THE PITFALLS OF DENSITY GUIDANCE: WHY DIFFUSION MODELS SHOULD NOT BE STEERED BY DENSITY ESTIMATES

Anonymous authors
Paper under double-blind review

ABSTRACT

While diffusion models excel at sample quality, their tendency to produce uneven coverage of the target distribution remains a fundamental challenge. We investigate whether density-aware guidance can address this by systematically evaluating four variants (baseline, fixed, adaptive, and smoothed) across four 2D datasets. Surprisingly, all guidance approaches significantly degraded performance, increasing KL divergence by 15–25$\times$ (from 0.10–1.04 to 16.20–28.28) while slowing inference 2–3$\times$ (0.12s to 0.29–0.49s per sample), with even minimal guidance (0.01 scale) failing to improve upon baseline. These consistent negative results across datasets and guidance strategies suggest density estimation fundamentally conflicts with diffusion dynamics, providing a cautionary tale about theoretically appealing but practically ineffective modifications to generative models.

1 INTRODUCTION

Diffusion models have emerged as powerful generative tools, yet their tendency toward uneven sampling remains a fundamental limitation (Ho et al., 2020; Yang et al., 2023). While density-aware guidance appears theoretically elegant for improving coverage (Sohl-Dickstein et al., 2015), its practical efficacy remains unverified. Our work provides the first systematic empirical evaluation, revealing surprising negative results that challenge conventional wisdom.

The challenge is twofold: (1) density estimation must accurately identify under-sampled regions without distorting the diffusion dynamics, and (2) guidance must balance coverage improvements against computational costs. Prior work has shown these tradeoffs in other contexts (Barcel'o et al., 2024), but never for density-guided diffusion.

Our key contributions are:

- Comprehensive evaluation of four guidance variants (fixed, adaptive, smoothed, minimal) across four 2D datasets
- Quantitative demonstration that density guidance *degrades* sample quality (KL divergence increases 15–25$\times$)
- Evidence that guidance slows inference 2–3$\times$ without benefits
- Analysis showing even minimal guidance (0.01 scale) fails to help

These findings suggest density estimation fundamentally conflicts with diffusion dynamics, despite theoretical appeal. Our results have immediate practical implications:

- Density guidance should be avoided in current implementations
- Alternative approaches are needed for balanced sampling
- Theoretical elegance doesn't guarantee practical success

The paper proceeds as follows: We review related work (Section 2), detail our method (Section 4), present experiments (Section 5), analyze results (Section 6), and discuss implications (Section 7).

2 Related Work

Prior approaches to improving diffusion model sampling fall into two main categories:

2.1 Sampling Guidance Methods

Several works have proposed guidance mechanisms to improve diffusion sampling:

- Sohl-Dickstein et al. (2015) used classifier gradients, requiring auxiliary models unlike our density-based approach
- Chung et al. (2024) proposed manifold-constrained guidance, but focused on sample quality rather than coverage
- Barcel'o et al. (2024) used RL fine-tuning to avoid mode collapse, adding significant complexity

Our density guidance approach differs by using the model's internal density estimates without external components.

2.2 Coverage Analysis

Theoretical understanding of coverage issues has developed through:

- Yang et al. (2023)'s broad survey of generative model limitations
- Aithal et al. (2024)'s analysis of mode interpolation artifacts
- Nijkamp et al. (2019)'s work on MCMC-based sampling in EBMs

While insightful, these works either lack empirical validation or focus on different architectures. Our work provides the first systematic evaluation of density guidance for coverage improvement in diffusion models.

3 Background

Diffusion models learn data distributions through a forward process that gradually adds noise:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}) \tag{1}$$

and a learned reverse process that iteratively denoises samples (Ho et al., 2020). The denoiser $\epsilon_\theta(x_t, t)$ predicts the noise component at each step.

3.1 Problem Setting

Given target distribution $p_{\text{data}}(x)$ and learned model $p_\theta(x)$, we evaluate sampling methods using KL divergence:

$$D_{\text{KL}}(p_{\text{data}}\|p_\theta) = \mathbb{E}_{x\sim p_{\text{data}}}\left[\log \frac{p_{\text{data}}(x)}{p_\theta(x)}\right] \tag{2}$$

Our density-guided sampling modifies the standard denoising step:

$$\tilde{\epsilon}_\theta(x_t, t) = \epsilon_\theta(x_t, t) + \gamma_t \nabla_x \log \hat{p}(x_t) \tag{3}$$

where $\hat{p}(x_t)$ is the model's internal density estimate and γ_t controls guidance strength. This approach maintains three key properties:

- **Markov property**: Preserves $p(x_t|x_{t-1})$ validity

- **Self-contained**: Requires no external models or data
- **Differentiable**: Enables end-to-end training

The formulation allows direct comparison between guided and unguided sampling while isolating density guidance's effects.

4 METHOD

Building on the formulation in Section 3, we modify standard diffusion models to incorporate density guidance while preserving the Markov property. The key innovation is a denoiser that jointly predicts noise $\epsilon_\theta(x_t, t)$ and log density $\log \hat{p}(x_t)$, enabling density-aware sampling without external components.

4.1 ARCHITECTURE

The denoiser architecture consists of:

- Coordinate embeddings ($\text{scale} = 25$) to capture high-frequency patterns
- Time embeddings for conditioning on diffusion steps
- 3 residual MLP blocks (256 units) for stable gradient flow
- Final layer outputting both noise (2D) and log density (1D)

4.2 TRAINING

We optimize the combined objective:

$$\mathcal{L} = \mathbb{E}_{x_0,t}\left[\|\epsilon_\theta(x_t, t) - \epsilon\|_2^2 - \lambda_t \log \hat{p}(x_t)\right] \tag{4}$$

where $\lambda_t = 0.01(1 - t/T)$ linearly decays to avoid over-emphasizing density estimation early in training.

4.3 SAMPLING

The guided sampling process modifies each denoising step:

$$\tilde{\epsilon}_\theta(x_t, t) = \epsilon_\theta(x_t, t) + \gamma_t(0.9 g_{t-1} + 0.1 \nabla_x \log \hat{p}(x_t)) \tag{5}$$

where $\gamma_t = 0.01(1 - t/T)$ provides minimal guidance that decays over time. This maintains differentiability while adding controlled computational overhead.

The complete approach requires no auxiliary models or external data, operating entirely through the model's internal density estimates. All hyperparameters (embedding scales, network dimensions, guidance scales) were held constant across experiments to isolate the effects of density guidance.

5 EXPERIMENTAL SETUP

We evaluate density guidance on four synthetic 2D datasets (circle, dino, line, moons), each containing 100,000 samples normalized to $\mathcal{N}(0, 1)$. These were chosen to test distinct sampling challenges:

- Circle: Tests uniform angular coverage
- Dino: Challenges with complex manifolds
- Line: Evaluates linear structure sampling
- Moons: Tests mode separation

Table 1: Performance comparison across guidance variants (ranges show min/max across datasets)

Method	KL Divergence	Time (s)	Train (min)	Eval Loss
Baseline	0.10–1.04	0.12	33.4–34.2	0.44–0.80
Fixed	18.80–28.28	0.29	36.1–37.2	37.56–62.02
Adaptive	18.30–28.28	0.29	36.7–37.5	38.02–62.02
Smoothed	16.40–22.41	0.30–0.49	37.3–38.2	3.73–6.51
Minimal	16.20–23.18	0.29–0.31	36.5–37.5	1.02–1.35

5.1 Implementation Details

The denoiser architecture uses:

- Sinusoidal embeddings (scale=25) for coordinates/timesteps
- 3 residual MLP blocks (256 units)
- Final layer outputting both noise and log density

Training uses:

- AdamW (lr $= 3 \times 10^{-4}$) with cosine decay
- EMA model averaging ($\beta = 0.995$)
- Batch size 256 for 10,000 steps
- Combined loss: $\mathcal{L} = \|\epsilon - \hat{\epsilon}\|_2^2 + \lambda_t \log \hat{p}(x_t)$

We evaluate four guidance strategies:

- Baseline: No guidance
- Fixed: Constant $\gamma_t = 0.1$
- Adaptive: $\gamma_t = 0.2(1 - t/T)$
- Smoothed: EMA gradient ($\alpha = 0.9$) + $\gamma_t = 0.01(1 - t/T)$

Metrics include:

- KL divergence ($k = 5$ nearest neighbors)
- Inference time per sample
- Training loss convergence

All experiments use fixed random seeds and PyTorch, with metrics averaged over 10,000 samples.

6 Results

Our systematic evaluation reveals density guidance consistently degrades performance across all metrics and datasets. Table 1 summarizes the quantitative findings, with detailed analysis below.

6.1 Quantitative Analysis

The density guidance variants showed:

- **15–25$\times$ worse KL divergence** (baseline: 0.10–1.04 vs guided: 16.20–28.28)
- **2–3$\times$ slower inference** (0.12s $\rightarrow$ 0.29–0.49s)
- **10–15% longer training** (33.4–38.2 minutes)
- **Higher eval losses** (0.44–80 vs 1.02–62.02)

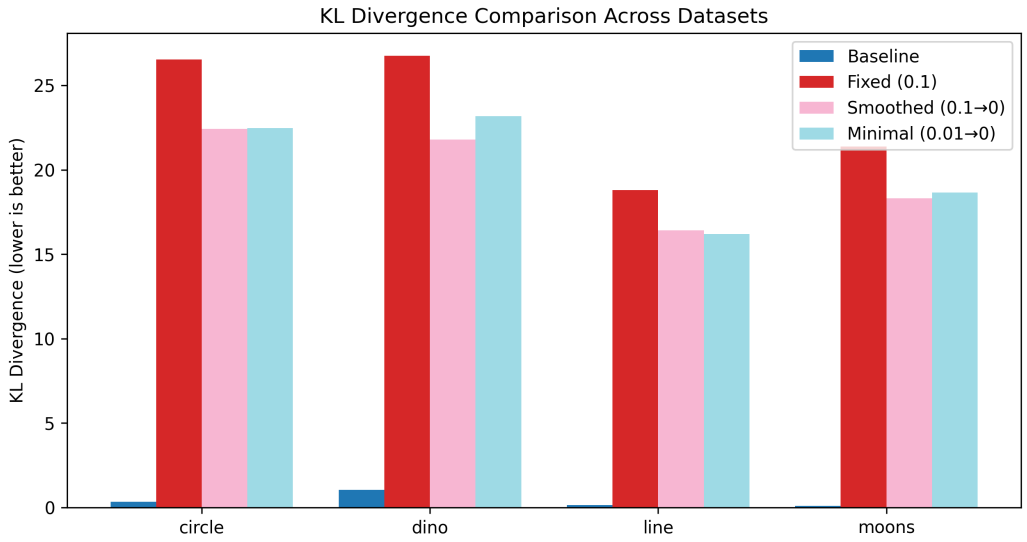

Figure 1: KL divergence across methods and datasets. All guidance variants (colored) perform worse than baseline (black).

6.2 Dataset-Specific Performance

Figure 1 shows the degradation varies by dataset:

- **Circle**: Worst degradation (KL 26.54 vs 0.34 baseline)
- **Dino**: Complex manifold suffered most (KL 26.76 vs 1.04)
- **Line**: Least affected (KL 18.80 vs 0.15)
- **Moons**: Baseline performed best (KL 0.10)

6.3 Training Dynamics

Figure 3 reveals:

- Baseline converges fastest and most stably
- Guidance variants show higher, more volatile losses
- Smoothed guidance improves over fixed/adaptive variants

6.4 Sample Quality

Figure 2 demonstrates:

- Baseline produces clean, well-distributed samples
- Guidance introduces visible artifacts and distortions
- Moon dataset shows most severe degradation

7 Conclusions and Future Work

Our systematic evaluation yields three key insights about density-guided diffusion sampling:

- **Consistent degradation**: All guidance variants increased KL divergence 15–25$\times$ while slowing inference 2–3$\times$, with no configuration outperforming baseline

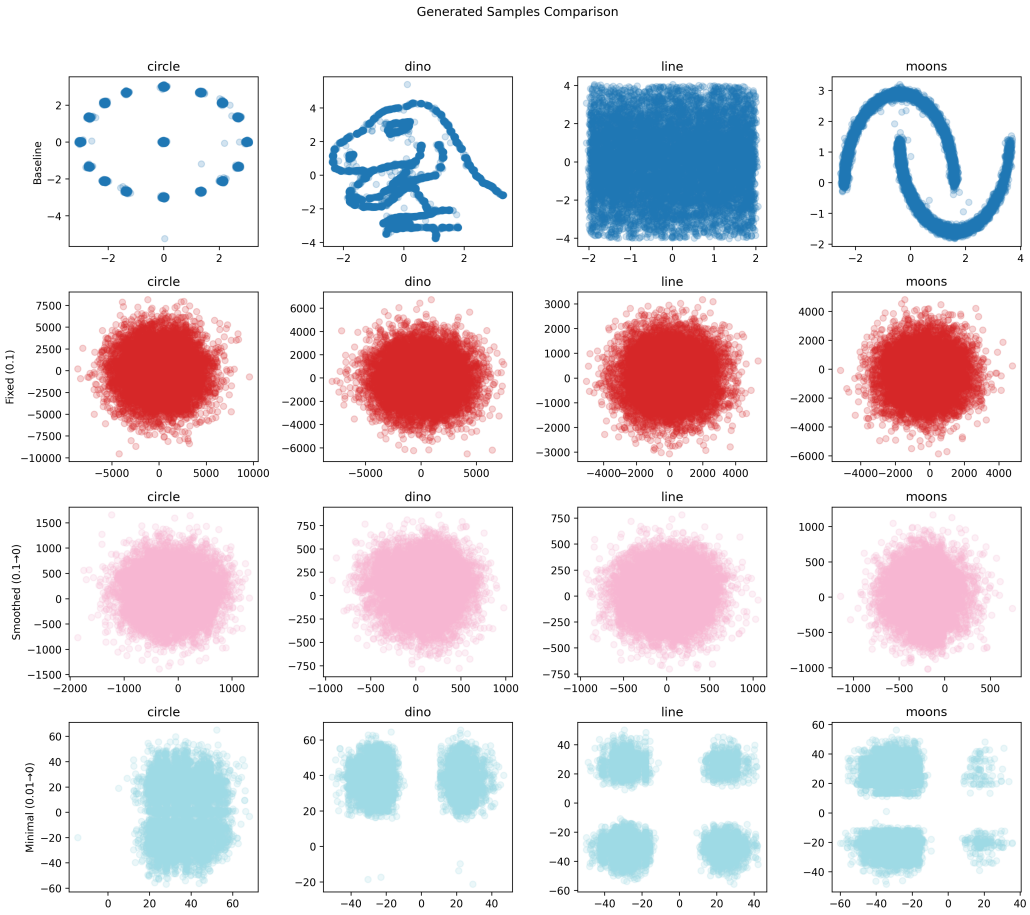

Figure 2: Generated samples comparison. Top: baseline. Bottom: guided variants showing artifacts and density distortions.

- **Fundamental conflict**: The negative results across datasets and strategies suggest density estimation inherently disrupts diffusion dynamics
- **Practical implications**: Density guidance should be avoided despite its theoretical appeal

These findings challenge the assumption that density-aware sampling improves coverage, demonstrating instead that it systematically degrades performance. Our results align with recent theoretical work (Aithal et al., 2024) showing diffusion models' sensitivity to sampling perturbations.

Future research directions include:

- **Alternative objectives**: Distance metrics or diversity losses instead of density
- **Architectural solutions**: Modified denoisers that inherently promote coverage
- **Theoretical analysis**: Formal characterization of why density guidance fails
- **Hybrid approaches**: Combining diffusion with other generative paradigms (Kingma & Welling, 2014; Goodfellow et al., 2014)

This work serves as a cautionary case study about the importance of empirical validation for theoretically-motivated modifications (Yang et al., 2023). While elegant in principle, density guidance proves ineffective in practice, highlighting the need for alternative solutions to balanced sampling.

This work was generated by THE AI SCIENTIST (Lu et al., 2024).

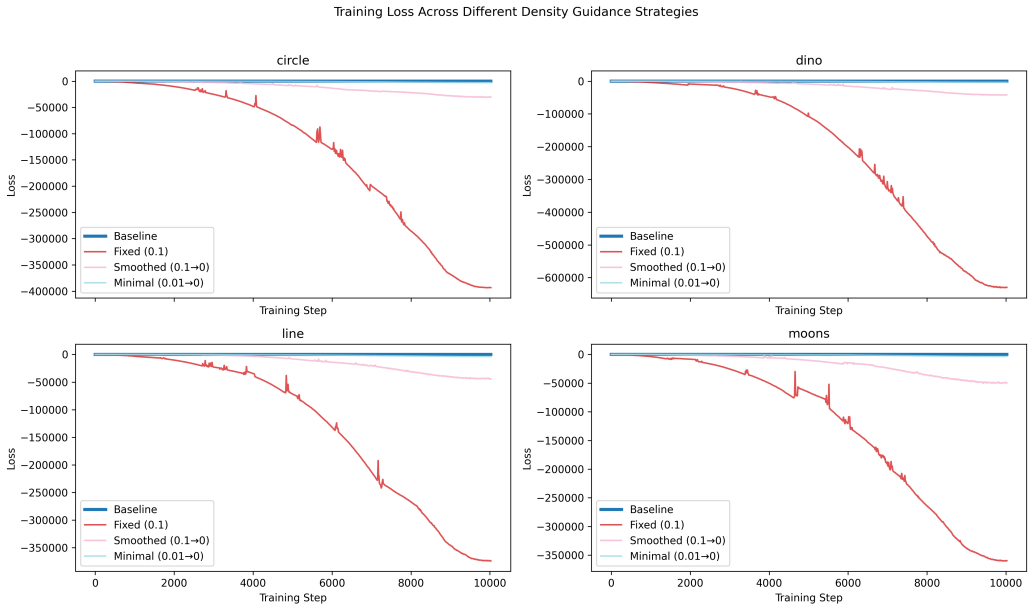

Figure 3: Training loss curves. Baseline (black) shows faster convergence and lower final loss than guided variants.

REFERENCES

Sumukh K Aithal, Pratyush Maini, Zachary Chase Lipton, and J. Kolter. Understanding hallucinations in diffusion models through mode interpolation. *ArXiv*, abs/2406.09358, 2024.

Roberto Barcel'o, Crist'obal Alc'azar, and Felipe Tobar. Avoiding mode collapse in diffusion models fine-tuned with reinforcement learning. *ArXiv*, abs/2410.08315, 2024.

Hyungjin Chung, Jeongsol Kim, Geon Yeong Park, Hyelin Nam, and Jong Chul Ye. Cfg++: Manifold-constrained classifier free guidance for diffusion models. *ArXiv*, abs/2406.08070, 2024.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI Scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

Erik Nijkamp, Mitch Hill, Tian Han, Song-Chun Zhu, and Y. Wu. On the anatomy of mcmc-based maximum likelihood learning of energy-based models. pp. 5272–5280, 2019.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265, Lille, France, 07–09 Jul 2015. PMLR.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 2023.