

# Analytic Score Distillation: Refining Latent Score-Based Generative Models via the Laplace Score Identity

Technical Report

January 19, 2026

## Abstract

Standard Latent Diffusion Models (LDMs) and Latent Score-based Generative Models (LSGMs) typically train the score network using Denoising Score Matching (DSM) with stochastic noise targets (the Tweedie identity). This approach implicitly treats the latent representations of training data as Dirac deltas, ignoring the probabilistic uncertainty explicitly predicted by the Variational Autoencoder (VAE). We propose **Analytic Score Distillation**, a training methodology that replaces the stochastic DSM target with the **Laplace Score Identity (LSI)**. By formalizing the latent space as a **Gaussian Field**—a continuous mixture of encoder posteriors—we derive a Rao-Blackwellized, zero-variance estimator for the score. We present a joint training protocol for the VAE and Score Network that utilizes LSI to align the geometric curvature of the encoder with the diffusion dynamics. We show that this approach is a strict mathematical refinement of the LSGM objective, effectively removing the Monte Carlo noise variance from the training signal.

## 1 Introduction

The current paradigm for Latent Diffusion involves a two-stage or joint process: mapping data  $\mathbf{x}$  to a latent  $\mathbf{z}$  via an encoder  $q_\phi(\mathbf{z}|\mathbf{x})$ , and training a score network  $s_\theta(\mathbf{z}_t, t)$  to reverse a diffusion process on  $\mathbf{z}$ .

Existing methods, including the seminal LSGM [1], rely on the standard Denoising Score Matching (DSM) objective:

$$\mathcal{L}_{\text{DSM}} = \mathbb{E}_{\mathbf{z}_0 \sim q_\phi, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} [\|s_\theta(\mathbf{z}_t, t) - (-\epsilon/\sigma_t)\|^2] \quad (1)$$

While effective, this objective estimates the score using a single realization of noise  $\epsilon$ . It ignores the fact that in a VAE,  $\mathbf{z}_0$  is not a point, but a distribution  $\mathcal{N}(\boldsymbol{\mu}_\phi, \boldsymbol{\Sigma}_\phi)$ .

This report introduces **LSI Score Distillation**, which substitutes the noisy target  $\epsilon$  with the exact analytic score of the diffused encoder posterior. We demonstrate that this substitution minimizes the variance of the training gradient and provides a stronger geometric signal to the encoder during joint training.

## 2 The Gaussian Field Assumption

In a VAE, the aggregate posterior (the "prior" for the diffusion model) is defined as:

$$q(\mathbf{z}) = \int p_{\text{data}}(\mathbf{x}) q_\phi(\mathbf{z}|\mathbf{x}) d\mathbf{x} \approx \frac{1}{N} \sum_{i=1}^N \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \quad (2)$$

where  $\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i$  are the encoder outputs for data point  $\mathbf{x}_i$ . We term this structure a **Gaussian Field**.

When this field undergoes the forward Ornstein-Uhlenbeck (OU) process defined by  $d\mathbf{z}_t = -\frac{1}{2}\beta(t)\mathbf{z}_t dt + \sqrt{\beta(t)}d\mathbf{w}_t$ , every individual component diffuses analytically. The transition parameters from  $t = 0$  to  $t$  are:

$$\alpha_t = e^{-\frac{1}{2} \int_0^t \beta(s) ds} \quad (3)$$

$$\sigma_t^2 = 1 - \alpha_t^2 \quad (4)$$

Crucially, the diffused posterior for a specific data point  $\mathbf{x}$  remains Gaussian:

$$q(\mathbf{z}_t | \mathbf{x}) = \mathcal{N}(\mathbf{z}_t; \boldsymbol{\mu}_t(\mathbf{x}), \boldsymbol{\Sigma}_t(\mathbf{x})) \quad (5)$$

where  $\boldsymbol{\mu}_t(\mathbf{x}) = \alpha_t \boldsymbol{\mu}_\phi(\mathbf{x})$  and  $\boldsymbol{\Sigma}_t(\mathbf{x}) = \alpha_t^2 \boldsymbol{\Sigma}_\phi(\mathbf{x}) + \sigma_t^2 \mathbf{I}$ .

### 3 Latent Score Identity and Score Regularity

We consider a latent representation induced by a variational encoder with diagonal covariance,

$$q_\phi(z | x) = \mathcal{N}(z; \mu_\phi(x), \Sigma_\phi(x)), \quad \Sigma_\phi(x) = \text{diag}(\sigma_{\phi,i}^2(x)), \quad (6)$$

and define the aggregate latent distribution

$$p_0(z) = \int q_\phi(z | x) p_{\text{data}}(x) dx, \quad (7)$$

which forms a *Gaussian field*, i.e., a mixture of Gaussians indexed by data points.

We evolve this field under an Ornstein–Uhlenbeck (OU) forward diffusion,

$$z_t = \alpha_t z_0 + \sigma_t \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, I), \quad (8)$$

which preserves Gaussianity at the component level. Each component evolves as

$$p(z_t | x) = \mathcal{N}(z_t; \mu_t(x), \Sigma_t(x)), \quad \mu_t(x) = \alpha_t \mu_\phi(x), \quad \Sigma_t(x) = \alpha_t^2 \Sigma_\phi(x) + \sigma_t^2 \mathbf{I}. \quad (9)$$

#### 3.1 Latent Score Identity

The score of the marginal latent distribution  $p_t(z)$  admits the Latent Score Identity (LSI),

$$s^*(z_t, t) = \nabla_{z_t} \log p_t(z_t) = \mathbb{E}_{x \sim p(x | z_t, t)} [-\Sigma_t(x)^{-1}(z_t - \mu_t(x))], \quad (10)$$

where the posterior over mixture components is

$$p(x | z_t, t) \propto p(z_t | x) p_{\text{data}}(x). \quad (11)$$

Equation (10) expresses the true score as the conditional expectation of component-wise linear scores. This identity yields an unbiased, minimum-variance training target for score estimation in Gaussian fields.

#### 3.2 Irreducible Variance of LSI

Let

$$s_x(z_t, t) := -\Sigma_t(x)^{-1}(z_t - \mu_t(x)). \quad (12)$$

For any estimator  $s_\theta(z_t, t)$ , the population LSI objective satisfies

$$\mathbb{E}_{x|z_t,t} \|s_\theta - s_x\|^2 = \|s_\theta - s^*\|^2 + \text{tr}(\text{Cov}(s_x | z_t, t)). \quad (13)$$

When  $s_\theta = s^*$ , the residual equals

$$\mathcal{V}_{\text{LSI}}(z_t, t) = \text{tr}(\text{Cov}(s_x | z_t, t)), \quad (14)$$

which we refer to as the *irreducible LSI variance*. This quantity measures local mixture disagreement and vanishes only when the latent field is locally unimodal.

#### 3.3 Score Jacobian and Lipschitz Regularity

The Jacobian of the true score field admits the exact decomposition

$$\nabla_{z_t} s^*(z_t, t) = -\mathbb{E}_{x|z_t,t} [\Sigma_t(x)^{-1}] + \text{Cov}(s_x | z_t, t). \quad (15)$$

Taking operator norms and using positivity of both terms,

$$\|\nabla_{z_t} s^*(z_t, t)\|_{\text{op}} \leq \text{tr}(\mathbb{E}[\Sigma_t(x)^{-1} | z_t, t]) + \text{tr}(\text{Cov}(s_x | z_t, t)). \quad (16)$$

Equation (16) provides a pointwise upper bound on the Lipschitz constant of the score field. Importantly:

- The second term coincides exactly with the irreducible LSI variance (14).
- The first term captures *stiffness* induced by small eigenvalues of the diffused covariance  $\Sigma_t(x)$ .

Thus, the same quantities that govern the statistical difficulty of score estimation also control the dynamical stiffness of the reverse-time flow.

### 3.4 Implications for Integrability and Robust Decoding

The bound (16) has two immediate consequences.

First, it controls *few-step integrability*: numerical error of any reverse-time integrator scales with  $\|\nabla s^*\|$ , implying that large irreducible variance or small eigenvalues of  $\Sigma_t$  necessitate prohibitively small step sizes.

Second, it controls *robustness of decoding*. Errors induced by imperfect score estimation or finite-step integration perturb the terminal latent distribution. When the score field is stiff, these perturbations are amplified, and nonlinear decoders trained on thin latent supports exhibit severe sensitivity. By contrast, controlling both terms in (16) yields latent representations whose decoding remains stable under transport error.

This observation motivates jointly shaping the latent representation to balance reconstruction fidelity against score regularity.

## 4 Joint Training with Score Regularity

We jointly train the encoder-decoder pair  $(E_\phi, D_\psi)$  and a score network  $s_\theta$  using objectives that balance reconstruction quality with statistical and dynamical regularity of the induced score field.

### 4.1 Objectives

The joint loss consists of four components:

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda_{\text{perc}} \mathcal{L}_{\text{perc}} + \lambda_{\text{KL}} \mathcal{L}_{\text{KL}} + \lambda_{\text{LSI}} \mathcal{L}_{\text{LSI}} + \lambda_{\text{stiff}} \mathcal{L}_{\text{stiff}}. \quad (17)$$

**Reconstruction and perceptual loss.**

$$\mathcal{L}_{\text{rec}} = \mathbb{E}_x \|D_\psi(E_\phi(x)) - x\|^2, \quad (18)$$

with an optional perceptual term  $\mathcal{L}_{\text{perc}}$ .

**Latent energy regularization.** We employ a modified KL term,

$$\mathcal{L}_{\text{KL}} = \frac{1}{2} \mathbb{E}_x [\|\mu_\phi(x)\|^2 + \text{tr}(\Sigma_\phi(x))], \quad (19)$$

which controls the aggregate second moment of the latent distribution without enforcing per-sample isotropy.

**LSI score matching.** The score network is trained using the LSI objective,

$$\mathcal{L}_{\text{LSI}} = \mathbb{E}_{t,x,z_t} \|s_\theta(z_t, t) + \Sigma_t(x)^{-1}(z_t - \mu_t(x))\|^2, \quad (20)$$

optionally weighted in time to interpolate between score- and noise-parameterizations.

**Stiffness regularization.** Motivated by the Lipschitz bound (16), we introduce a stiffness penalty

$$\mathcal{L}_{\text{stiff}} = \mathbb{E}_{t,x} [\text{tr}(\Sigma_t(x)^{-1})], \quad (21)$$

where  $\Sigma_t(x) = \alpha_t^2 \Sigma_\phi(x) + \sigma_t^2 I$ .

This term directly penalizes small eigenvalues of the diffused latent covariance, preventing unnecessary concentration of the latent representation and reducing dynamical stiffness of the reverse diffusion.

## 4.2 Interpretation

The combined effect of  $\mathcal{L}_{\text{LSI}}$  and  $\mathcal{L}_{\text{stiff}}$  is to control both terms in the Lipschitz bound (16):

- $\mathcal{L}_{\text{LSI}}$  reduces mixture-induced curvature by minimizing irreducible score variance.
- $\mathcal{L}_{\text{stiff}}$  prevents thin latent directions that would otherwise force small integration steps.

Together, these losses encourage latent Gaussian fields that are simultaneously:

1. statistically learnable (low-variance score targets),
2. dynamically integrable (bounded score Jacobian),
3. robust to transport error under decoding.

This joint training strategy replaces heuristic prior matching with a representation-aware regularization tailored to the generative transport.

## 4.3 Gradient Flow and Geometric Alignment

Crucially, we backpropagate  $\mathcal{L}_{\text{LSI}}$  into the \*\*Encoder parameters\*\*  $\phi$ . The LSI target depends explicitly on the encoder’s variance  $\Sigma_\phi(\mathbf{x})$ :

$$\text{Target} \propto \frac{1}{\alpha_t^2 \Sigma_\phi + \sigma_t^2} (\dots) \quad (22)$$

This provides a deterministic gradient path for the encoder to adjust its variance. The encoder learns to shape the latent “bubbles” such that their analytic curvature matches the vector field learned by the Score Network. This enforces a geometric consistency that standard DSM (which sees variance only implicitly via sampling noise) lacks.

---

### Algorithm 1 LSI Co-Training Loop

---

Given data batch  $\mathbf{x}$

1. Encoder:  $\mu, \log \Sigma \leftarrow E_\phi(\mathbf{x})$
  2. Reparameterize:  $\mathbf{z}_0 = \mu + \sqrt{\Sigma} \cdot \epsilon_0$
  3. Diffuse: Sample  $t, \epsilon$ . Compute  $\mathbf{z}_t = \alpha_t \mathbf{z}_0 + \sigma_t \epsilon$ .
  4. **Compute LSI Target:**  
 $\Sigma_t = \alpha_t^2 \exp(\log \Sigma) + \sigma_t^2 \mathbf{I}$   
 $\epsilon_{\text{LSI}} = \sigma_t \Sigma_t^{-1} (\mathbf{z}_t - \alpha_t \mu) \quad (\text{equivalently } \hat{s}_{\text{LSI}} = -\epsilon_{\text{LSI}} / \sigma_t)$
  5. Network Prediction:  $\hat{\epsilon}_\theta = \epsilon_\theta(\mathbf{z}_t, t)$
  6. Loss:  
 $\mathcal{L} = \mathcal{L}_{\text{rec}}(\mathbf{x}, D_\theta(\mathbf{z}_0)) + \lambda_{\text{KL}} \tilde{\mathcal{L}}_{\text{KL}} + \lambda_{\text{score}} \|\hat{\epsilon}_\theta - \epsilon_{\text{LSI}}\|^2$
  7. Update  $\phi, \theta$  via SGD.
- 

## 5 Refining LSGM: The Exact Connection

The Latent Score-based Generative Model (LSGM) [1] is the primary antecedent to this work. LSGM established that the diffusion loss upper-bounds the negative ELBO, motivating joint training. In our current implementation we deviate from a strict ELBO interpretation by using  $\tilde{\mathcal{L}}_{\text{KL}}$  (trace-only; no  $-\log \det$  term) as the prior regularizer; the encoder covariance scale is therefore determined primarily through the interaction between  $\mathcal{L}_{\text{LSI}}$  and this mild second-moment penalty.

**The Gap in LSGM:** In their implementation, Vahdat et al. utilized the standard Denoising Score Matching objective (Eq. 3 in their paper). While they derived the analytic cross-entropy in their Appendix, they relied on the stochastic estimator for training, likely due to the complexity of their hierarchical NVAE posterior.

**The LSI Refinement:** Our method can be viewed as \*\*LSGM with Rao-Blackwellization\*\*.

1. **Variance Reduction:** We replace the noisy  $\epsilon$  target with the deterministic  $\hat{s}_{\text{LSI}}$ . This removes the noise variance from the training signal, leading to faster convergence and stable gradients at small  $t$ .

Method	Recon FID ↓	FID (RK4@10) ↓	SW2 (RK4@10) ↓
(c) Two-stage Tweedie LDM (reg. KL)	6.51	13.45	0.002627
Two-stage LSI LDM (reg. KL)	6.30	12.18	0.001598
(a) LSI co-train + modified KL	9.33	9.39	0.000619
(b) Tweedie on LSI-shaped latent	9.33	9.26	0.000871
(d) Tweedie co-train + reg. KL	5.92	113.12	0.000090

Table 1: **Ablation summary (focus: RK4@10).** Co-training with LSI produces the best generative FID (9.26–9.39) and dramatically improves latent matching ( $SW2 \approx 6e-4$ – $9e-4$ ) relative to two-stage training ( $SW2 \approx 1.6e-3$ – $2.6e-3$ ). Tweedie co-training with regular KL yields extremely small SW2 but catastrophic FID, consistent with a latent-scale/variance collapse failure mode (see discussion).

2. **Explicit Variance Utilization:** LSGM encoders affect the score loss only by changing where samples  $z_0$  land. LSI encoders affect the score loss by explicitly changing the denominator of the target score. This creates a stronger feedback loop for variance estimation.
3. **The "Information" Argument:** LSGM effectively treats the latent space as a collection of Dirac deltas (samples) during the score matching step. LSI treats the latent space as a collection of Gaussian distributions. Since the VAE *is* a Gaussian model, LSI is the "structurally correct" estimator for this regime.

## 6 Results

### 6.1 Experimental setup and evaluation metrics

We evaluate five training regimes (two-stage baselines and co-training ablations) on FashionMNIST. All diffusion metrics below use the **probability-flow ODE** sampled with **RK4** using **10 steps**, and are computed on the full test set using fixed evaluation banks (fixed latent-noise bank for generation and fixed posterior-noise banks for aggregated-posterior sampling).

#### Metrics.

- **FID** (↓): image-space sample quality computed between generated samples and real test images.
- **SW2** (↓): sliced Wasserstein-2 distance in *latent space* between aggregated posterior samples  $z \sim q_\phi(z | x)$  and generated latents  $\hat{z} \sim p_\theta(z)$  (flattened); lower indicates a closer match of the *latent prior* to the encoder’s aggregated posterior.
- **Recon FID** (↓): FID of reconstructions obtained by sampling  $z \sim q_\phi(z | x)$  and decoding; this measures encoder/decoder reconstruction quality, *not* the quality of the learned prior.

### 6.2 Ablations

We compare:

- (c) **Two-stage Tweedie LDM**: train VAE (regular KL), then train diffusion prior in latent space with standard DSM/Tweedie noise prediction.
- **Two-stage LSI LDM**: same as (c) but train the prior using the analytic LSI target.
- (a) **LSI co-training**: jointly train VAE + prior with LSI supervision and the modified KL regularizer (trace/second-moment only).
- (b) **Tweedie on LSI-shaped latent**: use the co-trained VAE from (a) (same latent representation), but train the diffusion prior with Tweedie/DSM.
- (d) **Tweedie co-training + regular KL**: jointly train VAE + prior using Tweedie/DSM with gradients flowing to the encoder through  $z_t$ , while keeping the *regular* analytic KL.

### 6.3 Key findings and implications

- 1) **Representation dominates sampler quality.** Comparing (c) to (a)/(b), co-training improves FID by a large margin ( $13.45 \rightarrow 9.3$ -ish) while simultaneously reducing latent mismatch SW2 by roughly  $3\times\text{--}4\times$ . This supports the interpretation that **learning a diffusion-friendly latent representation** is the main lever on sampler quality in this regime.
- 2) **LSI is the right way to learn that representation.** Two-stage training already benefits from LSI at the score-head level (12.18 vs. 13.45 FID and SW2 0.001598 vs. 0.002627), but the large gain appears when LSI is used *as the co-training signal* (a), where it provides a stable, geometry-aware gradient path into the encoder variance via  $\Sigma_t^{-1}$ . This appears to be what enables the latent space to become both *more Gaussian-transport-friendly* and *more learnable by a low-step sampler*.
- 3) **Estimator quality matters, but is secondary here.** Holding the representation fixed (same VAE from (a)), swapping the score estimator from LSI to Tweedie yields only a small degradation in latent distributional metrics (SW2 0.000619  $\rightarrow$  0.000871) and essentially no degradation in FID (9.39 vs. 9.26, within noise). This suggests **the dominant improvement is representational**, with a smaller contribution from reduced-variance score supervision.
- 4) **Tweedie co-training (d) is brittle and can optimize the “wrong” latent objective.** Method (d) achieves the *best* latent SW2 (9e-5) yet the *worst* FID (113), producing partially structured but heavily corrupted samples. This decoupling (very low SW2 + terrible FID) is consistent with a failure mode where the joint objective encourages a pathological latent scaling (e.g., collapsing posterior variance / shrinking the latent support), making it easy for the diffusion prior to match the (collapsed) aggregated posterior in SW2 while breaking semantic alignment needed for image-space quality. In contrast, LSI co-training avoids this by using a deterministic, variance-aware target whose dependence on  $\Sigma_t$  makes encoder variance a *controlled* degree of freedom rather than an adversarial loophole.

**Compute note.** These results are obtained with a small latent (2 channels,  $8 \times 8$  spatial) and a lightweight training budget (30 epochs;  $\sim$ 15 minutes on our hardware), whereas LSGM-scale setups typically train for hours on larger backbones. Despite the small budget, LSI co-training reaches strong FID in this controlled setting, indicating that the signal is highly sample-efficient.

## 7 Conclusion

By acknowledging that the latent space of a VAE is a Gaussian Field, we replace the stochastic approximation of standard diffusion training with the exact analytic solution provided by the Laplace Score Identity. When applied to joint training, LSI Score Distillation acts as a high-fidelity geometric regularizer, strictly superseding the standard DSM objective used in prior works like LSGM by eliminating Monte Carlo variance from the supervision signal.

## References

- [1] Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. In *Advances in Neural Information Processing Systems*, volume 34, pages 11287–11302, 2021.