoc acceleration includes Denoising Diffusion *Implicit* Models (DDIM) [12] and high-order ODE solvers such as the Diffusion Probabilistic Model Solver (DPM-Solver) [13]. Alternative training paradigms learn vector fields directly via Flow Matching [14] or Rectified Flow [15], and Consistency Models impose algebraic relations across noise levels to enable one- or few-step generation [6]. *Our approach is complementary and orthogonal:* rather than proposing a new solver or a heuristic consistency constraint, we improve the *statistical estimator of the score field itself* at a fixed (y, t). This estimator drops into any reverse SDE/ODE integrator or consistency/distillation pipeline and, by provably lowering pointwise variance through optimal blending, enables comparable quality with fewer steps.

Score PDE and physics-informed approaches. Several works leverage the score Fokker–Planck (FP) equation to regularize denoising score matching (e.g., FP-Diffusion) [16], and Score-PINNs minimize the residual of the score PDE directly [17]. Mean-field/control formulations similarly connect sampling to forward PDEs [18]. In contrast, we work *semigroup-first*: for the Ornstein–Uhlenbeck (OU) flow we derive *exact finite-time identities* that yield well-conditioned supervision at small times, and we make *variance* a first-class quantity by proving (linear–Gaussian) or motivating (single-basin) anti-alignment of Monte Carlo errors. Empirically, when the same reverse integrator is used, our variance-minimizing blend attains the target quality in fewer steps because the local score estimates have lower risk at the points where they are consumed. Beyond OU, we formulate the same viewpoint for *affine diffusions* via gradient–semigroup commutation; OU serves as the canonical worked example.

Semigroup structure and theory. A growing body of theory emphasizes semigroup and commutation properties for score flows. For instance, Malliavin–Gamma calculus tools clarify gradient–semigroup commutation and the correctness of reverse-time drifts in infinite-dimensional settings [19]. Our results are finite-dimensional and estimator-centric: we exploit semigroup transport to relate scores across time (CSEM) and to justify variance-optimal blending, with exact results in the linear–Gaussian case and heuristic extensions under single-basin posteriors for more general affine diffusions.

Kernel-based score estimation and samplers. Nonparametric score estimation has deep connections to RKHS methods, including kernel exponential families and kernelized score matching [20–24]. Kernelized samplers such as SVGD and KSD/MMD flows transport particles using functionals of the *target* score or discrepancy [25–29]. Our contribution is different in aim and usage: we do not fit a parametric density nor assume access to the target score. Instead, we construct *kernel-weighted, nonparametric* estimators of the *time-marginal* score that are PDE-exact for OU

(via Tweedie and CSEM) and then combine them by variance-optimal blending. The benefit is statistical—lower risk at the query (y, t)—and thus portable across samplers.

Bayesian inverse problems with score priors. Score/diffusion priors are increasingly used for posterior inference in imaging and the sciences [30–32]. Our framework contributes a variance-aware, semigroup-grounded *estimator* that, via a simple likelihood tilt of SNIS weights, converts prior estimators into posterior ones without altering the reverse integrator. This keeps the efficiency gains of the blended score while changing only the weighting.

Scope and limitations. We present *exact* anti-correlation and identities in the linear–Gaussian (OU) setting and a *heuristic* extension under single-basin dominance for more general priors; the affine-diffusion formulation follows from gradient–semigroup commutation but lacks a universal closed form. Our guarantees are *variance-only* at fixed (y, t); we do not claim Loewner-order negativity of the full cross-covariance outside commuting cases. The method is estimator-centric rather than architecture-centric, and its empirical acceleration arises from lower local risk when plugged into standard few-step reverse integrators.

3 Theory: From Exact Identities to Optimal Estimators

3.1 Score-Based Sampling with the Ornstein-Uhlenbeck Process

In the following, we use the typical notation in that random variables are denoted by capital letters, while lowercase letters are for their values.

Score-based generative models first define a "forward process" that corrupts data with noise over a pseudo-time variable t. We focus on the Ornstein-Uhlenbeck (OU) process, defined by the Stochastic Differential Equation (SDE):

$$\mathrm{d}X_t = -X_t \mathrm{d}t + \sqrt{2}\mathrm{d}W_t, \quad X_0 \sim p_0 \tag{1}$$

where $X_0 := X_{t=0}$ is distributed by a data distribution p_0.

OU forward transition, kernel, and posterior. The OU SDE in Eq. (1) has the closed-form forward update

$$X_t \;=\; e^{-t} X_0 \;+\; \sigma_t\, \varepsilon, \qquad \sigma_t^2 \;=\; 1 - e^{-2t}, \quad \varepsilon \sim \mathcal{N}(0, I). \tag{2}$$

We denote the (Gaussian) **transition kernel** by

$$K_t(y \mid x) \;=\; \mathcal{N}\big(y;\, e^{-t}x,\, \sigma_t^2 I\big). \tag{3}$$

The time-t marginal is then the convolution

$$p_t(y) \;=\; \int K_t(y \mid x)\, p_0(x)\, dx, \tag{4}$$

and the corresponding **OU posterior** over initial states is

$$\pi_t(x \mid y) \;=\; \frac{p_0(x)\, K_t(y \mid x)}{p_t(y)}. \tag{5}$$

In particular, for any test function f we have $\mathbb{E}[f(X_0) \mid X_t{=}y] = \int f(x)\, \pi_t(x \mid y)\, dx$. We will use the shorthand "$\pi_t(x_0 \mid y)$" throughout to denote the OU posterior (also listed in the notation table).

As t increases, the distribution of X_t, denoted by $p_t(x)$, smoothly approaches a standard normal distribution. The generative task is to reverse this process. This is possible by solving the corresponding time-reversal SDE:

$$\mathrm{d}X_t = [X_t + 2s(X_t, t)]\mathrm{d}t + \sqrt{2}\mathrm{d}\bar{W}_t \tag{6}$$

where $\mathrm{d}t$ is a positive time step for the backward process and $s(x,t) = \nabla_x \log p_t(x)$ is the **score function**. If we can accurately estimate the score function, we can reverse the diffusion to generate new data. This is the premise of all Denoising Score Matching (DSM) Generative models.

3.2 The Tweedie Identity and Denoising Score Matching

A foundational result, Tweedie's formula [33, 34], provides an exact expression for the OU score function in terms of a conditional expectation over the initial data:

$$s(y,t) = -\frac{1}{1-e^{-2t}}\mathbb{E}[y - e^{-t}X_0 | X_t = y],$$

The conditional expectation is taken with respect to the OU posterior $\pi_t(x_0 \mid y)$ defined in Eq. (5). Given a reference set of particles $\{x_i\}_{i=1}^N \sim p_0$, we can form a nonparametric Tweedie estimator for the score using self-normalized importance sampling (SNIS) [35, 36] as

$$\hat{s}_{\mathrm{TWD}}(y,t) = -\frac{1}{1-e^{-2t}}\sum_{i=1}^{N}\tilde{w}_i(y,t)(y - e^{-t}x_i), \tag{7}$$

where the normalized weights $\tilde{w}_i$ are derived from the OU transition kernel and are given as

$$w_i(y,t) \propto \exp\left(-\frac{\|y - e^{-t}x_i\|^2}{2(1-e^{-2t})}\right), \quad \tilde{w}_i = \frac{w_i}{\sum_{j=1}^N w_j}.$$

(Equivalently, $\tilde{w}_i$ are the normalized importance weights induced by $\pi_t(\cdot \mid y)$ in Eq. (5).) The Tweedie estimation in Eq. (7) motivates a direct training objective for a score neural network $s_\theta(x,t)$, with θ denoting weights and biases of the neural network, often called Score Matching [20, 33, 37]. In particular, the population loss minimizes the mean squared error between the network's score prediction and the score target derived from a data sample x_0:

$$\mathcal{L}_{\mathrm{TW}} = \mathbb{E}_{t,x_0}\left[\left\| s_\theta(x_t,t) - \left(-\frac{x_t - e^{-t}x_0}{1-e^{-2t}}\right)\right\|^2\right] \tag{8}$$

where $x_t = e^{-t}x_0 + \sqrt{1-e^{-2t}}\varepsilon$, a single step sample path from x_0 with ε as a Gaussian sample [38–41]. Note that for numerical stability, this objective is often implemented by reparameterizing the network to predict the noise ε directly [1]. This reparmatrization yields the standard loss for denoising score matching (DSM) [37].

3.3 The CSEM Identity: A Foundational View of Score Dynamics

Most existing diffusion models rely on the score matching loss (Eq. (8)) and often overlook the underlying dynamics of the score function itself. Instead of approximating the score's evolution numerically, our work is founded on an exact analytical solution that describes its finite-time evolution, which we introduce for the first time in the following lemma (the proof for a general case is presented in Appendix A).

Lemma 3.1 (Conditional Score Expectation Matching (CSEM) identity)**.** *Let $X_0 \sim p_0$, $\varepsilon \sim \mathcal{N}(0, I_d)$ be independent, and for $t > 0$ the exact sample path of the OU process (Eq. (1)) be*

$$X_t = e^{-t} X_0 + \sigma_t \varepsilon, \qquad \sigma_t^2 = 1 - e^{-2t}.$$

*Denote p_t for the probability density of X_t, $s_0(x) = \nabla \log p_0(x)$, and $s(y,t) = \nabla_y \log p_t(y)$. Then the **Conditional Score Expectation Matching (CSEM)** identity*

$$\boxed{s(y,t) = e^t \mathbb{E}[s_0(X_0) | X_t = y]} \tag{9}$$

holds.

This CSEM identity, new to the score-based modeling literature, follows directly from the gradient-semigroup commutation property of the OU process. Its significance is threefold. First, it is an **exact, non-asymptotic** relationship, holding for any finite $t > 0$. Second, it is **non-parametric**, connecting the score field directly to the initial data without requiring a functional ansatz. Finally, it provides the crucial bridge from the complex score evolution in a fixed Eulerian frame to a much simpler, linear evolution in a **Lagrangian frame**. This shift in perspective is the foundation of the **Lagrangian Score Formalism** that we propose in this paper.

Instead of viewing the score as a static field $s(x,t)$ to be solved at all points (x,t), the CSEM identity Eq. (9) allows us to see it as a dynamic object attached to each particle. In this particle frame, the complex score dynamics simplify dramatically: the score is propagated forward in time by exponential growth times the conditional expectation of the initial score given the particle's future location. The challenge of score estimation is thus transformed to evaluating this conditional expectation.

This identity provides a powerful recipe for score estimation. Given a set of reference particles at initial time $\{x_i, s_0(x_i)\}_{i=1}^N$, we can form a corresponding non-parametric **CSEM estimator** by replacing the conditional expectation with a self-normalized importance sampling (SNIS) average:

$$\hat{s}_{\mathrm{CSEM}}(y,t) = e^t \sum_{i=1}^N \tilde{w}_i(y,t) s_0(x_i). \tag{10}$$

This CSEM estimator for $s(y,t)$ is applicable when the initial score $s_0(x) = \nabla_x \log p_0(x)$ is either known analytically or can be well-approximated, and is sufficiently regular such that the estimator variance is controlled. This setting describes many problems in scientific computing. For instance, in molecular dynamics, s_0 can be computed from a known potential function [42, 43], and in PDE-constrained inverse problems, it can be computed efficiently using adjoint methods [44–52]. However, existing diffusion-based approaches to these problems have generally not leveraged this readily available information.

Unifying Principle: Gradient-Semigroup Commutation. This connection is a consequence of a principle known as **Gradient-Semigroup Commutation (GSC)**. Let P_t be the semigroup operator that evolves the initial density, $p_t = P_t p_0$. The score $s(y,t) = \nabla_y \log p_t(y)$ can be seen from two equivalent viewpoints. The Tweedie perspective can be thought of as first annealing the *measure* and then taking the gradient: $\nabla_y \log(P_t p_0)$. In contrast, the CSEM identity is equivalent to first finding the initial score (a vector field, or *flow*) $s_0 = \nabla_x \log p_0$, and then annealing this flow. The GSC principle is precisely the statement that these two operations commute up to a known factor:

$$\nabla_y \log(P_t p_0) = e^t \mathbb{E}[\nabla_x \log p_0(X_0) \mid X_t = y].$$

In short, blurring the density then taking the gradient is equivalent to taking the gradient then blurring the resulting vector field. This duality provides the theoretical underpinning for both the Tweedie and CSEM estimators. The CSEM identity extends from the OU process to **all affine SDEs** with time-varying linear drift and diffusion; see Appendix A for a complete statement, proof, and worked examples for common SDEs used in generative modeling.

3.4 Optimal Blending of Complementary Estimators

The Tweedie estimator Eq. (7) and the CSEM estimator Eq. (10) converge to the same true score but have two important complementary finite-sample properties. In particular, Section 3.4.1 discusses their opposite growth and decay, and Section 3.4.2 shows that they are negatively correlated. In Section 3.4.3, we exploit the negative correlation to provide a variance-minimal optimal convex blending of the two estimator

3.4.1 Opposite growth and decay of the two estimators

Their Monte Carlo variances scale differently with t, in a complementary fashion. Indeed, from Eq. (10) and Eq. (7) it is easy to see that

$$\mathrm{Var}[\hat{s}_{\mathrm{CSEM}}] \propto \frac{e^{2t}}{N}, \quad \mathrm{Var}[\hat{s}_{\mathrm{TWD}}] \propto \frac{1}{N(1-e^{-2t})^2},$$

and these scalings are illustrated in Fig. 1 as a function of time. As can be seen, the Tweedie variance grows exponentially as $t \to 0$, while the CSEM counterpart does so as t increases. We shall exploit this complementary behavior to provide an optimal blend score estimator that is much more accurate than either of them.

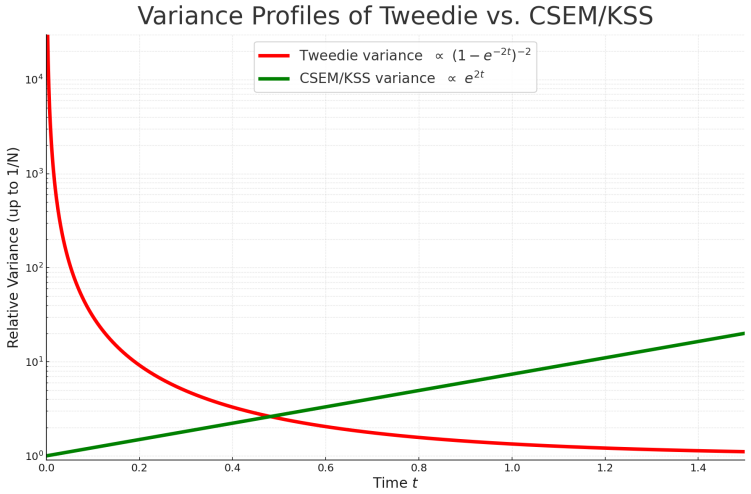

Figure 1: Relative variance of the Tweedie and CSEM non-parametric score estimators as a function of time t. The former has low variance at large t but diverges at $t = 0$, while the latter has low variance at small t but grows exponentially.

3.4.2 Negative correlation of the two estimators

We formalize when and why the Monte Carlo errors of the CSEM and Tweedie estimators are negatively aligned. In the *linear–Gaussian* case this anti-correlation is *exact* and purely algebraic;

beyond Gaussian settings we state a *heuristic* implication of OU single–basin dominance and defer all details to Appendix §B.

Proposition 3.2 (Gaussian case: exact anti-correlation)**.** *Assume* $p_0 = \mathcal{N}(\mu, \Sigma)$ *with* $\Sigma \succ 0$*. For a given* (y, t)*, let*

$$\widehat{A} := \sum_{i=1}^{N} \tilde{w}_i \, x_i, \qquad A^\star := \mathbb{E}[X_0 \mid X_t = y], \qquad \Delta := \widehat{A} - A^\star.$$

Let $\widehat{s}_{\mathrm{CSEM}}(y,t)$ *and* $\widehat{s}_{\mathrm{TWD}}(y,t)$ *be the nonparametric CSEM and Tweedie estimators, and let* $s(y,t)$ *denote the true time-*t *score. Their errors*

$$\varepsilon_{\mathrm{C}} := \widehat{s}_{\mathrm{CSEM}} - s(y,t), \qquad \varepsilon_{\mathrm{T}} := \widehat{s}_{\mathrm{TWD}} - s(y,t)$$

satisfy the deterministic *identities*

$$\varepsilon_{\mathrm{C}} = -e^{t} \, \Sigma^{-1} \, \Delta, \qquad \varepsilon_{\mathrm{T}} = \frac{e^{-t}}{1 - e^{-2t}} \, \Delta, \tag{11}$$

and hence the inner product is deterministically nonpositive:

$$\varepsilon_{\mathrm{T}}^{\top} \varepsilon_{\mathrm{C}} = -\frac{1}{1 - e^{-2t}} \, \Delta^{\top} \Sigma^{-1} \Delta \leq 0, \tag{12}$$

with equality iff $\Delta = 0$*. Taking expectations yields*

$$\mathbb{E}\big[\varepsilon_{\mathrm{T}}^{\top} \varepsilon_{\mathrm{C}}\big] = -\frac{1}{1 - e^{-2t}} \, \mathbb{E}\big[\Delta^{\top} \Sigma^{-1} \Delta\big] \leq 0, \qquad \operatorname{tr} \operatorname{Cov}(\varepsilon_{\mathrm{C}}, \varepsilon_{\mathrm{T}}) = -\frac{1}{1 - e^{-2t}} \, \mathbb{E}\big[\Delta^{\top} \Sigma^{-1} \Delta\big] \leq 0. \tag{13}$$

Caveat. We do *not* claim a general Loewner–order statement $\operatorname{Cov}(\varepsilon_{\mathrm{C}}, \varepsilon_{\mathrm{T}}) \preceq 0$ without a commutation/simultaneous–diagonalization condition; see the note in Appendix §B. The scalar statement (12)–(13) provides a theoretical justification explaining why the blending approach works well.

Tan2Tan *Revise heuristic section when done with other parts of the manuscript*

Beyond Gaussian (heuristic, single–basin regime). When the OU posterior $\pi_t(x \mid y)$ in §5 is dominated by a single local basin, a second–order (local quadratic) reduction of $\log p_0$ implies that the leading SNIS fluctuations of the CSEM and Tweedie estimators are approximately opposite along the same random direction, yielding a *negative expected inner product* $\mathbb{E}\big[\varepsilon_{\mathrm{T}}^{\top} \varepsilon_{\mathrm{C}}\big] \lesssim 0$. Assumptions, the local quadratic reduction, and scope/limitations are detailed in Appendix §B. Figure 2 numerically confirms that this negative correlation persists beyond the linear–Gaussian case.

Summary. In the linear–Gaussian case the anti-alignment (12) holds pathwise for every realized SNIS fluctuation Δ; taking expectations merely re-expresses it at trace level. Beyond Gaussian, our claim is restricted to *expected* inner products under single-basin conditions.

3.4.3 Optimal blending as variance minimization

Given the complementary growth/decay of the variance profiles and, more importantly, the negative correlation of the CSEM Eq. (10) and Tweedie Eq. (7) estimators, we propose to blend them in a convex fashion:

$$\hat{s}_{\mathrm{BLEND}}(\lambda) = \lambda \, \hat{s}_{\mathrm{TWD}} + (1 - \lambda) \, \hat{s}_{\mathrm{CSEM}}. \tag{14}$$

The question is how to choose λ so that the blend $\hat{s}_{\text{BLEND}}(\lambda)$ retains the best features of both estimators and improves accuracy. Since the variances and correlation depend on (y,t) (Figs. 1 and 2), we choose $\lambda(y,t)$ to minimize the (conditional) variance of the blend at (y,t):

$$\lambda^*(y,t) \;\in\; \arg\min_{\lambda\in\mathbb{R}} J(\lambda;y,t), \qquad J(\lambda;y,t) := \mathbb{E}\Big[\big\|\lambda\,\varepsilon_T + (1-\lambda)\,\varepsilon_C\big\|^2\Big],$$

where $\varepsilon_T := \hat{s}_{\text{TWD}} - s$ and $\varepsilon_C := \hat{s}_{\text{CSEM}} - s$. Define

$$V_T := \mathbb{E}\|\varepsilon_T\|^2, \qquad V_C := \mathbb{E}\|\varepsilon_C\|^2, \qquad C := \mathbb{E}\langle\varepsilon_C, \varepsilon_T\rangle.$$

Then

$$J(\lambda) = \lambda^2 V_T + (1-\lambda)^2 V_C + 2\lambda(1-\lambda)C,$$

with curvature $V_T + V_C - 2C = \mathbb{E}\|\,\varepsilon_T - \varepsilon_C\,\|^2 \geq 0$. Whenever $V_T + V_C - 2C > 0$, the unique minimizer is

$$\boxed{\lambda^*(y,t) = \frac{V_C - C}{V_C + V_T - 2C}}, \qquad \lambda^*_{\text{clip}} = \min\{1, \max\{0, \lambda^*\}\}. \tag{15}$$

Optimal risk and guarantees (theory). Write $a := V_T$, $b := V_C$, and $c := C$. At the optimum,

$$J(\lambda^*) \;=\; \frac{ab - c^2}{a + b - 2c}. \tag{16}$$

Two immediate consequences (used later) are:

$$J(\lambda^*) \;\leq\; \min\{a,b\}, \qquad a - J(\lambda^*) = \frac{(a-c)^2}{a+b-2c}, \quad b - J(\lambda^*) = \frac{(b-c)^2}{a+b-2c}. \tag{17}$$

Hence the optimal blend is never worse (in MSE) than the better constituent as long as $a+b-2c > 0$, with strict improvement unless $c = a$ or $c = b$ (perfect alignment with one estimator). Moreover, the weight is interior iff

$$0 < \lambda^* < 1 \quad\Longleftrightarrow\quad c < \min\{a,b\}, \tag{18}$$

which holds whenever $c \leq 0$ (since $a, b > 0$). Equivalently, writing $a = \sigma_T^2$, $b = \sigma_C^2$, and $c = \rho\,\sigma_T\sigma_C$,

$$\lambda^* = \frac{\sigma_C^2 - \rho\,\sigma_T\sigma_C}{\sigma_T^2 + \sigma_C^2 - 2\rho\,\sigma_T\sigma_C}, \qquad J(\lambda^*) = \frac{\sigma_T^2\,\sigma_C^2\,(1-\rho^2)}{\sigma_T^2 + \sigma_C^2 - 2\rho\,\sigma_T\sigma_C}. \tag{19}$$

On variance vs. MSE. If both estimators are unbiased for s (or share the same bias vector so that $\mathbb{E}[\varepsilon_T] = \mathbb{E}[\varepsilon_C]$), minimizing $J(\lambda)$ is equivalent to minimizing the MSE of $\hat{s}_{\text{BLEND}}(\lambda)$. We adopt this variance-only rule in our claims; bias-aware corrections are omitted.

SNIS plug-in (definitions only). With SNIS weights $\tilde{w}_i$ targeting $\pi_t(\cdot \mid y)$, define

$$a_i := e^t s_0(x_i), \qquad b_i := -\frac{1}{1-e^{-2t}}\,(y - e^{-t}x_i), \qquad \hat{s}_{\text{CSEM}} = \sum_i \tilde{w}_i a_i, \quad \hat{s}_{\text{TWD}} = \sum_i \tilde{w}_i b_i,$$

and centered contributions $\delta a_i = a_i - \hat{s}_{\text{CSEM}}$, $\delta b_i = b_i - \hat{s}_{\text{TWD}}$. The standard SNIS plug-in estimates

$$\hat{V}_C = \frac{\sum_i \tilde{w}_i^2 \|\delta a_i\|^2}{1 - \sum_i \tilde{w}_i^2}, \quad \hat{V}_T = \frac{\sum_i \tilde{w}_i^2 \|\delta b_i\|^2}{1 - \sum_i \tilde{w}_i^2}, \quad \hat{C} = \frac{\sum_i \tilde{w}_i^2 \langle\delta a_i, \delta b_i\rangle}{1 - \sum_i \tilde{w}_i^2},$$

Plugging these into Eq. (15) yields the approximate blend weight $\widehat{\lambda^*}\,(y,t)$, which forms the core of our non-parametric sampling Algorithm 1.

Algorithm 1 Reverse Sampling optimal blend score

Input: Initial sampling particles $\{y_j(T)\}_{j=1}^M \sim \mathcal{N}(0, I_d)$, time grid $T = t_K > \cdots > t_0 = 0$, reference data $\{x_i, s_0(x_i)\}_{i=1}^N$, with $x_i \sim p_0$.
for $k = K-1, \ldots, 0$ **do**
  Let current time be t_{k+1} and target time be t_k.
  for $j = 1, \ldots, M$ **do**
    Compute the optimal score $\hat{s}_{\text{BLEND}}\left(\widehat{\lambda^*}\left(y_j(t_{k+1}), t_{k+1}\right)\right)$ in Eq. (14) for particle y_j.
    Update particle $y_j(t_k)$ using chosen SDE integrator with $\hat{s}_{\text{BLEND}}\left(\widehat{\lambda^*}\left(y_j(t_{k+1})\right)\right)$.
  end for
end for
Output: Final samples $\{y_j(0)\}_{j=1}^M$.

Parametric Extension (see Appendix C). While our main results focus on the nonparametric blended estimator, we also provide a proof-of-concept parametric extension in Appendix C. A critic–gate architecture distills the blended score into a single neural network, removing the need for a reference set at inference. This establishes feasibility and conceptual priority for amortized variants, while scaling to large data is deferred to future work.

3.5 A Learned Proxy for the Initial Score

The CSEM estimator (Section 1.3) provides a low-variance signal at small diffusion times but requires access to the initial score $s_0(x) = \nabla_x \log p_0(x)$ at data points. In many settings we only have a reference set $X = \{x_i\}_{i=1}^N \sim p_0$. We therefore construct a *learned local score proxy* $\hat{s}_0(x)$ directly from data by (i) fitting local Gaussian approximations around anchor points via kNN with adaptive bandwidths, and (ii) *recomputing* a compact Gaussian–mixture score in the query's neighborhood to capture curvature and multimodality. The goal is geometric rather than global density fitting: we estimate *local tangent gradients* needed by CSEM at small t, complementary to the coarse transport signal provided by Tweedie at larger t.

3.5.1 Local Gaussian approximation from kNN (general scheme)

For each anchor x_i, let $\mathcal{N}_k(i)$ be the indices of its k nearest neighbors under the ambient metric. Define an adaptive bandwidth as the squared distance to the k-th neighbor:

$$h_i^2 \;=\; \max_{j \in \mathcal{N}_k(i)} \|x_i - x_j\|^2,$$

and weights

$$w_{ij} \;\propto\; \exp\Big(-\frac{\|x_i - x_j\|^2}{2h_i^2}\Big), \qquad j \in \mathcal{N}_k(i).$$

The locally weighted mean is

$$\mu_i \;=\; \frac{\sum_{j \in \mathcal{N}_k(i)} w_{ij} x_j}{\sum_{j \in \mathcal{N}_k(i)} w_{ij}}.$$

We then choose a structured estimate for the local covariance Σ_i (two options below) and define the anchor score proxy as the exact Gaussian score

$$\hat{s}_0(x_i) \;=\; \Sigma_i^{-1}\,(\mu_i - x_i). \tag{20}$$

Diagonal covariance (compute-robust baseline; kept for scaling and NN supervision). We retain a diagonal proxy because it is *computationally light*, *memory-efficient*, and *statistically stable* when d and N_{ref} are large—precisely the regime implicated by neural distillation (cf. App. C). Empirically, in the large-data limit relevant for training networks, the diagonal proxy can outperform hand-tuned low-rank SVD scores as a teacher signal (see App. C), even though it is not our strongest non-parametric estimator at moderate N_{ref} (where LR+D excels).
The construction estimates per-coordinate variances with a local ridge:

$$\mathrm{var}_i^{\ell} \;=\; \frac{\sum_{j\in\mathcal{N}_k(i)} w_{ij}\,\big(x_j^{\ell}-\mu_i^{\ell}\big)^2}{\sum_{j\in\mathcal{N}_k(i)} w_{ij}}, \qquad \tau_i \;=\; \gamma\cdot\frac{1}{d}\sum_{\ell=1}^{d}\mathrm{var}_i^{\ell},$$

$$\tilde{\Sigma}_i^{\mathrm{diag}} \;=\; \mathrm{diag}\big(\mathrm{var}_i^1+\tau_i,\dots,\mathrm{var}_i^d+\tau_i\big), \qquad \hat{s}_0^{\mathrm{diag}}(x_i) \;=\; \big(\tilde{\Sigma}_i^{\mathrm{diag}}\big)^{-1}(\mu_i-x_i).$$

Complexity (per anchor): diagonal moments are $O(kd)$ time and $O(d)$ memory; the inverse is trivial. This scaling makes diagonal proxies suitable for very large N_{ref} and high-dimensional latents where repeated fits are required for NN training.

Low-rank SVD + diagonal tail (LR+D; primary for non-parametric evaluation). Form weighted neighbor residuals relative to μ_i and compute a rank-r SVD to estimate the principal subspace and its energies; model discarded energy by a *diagonal tail*. The covariance model is

$$\Sigma_i^{\mathrm{LR+D}} \;=\; V_i\Lambda_i V_i^{\top} \;+\; \mathrm{diag}\big(\tau_{i,1},\dots,\tau_{i,d}\big),$$

where $V_i\in\mathbb{R}^{d\times r}$ contains top singular directions, $\Lambda_i=\mathrm{diag}(\lambda_{i,1},\dots,\lambda_{i,r})$ the corresponding variances, and $(\tau_{i,\ell})$ collects per-coordinate tail energy (with small floors/clipping for stability).

The proxy is

$$\hat{s}_0^{\mathrm{LR+D}}(x_i) \;=\; \big(\Sigma_i^{\mathrm{LR+D}}\big)^{-1}(\mu_i-x_i),$$

computed efficiently via Woodbury. *Complexity (per anchor):* $O(kdr)$ time for subspace estimation (or $O(kd\min\{d,k\})$ if a dense SVD is used), memory $O(dr)$. This choice captures anisotropic curvature along data-tangent directions and is our *primary* proxy for the moderate N_{ref} non-parametric experiments in the main text.

Algorithm 2 Local Gaussian proxy at anchors (diagonal or LR+D)

Input: $X=\{x_i\}_{i=1}^N\subset\mathbb{R}^d$, neighbor count k, ridge/regularization hyperparameters; rank r for LR+D.
for $i=1,\dots,N$ **do**
  Find $\mathcal{N}_k(i)$ (k-NN of x_i); set $h_i^2\leftarrow\max_{j\in\mathcal{N}_k(i)}\|x_i-x_j\|^2$ and $w_{ij}\propto\exp(-\|x_i-x_j\|^2/(2h_i^2))$.
  Compute $\mu_i\leftarrow\big(\sum_{j\in\mathcal{N}_k(i)} w_{ij}x_j\big)/\big(\sum_{j\in\mathcal{N}_k(i)} w_{ij}\big)$.
  If mode="diag": form $\tilde{\Sigma}_i^{\mathrm{diag}}$ with per-coordinate variances + ridge; set $\Sigma_i\leftarrow\tilde{\Sigma}_i^{\mathrm{diag}}$.
  If mode="LR+D": compute rank-r SVD subspace (V_i,Λ_i) and diagonal tail $(\tau_{i,\ell})$; set $\Sigma_i\leftarrow\Sigma_i^{\mathrm{LR+D}}$.
  Store (μ_i,Σ_i) and anchor score $\hat{s}_0(x_i)=\Sigma_i^{-1}(\mu_i-x_i)$.
end for
Output: $\{(\mu_i,\Sigma_i)\}_{i=1}^N$ and $\{\hat{s}_0(x_i)\}_{i=1}^N$.

Positioning and reporting policy. *Main text:* we report LR+D as the primary proxy (strongest at moderate N_{ref}) and include the *diagonal* baseline for transparency and scale realism (lighter line style in plots). *Appendix:* we provide large-N_{ref} sweeps and NN-teacher ablations (critic–gate) where the diagonal proxy is competitive or superior as a supervision signal (App. C). This separation avoids conflating two regimes: (i) estimator strength at moderate N_{ref} (favoring LR+D) versus (ii) stability and throughput at very large N_{ref} required for parametric distillation (favoring diagonal).

3.5.2 Recompute as a compact Gaussian–mixture score (k-mix)

A single local Gaussian can be biased in regions of high curvature or at crossings. We therefore *recompute* the proxy score at a query x by treating its neighborhood as a *mixture of local Gaussians*:

$$\text{select } \{i_m\}_{m=1}^{M} \text{ as the } k_{\text{mix}} \text{ nearest anchors to } x, \qquad M = k_{\text{mix}} \ll N.$$

For each selected anchor i_m, use its parameters $(\mu_{i_m}, \Sigma_{i_m})$ from §3.5.1 (Diagonal *or* LR+D). Form

$$q(x) \;=\; \sum_{m=1}^{M} \pi_m \, \mathcal{N}\big(x \mid \mu_{i_m}, \Sigma_{i_m}\big),$$

with simple priors π_m (e.g., softmaxed proximity weights). The global mixture score is

$$\nabla_x \log q(x) \;=\; \sum_{m=1}^{M} \underbrace{\frac{\pi_m \, \mathcal{N}(x \,|\, \mu_{i_m}, \Sigma_{i_m})}{\sum_{j=1}^{M} \pi_j \, \mathcal{N}(x \,|\, \mu_{i_j}, \Sigma_{i_j})}}_{\tilde{w}_m(x)} \; \Sigma_{i_m}^{-1}\big(\mu_{i_m} - x\big), \tag{21}$$

evaluated numerically via log–sum–exp for stability. This k-mix recomputation explicitly incorporates local multimodality and curvature while keeping compute bounded. In practice it is *crucial* for strong performance with LR+D in thin tubes and at manifold crossings.

Algorithm 3 Recompute (k-mix) mixture score at query x

Input: query x, anchor params $\{(\mu_i, \Sigma_i)\}_{i=1}^{N}$, k_{mix}.
Find indices of the k_{mix} nearest anchors to x: $\{i_m\}_{m=1}^{M}$.
for $m = 1, \dots, M$ **do**
    $\ell_m \leftarrow \log \pi_m \;-\; \frac{1}{2}\big[(x - \mu_{i_m})^\top \Sigma_{i_m}^{-1}(x - \mu_{i_m}) + \log\det(2\pi\Sigma_{i_m})\big]$.
end for
$a \leftarrow \max_m \ell_m$; $\tilde{w}_m \leftarrow \exp(\ell_m - a) \Big/ \sum_{j=1}^{M} \exp(\ell_j - a)$.
Return $\hat{s}_0^{\text{recomp}}(x) \leftarrow \sum_{m=1}^{M} \tilde{w}_m \, \Sigma_{i_m}^{-1}(\mu_{i_m} - x)$.

Remark. The k-mix recomputation in (21) accepts either diagonal or LR+D anchors. Even with diagonal anchors, recomputation substantially mitigates single-Gaussian bias in high-curvature regions, while remaining $O(k_{\text{mix}} d)$ per query.

3.5.3 Blending with the learned proxy

With $\hat{s}_0(x)$ given by the recomputed mixture score (default) or the single–component proxy, we proceed exactly as in Section 3.4.3: substitute $\hat{s}_0$ for the CSEM term while Tweedie remains unchanged. The variance–minimizing blend then arbitrates automatically across t, leveraging CSEM's local geometry at small t and Tweedie's coarse transport at large t.

3.5.4 Convergence (summary) and implications

Under standard smoothness/positivity assumptions and classical kNN bandwidth scaling ($k\to\infty$, $k/N\to 0$), the single–component local Gaussian proxy is a consistent estimator of $s_0(x)$. A textbook bias–variance tradeoff yields

$$\Big(\mathbb{E}\,\|\,\hat{s}_0(x) - s_0(x)\|_2^2\Big)^{1/2} \;=\; \mathcal{O}\big(N^{-\frac{2}{d+4}}\big) \qquad \text{for} \qquad k \asymp N^{\frac{4}{d+4}},$$

up to curvature–dependent constants and the chosen covariance structure. The LR+D choice reduces bias in anisotropic neighborhoods; the k-mix recomputation further mitigates single–mode bias near crossings by recovering the mixture score (21). Because $\|\hat{s}_0 - s_0\|_2 \to 0$ as $N\to\infty$ (under mild neighborhood–shrinkage for the mixture), the CSEM term built from $\hat{s}_0$ remains consistent at small diffusion times, and the blended estimator inherits the ground truth scores behavior at scale.

3.6 Application to Bayesian Inverse Problems

We adapt our framework to posterior sampling in inverse problems. Given a prior $p_0(x)$, likelihood $L(y_{\text{obs}} \mid x)$, and observation y_{obs}, the posterior is

$$\pi(x) \;\propto\; p_0(x)\, L(y_{\text{obs}} \mid x). \tag{22}$$

As is standard in inverse problems [53, 54], it is typically easy to sample the prior p_0 but not the posterior. We therefore reuse the same prior reference set $\{x_i\}_{i=1}^N$ and *tilt* weights by the likelihood.

Assumption (likelihood independent of diffusion noise). The observation y_{obs} depends on the unknown X_0 but is independent of the forward OU corruption noise. This includes the common linear–Gaussian case $y_{\text{obs}} = HX_0 + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \Sigma_y)$, and more generally any $L(y_{\text{obs}} \mid x_0)$ that does not involve the OU noise used to generate X_t.

OU kernel notation (as in §2). We write the OU transition kernel as $K_t(y \mid x) = \mathcal{N}\big(y;\, e^{-t}x, \sigma_t^2 I\big)$ with $\sigma_t^2 = 1 - e^{-2t}$, and the t-marginal as $p_t(y) = \int K_t(y \mid x)\, p_0(x)\, dx$.

Proposition 3.3 (Weight–tilting suffices for posterior smoothing and scores)**.** *Under the assumption above, the posterior-smoothed density at time t is*

$$p_t^{post}(y) \;\propto\; \int K_t(y \mid x)\, p_0(x)\, L(y_{obs} \mid x)\, dx, \tag{23}$$

and any score identity that conditions on $X_t{=}y$ (e.g., Tweedie or CSEM/OU) transfers verbatim to the posterior by the single replacement

$$w_i \;\longmapsto\; w_i^{post} \;\propto\; w_i\, L(y_{obs} \mid x_i)$$

in the self-normalized importance weights. Concretely, the transition-normalized mixture weights become

$$\alpha_i(y, t; y_{obs}) \;=\; \frac{w_i\, L(y_{obs} \mid x_i)\, K_t(y \mid x_i)}{\sum_j w_j\, L(y_{obs} \mid x_j)\, K_t(y \mid x_j)}, \tag{24}$$

and for any kernelized score estimator of the form $\hat{s}_t(y) = \sum_i \alpha_i(y,t)\, \Psi_t(y, x_i)$, the posterior estimator is

$$\hat{s}_t^{post}(y) \;=\; \sum_{i=1}^{N} \alpha_i(y, t; y_{obs})\, \Psi_t(y, x_i). \tag{25}$$

Sketch. *Bayes' rule gives $p(x \mid X_t{=}y, y_{obs}) \propto p_0(x)\, L(y_{obs} \mid x)\, K_t(y \mid x)$; plug into the same conditional-expectation identity and approximate by SNIS using the prior reference set.*

Posterior score estimators (Tweedie, CSEM) and blend. Using (24) and the OU forms,

$$s_0^{\text{post}}(x) := \nabla_x \log \pi(x) \;=\; s_0(x) + \nabla_x \log L(y_{\text{obs}} \mid x),$$

$$\hat{s}^{\text{post}}_{\text{TWD}}(y,t) = -\frac{1}{1-e^{-2t}} \sum_{i=1}^{N} \alpha_i(y,t;y_{\text{obs}}) \left(y - e^{-t} x_i\right),$$

$$\hat{s}^{\text{post}}_{\text{CSEM}}(y,t) = e^{t} \sum_{i=1}^{N} \alpha_i(y,t;y_{\text{obs}})\, s_0^{\text{post}}(x_i),$$

and the SNIS variance-minimizing blend weight computed with the same $\alpha_i(y,t;y_{\text{obs}})$ yields

$$\hat{s}^{\text{post}}_{\text{BLEND}}(y,t) = (1-\lambda^{\text{post}}_{\text{snis}})\, \hat{s}^{\text{post}}_{\text{CSEM}}(y,t) + \lambda^{\text{post}}_{\text{snis}}\, \hat{s}^{\text{post}}_{\text{TWD}}(y,t).$$

The anti-correlation mechanism persists after tilting (empirically strongest at intermediate t as in Fig. 2); theoretically the single-basin heuristic carries over with s_0 replaced by s_0^{post}.

Practical note. Computing $s_0^{\text{post}}(x_i)$ only requires $\nabla_x \log L(y_{\text{obs}} \mid x_i)$, available by adjoints in the linear–Gaussian case and by automatic differentiation in general. All other components—per-particle signals, plug-in variances, and the Heun PC integrator—remain unchanged.

4 Results

We present numerical experiments designed to validate the theoretical claims of our framework. We begin with *toy* targets built from low-dimensional Gaussian mixtures embedded in a higher-dimensional ambient space. A concise description of the construction appears below; full details and exact hyperparameters are deferred to App. D. These experiments isolate estimator quality while controlling geometry and conditioning. We then illustrate a practical downstream task (MNIST deblurring). The goal throughout is not to maximize benchmark scores via aggressive tuning, but to demonstrate that our *blended* score estimator yields a statistically stronger supervision signal and better sampling behavior than Tweedie alone.

Experimental setup (toy GMMs). We generate targets using the utilities `get_gmm_funcs` and `run_comparison` in our released notebook (`non_param_sampler.ipynb`). Unless stated otherwise, **all quantitative loss curves in this section are computed on the *6D helix GMM*** shown in Fig. 3, which realizes a smooth, low-dimensional manifold with anisotropic, locally aligned component covariances. **Integrator.** In *all* sampling tests (toy and MNIST), we integrate the reverse-time dynamics with the *second–order Heun predictor–corrector (PC) solver* (standard choice): each step uses an explicit Euler *predictor* followed by a slope-averaged *corrector*. The same time grid and solver are used for Tweedie, CSEM-only, and Blend to ensure comparability. Samplers are evaluated along a log-spaced diffusion time grid $t \in [t_{\min}, t_{\max}]$. For CSEM we estimate conditional expectations via SNIS with an ESS threshold (see footnote in §4); we report median-of-means over independent batches. The number of reference samples N_{ref} is varied to produce the curves in Figs. 6–5. Throughout the visualization figures, we denote by $d_1, d_2, \ldots$ the principal directions returned by PCA fitted to the relevant target (prior or posterior); 2D histograms are shown in the planes (d_i, d_j). Implementation details, including the exact helix parameterization, mixture weights (uniform), covariance construction (tangent/normal scaling), bandwidth grids, and the t-grid, are provided in App. D.

Anti-correlation across time. A central pillar of our theory is that the Monte Carlo errors of the CSEM and Tweedie estimators are *negatively correlated* (see Section 3.4.2). We now numerically verify this on the **6D helix GMM** in Fig. 2, which plots the correlation as a function of diffusion time t. Specifically, writing

$$\varepsilon_T(y,t) := \hat{s}_{\mathrm{TWD}}(y,t) - s(y,t), \qquad \varepsilon_C(y,t) := \hat{s}_{\mathrm{CSEM}}(y,t) - s(y,t),$$

We define the correlation coefficient between s_{TWD} and s_{CSEM} as

$$\rho(t) \;=\; \frac{\mathbb{E}_{y\sim p_t}\big[\langle \varepsilon_T(y,t),\, \varepsilon_C(y,t)\rangle\big]}{\sqrt{\mathbb{E}\|\varepsilon_T(y,t)\|^2}\,\sqrt{\mathbb{E}\|\varepsilon_C(y,t)\|^2}},$$

which can be computed using Monte Carlo over $y \sim p_t$. The *oracle* Tan2Alois *again, need to change oracle to ground truth everywhere including figures* curve uses the exact s_0 inside CSEM (and Tweedie where applicable), while the *proxy* curve replaces s_0 by the learned local-Gaussian proxy Tan2Alois *we have diagonal and LR proxies there. Which one you used for this? Need to describe* from Section 3.5; the definition of $\rho(t)$ always compares to the true $s(y,t)$ of the target.[1] A distinct sweet spot emerges around $t \approx 10^{-3}$ where anti-correlation is strongest: for very small t the ESS Tan2Alois *define ESS* collapses; for large t the conditional distribution Eq. (5) becomes multi-modal and the geometric anti-alignment weakens. This intermediate sweet spot region, close to single-basin geometry and high ESS, maximizing negative correlation is where the blending is most beneficial. While the proxy curve shows milder anti-correlation than the oracle, it remains sufficiently negative to enable substantial variance cancellation in the blended estimator (see also Section 3.4.2).

[1] We drop time points with low importance-sampling quality: $\mathrm{ESS} = 1/\sum_i \tilde{w}_i^2 < \tau_{\mathrm{ESS}}$. Tan2Alois *define and say what value of* τ_{ESS} *is.*

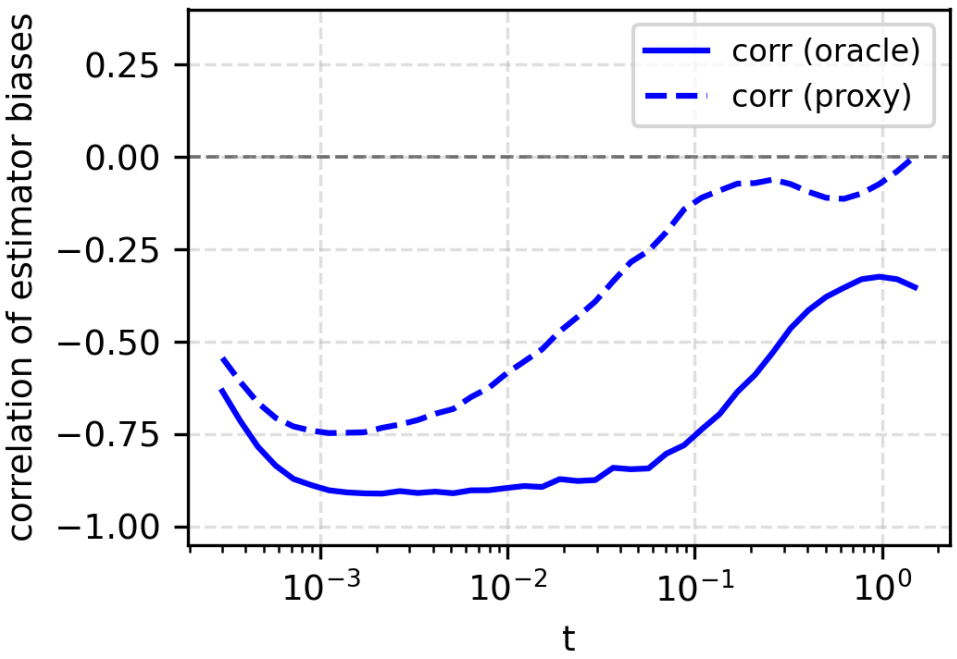

Figure 2: **Anti-correlation between CSEM estimator Eq. (10) and Tweedie estimator Eq. (7)** on the **6D helix GMM**. Negative correlation peaks near $t \approx 10^{-3}$, balancing high ESS with nearly single-basin geometry. The proxy curve (dashed) preserves the anti-correlation effect, enabling the blended estimator Eq. (14) to reduce variance.

Mass vs. gradient fidelity (6D helix GMM). To assess complementary aspects of sample quality on the **same 6D helix GMM**, we report three metrics:

- **Maximum Mean Discrepancy (MMD) with radial basis function kernel.** Let P and Q be two probability distribution and $k_\sigma(x, y) = \exp\big(-\|x-y\|^2/(2\sigma^2)\big)$. MMD is defined as

$$\mathrm{MMD}^2(P, Q) := \mathbb{E}_{x,x'\sim P}\big[k_\sigma(x, x')\big] - 2\,\mathbb{E}_{x\sim P, y\sim Q}\big[k_\sigma(x, y)\big] + \mathbb{E}_{y,y'\sim Q}\big[k_\sigma(y, y')\big],$$

  using a small bandwidth grid and median heuristic [55]. Tan2Alois *Need to define*

- **Kernel Stein Discrepancy (KSD) with inverse multiquadric kernel).** With $k(x, y) = (c^2 + \|x-y\|^2)^\beta$ for $\beta \in (-1, 0)$, the KSD for score s is given by

$$\mathrm{KSD}^2(Q, s) := \mathbb{E}_{y,y'\sim Q}\big[u_s(y, y')\big], \quad u_s(y, y') = \langle s(y), s(y')\rangle k + \langle s(y), \nabla_{y'} k\rangle + \langle s(y'), \nabla_y k\rangle + \operatorname{tr} \nabla_y \nabla_{y'} k,$$

  estimated by the U-statistic [26, 56].

- **Root Mean Square Error (RMSE).** We define $\mathrm{RMSE} := \big(\mathbb{E}_{y\sim p_t}\|\hat{s}(y, t) - s(y, t)\|^2\big)^{1/2}$, where the ground truth s available for the GMM, averaged over a log-spaced grid of t values on $[.0005, 1.5]$, coinciding with the t values queried during sampling.

Fig. 6, Fig. 4, and Fig. 5 show that the *blended* estimator consistently matches or improves global mass placement (MMD) while clearly outperforming Tweedie in gradient fidelity (KSD) and score accuracy (RMSE). In these figures, we refer to the *oracle floor* as the KSD/MMD obtained using samples from the actual ground truth target distribution. To keep the axes readable, we plot *Tweedie* and *Blend* in the main curves and provide a qualitative CSEM-only comparison in Fig. 3; full CSEM-only curves (log-scale and zoomed insets) are given in App. D for transparency.

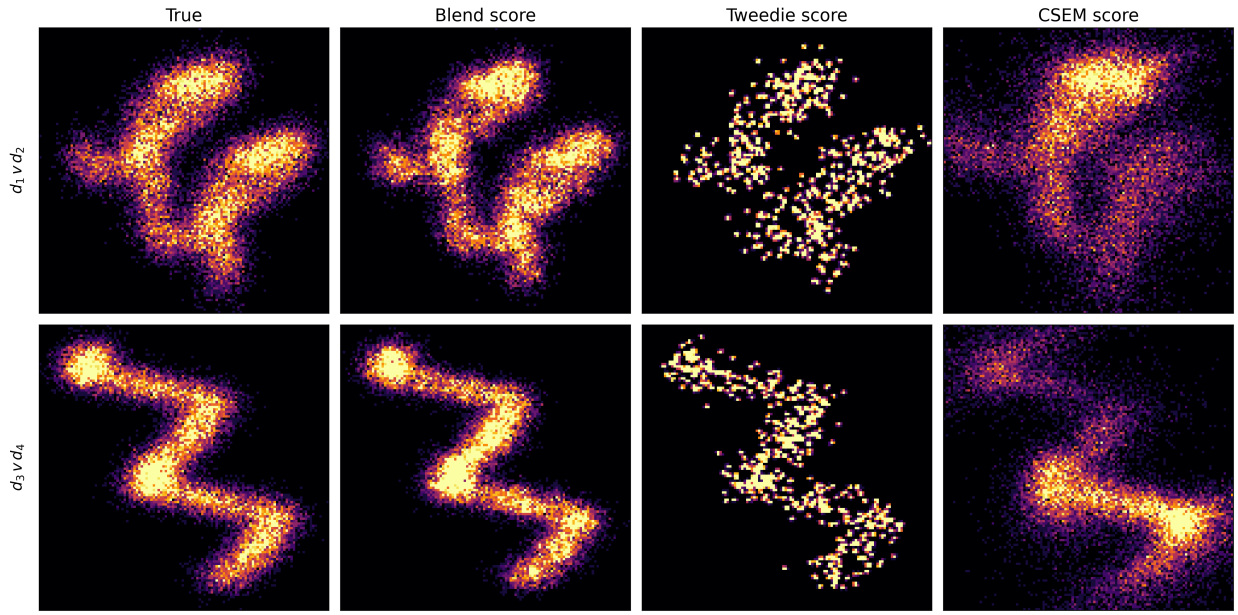

Figure 3: **Qualitative comparison on the *6D helix GMM*** (the same target used for the quantitative metrics in Figs. 6–5). Columns: *True*, *Blend*, *Tweedie*, *CSEM-only*; rows: selected 2D projections (d_1, d_2), (d_3, d_4), (d_5, d_6). CSEM-only is diffuse and not competitive as a stand-alone estimator, but its errors are complementary to Tweedie's; the variance-optimal blend leverages this to recover sharper structure.

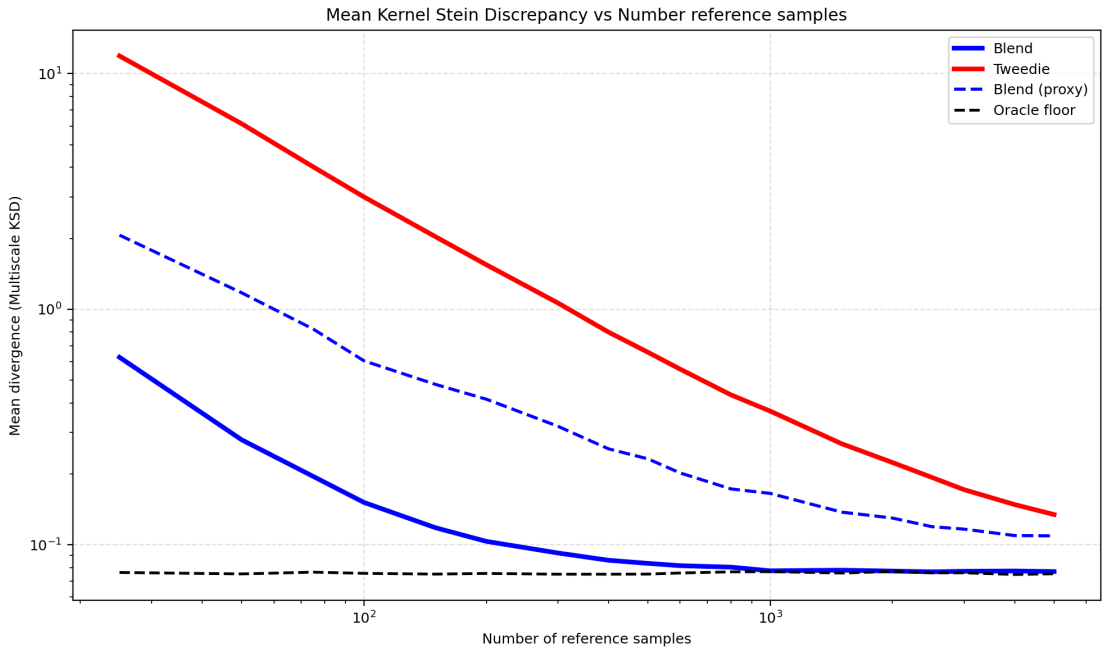

Figure 4: **KSD vs. number of references (lower is better)** on the **6D helix GMM**. The blended estimator rapidly approaches the oracle floor with only a few hundred references; Tweedie requires far more and does not reach the floor in this setting.

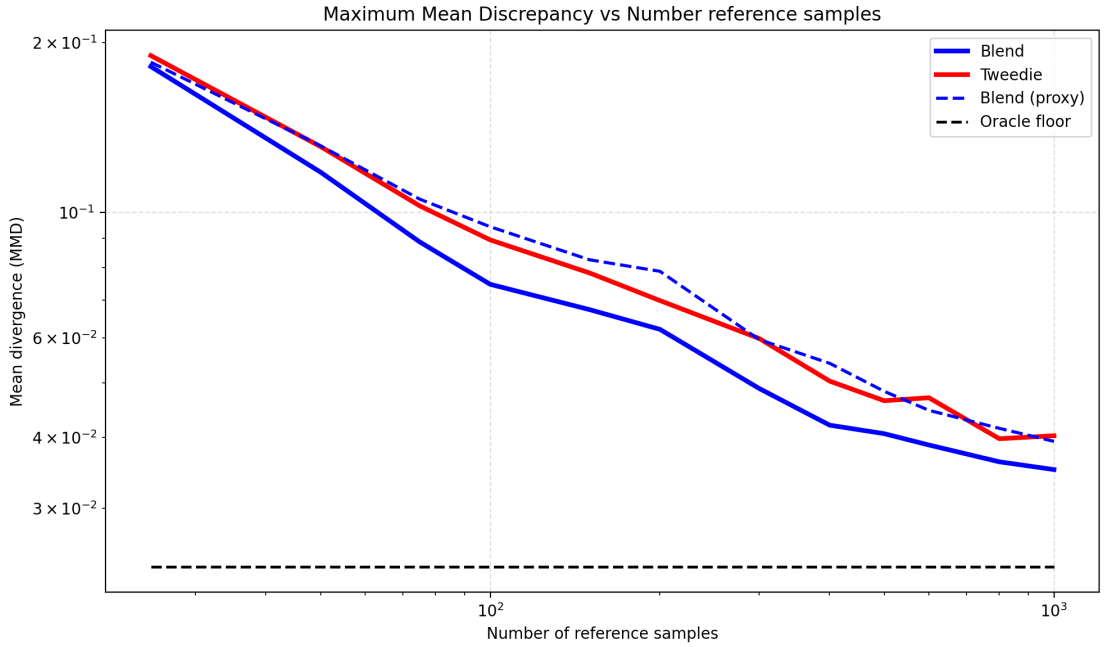

Figure 5: **MMD vs. number of references (lower is better)** on the **6D helix GMM**. Blend matches or slightly improves global mass placement relative to Tweedie while delivering much better gradient fidelity (cf. Fig. 4).

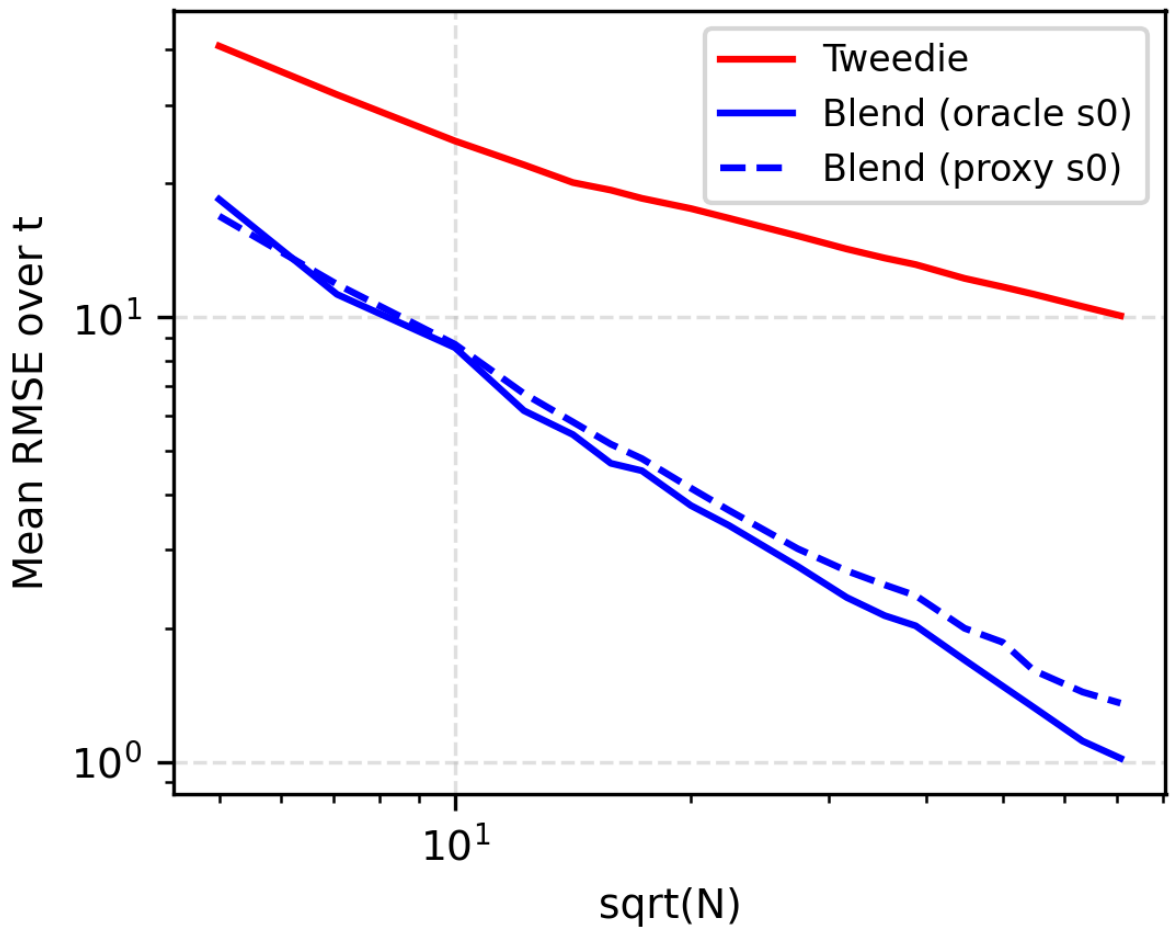

Figure 6: **Score RMSE vs. $\sqrt{N_{\text{ref}}}$ (mean over t)** on the **6D helix GMM**. Ground truth and proxy blends track closely and are uniformly below Tweedie, demonstrating more efficient use of references.

Where to find exact settings. All concrete values (number of components, helix pitch/radius, covariance anisotropy, bandwidth grids, t-grid, and SNIS batch sizes) are provided in App. **??** and in the accompanying notebook; see `get_gmm_funcs` and `run_comparison`.

4.1 Prior sampling: 6D helix GMM

We test the ability of the samplers to capture a complex, low-dimensional manifold embedded in a higher-dimensional space. The target is a 6D Gaussian Mixture Model (GMM) whose intrinsic structure is a 3D helix. For visualization, we project onto three orthogonal planes (d_1, d_2), (d_3, d_4), and (d_5, d_6), where $d_1, \ldots, d_6$ denote the first six principal directions obtained by PCA fit to the *target* distribution (fixed once for all methods). This axis selection highlights high-variance structure and does not *a priori* advantage Blend over Tweedie (or vice versa). We show three samplers: *Blend (true)*—which uses the exact target score s in the blend; *Blend (proxy)*—which replaces s_0 by LR+D (low-rank-plus-diagonal) local Gaussian score proxies fit directly to the raw reference data; and *Tweedie*. A qualitative CSEM-only panel appears separately in Fig. 3.

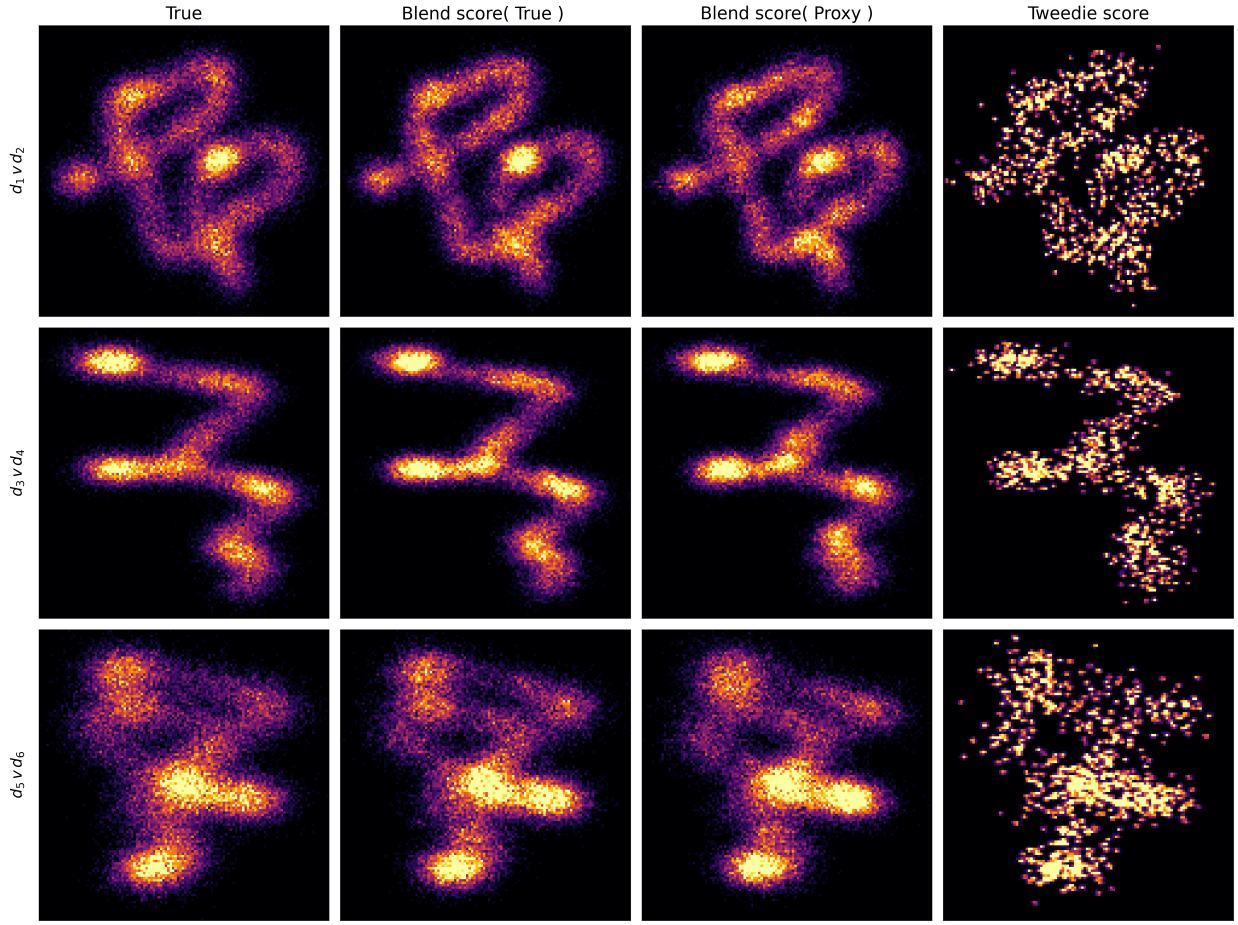

Figure 7: **Prior sampling (6D helix GMM, N=500).** Projected histograms in PCA planes (d_1, d_2), (d_3, d_4), (d_5, d_6) (principal directions fitted to the *target*). *Blend (true)* uses the exact target score; *Blend (proxy)* uses LR+D local Gaussian score proxies fit to the raw data; *Tweedie* is the standard nonparametric baseline. Both blends accurately track the nonlinear helical geometry across projections, while Tweedie exhibits fragmentation or over-smoothing depending on the marginal.

4.2 Posterior sampling: 12D GMM, rank-1 likelihood

We next evaluate a Bayesian inference task where a 12D GMM prior (a 3D manifold embedded in $\mathbb{R}^{12}$) is constrained by a rank-1 Gaussian likelihood. For visualization, planes are chosen by PCA on the *posterior* mass (approximated via importance-reweighted prior samples), yielding the principal directions $d_1, \ldots$ used to form the three shown projections (d_1, d_2), (d_2, d_5), and (d_1, d_6). As before, this axis choice depends only on the target and does not advantage any method.

We compare *Blend (true)*, *Blend (proxy)* (LR+D local Gaussian proxies fit to the raw data), and *Tweedie*. Both blends concentrate mass where the prior manifold intersects the likelihood level sets (white contours), while Tweedie degrades under localization pressure, producing noisy/disconnected samples.

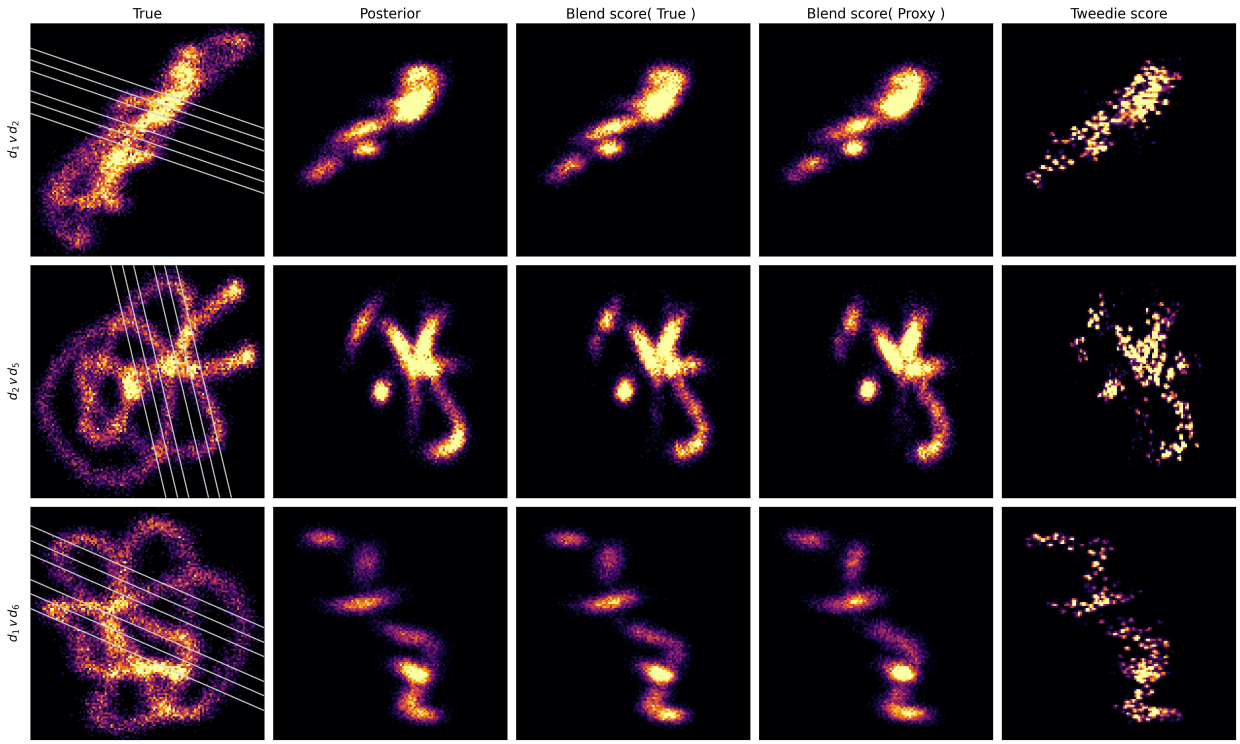

Figure 8: **Posterior sampling (12D, N=1000).** Projected histograms in PCA planes (d_1, d_2), (d_2, d_5), (d_1, d_6) (principal directions fitted to the *posterior* via importance-weighted prior samples). *Blend (true)* uses the exact target score; *Blend (proxy)* uses LR+D local Gaussian score proxies fit to the raw data; *Tweedie* is the baseline. White contours indicate likelihood level sets. Both blends capture the localized posterior manifold, while Tweedie yields noisy, fragmented samples.

4.3 MNIST deblurring, D=15 (PCA), LR+D proxy, N=18,000

We conclude with a practical inverse problem: linear deblurring on MNIST in a D=15 PCA latent space using the LR+D proxy. **Setup.** PCA is fit on the MNIST training set and we work in the first 15 principal directions. The blur operator is a 9×9 Gaussian kernel with standard deviation σ_{blur}=1.5 applied in image space; observations are corrupted with i.i.d. Gaussian noise of level σ_{obs}=0.1 (pixels scaled to $[0, 1]$). Priors over latents are the empirical PCA distribution; posterior *reference* contours are obtained by importance reweighting N=65,000 prior samples with the Gaussian likelihood (IS). Sampling uses the same **Heun PC (second order)** integrator and log-spaced time grid as in the toy GMMs. **Blend (proxy)** uses LR+D local Gaussian score proxies fit directly to the raw PCA codes; **Blend (true)** refers to experiments that use the exact target score when available (only in toy settings, not MNIST).

Metrics (MNIST). In addition to MMD and SW2 to the IS reference, we report:

- **PSNR (dB).** For reconstructions $\hat{x}$ of ground-truth images x (both in $[0, 1]$),

$$\text{PSNR} = 10 \log_{10}\left(\frac{1}{\text{MSE}(x, \hat{x})}\right), \quad \text{MSE} = \tfrac{1}{n}\sum_{p}(x_p - \hat{x}_p)^2,$$

  averaged over the test set.

- **Coverage (%)** of the posterior HPD region. Let c_α be the threshold defining the IS-estimated highest posterior density region $\Omega_\alpha = \{z : p_{\text{IS}}(z) \geq c_\alpha\}$ with mass α=0.999. Coverage is the fraction of method samples falling in Ω_α (higher is better).

- $\mathbb{E}[\log p_{\text{KDE}}]$**.** Average log-density of method samples under a Gaussian KDE fit to IS posterior samples in the 15-D latent space (bandwidth by Scott's rule). Higher indicates better alignment with the IS posterior mode structure, though this metric can favor diffuse samplers.

MCMC baseline (MCMC–GMM). We include a Metropolis–Adjusted Langevin (MALA) baseline that targets the posterior defined by a uniform-weight GMM fit to the PCA latents (prior surrogate) times the Gaussian likelihood. Proposals are preconditioned by the empirical latent covariance; stepsizes are tuned to 0.5–0.6 acceptance. Unless noted, we run 5 chains with 10k iterations, 2k burn-in and thinning by 10. This baseline is informative but inherits bias from the GMM prior surrogate and exhibits mixing limitations in narrow posterior regions.

The quantitative metrics in Table 2 show the blended proxy estimator achieving superior performance across the board, notably in PSNR (image quality), Coverage (posterior support), MMD, and sliced Wasserstein (SW2) distance to the IS reference. The Tweedie sampler wins only on the KDE log-likelihood metric, which—as discussed—can reward overly diffuse samples.

This superiority is visually confirmed in Figs. 9 and 10. Figure 9 overlays posterior samples against IS contours (mass levels $0.4, 0.8, 0.95, 0.995, 0.999$) in the (d_1, d_2) PCA plane, where the blended posterior aligns closely with the IS support while Tweedie is diffuse. Figure 10 shows reconstructions: *Blend (Proxy)* produces sharper means and more coherent samples that respect the contours, whereas *Tweedie only* visibly degrades.

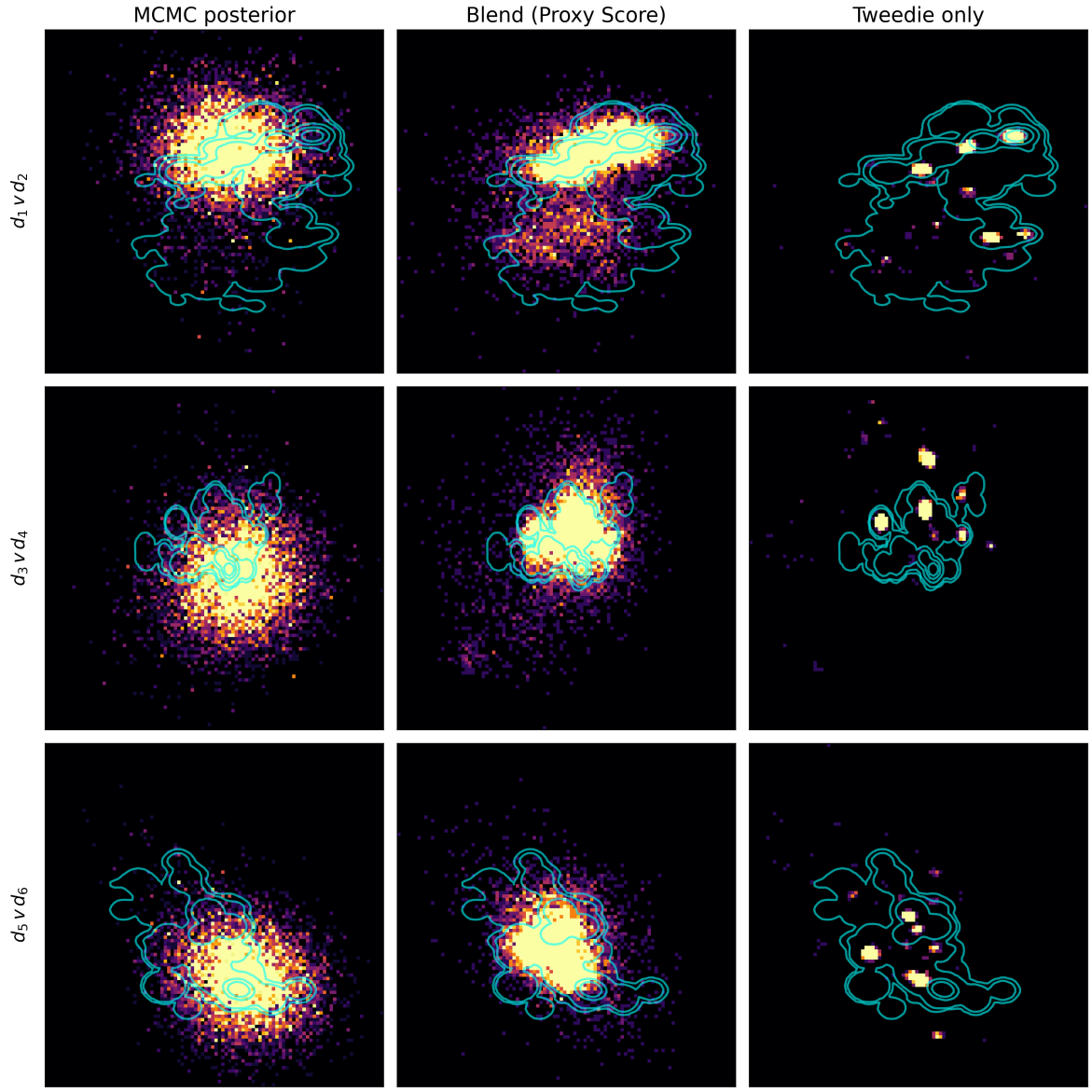

Figure 9: **MNIST: likelihood pushes prior mass (PCA plane (d_1, d_2)).** Cyan IS contours denote HPD levels of the IS posterior. The blended posterior sample distribution aligns with these contours substantially better than Tweedie, which is more diffuse and misplaced.

Method	PSNR (dB) ↑	Pred ≈ 1	Coverage (%) ↑	$\mathbb{E}[\log p_{\text{KDE}}]$ ↑	MMD→IS ↓	SW2→IS ↓
Blend (Proxy)	**28.16**	1.033	**99.9**	-22.094	$\mathbf{1.225 \times 10^{-1}}$	**0.290**
Tweedie only	25.25	1.036	73.7	**-21.618**	3.378×10^{-1}	0.325
MCMC–GMM	24.79	**1.028**	99.6	-23.509	1.725×10^{-1}	0.328

Table 2: **MNIST deblurring metrics (combined).** *Blend (Proxy)* is the top performer in nearly every metric, indicating higher image quality (PSNR) and better posterior fidelity (Coverage, MMD, SW2). Arrows indicate direction of improvement. All samplers use the same Heun PC (second order) solver and time grid.

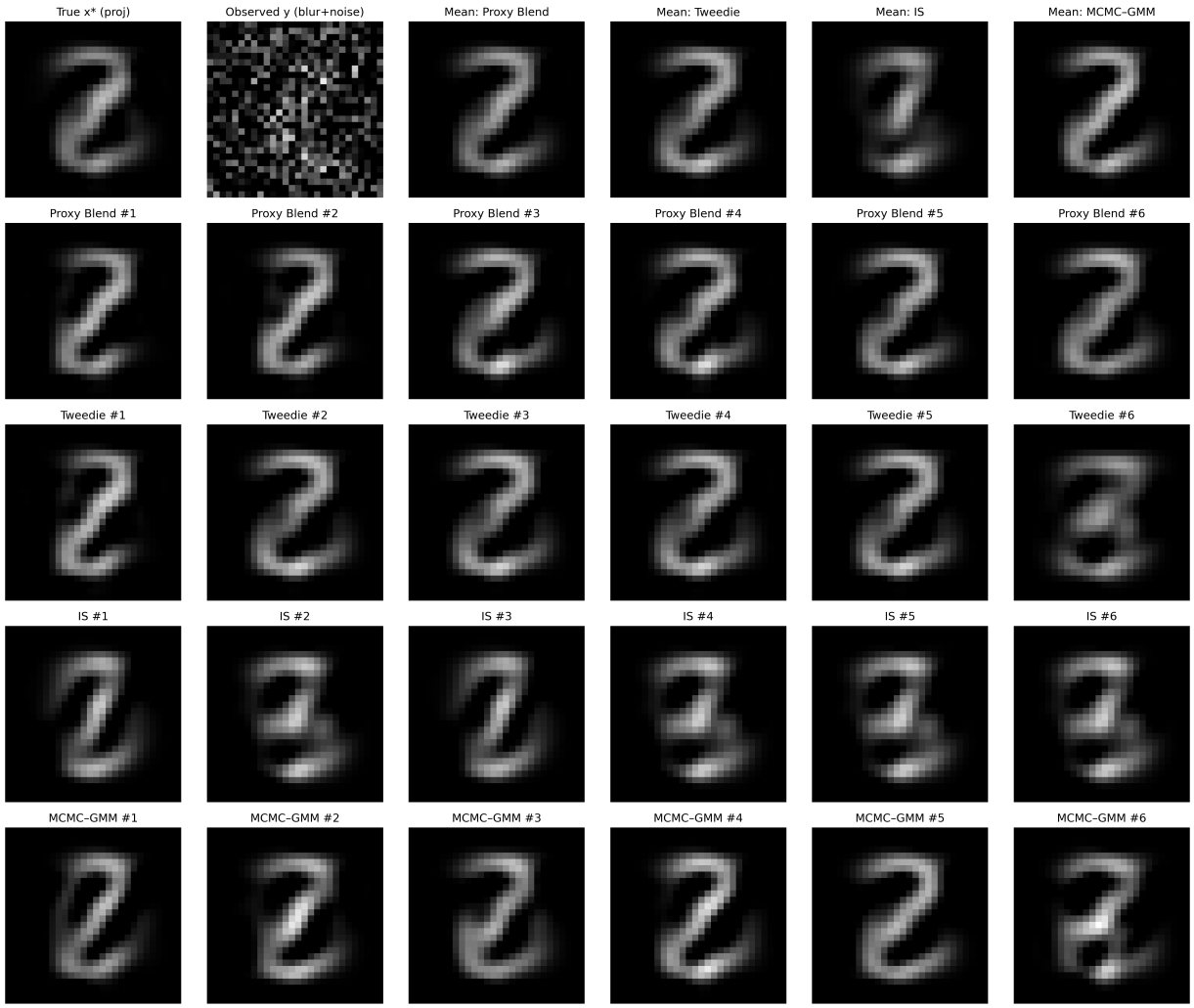

Figure 10: **MNIST deblurring,** N=18,000 **references (LR+D proxy).** Visual comparison: *Blend (Proxy)* produces sharper posterior means and more coherent individual samples than *Tweedie only*. Samples align with IS contours (cyan; HPD mass levels 0.4 to 0.999). All samplers use the Heun PC (second order) integrator.

Discussion

Estimator–centric takeaway. This paper advances *statistical estimation of score fields* for diffusion/flow models. Rather than proposing a new reverse solver, we improve the *local score signal* at a queried (y, t) by combining two complementary, semigroup-native estimators (Tweedie and CSEM) via a variance-minimizing convex blend. The Ornstein–Uhlenbeck (OU) case is our canonical worked example because it affords closed-form identities, but the CSEM viewpoint itself extends to *affine diffusions* through gradient–semigroup commutation; our analysis and claims are framed with this generality in mind.

Local curvature estimation as a unifying principle. A practical route to high-quality scores is to estimate *local curvature of* $\log p_0$ in neighborhoods selected by the forward semigroup and to

transport this information to time t via CSEM. In this work we instantiate curvature by a *local Gaussian approximation* (diagonal/SVD/PCA variants), which supplies a stable proxy for the initial score and its local geometry and yields the nonparametric CSEM estimator. The principle is broader than Gaussian surrogates: curvature information is *natively available* in curvature-aware embeddings such as VAE latents with per-code covariance (or Fisher-metric surrogates). CSEM then acts as the transport rule that turns such local geometry into a time-t score estimate.

Variance-optimal blending: what it guarantees. Section 3.4.2 established that the Monte Carlo errors of Tweedie and CSEM are *negatively aligned* (exactly in the linear–Gaussian setting; in expectation under single-basin posteriors). Consequently, the convex blend in §3.4.3 with weight

$$\lambda^*(y,t) = \frac{V_C - C}{V_C + V_T - 2C}$$

achieves the minimum pointwise risk

$$J(\lambda^*) = \frac{V_T V_C - C^2}{V_T + V_C - 2C} \;\le\; \min\{V_T, V_C\},$$

with a strictly interior weight $0 < \lambda^* < 1$ whenever $C < \min\{V_T, V_C\}$. These statements are *estimator-level* and agnostic to the downstream reverse integrator: improving local risk where the integrator consumes the score translates into fewer steps to reach a target quality.

From estimator to neural samplers: a low-variance teacher. The blended estimator serves as a *plug-in teacher* for DSM/consistency/flow networks. One may distill $(y,t) \mapsto \hat{s}_{\mathrm{BLEND}}(y,t)$ by MSE (or by a fixed PSD inner product) to supervise a parametric score or velocity field; the lower variance at a fixed reference budget yields a stronger target and better sample-efficiency for the student. The same teacher can be used for consistency-style objectives (cross-time targets) or flow/rectified-flow training (drift/velocity supervision), without architectural changes.

Beyond OU. CSEM expresses a conditional-expectation transport of scores along the forward semigroup. While OU provides the cleanest finite-time identity and admits exact proofs of anti-correlation, the conceptual apparatus—semigroup transport plus variance-optimal blending—extends to affine diffusions under gradient–semigroup commutation. Our empirical and theoretical posture therefore treats OU as a standard canonical test case, not as a limitation of the framework.

Reframing the sample set. Rather than a "Lagrangian mesh," we view the reference set as *moving probes of geometry*: Tweedie queries them for global mass placement, while CSEM—via local curvature proxies—extracts fine structure. The blend reconciles these two information sources into the single signal the reverse integrator actually needs.

Limitations and open problems. (i) The expected anti-alignment beyond linear–Gaussian relies on single-basin dominance; behavior near decision boundaries (a set of small p_t-probability for small windows) is weaker. (ii) Our guarantees are variance-level at fixed (y,t); we do not claim Loewner-order negativity of the full cross-covariance without commutation. (iii) SNIS can be fragile under extreme concentration; standard remedies (tempering/mixtures) are orthogonal to our estimator and could be layered on. (iv) Extending curvature estimation from local Gaussians to curvature-aware embeddings (e.g., VAE latents, Fisher metrics) at high dimension, and developing operator-valued/diagonal gates that remain well-regularized, are promising directions.

Broader implications. An estimator-first lens suggests hybrid pipelines: pretrain or regularize neural samplers with the low-variance blended teacher; deploy the same estimator in Bayesian inverse problems by likelihood tilting of the importance weights; and use the affine-diffusion generalization to port the approach to alternative forward processes. Across these settings, the common thread is the same: *better local score estimation* is a portable lever for fewer steps and higher fidelity.

5 Conclusion and Future Work

We reframed score learning as a *statistical estimation* problem at a queried (y, t) and introduced a semigroup-native estimator that combines two complementary signals. The key ingredient is the Conditional Score Expectation Matching (CSEM) principle, which transports score information across time through the forward semigroup. Using the Ornstein–Uhlenbeck (OU) flow as a canonical worked example, we derived an exact finite-time identity, constructed a nonparametric CSEM estimator, and paired it with the classical Tweedie estimator. We proved (linear–Gaussian) and motivated (single-basin) negative alignment of their Monte Carlo errors, yielding a *variance-minimizing convex blend* with a closed-form weight $\lambda^*(y, t)$. A simple SNIS plug-in provides the quantities needed for λ^*, while local Gaussian proxies supply stable curvature information when ground truth s_0 is unavailable. The same estimator extends to posterior inference by a one-line likelihood tilt of the SNIS weights. Although OU affords closed forms, the CSEM viewpoint itself extends to *affine diffusions* via gradient–semigroup commutation; our claims and constructions are formulated with this generality in mind.

Future work. We highlight four directions that build directly on this estimator-centric perspective:

- **Amortized Critic–Gate distillation.** Instead of supervising a network with a pre-averaged estimator, we *distill the statistical signal* produced by the Critic–Gate mechanism itself: the gate predicts a blend of Tweedie- and CSEM-derived per-particle contributions at (y, t), and the critic amortizes this signal into a single fast student usable by DSM/consistency/flow samplers. This preserves the variance advantages of blending while avoiding estimator collapse, and it naturally supports extensions such as diagonal or low-rank (operator-leaning) gates, mild reweighting of the loss (e.g., W-weighted), and light cross-time coupling for consistency-style training.

- **Curvature-aware embeddings at scale.** We will push beyond Gaussian local proxies by using embeddings with per-code geometry (e.g., VAE latents that expose local covariance/curvature). The goal is to quantify how embedding quality shapes anti-alignment, the distribution of $\lambda^*(y, t)$, and the few-step sample quality on high-dimensional image benchmarks.

- **Robust and scalable importance sampling.** For concentrated posteriors or large t, we will combine the estimator with practical remedies (tempering, mixture proposals, multi-proposal SNIS) and study their effect on plug-in variance estimates and the stability of $\lambda^*(y, t)$ in training and inference.

Across these directions the objective remains the same: deliver a *superior local score signal*—amortized by the critic and modulated by the gate—that downstream samplers and neural students can exploit for higher fidelity with fewer steps.

A CSEM Identity for Linear/Affine Gaussian SDEs

Setup and Notation

We consider the time-inhomogeneous linear/affine SDE on $\mathbb{R}^d$:

$$dX_t = A(t)X_t dt + b(t)dt + G(t)dW_t, \qquad X_0 \sim p_0,$$

where $A(t) \in \mathbb{R}^{d\times d}$, $b(t) \in \mathbb{R}^d$, $G(t) \in \mathbb{R}^{d\times r}$ are measurable and locally bounded, and W_t is an r-dimensional standard Brownian motion.

Fundamental matrix. The *fundamental matrix* $\Phi(t,s) \in \mathbb{R}^{d\times d}$ of the linear ODE $\dot{Z}(t) = A(t)Z(t)$ is the unique matrix function satisfying

$$\partial_t \Phi(t,s) = A(t)\Phi(t,s), \qquad \Phi(s,s) = I_d.$$

For the time-homogeneous case where $A(t) \equiv A$, the fundamental matrix is $\Phi(t,s) = e^{A(t-s)}$.

Transition Mean and Covariance. The solution to the SDE is a Gaussian process. The transition kernel $K_t(y \mid x)$ is Gaussian, $\mathcal{N}(y; \Phi(t,0)x + m(t), \Gamma(t))$, where the mean offset $m(t)$ and covariance $\Gamma(t)$ are given by:

$$m(t) := \int_0^t \Phi(t,\tau)b(\tau)d\tau, \qquad \Gamma(t) := \int_0^t \Phi(t,\tau)G(\tau)G(\tau)^\top \Phi(t,\tau)^\top d\tau.$$

We denote the score of the time-t marginal density $p_t(y)$ as $s(y,t) := \nabla_y \log p_t(y)$ and the initial score as $s_0(x) := \nabla_x \log p_0(x)$. We assume that for each $t > 0$, the transition is nondegenerate ($\Gamma(t)$ is positive definite) and that boundary terms vanish during integration by parts.

The CSEM Identity: General Statement and Proof

Theorem A.1 (CSEM for Linear/Affine SDEs)**.** *For any affine SDE satisfying the conditions above, for every $t > 0$ and $y \in \mathbb{R}^d$, the score function is given by:*

$$\boxed{s(y,t) = \Phi(t,0)^{-\top}\mathbb{E}\left[s_0(X_0) \mid X_t = y\right].} \tag{26}$$

Proof. The Gaussian transition kernel is $K_t(y \mid x) \propto \exp\left(-\frac{1}{2}\left\|y - (\Phi(t,0)x + m(t))\right\|^2_{\Gamma(t)^{-1}}\right)$. Taking gradients with respect to y and x yields the cross-derivative identity:

$$\nabla_y K_t(y \mid x) = -\Phi(t,0)^{-\top}\nabla_x K_t(y \mid x). \tag{27}$$

The score is $s(y,t) = \frac{\nabla_y p_t(y)}{p_t(y)}$. Differentiating $p_t(y) = \int K_t(y \mid x)p_0(x)dx$ under the integral sign and applying the identity (27), we get:

$$\nabla_y p_t(y) = \int \nabla_y K_t(y \mid x)p_0(x)dx = -\Phi(t,0)^{-\top}\int \nabla_x K_t(y \mid x)p_0(x)dx.$$

Integrating the right-hand side by parts with respect to x gives:

$$\int \nabla_x K_t(y \mid x)p_0(x)dx = -\int K_t(y \mid x)\nabla_x p_0(x)dx = -\int K_t(y \mid x)p_0(x)s_0(x)dx.$$

Substituting this back, we have:

$$\nabla_y p_t(y) = \Phi(t,0)^{-\top}\int K_t(y \mid x)p_0(x)s_0(x)dx = \Phi(t,0)^{-\top}p_t(y)\mathbb{E}[s_0(X_0) \mid X_t = y].$$

Dividing by $p_t(y)$ yields the general CSEM identity. □

Relation to the OU-specific case. The standard OU process (Eq. (1)) corresponds to $A(t) \equiv -I_d$. In this case, the fundamental matrix is $\Phi(t,0) = e^{-t}I_d$. Substituting this into the general identity (Eq. (26)) gives:

$$s(y,t) = (e^{-t}I_d)^{-\top}\mathbb{E}[s_0(X_0) \mid X_t = y] = e^{t}\mathbb{E}[s_0(X_0) \mid X_t = y],$$

which is exactly the identity presented in the main text in Eq. (9).

Equivalent "Tweedie-L" Form. The same calculus also yields an equivalent identity related to Tweedie's formula:

$$\boxed{s(y,t) = -\Gamma(t)^{-1}\left(y - \Phi(t,0)\mathbb{E}[X_0 \mid X_t = y] - m(t)\right).}$$

Worked Examples for Common Generative SDEs

The general identity applies to all common linear SDEs used in generative modeling.

Variance-Preserving (VP) SDE. For $dX_t = -\frac{1}{2}\beta(t)X_t dt + \sqrt{\beta(t)}dW_t$, we have $\Phi(t,0) = \alpha(t)I$ where $\alpha(t) = \exp(-\frac{1}{2}\int_0^t \beta(u)du)$. The CSEM identity is:

$$s(y,t) = \alpha(t)^{-1}\mathbb{E}[s_0(X_0) \mid X_t = y].$$

Variance-Exploding (VE) SDE. For $dX_t = g(t)dW_t$, we have $\Phi(t,0) = I$. The CSEM identity is:

$$s(y,t) = \mathbb{E}[s_0(X_0) \mid X_t = y].$$

Anisotropic OU / Whitening SDE. For $dX_t = AX_t dt + GdW_t$ with constant matrices A and G, we have $\Phi(t,0) = e^{At}$. The CSEM identity is:

$$s(y,t) = e^{-A^\top t}\mathbb{E}[s_0(X_0) \mid X_t = y].$$

B OU single–basin dominance and CSEM–Tweedie anti–correlation (heuristic detail)

Setup (OU conditioning). For the OU forward dynamics with unit stationary covariance,

$$X_t \;=\; e^{-t}X_0 + \sqrt{1-e^{-2t}}\,Z, \qquad Z \sim \mathcal{N}(0,I),$$

the marginal can be written

$$p_t(y) = \int p_0(x_0)\,\varphi\big(y;\, e^{-t}x_0,\, (1-e^{-2t})I\big)\,dx_0,$$

and conditioning on (y,t) gives the posterior over x_0,

$$\pi_t(x_0 \mid y) \;\propto\; p_0(x_0)\,\exp\Big(-\frac{\|x_0 - \mu_t(y)\|^2}{2\sigma_t^2}\Big), \qquad \mu_t(y) = e^t y, \quad \sigma_t^2 = e^{2t}-1.$$

Write $\ell_t(x_0 \mid y) = \log p_0(x_0) - \frac{1}{2\sigma_t^2}\|x_0 - \mu_t(y)\|^2$.

Lemma B.1 (Strong concavity under a local Hessian bound)**.** *Assume* $\log p_0 \in C^2$ *and* $\nabla^2 \log p_0(x) \preceq MI$ *on a region containing the mass of* $\pi_t(\cdot \mid y)$*. If* $\sigma_t^2 < 1/M$ *(equivalently* $t < \frac{1}{2}\log(1 + 1/M)$*), then*

$$\nabla^2 \ell_t(x_0 \mid y) \;=\; \nabla^2 \log p_0(x_0) - \sigma_t^{-2} I \;\preceq\; (M - \sigma_t^{-2}) I \;\prec 0,$$

so $\ell_t(\cdot \mid y)$ *is globally strongly concave on that region and* $\pi_t(\cdot \mid y)$ *is unimodal with a unique local mode.*

[Separated mixtures] If p_0 is multimodal (e.g. a well-separated GMM), then for small windows σ_t the Gaussian factor selects a single basin except when $\mu_t(y)$ lies within an $O(\sigma_t)$ tube of a decision boundary. In that case the responsibilities satisfy

$$\gamma_k(y,t) \;\propto\; \pi_k \mathcal{N}\big(\mu_t(y);\; \mu_k,\; \Sigma_k + \sigma_t^2 I\big),$$

so the nearest component dominates and the complement mass decays with separation. The p_t–probability of landing in the $O(\sigma_t)$ boundary tube is itself $O(\sigma_t)$ by OU smoothing.

Proposition B.2 (Heuristic reduction and sign of the expected inner product)**.** *Suppose for fixed* (y,t) *that* $\pi_t(\cdot \mid y)$ *is dominated by a single local basin with mode* μ_t *and positive-definite Hessian* $H_t \equiv -\nabla^2 \log p_0(\mu_t) \succ 0$*. Then on that basin*

$$s_0(x) = \nabla \log p_0(x) \;\approx\; -H_t(x - \mu_t),$$

and at the population level

$$A^\star(y,t) = \mathbb{E}[X_0 \mid X_t = y], \qquad B^\star(y,t) = \mathbb{E}[s_0(X_0) \mid X_t = y] \;\approx\; -H_t\big(A^\star(y,t) - \mu_t\big).$$

Let $\widehat{A} = \sum_i \bar{w}_i X_i$ *and* $\widehat{B} = \sum_i \bar{w}_i s_0(X_i)$ *be self-normalized IS estimators targeting* $\pi_t(\cdot \mid y)$*, and define* $\Delta = \widehat{A} - A^\star$*. By the delta method for ratios of means,*

$$\widehat{B} - B^\star \;\approx\; -\,H_t\,\Delta.$$

For the OU-scaled scores

$$\widehat{s}_{\mathrm{CSEM}} = e^t \widehat{B}, \qquad \widehat{s}_{\mathrm{TWD}} = \frac{\widehat{A} - e^{-t} y}{1 - e^{-2t}},$$

the leading errors satisfy

$$\varepsilon_{\mathrm{C}} \;:=\; \widehat{s}_{\mathrm{CSEM}} - s \;\approx\; -\,e^t\,H_t\,\Delta, \qquad \varepsilon_{\mathrm{T}} \;:=\; \widehat{s}_{\mathrm{TWD}} - s \;\approx\; \frac{1}{1 - e^{-2t}}\,\Delta,$$

hence

$$\mathbb{E}\big[\varepsilon_{\mathrm{T}}^\top \varepsilon_{\mathrm{C}}\big] \;\approx\; -\,\frac{e^t}{1 - e^{-2t}}\,\mathbb{E}\big[\Delta^\top H_t\,\Delta\big] \;=\; -\,\frac{e^t}{1 - e^{-2t}}\;\mathrm{tr}\big(H_t\,\mathbb{E}[\Delta\Delta^\top]\big) \;\le\; 0,$$

with strict negativity whenever $\mathrm{Var}(\Delta)$ *has support on any eigenvector of* H_t*.*

[Scalar projections] For any $u \in \mathbb{R}^d$, $E_K := u^\top(\widehat{s}_{\mathrm{CSEM}} - s)$ and $E_T := u^\top(\widehat{s}_{\mathrm{TWD}} - s)$ satisfy $\mathrm{Cov}(E_K, E_T) \approx -\frac{e^t}{1-e^{-2t}}\, u^\top H_t\, \mathrm{Var}(\Delta)\, u \le 0$.

Loewner caveat (where matrix negativity can fail). The product form $H_t\,\mathbb{E}[\Delta\Delta^\top]$ is PSD, but the exact Gaussian cross-covariance involves a product of PSD factors that need not commute; without simultaneous diagonalization one cannot assert $\mathrm{Cov}(\varepsilon_{\mathrm{C}}, \varepsilon_{\mathrm{T}}) \preceq 0$ in general. Our claims here are restricted to *negative expected inner products* (trace level) in the single–basin regime.

Proof of Proposition 3.2 (linear–Gaussian case)

Let $p_0 = \mathcal{N}(\mu, \Sigma)$ and write $V_t := e^{-2t}\Sigma + (1 - e^{-2t})I$. Standard Gaussian conditioning yields

$$A^\star(y,t) = \mathbb{E}[X_0 \mid X_t = y] = \mu + e^t \Sigma V_t^{-1}\big(y - e^{-t}\mu\big), \qquad B^\star(y,t) = \mathbb{E}[s_0(X_0) \mid X_t = y] = \Sigma^{-1}\big(\mu - A^\star(y,t)\big).$$

For any self-normalized IS $\widehat{A} = \sum_i \tilde{w}_i x_i$ targeting $\pi_t(\cdot \mid y)$, set $\Delta = \widehat{A} - A^\star$. Then

$$\widehat{s}_{\mathrm{TWD}} - s = \frac{\widehat{A} - A^\star}{1 - e^{-2t}} = \frac{e^{-t}}{1 - e^{-2t}}\,\Delta, \qquad \widehat{s}_{\mathrm{CSEM}} - s = e^t\big(\widehat{B} - B^\star\big) = -e^t\Sigma^{-1}\Delta,$$

which are the identities in (11). Taking their inner product gives (12), and taking expectations gives (13).

Scope and limitations. Proposition B.2 is heuristic: its accuracy requires (i) a small enough OU window so that Lemma B.1 (or Remark B) ensures single–basin dominance, and (ii) that the second–order expansion of $\log p_0$ around the local mode controls the posterior concentration set. Near decision boundaries (a p_t–probability $O(\sigma_t)$ set) the effect can weaken, but these regions vanish as $t \downarrow 0$ or as inter-basin separation grows. Empirically we observe broadly negative correlations consistent with the above reduction.

C Parametric Distillation via a Critic–and–Gate Network (Proof-of-Concept)

Goal. The main paper develops a nonparametric, variance–optimal blended score estimator $\hat{s}_{\mathrm{BLEND}}(y,t)$ that combines the CSEM and Tweedie identities. For deployment without a reference set at test time, we provide a *parametric distillation* that amortizes this blended estimator into a single neural score model. This appendix presents the minimal ingredients: the setup, a learning objective whose efficacy follows from a law-of-total-variance decomposition, and a single 48-D GMM experiment (qualitative panel + quantitative table).

C.1 Setup and Learning Objective

Let $x_0 \sim p_0$, $\xi \sim \mathcal{N}(0, I)$, and for OU forward dynamics define $y = e^{-t}x_0 + \sigma_t \xi$ with $\sigma_t^2 = 1 - e^{-2t}$. Denote the per-particle signals (from Sections 3.2–3.3) by

$$a(x_0, t) \;=\; e^{t}\, s_0(x_0), \qquad b(x_0, y, t) \;=\; -\sigma_t^{-2}\big(y - e^{-t}x_0\big)\,.$$

A **gate** $g(y, t; \psi) \in [0, 1]$ produces a blended per-particle signal

$$z_g(x_0; y, t) \;=\; \big(1 - g(y, t; \psi)\big)\, a(x_0, t) \;+\; g(y, t; \psi)\, b(x_0, y, t),$$

and a **critic** $q(y, t; \omega)$ predicts the final score (a function of (y, t) only). We train (ψ, ω) by minimizing the population MSE

$$\mathcal{L}(\psi, \omega) \;=\; \mathbb{E}_{x_0,\xi,t}\Big[\,\| \, z_g(x_0; y, t) \;-\; q(y, t; \omega) \, \|_2^2\Big]\,. \tag{28}$$

C.2 Why the Objective Works: Law of Total Variance

Condition on a fixed (y, t) and let the conditional (posterior) law of x_0 given (y, t) be $\pi(\cdot \mid y, t)$. Abbreviate $z_g = z_g(x_0; y, t)$ and $q = q(y, t; \omega)$. Applying the law of total variance yields the pointwise decomposition

$$\mathbb{E}\big[\|z_g - q\|_2^2 \,\big|\, y, t\big] \;=\; \underbrace{\mathrm{Var}_\pi(z_g)}_{\text{depends only on } g} \;+\; \Big\| \underbrace{\mathbb{E}_\pi[z_g]}_{\text{posterior mean}} \;- q\Big\|_2^2 \tag{29}$$

Taking total expectation over (y, t) shows that minimizing (28) enforces two complementary roles:

1. **Critic as posterior mean.** For any fixed gate g, the inner minimum of (29) is attained at
$$q^\star(y, t) \;=\; \mathbb{E}_\pi[z_g(x_0; y, t)]\,,$$
i.e., the critic learns the *MSE-optimal* blended score at (y, t).

2. **Gate as variance minimizer.** Substituting $q^\star$ back into (29) leaves the gate with the objective $\mathrm{Var}_\pi(z_g)$, so g is driven to the *variance-minimizing* blend at each (y, t).

Connection to the nonparametric SNIS blend. Write $a = s_{\mathrm{CSEM}}$ and $b = s_{\mathrm{TWD}}$. For scalar blending, $z_\lambda = (1-\lambda)a + \lambda b$, the variance $\mathrm{Var}_\pi(z_\lambda)$ is minimized by $\lambda^* = \dfrac{\mathrm{Var}[a] - \mathrm{Cov}[a, b]}{\mathrm{Var}[a] + \mathrm{Var}[b] - 2\,\mathrm{Cov}[a, b]}$ computed under $\pi(\cdot \mid y, t)$. The nonparametric SNIS plug-in estimator approximates this $\lambda^*(y, t)$ from posterior samples. Equation (29) shows that the parametric critic–and–gate reproduces the same population objective: the critic learns the posterior mean of the *current* blend while the gate searches over λ to minimize its conditional variance. Thus, the learned $g(y, t; \psi)$ amortizes $\lambda^*(y, t)$ and $q(y, t; \omega)$ amortizes the resulting blended score, yielding a direct parametric distillation of the nonparametric rule.

C.3 Proof-of-Concept Experiment (48-D GMM)

We evaluate on a d=48 Gaussian mixture with highly curved, filamentary structure, using a 10-step reverse-OU sampler. Critic Gate is trained using diagonal covariance proxy scores(3.5) learned from data alone. Figure 11 shows qualitative projections: the distilled critic preserves filament geometry much better than a DSM baseline. Table 3 lists quantitative metrics at 15 steps; our critic gate score distillation outpeforms the DSM baselines accross all divergence metrics.

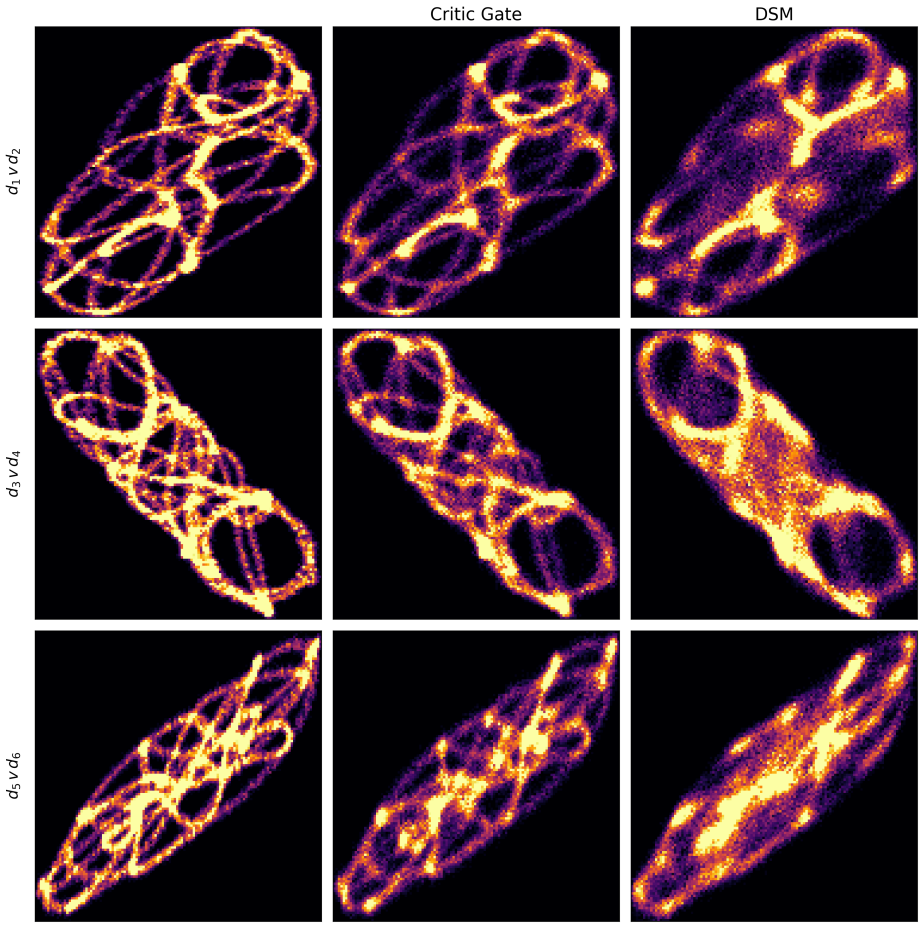

Figure 11: **Critic–and–Gate distillation on 48-D GMM (10 steps).** Qualitative density projections: left column (truth), middle (Critic–Gate), right (DSM). The distilled critic, trained by (28), recovers thin filamentary sets that DSM blurs.

Table 3: **Quantitative comparison at 15 steps.** Metrics on matched samples; lower is better.

Metric	DSM	Critic–Gate (ours)	Floor
MMD@15	3.732×10^{-2}	$\mathbf{2.507 \times 10^{-2}}$	2.053×10^{-2}
W2@15	5.586×10^{-2}	$\mathbf{3.865 \times 10^{-2}}$	2.515×10^{-2}
M-KSD@15	4.727×10^{2}	$\mathbf{1.044 \times 10^{2}}$	1.590×10^{1}

Scope. This appendix establishes feasibility and preserves *conceptual priority* for the parametric extension. Scaling to large real datasets and ablations is deferred to a future companion work.

D Reproducibility

Environment. All experiments run in Python with `NumPy` and `CuPy` (GPU) in double precision (`float64`); plots use `Matplotlib`; `scikit-learn` is used only for PCA visualizations. Core array ops are vectorized and CuPy-backed with transparent CPU fallback when inputs are on host memory.

OU semigroup and reverse-time dynamics. We use the Ornstein–Uhlenbeck (OU) forward corruption

$$X_t \;=\; e^{-t}X_0 \;+\; \sqrt{1-e^{-2t}}\,\varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I),$$

and integrate the reverse-time dynamics with a **second–order Heun predictor–corrector (PC) solver** for *all* samplers and datasets (toy and MNIST). Each step performs an explicit Euler *predictor* followed by a slope-averaged *corrector* on the same time grid; probability-flow/ODE variants reuse the drift without stochastic terms. We count **NFE** as drift evaluations; Heun PC uses 2 NFE per step.

Time grids. Unless stated otherwise, time grids are log- or power-spaced and shared across methods:

$$t_0 = T_{\text{end}}\,, \quad t_K = T_{\text{target}}, \quad t_{k+1} = \text{power\_grid}(t_k;\gamma) \text{ or log-spaced},$$

with typical values $T_{\text{end}} = 1.5$, $T_{\text{target}} \in \{10^{-3}, 5\cdot 10^{-4}\}$, $K \le 20$. We report few-step regimes consistent with the figures in the main text.

Targets and priors (toy GMMs). We construct Gaussian mixture targets via `get_gmm_funcs` (notebook `non_param_sampler.ipynb`). The **6D helix GMM** used for all quantitative curves in §4 has uniformly weighted components placed along a smooth 3D helical curve embedded in $\mathbb{R}^6$. Each component covariance is anisotropic: larger variance along the local tangent direction, smaller in the normal subspace (tangent/normal alignment is computed from the helix geometry). A second target is a **12D GMM** whose intrinsic support is a 3D manifold embedded in $\mathbb{R}^{12}$ (used for the rank-1 likelihood posterior experiment). Exact means/covariances, helical pitch/radius, and mixture counts are emitted by the generator and logged with the run configuration.

Visualization planes (PCA convention). In all histogram figures, we denote by $d_1, d_2, \ldots$ the principal directions returned by PCA fit to the relevant *target* distribution (prior) or to an IS-approximated *posterior*. We display 2D histograms in planes (d_i, d_j). This choice highlights high-variance structure and does not *a priori* advantage any method.

Nonparametric score estimators. At time t, we evaluate:

- **Tweedie:** $\hat{s}_{\text{TWD}}(y,t)$ from the OU Tweedie identity using denoising residuals.
- **CSEM:** $\hat{s}_{\text{CSEM}}(y,t)$ via Conditional Score Expectation Matching using an OU-consistent importance model over references x_i with unnormalized logits

  $$\ell_i(y,t) = -\tfrac{1}{2(1-e^{-2t})}\,\|y - e^{-t}x_i\|^2 + \log L(x_i),$$

  where $\log L(x)$ is an optional likelihood tilt; weights are the softmax of $\{\ell_i\}$.
- **Variance–optimal blend:** $\hat{s}_{\text{blend}}(y,t) = (1-\omega)\hat{s}_{\text{CSEM}} + \omega\hat{s}_{\text{TWD}}$, with $\omega = \omega(y,t)$ chosen to approximately minimize pointwise variance using SNIS moment estimates (the `snis_blend` routine).

For CSEM/Tweedie with likelihoods, we optionally use the "posterior-at-0" correction $s_0^{\text{eff}}(x) = s_0(x) + \nabla_x \log L(x)$ when $\nabla \log L$ is available; otherwise $\log L$ only tilts the logits ℓ_i.

LR+D proxy for the initial score. Blend (proxy) replaces s_0 with local Gaussian score proxies fit directly to the raw reference data using a low-rank-plus-diagonal precision model:

- k-NN neighborhood per query (default $k \approx 4D$, clipped to $N_{\text{ref}} - 1$).
- Local PCA rank r (default $r \leq 8$); Woodbury inversion for the rank-r part, plus a diagonal tail.
- Ridge on tail eigenvalues with a "tail-mean" (default) or "trimmed-mean" rule; optional eigenvalue clip $\lambda_i \leq \text{lam\_clip\_mult} \cdot \tau$.
- Optional score whitening for Tweedie (recompute-white mode) for stability; disabled for CSEM.

Blend (true) uses the exact target score wherever available (toy GMMs only).

SNIS quality control. We compute the effective sample size $\text{ESS} = 1/\sum_i \tilde{w}_i^2$ from normalized weights $\tilde{w}_i$. For correlation and variance diagnostics, time points with $\text{ESS} < \tau_{\text{ESS}}$ are dropped (default $\tau_{\text{ESS}} = 50$). All reported numbers use median-of-means over independent batches.

Samplers and common API. We expose `tweedie_sampler`, `CSEM_sampler`, and `blend_sampler` with a shared signature

$$(\texttt{N\_part}, \texttt{N\_ref}, \texttt{time\_pts}, \texttt{batch\_size}, \texttt{sampler\_func}, \ldots),$$

and a common Heun PC integrator on a shared time grid. Typical settings used in figures: $\texttt{N\_ref} \in \{10^3, 2 \cdot 10^3\}$, $\texttt{N\_part} \in \{2 \cdot 10^3, 5 \cdot 10^3\}$, $\texttt{batch\_size} = 10^3$.

MNIST deblurring setup (PCA D=15). PCA is fit on the MNIST training set; inference is carried out in the first D=15 principal directions. The blur operator is a 9×9 Gaussian kernel with $\sigma_{\text{blur}} = 1.5$ in image space; observations include i.i.d. Gaussian noise with $\sigma_{\text{obs}} = 0.1$ (images scaled to $[0, 1]$). The IS *reference posterior* is formed by importance-reweighting $N = 65{,}000$ prior samples with the Gaussian likelihood; HPD contour levels shown in figures are $\{0.4, 0.8, 0.95, 0.995, 0.999\}$. Sampling uses the same Heun PC solver and time grid as the toy experiments. Unless noted, the reported MNIST results use **Blend (proxy)** with LR+D local Gaussian score proxies.

MCMC baseline (MCMC–GMM). We include a Metropolis–Adjusted Langevin (MALA) sampler that targets a posterior defined by a uniform-weight GMM surrogate prior (fit to the PCA latents) times the same Gaussian likelihood. Proposals are preconditioned by the latent covariance; stepsizes are tuned to 0.5–0.6 acceptance. We run 5 chains, 10k iterations, 2k burn-in, thinning by 10, unless otherwise specified. This baseline is informative but inherits bias from the GMM surrogate and mixes slowly in narrow posterior regions.

Component	Setting (typical values)
Time grid	T_{end}=1.5, $T_{\text{target}} \in \{10^{-3}, 5\times10^{-4}\}$; ≤ 20 steps; log/power spacing
Integrator	**Heun predictor–corrector (2nd order)**; *2 NFE/step* (all methods share grid/solver)
References/particles	`N_ref` $\in [10^3, 2\times10^3]$; `N_part` $\in \{2000, 5000\}$; `batch_size`=10^3
LR+D proxy	$k \approx 4D$ (clip); rank $r \leq 8$; tail ridge (tail-mean); optional λ-clip; Tweedie whitening optional
SNIS/ESS	softmax logits with likelihood tilt; drop t if ESS < 50; median-of-means aggregation
Metrics	KSD (IMQ, β=$-1/2$, c^2=0.5), MMD (RBF, 5 scales), SW2 (512 slices), PSNR, Coverage, $E[\log p_{\text{KI}}$
Visualization	PCA planes (d_i, d_j) from target/posterior; same planes for all methods per figure
Precision/Device	`float64`; GPU via `CuPy` with CPU fallback

Table 4: Key hyperparameters and defaults across experiments.

Metrics and diagnostics. We adopt common metrics with fixed configurations to ensure comparability:

- **Score RMSE** (toy only): $\left(\mathbb{E}_{y\sim p_t}\|\hat{s}(y,t) - s(y,t)\|^2\right)^{1/2}$, averaged over a log-spaced grid of t.
- **KSD** (IMQ kernel): $k(x,y) = (c^2 + \|x-y\|^2)^\beta$ with $\beta \in (-1, 0)$ (default $\beta = -\frac{1}{2}$), $c^2 = 0.5$; unbiased U-statistic estimator on blocks (`chunk_size`=4096).
- **MMD** (RBF): bandwidth grid around the median heuristic; 5 scales in $[0.5, 2.0]$ relative to the median distance; unbiased estimator.
- **SW2→IS** (MNIST): sliced 2-Wasserstein distance to the IS posterior using 512 random one-D projections with antithetic pairs.
- **PSNR** (MNIST): $10\log_{10}\left(1/\text{MSE}(x, \hat{x})\right)$ on reconstructions $\hat{x}$ from posterior samples.
- **Coverage** (MNIST): fraction of samples within the IS-estimated HPD region $\Omega_{0.999} = \{z : p_{\text{IS}}(z) \geq c\}$.
- **E**$[\log \mathbf{p}_{\text{KDE}}]$ (MNIST): average log-density of method samples under a Gaussian KDE fit to IS samples in 15-D latent space (bandwidth by Scott's rule).

We additionally report ESS and drop t with ESS $< \tau_{\text{ESS}}$ for correlation/variance plots.

Baselines and ablations. All samplers use the same Heun PC solver and time grid. We ablate: (a) CSEM vs Tweedie vs Blend, (b) uniform vs SNIS blending, (c) ground truth vs proxy s_0 (toy), (d) time-grid schedules (linear vs power/log), and (e) optional Tweedie score whitening.

Randomness and seeds. A `seed` argument controls all RNGs; figures use fixed seeds per panel and report mean ± 95% CI over $R = 5$ seeds unless noted.

Compute. Results are produced on a single GPU via `CuPy`. Long-range pairwise operations (KSD/MMD) are block-processed (`chunk_size`=4096) to bound memory.

Code and data. We will release code, configs, and scripts to reproduce all figures (including notebook cells invoking `get_gmm_funcs`, `run_comparison`, and the MNIST adapters). Results average R=5 seeds with mean ± 95% CI.

References

[1] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. URL https://proceedings.neurips.cc/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

[2] Y. Song, J. Sohl-Dickstein, D.P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations, 2021. PLACEHOLDER.

[3] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[4] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2022. URL https://arxiv.org/abs/2210.02747. NeurIPS 2023 (poster).

[5] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2209.03003.

[6] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning (ICML)*. PMLR, 2023.

[7] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[8] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019. doi: 10.18653/v1/P19-1355. URL https://aclanthology.org/P19-1355/.

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 6840–6851, 2020.

[10] 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 (ICLR)*, 2021.

[11] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning (ICML)*, pages 8162–8171. PMLR, 2021.

[12] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021.

[13] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[14] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[15] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

[16] Chieh-Hsin Lai, Yuhta Takida, Naoki Murata, Toshimitsu Uesaka, Yuki Mitsufuji, and Stefano Ermon. FP-Diffusion: Improving score-based diffusion models by enforcing the underlying score fokker–planck equation. In *International Conference on Machine Learning (ICML)*. PMLR, 2023.

[17] Zheyuan Hu, Zhongqiang Zhang, George Em Karniadakis, and Kenji Kawaguchi. Score-based physics-informed neural networks for high-dimensional fokker–planck equations. *arXiv preprint arXiv:2402.07465*, 2024.

[18] Mo Zhou, Stanley Osher, and Wuchen Li. Simulating fokker–planck equations via mean field control of score-based normalizing flows. *arXiv preprint arXiv:2506.05723*, 2025.

[19] Giacomo Greco. A malliavin–gamma calculus approach to score-based diffusion generative models for random fields. *arXiv preprint arXiv:2505.13189*, 2025.

[20] Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.

[21] Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Aapo Hyvärinen, and Revant Kumar. Density estimation in infinite dimensional exponential families. *Journal of Machine Learning Research*, 18(57):1–59, 2017.

[22] Michael Arbel and Arthur Gretton. Kernel conditional exponential family. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 84 of *Proceedings of Machine Learning Research*, pages 1337–1346. PMLR, 2018.

[23] Li K. Wenliang, Danica J. Sutherland, Heiko Strathmann, and Arthur Gretton. Learning deep kernels for exponential family densities. In *International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pages 6737–6746. PMLR, 2019.

[24] Yuhao Zhou, Jiaxin Shi, and Jun Zhu. Nonparametric score estimators. In *International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pages 11513–11523. PMLR, 2020.

[25] Q. Liu and D. Wang. Stein variational gradient descent, 2016. PLACEHOLDER.

[26] Qiang Liu, Jason Lee, and Michael Jordan. A kernelized stein discrepancy for goodness-of-fit tests. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, volume 48 of *Proceedings of Machine Learning Research*, pages 276–284. PMLR, 2016.

[27] Krzysztof Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, volume 48 of *Proceedings of Machine Learning Research*, pages 2606–2615. PMLR, 2016.

[28] Anna Korba, Pierre-Cyril Aubin-Frankowski, Szymon Majewski, and Pierre Ablin. Kernel stein discrepancy descent. In *International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*. PMLR, 2021.

[29] Michael Arbel, Anna Korba, Adil Salim, and Arthur Gretton. Maximum mean discrepancy gradient flow. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[30] Berthy T. Feng, Jamie Smith, Michael Rubinstein, Huiwen Chang, Katherine L. Bouman, and William T. Freeman. Score-based diffusion models as principled priors for inverse imaging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[31] Alexandre Adam, Adam Coogan, Nikolay Malkin, Ronan Legin, Laurence Perreault-Levasseur, Yashar Hezaveh, and Yoshua Bengio. Posterior samples of source galaxies in strong gravitational lenses with score-based priors. *arXiv preprint arXiv:2211.03812*, 2022.

[32] Ronan Legin, Alexandre Adam, Yashar Hezaveh, Laurence Perreault-Levasseur, D. Zhang, Francisco Villaescusa-Navarro, Shirley Ho, Siamak Ravanbakhsh, and Yoshua Bengio. Posterior sampling of the initial conditions of the universe from large-scale structure surveys with score-based generative models. *arXiv preprint arXiv:2304.03788*, 2023.

[33] Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. doi: 10.1198/jasa.2011.tm11181.

[34] Herbert E. Robbins. An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 157–163. University of California Press, Berkeley, CA, 1956.

[35] A. Owen. Monte carlo theory, methods and examples, 2013. PLACEHOLDER.

[36] C. Robert and G. Casella. Monte carlo statistical methods, 2004. PLACEHOLDER.

[37] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011. doi: 10.1162/NECO_a_00142.

[38] Daniel T. Gillespie. *Markov Processes: An Introduction for Physical Scientists*. Academic Press, Boston, 1st edition, 1991. ISBN 978-0122839559.

[39] Daniel T. Gillespie. The mathematics of brownian motion and johnson noise. *American Journal of Physics*, 64(3):225–240, 1996. doi: 10.1119/1.18210.

[40] N. Jeremy Kasdin. Discrete simulation of colored noise and stochastic processes and $1/f^{\alpha}$ power law noise generation. *Proceedings of the IEEE*, 83(5):802–827, 1995. doi: 10.1109/5.381848.

[41] Daniel T. Gillespie. Exact numerical simulation of the ornstein–uhlenbeck process and its integral. *Physical Review E*, 54(2):2084–2091, Aug 1996. doi: 10.1103/PhysRevE.54.2084. URL https://link.aps.org/doi/10.1103/PhysRevE.54.2084.

[42] Michael P. Allen and Dominic J. Tildesley. *Computer Simulation of Liquids*. Oxford University Press, 2 edition, 2017. ISBN 978-0198803201.

[43] Daan Frenkel and Berend Smit. *Understanding Molecular Simulation: From Algorithms to Applications*. Academic Press, 2 edition, 2001. ISBN 978-0122673511.

[44] Tan Bui-Thanh and Mark Andrew Girolami. Solving large-scale PDE-constrained Bayesian inverse problems with Riemann manifold Hamiltonian Monte Carlo. *Inverse Problems*, Special Issue:114014, 2014. http://users.ices.utexas.edu/%7Etanbui/PublishedPapers/RMHMC.pdf.

[45] Shiwei Lan, Tan Bui-Thanh, Mike Christie, and Mark Girolami. Emulation of higher-order tensors in manifold monte carlo methods for bayesian inverse problems. *Journal of Computational Physics*, 308:81 – 101, 2016. ISSN 0021-9991. doi: https://doi.org/10.1016/j.jcp.2015.12.032. URL http://www.sciencedirect.com/science/article/pii/S0021999115008517.

[46] Tan Bui-Thanh and Quoc P. Nguyen. FEM-based discretization-invariant mcmc methods for pde-constrained bayesian inverse problems. *Inverse Problems and Imaging*, 10(4):943–975, 2016. ISSN 1930-8337. doi: 10.3934/ipi.2016028. URL http://aimsciences.org/journals/displayArticlesnew.jsp?paperID=13201. http://users.ices.utexas.edu/%7Etanbui/PublishedPapers/FEMBayesian.pdf.

[47] Tan Bui-Thanh and Omar Ghattas. Analysis of the Hessian for inverse scattering problems. Part I: Inverse shape scattering of acoustic waves. *Inverse Problems*, 28(5):055001, 2012. doi: 10.1088/0266-5611/28/5/055001. http://users.ices.utexas.edu/%7Etanbui/PublishedPapers/CompactI.pdf.

[48] Tan Bui-Thanh and Omar Ghattas. Analysis of the Hessian for inverse scattering problems. Part II: Inverse medium scattering of acoustic waves. *Inverse Problems*, 28(5):055002, 2012. doi: 10.1088/0266-5611/28/5/055002. http://users.ices.utexas.edu/%7Etanbui/PublishedPapers/CompactII.pdf.

[49] Tan Bui-Thanh and Omar Ghattas. Analysis of the Hessian for inverse scattering problems. Part III: Inverse medium scattering of electromagnetic waves. Inverse Problems and Imaging, 2013. http://users.ices.utexas.edu/%7Etanbui/PublishedPapers/EM3Dmedium.pdf.

[50] Tan Bui-Thanh, Omar Ghattas, James Martin, and Georg Stadler. A computational framework for infinite-dimensional Bayesian inverse problems Part I: The linearized case, with application to global seismic inversion. *SIAM Journal on Scientific Computing*, 35(6):A2494–A2523, 2013. doi: 10.1137/12089586X. http://users.ices.utexas.edu/%7Etanbui/PublishedPapers/InfiniteBayesianSisc13.pdf.

[51] Tan Bui-Thanh, Carsten Burstedde, Omar Ghattas, James Martin, Georg Stadler, and Lucas C. Wilcox. Extreme-scale UQ for Bayesian inverse problems governed by PDEs. In *SC12: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 2012. Gordon Bell Prize finalist, http://users.ices.utexas.edu/%7Etanbui/PublishedPapers/sc12.pdf.

[52] Tan Bui-Thanh and Omar Ghattas. A scalable MAP solver for Bayesian inverse problems with Besov priors. *Inverse Problems and Imaging*, 9(1):27–53, 2015. http://users.ices.utexas.edu/%7Etanbui/PublishedPapers/BesovMAP.pdf.

[53] Andrew M. Stuart. Inverse problems: A Bayesian perspective. *Acta Numerica*, 19:451–559, 2010. doi: doi:10.1017/S0962492910000061.

[54] Jari Kaipio and Erkki Somersalo. *Statistical and Computational Inverse Problems*, volume 160 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2005.

[55] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

[56] Jackson Gorham and Lester Mackey. Measuring sample quality with stein's method. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.