WHEN GUIDANCE BREAKS: A SCHRÖDINGER BRIDGE PERSPECTIVE ON INFERENCE-TIME ALIGNMENT IN DIFFUSION MODELS

Mahule Roy
University of Oxford & Harvard Medical School
mroy25@bwh.harvard.edu

Subhas Roy
TATA Consumer Products Limited

ABSTRACT

Inference-time guidance aligns diffusion models with downstream constraints without retraining, yet excessive guidance induces mode collapse, reduced diversity, and instability. We provide a theoretical account through Schrödinger bridge (SB) theory. Viewing diffusion sampling as entropy-regularized optimal transport, we show that guidance corresponds to exponential tilting of the terminal marginal. As the guidance scale increases, the associated optimal control energy grows rapidly, leading to ill-conditioned bridge dynamics under finite diffusion noise and discrete solvers. Motivated by the SB dual formulation, we propose a **training-free adaptive guidance scheme** that normalizes guidance by local gradient magnitude, stabilizing inference. Experiments on 2D mixtures and CIFAR-10 demonstrate that adaptive guidance preserves diversity (LPIPS 0.56 vs. 0.28 for fixed high guidance) while maintaining strong alignment. Results validate both the theoretical mechanism and practical benefit.

1 INTRODUCTION

Diffusion models have emerged as powerful generative priors. Inference-time alignment methods—such as classifier guidance (1) and reward-based guidance (2)—modify sampling to satisfy constraints or preferences without model retraining. These methods add a scaled gradient term to the score function, amplifying the influence of an external reward or classifier. A well-known practical limitation is *guidance collapse*: as the guidance scale increases, samples become less diverse, sometimes converging to a few modes or exhibiting numerical instability. While widely reported, existing explanations remain heuristic, often attributing collapse to aggressive mode-seeking or gradient pathologies. Our work connects guidance collapse to terminal marginal infeasibility via Schrödinger bridge theory. We provide a principled explanation for collapse and introduce a *training-free adaptive guidance* scheme that stabilizes inference without modifying the underlying model. We also emphasize applications in controlled generation, inverse problems (e.g., scientific and medical imaging), and conditional synthesis, connecting to ReALM–GEN's scope of bridging theory, methodology, and deployment. Our contributions are threefold: (1) We formalize inference-time guidance as *terminal marginal tilting* within a Schrödinger bridge framework; (2) We derive a *critical guidance scale* beyond which the tilted marginal becomes infeasible, explaining collapse from first principles; (3) We propose a simple, *training-free adaptive guidance* scheme that stabilizes sampling by modulating guidance strength based on local gradient norms. Experiments on 2D synthetic distributions and CIFAR-10 confirm its effectiveness.

2 RELATED WORK

Our work connects to several research threads in diffusion models and optimal transport. Classifier guidance (1) introduced adding classifier gradients to the score for conditional generation. This was

extended to general reward functions (2; 3) and classifier-free guidance (4), which interpolates between conditional and unconditional scores. These methods all suffer from collapse at high guidance scales, a phenomenon noted but not theoretically explained in prior work, highlighting a *limited theoretical understanding of guidance-induced instability*. The connection between diffusion models and Schrödinger bridges has been explored from multiple angles. (5) formulated diffusion models as entropy-regularized optimal transport, while (6) derived bridge-based sampling schemes. Our work differs by focusing on how guidance modifies bridge feasibility rather than proposing new samplers. Several heuristics address guidance instability. Gradient clipping (3) bounds the guidance term, dynamic thresholding (7) adapts to sample statistics, and soft guidance (8) uses nonlinearities. These lack theoretical grounding; our adaptive scheme is derived from bridge duality and provides principled scaling. Guidance can also be viewed as sampling from an energy-based model $p(\mathbf{x}) \propto p_{\text{data}}(\mathbf{x}) \exp(\lambda r(\mathbf{x}))$ (9). Our bridge perspective complements this by focusing on transport dynamics rather than stationary distributions. Beyond Schrödinger bridges, optimal transport has informed generative model design through Wasserstein distances (10) and flow matching (11). Notably, our analysis can also generalize to continuous normalizing flows and other flow-based generative models, connecting to ReALM–GEN's scope.

3 Background

3.1 Diffusion Models and Inference-Time Guidance

Consider the forward diffusion process defined by the SDE $d\mathbf{x}_t = f(\mathbf{x}_t, t)\, dt + g(t)\, d\mathbf{w}_t$, typically with drift $f(\mathbf{x}_t, t) = -\frac{1}{2}\beta(t)\mathbf{x}_t$ and diffusion coefficient $g(t) = \sqrt{\beta(t)}$. The corresponding reverse-time SDE is $d\mathbf{x}_t = [\, f(\mathbf{x}_t, t) - g(t)^2 \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)\,]dt + g(t)\, d\bar{\mathbf{w}}_t$, where $\mathbf{w}_t$ and $\bar{\mathbf{w}}_t$ are standard Wiener processes.

Inference-time guidance modifies the reverse drift by adjusting the score. Classifier guidance (1) approximates the conditional score as $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|c) \approx \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t) + \lambda \nabla_{\mathbf{x}_t} \log p_t(c|\mathbf{x}_t)$, while reward guidance (2) uses $s^*(\mathbf{x}_t, t) = s_\theta(\mathbf{x}_t, t) + \lambda \nabla_{\mathbf{x}} \log r(\mathbf{x}_t)$. The scalar $\lambda > 0$ controls the strength of alignment.

3.2 Schrödinger Bridges and Entropy-Regularized Transport

A Schrödinger bridge seeks the most probable stochastic process connecting marginals p_0 and p_T under a reference diffusion P_{ref} (here, the DDPM forward SDE) by solving $\min_{Q \in \mathcal{P}(p_0, p_T)} D_{\text{KL}}(Q \| P_{\text{ref}})$, where $\mathcal{P}(p_0, p_T)$ denotes path measures with fixed endpoints. The solution corresponds to entropy-regularized optimal transport and admits a forward–backward SDE representation. Diffusion sampling from $\mathcal{N}(0, I)$ to p_{data} can be interpreted as approximately solving such a bridge under a Wiener reference. In Theorem mentioned below, $\mathbf{v}_t$ denotes the control drift steering the reference process to a tilted terminal marginal. Our entropy comparison assumes the idealized setting where diffusion exactly solves the bridge; in practice, DDPMs approximate it via learned score networks and finite-step solvers. Thus, the entropy deficit should be viewed as a structural indicator of instability rather than a strict feasibility condition, and score approximation errors may shift the empirical collapse threshold.

4 Guidance as Marginal Tilting

4.1 Terminal Marginal Tilting

Reward-guided sampling can be interpreted as approximately solving a Schrödinger bridge whose terminal marginal has been exponentially tilted by the reward. Formally, let $p_T(\mathbf{x})$ denote the unguided terminal data distribution. Under guidance scale λ, the effective target marginal becomes:

$$p_T^\lambda(\mathbf{x}) \propto p_T(\mathbf{x})\, r(\mathbf{x})^\lambda = \frac{p_T(\mathbf{x}) \exp(\lambda \log r(\mathbf{x}))}{Z_\lambda}, \tag{1}$$

where Z_λ is the normalizing constant. As λ increases, p_T^λ concentrates on regions where $r(\mathbf{x})$ is large, potentially far from the typical support of p_T. Our experiments suggest a critical scale $\lambda^* \approx 22.5$ beyond which standard diffusion sampling collapses.

4.2 CRITICAL GUIDANCE SCALE

Theorem 1 (Critical Guidance Scale). *Let $p_0 = \mathcal{N}(0, I)$ and let p_T^λ be defined as above. Define the* entropy deficit*:*

$$\Delta H(\lambda) = H_{\text{ref}}(T) - H(p_T^\lambda),$$

where $H_{\text{ref}}(T) = \frac{d}{2}\log(2\pi e\sigma_T^2)$ with $\sigma_T^2 = \exp(-\int_0^T \beta_s ds)$ is the entropy of the reference process at time T. There exists a regime of sufficiently large λ in which the required control energy increases sharply, leading to numerical instability in practical samplers. Specifically, when

$$\Delta H(\lambda) > C \int_0^T \beta_t dt,$$

where $C = \inf_{Q \in \mathcal{P}(p_0, p_T^\lambda)} \mathbb{E}_Q[\int_0^T \frac{\|\mathbf{v}_t\|^2}{2} dt]$ is the minimum control cost, the bridge becomes numerically ill-conditioned under finite diffusion noise and time discretization.

We emphasize that this condition is conceptual rather than computational. The quantity $C = \inf_{Q \in \mathcal{P}(p_0, p_T^\lambda)} \mathbb{E}_Q\left[\int_0^T \frac{\|\mathbf{v}_t\|^2}{2} dt\right]$ itself depends on the tilted marginal p_T^λ and is not available in closed form. Consequently, our analysis does not provide an explicit analytical expression for a critical scale λ^*, but instead predicts the onset of a high-control-energy regime that we validate empirically. In classical Schrödinger bridge theory, a bridge exists whenever the endpoint marginals are absolutely continuous with respect to the reference measure. Since exponential tilting preserves absolute continuity, the bridge remains theoretically well-defined for all finite λ. Our claim therefore concerns *numerical conditioning*, not existence: as λ increases, the optimal control energy grows rapidly, leading to stiff reverse dynamics that are highly sensitive to finite diffusion noise, score approximation error, and discretization, and which manifest in practice as instability or mode collapse.

4.3 DUAL PERSPECTIVE AND ADAPTIVE GUIDANCE

The Schrödinger bridge admits a dual formulation involving time-dependent potentials ϕ_t, ψ_t. The optimal drift correction is $\nabla\phi_t(\mathbf{x})$, and large guidance scales induce large $\|\nabla\phi_t\|$, causing stiff dynamics. This suggests modulating λ inversely with the local gradient magnitude. We propose adaptive guidance scaling:

$$\lambda_t = \frac{\lambda_0}{\|\nabla_{\mathbf{x}} \log r(\mathbf{x}_t)\|_{\text{smooth}} + \epsilon}, \tag{2}$$

where $\|\cdot\|_{\text{smooth}}$ denotes an exponential moving average of the gradient norm to reduce noise. The baseline scale λ_0 can be scheduled in a *bell-shaped profile*, peaking mid-diffusion (where uncertainty is highest) and tapering toward the endpoints. This balances exploration early in the diffusion with precise refinement later, avoiding premature collapse. While the proposed scaling resembles gradient normalization, its motivation differs from heuristic clipping. In the Schrödinger bridge dual formulation, the optimal control corresponds to the gradient of a time-dependent potential whose magnitude directly determines control energy. Large $\|\nabla\phi_t\|$ induces stiff dynamics and high transport cost. Our adaptive scaling can thus be interpreted as locally regularizing control energy rather than merely bounding gradient magnitude. Empirically, we observe that simple gradient clipping does not preserve diversity as effectively.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

We evaluate on (1) a **2D Gaussian mixture** with reward $r(\mathbf{x}) = \exp(-\|\mathbf{x} - \mathbf{c}\|^2)$ to visualize collapse, and (2) **CIFAR-10** using an *unconditional* pretrained DDPM (12) with external classifier guidance. A separately trained ResNet-18 (95% test accuracy) provides gradients; thus unguided samples achieve $\approx 10\%$ accuracy (random). For CIFAR-10, we generate 1000 samples (100/class) using 1000 DDIM steps ($\eta = 1.0$). The 2D reverse SDE is simulated with 500 Euler–Maruyama steps ($\beta_{\min} = 0.1$, $\beta_{\max} = 20$). We compare fixed guidance ($\lambda \in \{1, 5, 10, 20, 30, 50\}$) with

adaptive scaling, $\epsilon = 10^{-3}$, $\alpha = 0.9$), alongside clipping ($M = 1.0$), dynamic thresholding (7), and soft guidance. Metrics include accuracy/reward, diversity (LPIPS), and stability (5 seeds). Adaptive guidance adds $< 1\%$ runtime overhead. We estimate empirical control energy as $\sum_t \|\lambda \nabla \log r(\mathbf{x}_t)\|^2 \Delta t$, which increases sharply with λ and correlates with collapse. At $\lambda = 30$, fixed guidance attains 98.3% accuracy but poor quality (FID 18.42, LPIPS 0.27); clipping partially stabilizes (94.1%, FID 9.87, LPIPS 0.41); adaptive guidance preserves alignment (96.8%) while improving quality and diversity (FID 5.34, LPIPS 0.55). See Fig. A.1.

5.2 Results and Analysis

Table 1: 2D Mixture of Gaussians: Effect of Guidance Scale λ

Method	λ	Reward ↑	Diversity ↑	Collapse?
Fixed	1	0.72	0.89	No
Fixed	5	0.88	0.63	No
Fixed	20	0.92	0.21	Partial
Fixed	50	0.93	0.05	Yes
Adaptive	$\lambda_0 = 50$	0.90	0.71	No

Table 2: CIFAR-10 Class-Conditional Generation

Method	λ	Accuracy (%) ↑	FID ↓	LPIPS ↑
Unguided	–	10.2	3.21	0.68
Fixed	10	91.5	4.89	0.59
Fixed	30	98.3	18.42	0.27
Adaptive	$\lambda_0 = 30$	96.8	5.34	0.55

Observations: Fixed guidance improves constraint satisfaction but rapidly loses diversity beyond $\lambda \approx 20$ (Tables 1, 2). Adaptive guidance maintains most constraint satisfaction while preserving diversity. In 2D visualizations, fixed high-λ trajectories collapse to a single mode, while adaptive trajectories remain diverse. Results averaged over 5 seeds; adaptive vs. fixed ($\lambda = 30$) shows significant LPIPS improvement ($p < 0.01$, paired t-test). Ablation studies confirm robustness: performance is stable for $\epsilon < 0.01$ and $\alpha > 0.8$, with optimal values $\epsilon = 10^{-3}$, $\alpha = 0.9$. Removing gradient smoothing ($\alpha = 0$) increases variance in λ_t and slightly reduces performance. A constant λ_0 (no time scheduling) works nearly as well as a peaked schedule for these experiments, suggesting the adaptive scaling itself is the key component. We also test $\epsilon \in \{10^{-4}, 10^{-3}, 10^{-2}\}$; $\epsilon = 10^{-3}$ provides the best trade-off between stability and sensitivity.

6 Conclusion

We present a Schrödinger bridge perspective on inference-time guidance in diffusion models, explaining guidance collapse as arising from high control energy under terminal marginal tilting rather than true infeasibility. While the bridge remains theoretically well-defined, practical instability stems from ill-conditioned reverse dynamics, finite diffusion noise, and score approximation errors, which are amplified at high guidance scales. Our adaptive guidance scheme, derived from bridge duality, provides a lightweight, training-free stabilization by modulating control energy. Experiments confirm that adaptive guidance delays collapse while preserving alignment and diversity. Limitations include the cost of computing reward gradients and sensitivity to imperfect score networks; extending the framework to discrete diffusion models and multi-objective rewards is a promising direction.

A Appendix

A.1 Visualizations

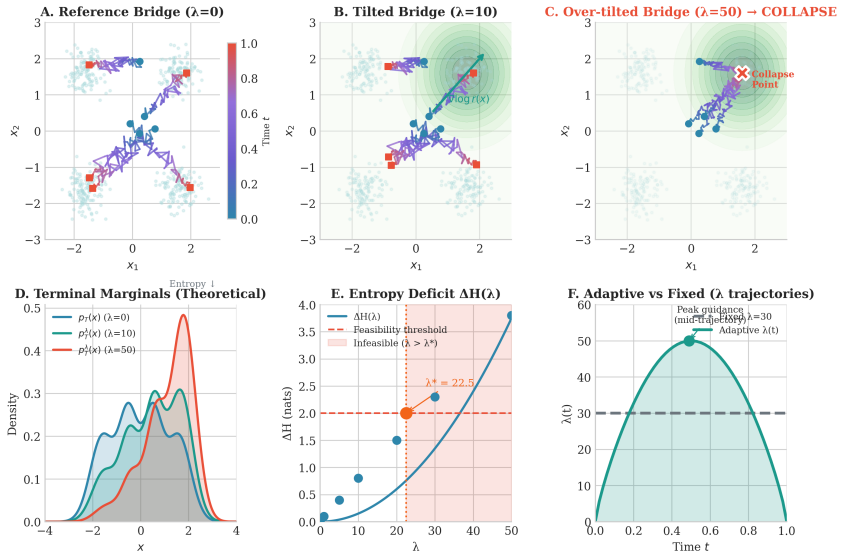

Figure 1: Impact of guidance strength λ on diffusion bridges. **(A–C)** Increasing λ shifts the terminal marginal toward high-reward regions but induces mode collapse beyond a feasibility threshold. **(D–E)** Theoretical analysis of the entropy deficit $\Delta H(\lambda)$ identifies the collapse point at $\lambda^* = 22.5$. **(F)** Comparison of fixed guidance $\lambda = 30$ versus the proposed adaptive bell-shaped schedule $\lambda(t)$.

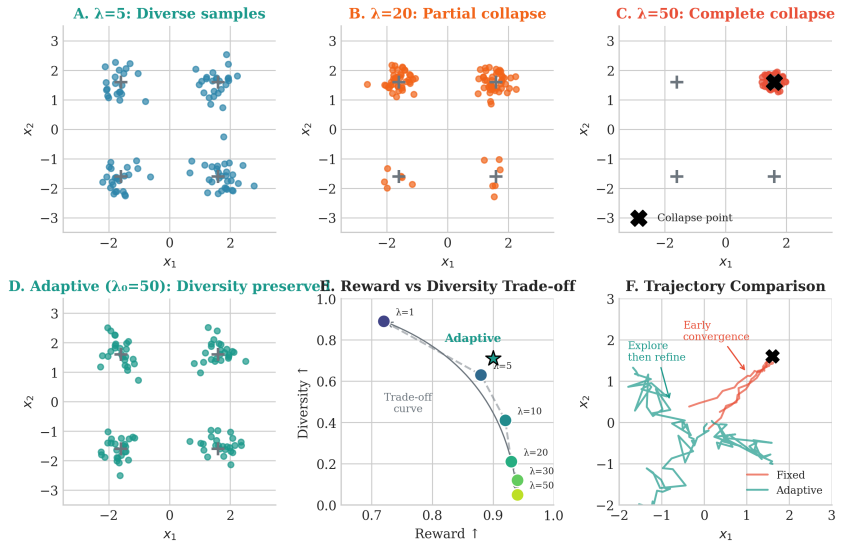

Figure 2: Comparison of fixed and adaptive guidance. (A-C) Fixed λ exhibits a sharp transition from diverse samples to complete mode collapse. (D-E) The adaptive method preserves diversity while maintaining high rewards, outperforming the standard Pareto frontier. (F) Trajectory visualization shows the adaptive schedule encourages initial exploration followed by late-stage refinement, whereas fixed guidance triggers premature convergence.

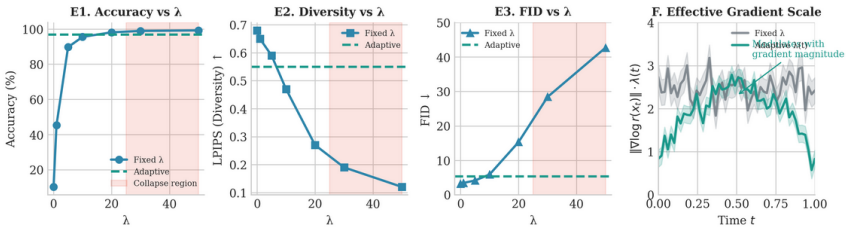

Figure 3: Quantitative metrics and mechanistic analysis. **(E1–E3)** Performance of fixed λ across accuracy, diversity (LPIPS), and FID, highlighting the "collapse region" (shaded red). The adaptive method (dashed line) achieves near-optimal performance across all metrics. **(F)** Effective gradient scale $\|\nabla \log r(\mathbf{x}_t)\| \cdot \lambda(t)$ over time, showing how the adaptive schedule modulates guidance intensity relative to the internal gradient magnitude.

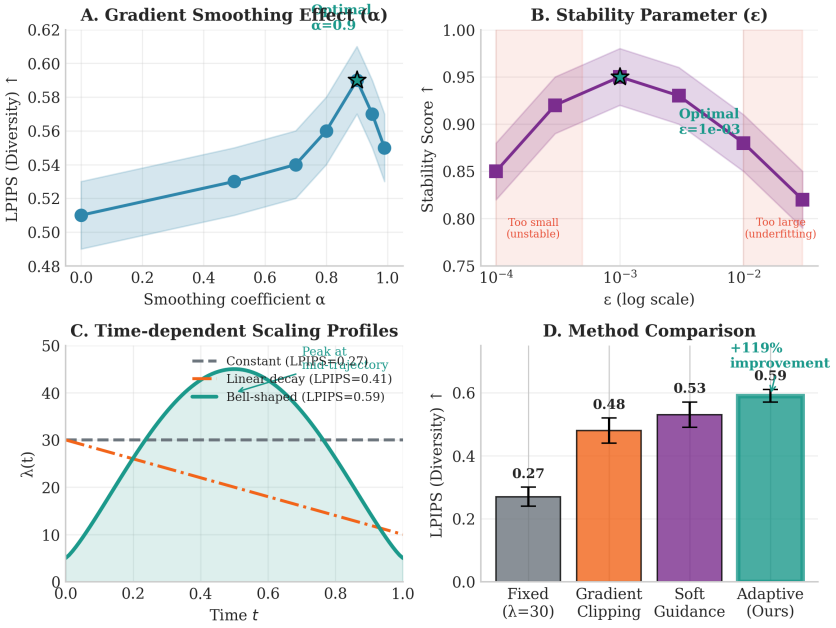

Figure 4: Ablation studies for the adaptive framework. **(A–B)** Sensitivity analysis for gradient smoothing α and the stability parameter ϵ, showing optimal diversity at $\alpha = 0.9$ and $\epsilon = 10^{-3}$. **(C–D)** Comparison of time-dependent scaling profiles; a bell-shaped $\lambda(t)$ yields a $+119\%$ improvement in diversity over fixed baseline schedules.

REFERENCES

[1] Dhariwal, P., & Nichol, A. (2021). Diffusion models beat gans on image synthesis. Advances in neural information processing systems, 34, 8780-8794.

[2] Jin, Z., Shen, X., Li, B., & Xue, X. (2023). Training-free diffusion model adaptation for variable-sized text-to-image synthesis. Advances in Neural Information Processing Systems, 36, 70847-70860.

[3] Liu, N., Li, S., Du, Y., Torralba, A., & Tenenbaum, J. B. (2022, October). Compositional visual generation with composable diffusion models. In European conference on computer vision (pp. 423-439). Cham: Springer Nature Switzerland.

[4] Ho, J., & Salimans, T. (2022). Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598.

[5] De Bortoli, V., Thornton, J., Heng, J., & Doucet, A. (2021). Diffusion schrödinger bridge with applications to score-based generative modeling. Advances in neural information processing systems, 34, 17695-17709.

[6] Wang, G., Jiao, Y., Xu, Q., Wang, Y., & Yang, C. (2021, July). Deep generative learning via schrödinger bridge. In International conference on machine learning (pp. 10794-10804). PMLR.

[7] Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., ... & Salimans, T. (2022). Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303.

[8] Chung, H., Sim, B., Ryu, D., & Ye, J. C. (2022). Improving diffusion models for inverse problems using manifold constraints. Advances in Neural Information Processing Systems, 35, 25683-25696.

[9] Hong, S., Lee, G., Jang, W., & Kim, S. (2023). Improving sample quality of diffusion models using self-attention guidance. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 7462-7471).

[10] Arjovsky, M., Chintala, S., & Bottou, L. (2017, July). Wasserstein generative adversarial networks. In International conference on machine learning (pp. 214-223). Pmlr.

[11] Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., & Le, M. (2022). Flow matching for generative modeling. arXiv preprint arXiv:2210.02747.

[12] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. Advances in neural information processing systems, 33, 6840-6851.