LET ATTENTION FLOW: ENERGY-DISSIPATING GUIDANCE FOR DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion models achieve remarkable generative performance by iteratively refining noisy samples through a sequence of denoising operations. An important component driving this refinement is the *attention mechanism* within the learned parameterization, which governs how information flows across spatial locations, semantic tokens and intermediate features. Existing training-free guidance methods often rely on score-space corrections or heuristic changes to internal activations, such as blurring attention maps. We instead guide diffusion by evolving attention through a Partial Differential Equation (PDE) inspired by Landau–de Gennes (LdG) dynamics. This energy-dissipating flow suppresses unstable attention patterns while preserving directional structure, providing a principled way to modulate the model's internal information flow without retraining. The resulting guidance integrates cleanly with modern text-to-image samplers, requires no architectural modification, and improves generation quality both qualitatively and quantitatively. On MS-COCO, our method improves over SDXL and prior attention-based guidance methods. Under 4-bit model quantization, it improves generation quality and outperforms the unguided FP16 SDXL baseline across *all* reported metrics.

1 INTRODUCTION

Diffusion models are now the default choice for image generation, from high-resolution text-to-image synthesis to controllable editing and image restoration (Ho et al., 2020; Song et al.; Karras et al., 2022; Rombach et al., 2022). We know that their overall performance depends not only on the learned score network, but also on how the reverse trajectory is guided at inference time. Classifier-free guidance (CFG) (Ho & Salimans, 2022) remains a dominant mechanism for improving fidelity and prompt alignment, yet recent work has emphasized that guidance is more than a fixed extrapolation rule: it can be interpreted as predictor-corrector sampling, trajectory control, or a training-free modification of the denoising dynamics (Bradley & Nakkiran, 2025; Sadat et al., 2025; Chen et al., 2025). This view is especially relevant for modern diffusion models, where users increasingly seek inference-time intervention or control that improves generation quality without retraining large models.

Together with score-space guidance, a growing body of work intervenes directly on internal representations of pre-trained diffusion networks. Self-attention guidance (SAG) (Hong et al., 2023), perturbed-attention guidance (PAG) (Ahn et al., 2024), smoothed energy guidance (SEG) (Hong, 2024), and more recent token-level perturbation methods (Rajabi et al., 2025) construct guidance signals by modifying features, attention maps, logits, or tokens during sampling. Separately, attention-control methods show that attention maps encode semantic alignment, object binding, and spatial organization in text-to-image models (Kim et al., 2025). These results point to a useful guiding principle: the denoising network already contains structured internal signals that can be shaped at *inference time*. However, most existing interventions choose the perturbation operator by trial-and-error such as blurring, masking, shuffling, or replacing attention. Our main goal is to ground these ideas in an energy formulation of the dynamics and derive the interventions from first principles.

The distinction above is important since the denoising network's attention module contains structured geometric information. The attention-weighted covariance of keys can be decomposed into isotropic and anisotropic components. A higher concentration of attention corresponds to strong *directional structure*, while diffuse attention corresponds to a more *isotropic state*. Since diffusion proceeds through many denoising steps, perturbing this structure meaningfully affects the trajectory itself.

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

Isotropic smoothing, including Gaussian blur used in prior attention-smoothing approaches (Hong, 2024), treat all directions uniformly and can attenuate useful anisotropic structure, producing over-regularization, flat texture, and weaker spatial coherence when guidance is strong.

From smoothing to energy flow. We therefore ask whether internal guidance can be formulated as a principled dynamics over attention. Physical systems with directional order, such as liquid crystals and nematic films, are commonly modeled by free-energy functionals whose gradient flows dissipate unstable configurations while preserving coherent anisotropic structure (De Gennes & Prost, 1993; Ball, 2017; Virga, 2018). This view also aligns with recent stochastic-thermodynamic perspectives on generative modeling, where free-energy dissipation and gradient-flow structure provide organizing principles for diffusion and related stochastic dynamics (Welling et al., 2026). Our goal is to bring this principle inside the denoising network: instead of guiding only in image or score space, we define an energy-dissipating flow directly over attention.

> *Can training-free diffusion guidance be improved by replacing heuristic attention perturbations with principled, energy-minimizing dynamics?*

This paper. We study the question by drawing on Landau–de Gennes (LdG) theory (De Gennes & Prost, 1993; Ball, 2017), a free-energy framework for fields with directional order. We treat the attention tensor as a spatially varying field and derive an LdG-inspired relaxation that penalizes excessively sharp or unstable configurations while preserving directional coherence. The resulting update provides physically interpretable controls over magnitude, nonlinear anisotropy suppression, and spatial regularity, replacing a single blur width with an energy-based structure-smoothness tradeoff. Our method is *training-free*, requires *no architectural modification*, and integrates as a *drop-in attention-space guidance step* for pre-trained text-to-image samplers, as shown in Figure 1. Empirically, LdG improves sample fidelity and alignment on SDXL, outperforms SEG on MS-COCO, and remains effective under 4-bit quantized SDXL, a setting motivated by memory-efficient and resource-constrained deployment of generative models (Li et al., 2023). Our results show that attention-space free-energy flows provide a principled and practical alternative to heuristic internal perturbations for guiding diffusion models.

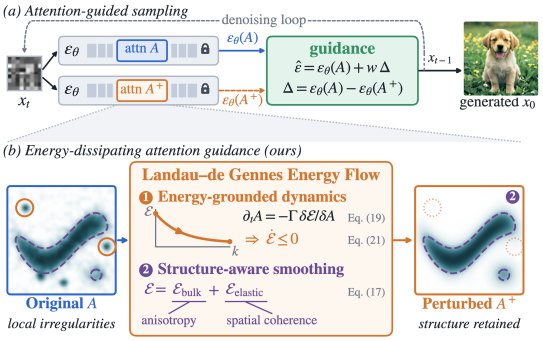

Figure 1: **Overview of LdG-style attention guidance**. (a) Prediction differences between original attention A and perturbed attention A^+ guide denoising in a frozen pre-trained diffusion model. (b) LdG-inspired bulk and elastic energies drive dissipative attention dynamics, suppressing local irregularities while preserving directional structure.

2 Preliminaries

Our approach is based on two line of ideas: continuum models of anisotropic media, particularly the Landau–de Gennes (LdG) theory (Ball, 2017), and attention-smoothing strategies that stabilize diffusion sampling. We start with the notation and concepts needed to connect these concepts.

Matrix-Valued Fields and Notation We denote by $A : \Omega \to \mathbb{R}^{d\times d}$ a spatially varying field defined over coordinates $x \in \Omega$ (image pixels or token positions). For a matrix M, $\mathrm{Tr}(M)$ denotes its trace and $M^\top$ its transpose. The Frobenius inner product is $\langle M, N\rangle_F = \mathrm{Tr}(M^\top N)$, with norm $\|M\|_F^2 = \langle M, M\rangle_F$. Spatial derivatives are taken component-wise: ∇A is the tensor of spatial gradients, $\|\nabla A\|^2$ is the sum of squared gradients over all entries, and $(\Delta A)_{ij} = \sum_{k=1}^{d} \partial_{x_k x_k} A_{ij}$.

Diffusion Models and the Role of Attention Diffusion models (Ho et al., 2020) construct samples by iteratively denoising a noisy variable x_t using a score network (Song et al.) $s_\theta(x_t, t)$. Modern architectures implement this with transformer layers (Vaswani et al., 2017), which allow long-range interaction. At each denoising step, given latent features $\{h_j\}$ at time t, for parameter $\tau > 0$, the model forms queries $q_i = W_q h_i$, keys $k_j = W_k h_j$, and values $v_j = W_v h_j$, and computes

$$A_{ij}(t) = \frac{\exp(\tau^{-1} q_i^\top k_j)}{\sum_\ell \exp(\tau^{-1} q_i^\top k_\ell)}, \tag{1}$$

which aggregates via $\mu_i(t) = \sum_j A_{ij}(t)v_j$. $A(t)$ represents the query-key affinity map for a given layer and diffusion timestep. It is sensitive to noise in the intermediate states x_t and can exhibit high-curvature, spatially irregular patterns that destabilize semantic consistency and sampling trajectories (Karras et al., 2022; Zhu et al., 2025). This is the object we will regularize to guide denoising.

Landau-de Gennes (LdG) PDE. In liquid-crystal theory (Ball, 2017), the mesoscopic state is the *alignment tensor* $Q(x) \in \mathbb{R}^{d\times d}$, typically symmetric and traceless. The canonical dissipative dynamics is

$$\partial_t Q \;=\; -\Gamma\, H, \qquad \Gamma > 0, \tag{2}$$

where H is *molecular field*. H is defined through the functional derivative of the free energy:

$$H \;:=\; -\left(\frac{\delta F_{\mathrm{LdG}}}{\delta Q} - \frac{I}{d}\,\mathrm{tr}\left(\frac{\delta F_{\mathrm{LdG}}}{\delta Q}\right)\right). \tag{3}$$

This variational form guarantees energy dissipation,

$$\frac{d}{dt}F_{\mathrm{LdG}}[Q(t)] \;=\; -\Gamma \int_\Omega \|H\|_F^2\, dx \;\le\; 0, \tag{4}$$

under standard boundary conditions. This monotone relaxation is the principle we want to preserve when acting on attention fields; additional details are described in Appendix A.4.

LdG free energy. The Landau–de Gennes free energy combines a *bulk* potential f_{bulk}, which encodes local thermodynamic preference, and an *elastic* term f_{elastic}, which encodes spatial regularity:

$$F_{\mathrm{LdG}}[Q] = \int_\Omega \Big(f_{\mathrm{bulk}}(Q) + f_{\mathrm{elastic}}(\nabla Q)\Big)\, dx, \tag{5}$$

$$f_{\mathrm{bulk}}(Q) = \frac{a}{2}\,\mathrm{Tr}(Q^2) - \frac{b}{3}\,\mathrm{Tr}(Q^3) + \frac{c}{4}\left(\mathrm{Tr}(Q^2)\right)^2, \quad f_{\mathrm{elastic}}(\nabla Q) = \frac{\lambda}{2}\|\nabla Q\|^2. \tag{6}$$

The anisotropic structure is regularized by an *energy minimizing* dynamics that preserves tensorial character of the order parameter while penalizing excessive local anisotropy and spatial roughness. We adopt this "order-parameter relaxation" strategy for attention maps inside diffusion models.

3 Diffusion Trajectory as a Dynamical System

Denoising diffusion sampling is an iterative process. Instabilities at intermediate noise levels can accumulate and amplify, degrading the final sample. Guidance of the reverse process requires stabilizing the entire trajectory, as illustrated in Figure 2. Recent work formalizes this *trajectory-level* intuition (Dockhorn et al.; Pan et al., 2024). For example, Cheng et al. (2025) formulates guided generation as sampling from a target distribution over trajectories. Let $U := \{x(t)\}_{t\in[0,T]}$ be a reverse diffusion trajectory. A trajectory-level energy $\mathcal{J}(U)$ penalizes instability along the path and induces $p^\star(U) \propto \exp(-\mathcal{J}(U))$, so trajectories that remain well-behaved throughout denoising receive higher probability mass. Sampling from $p^\star$ is achieved by augmenting the reverse drift with local corrections derived from this global objective. The key question becomes: *what should $\mathcal{J}$ encode?* Lyapunov stability theory (Sastry, 1999) provides a natural answer.

Lyapunov stability of Diffusion Trajectory. Consider the system $\dot{x}(t) = f\big(x(t), t\big)$, with state $x(t) \in \mathbb{R}^d$. A Lyapunov functional $V : \mathbb{R}^d \to \mathbb{R}_{\ge 0}$ certifies stability around reference state $x^\star$ when

$$V(x^\star) = 0, \qquad V(x) > 0\ (x \neq x^\star), \quad \dot{V}(x(t)) := \nabla V(x(t))^\top f(x(t), t) \le 0. \tag{7}$$

If the Lie derivative (Yano, 2020) satisfies $\dot{V} \le 0$, the dynamics *cannot* increase the energy along trajectories. This suggests *trajectory adjustment as control* (Haddad & Chellaboina, 2008). Introducing a control input $u(t)$ with input matrix $B(t)$,

$$\dot{x}(t) = f\big(x(t), t\big) + B(t)\, u(t), \tag{8}$$

one may select $u(t)$ to enforce decreasing V. The choice of $u(t) = -B(t)^\top \nabla V\big(x(t)\big)$, yields

$$\dot{V}(x(t)) = \nabla V(x(t))^\top f(x(t), t) - \big\|B(t)^\top \nabla V(x(t))\big\|^2. \tag{9}$$

Thus, this control provides a principled mechanism for enforcing trajectory-level energy dissipation. Appendix A.4 includes the corresponding stability interpretation in much more detail.

Remark 3.1 (Guided denoising as stabilized dynamics). The probability-flow ODE of diffusion, $\dot{x}_t = b_\theta(x_t, t) + u_t$, has the same controlled form. From a Lyapunov perspective, the guidance term acts as stabilizing feedback across noise scales, modifying the denoising trajectory rather than only the final sample. The question is: *where should this control act?* We propose to act on the attention tensor, replacing image-space control with attention-space energy minimization.

Stabilization via Attention Recall from §2 that the denoising network computes attention A_t as a function of current state, timestep, and conditioning via an external text prompt: $A_t = \mathcal{A}_\theta(x_t, t, \kappa)$. We propose to act on this attention tensor. Let $\mathcal{E}[A]$ be an energy functional that penalizes undesirable attention patterns. A stabilizing gradient descent step in attention space is

$$A_t^+ = A_t - \eta \nabla_A \mathcal{E}[A_t], \qquad \eta > 0, \quad (10)$$

which guarantees, for sufficiently small η, a monotone decrease of the energy. Since attention feeds into the pre-trained score network, this perturbation induces an implicit control on the denoising trajectory,

$$u_t \;\approx\; -\eta \frac{\partial b_\theta}{\partial A}(x_t, t)\, \nabla_A \mathcal{E}[A_t], \qquad (11)$$

where b_θ is the drift. Thus attention-space energy minimization gives image-space trajectory stabilization. Prior methods (Hong, 2024; Ahn et al., 2024) implicitly follow this template but uses ad-hoc perturbations without a coherent energy.

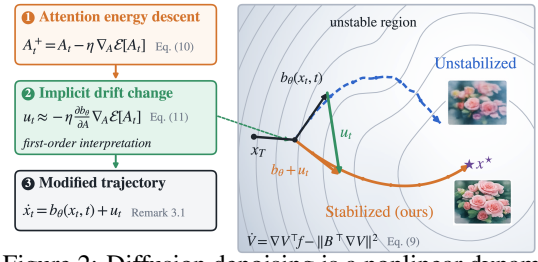

Figure 2: Diffusion denoising is a nonlinear dynamical system whose trajectories may encounter unstable regions, amplifying errors and degrading samples. Attention-space energy descent induces implicit trajectory control, promoting more stable guidance and coherent generation.

4 The Attention and LdG Correspondence

Next, let us treat attention as a spatially varying field and construct $\mathcal{E}$ carefully from first principles.

Perturbing and Analyzing Attention. Understanding how the attention tensor $A(x)$ responds to noise in the queries and intermediate latent representation x_t shows a link to LdG energy. The attention mechanism outputs a weighted mean with query q, keys k_j and $p_j(q)$ as softmax weights.

$$\mu(q) = \sum_{j=1}^{n} p_j(q)\, k_j, \qquad p_j(q) = \frac{\exp(\tau^{-1} q^\top k_j)}{\sum_\ell \exp(\tau^{-1} q^\top k_\ell)}, \qquad (12)$$

When q is perturbed as $q \mapsto q + \delta$, a Taylor expansion yields:

$$\mu(q + \delta) = \mu(q) + J(q)\delta + O(||\delta||^2), \qquad (13)$$

where $J(q) = \partial\mu/\partial q$ is the Jacobian. Differentiation of the softmax shows that the Jacobian is directly proportional to the covariance of keys under the attention weights, $J(q) = \frac{1}{\tau}\mathrm{Cov}_p(k)$. The reconstruction error/loss, $\ell(q) = \frac{1}{2}\|\mu(q) - q\|^2$, measures how well attention preserves local information. Under isotropic noise $\delta \sim \mathcal{N}(0, \sigma^2 I)$, the perturbed reconstruction error results in

$$\mathbb{E}_\delta \|\mu(q+\delta) - (q+\delta)\|^2 \approx \|\mu(q) - q\|^2 + \sigma^2 \Big[\tau^{-2}\mathrm{Tr}(\mathrm{Cov}_p(k)^2) - \tfrac{2}{\tau}\mathrm{Tr}(\mathrm{Cov}_p(k)) + d\Big]. \quad (14)$$

Decomposing the covariance into isotropic and anisotropic components by defining

$$Q = \mathrm{Cov}_p(k) - \frac{\mathrm{Tr}(\mathrm{Cov}_p(k))}{d} I, \qquad (15)$$

where Q is traceless and captures the anisotropic part of attention-weighted covariance. This leads to:

$$\mathrm{Tr}(\mathrm{Cov}_p(k)^2) = \mathrm{Tr}(Q^2) + \frac{1}{d}\mathrm{Tr}(\mathrm{Cov}_p(k))^2. \qquad (16)$$

Thus the error expansion includes (i) isotropic variance, (ii) anisotropic variance (via $\mathrm{Tr}(Q^2)$), and (iii) a spatial smoothing term (via d). This structure closely resembles the form that appears in the classic Landau–de Gennes free energy used to model anisotropic materials discussed in Section 2. This link motivates treating attention as an order-parameter field.

Attention as Order Parameter. The analysis above shows that attention geometry decomposes into *isotropic* and *anisotropic* components. The traceless tensor Q in (15) captures the *anisotropic*

part of the attention-weighted covariance. This object plays the role of a local nematic order parameter: $Q(q) = 0$ when attention is isotropic (uniform over keys), and $\|Q(q)\|$ grows when attention concentrates along preferred directions in key space. Diffusion models do not compute $Q(q)$ directly. Instead, they compute the attention tensor $A(x)$, whose rows define the attention weights used to aggregate keys at each spatial location or token x. Thereby Q is induced through the interaction between queries and keys. Indeed, the attention-weighted covariance and so, $Q(q)$ is a nonlinear functional of the attention weights encoded in $A(x)$.

This motivates a change of viewpoint. Rather than regularizing $Q(q)$ pointwise, we treat the attention tensor $A(x)$ itself as a *generalized order-parameter*. In this interpretation, $A(x)$ encodes *both* the magnitude and directional structure of attention, while its induced covariance recovers the nematic-like anisotropy captured by $Q(q)$. First, regions where attention collapses onto a small set of keys correspond to high-order, strongly anisotropic configurations. Second, diffused attention corresponds to locally isotropic configurations, similar to disordered phases. We discuss the relation between the order parameter field and score smoothing in Sec A.1.

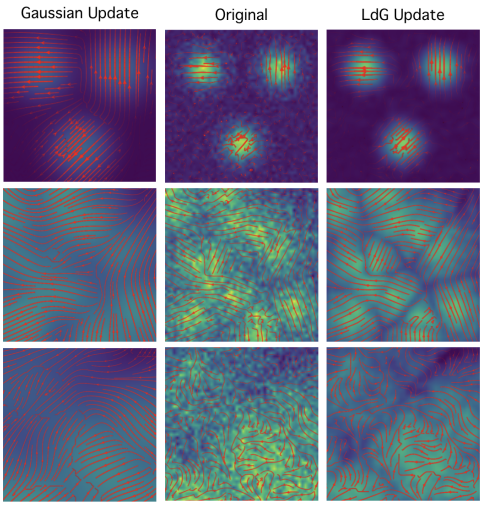

Figure 3: Vector fields under different smoothing strategies. Each row is for a different pattern. *Left:* Gaussian update, which applies isotropic diffusion *Center:* Original field, *Right:* Our LdG update, which selectively dissipates irregular fluctuations but preserves salient peaks, coherent flow lines, and anisotropic directions. Unlike Gaussian smoothing it respects the underlying directional geometry.

What does this viewpoint buy us? Thinking of $A(x)$ as a spatially varying order-parameter field enables a natural energy formulation. First, we can define a *bulk energy* that penalizes excessive magnitude and anisotropy of attention, similar to polynomial potential in LdG theory (Ball, 2017). Second, because $A(x)$ varies over spatial coordinates, we can define an *elastic energy* that penalizes rapid spatial variation, analogous to the spatial-regularity terms used in continuum theories of ordered media (De Gennes & Prost, 1993; Ball, 2017; Virga, 2018). Together, these give a free energy functional on attention whose gradient flow provides a principled, structure-preserving relaxation.

5 Anisotropic Relaxation of Attention Field

Energy formulation. Using the order-parameter view, we instantiate free energy on attention $A(x)$:

$$\mathcal{E}[A] = \int_\Omega \left[\frac{a}{2}\mathrm{Tr}(A^2) + \frac{c}{4}\mathrm{Tr}(A^2)^2 + \frac{\lambda}{2}\|\nabla A\|^2 \right] dx. \tag{17}$$

The quadratic term controls isotropic attention magnitude, while the quartic term penalizes sharp anisotropic concentration onto a small set of keys. The elastic term $\frac{\lambda}{2}\|\nabla A\|^2$ penalizes rapid spatial variation across neighboring positions. Thus, $\mathcal{E}$ defines a structure-smoothness tradeoff on attention, with separate controls for magnitude, concentration, and spatial coherence. We omit the cubic term in (6), since it has no direct interpretation for the attention fields.

$$\frac{\delta \mathcal{E}}{\delta A} = \big(a + c\,\mathrm{Tr}(A^2)\big)A^\top - \lambda \Delta A. \tag{18}$$

Here, the first term combines uniform damping from a with adaptive anisotropy suppression. The Laplacian term provides spatial smoothing. Together, we get a relaxation that reduces attention magnitude and sharpness while encouraging spatial coherence. Taking steepest descent in this energy, $\mathcal{E}[A]$ gives the gradient flow

$$\frac{\partial A}{\partial t} = -\Gamma \frac{\delta \mathcal{E}}{\delta A}, \qquad \Gamma > 0. \tag{19}$$

Substituting the functional derivative (18) gives the PDE

$$\frac{\partial A}{\partial t} = \Gamma \Big[\lambda \Delta A - a A^\top - c\,\mathrm{Tr}(A^2)A^\top \Big], \tag{20}$$

which is the LdG inspired gradient flow adapted to our attention tensor. In (20), $\lambda\Delta A$ provides elastic smoothing, while $-aA^\top$ and $-c\,\mathrm{Tr}(A^2)A^\top$ damp large and concentrated attention responses. Consequently, unstable high-curvature attention patterns are relaxed while coherent spatial structure is retained. The flow dissipates energy, with the guarantee

$$\frac{d}{dt}\mathcal{E}[A(t)] = -\Gamma \left\| \frac{\delta \mathcal{E}}{\delta A} \right\|^2 \leq 0. \tag{21}$$

Discretization. For integration into diffusion sampling, we discretize it using an explicit forward Euler step with effective step size $\eta > 0$ (absorbing Γ):

$$A^{(k+1)} = A^{(k)} + \eta \left.\frac{\partial A}{\partial t}\right|_{A=A^{(k)}}, A^{(k+1)} = A^{(k)} + \eta\Big[\lambda\Delta A^{(k)} - a(A^{(k)})^\top - c\,\mathrm{Tr}\big((A^{(k)})^2\big)(A^{(k)})^\top\Big] \tag{22}$$

A small number of such steps applied to the attention tensor at each denoising iteration relaxes high-curvature, unstable attention modes while preserving semantic structure, Figure 3. This is our complete update rule which requires no retraining or architecture change.

LdG relaxation under quantization. Low-bit quantization perturbs attention logits through query, key, and value. We use LdG relaxation to correct perturbed attention fields at inference time, without updating the quantized weights. Consider a transformer attention head in a full-precision model,

$$S = \frac{QK^\top}{\sqrt{d_h}}, \qquad P = \mathrm{softmax}(S + M), \qquad O = PV, \tag{23}$$

where $S \in \mathbb{R}^{L\times L}$ denotes the pre-softmax attention logits, M is an additive attention mask, and d_h is the head dimension. Quantizing the weights or activations perturbs the query, key, and value projections. Let Q_q, K_q, V_q denote the quantized counterparts. The attention logits then become, to first order,

$$S_q = \frac{Q_q K_q^\top}{\sqrt{d_h}} = S + \Delta S_q, \qquad \Delta S_q \approx \frac{\Delta Q K^\top + Q\Delta K^\top}{\sqrt{d_h}}. \tag{24}$$

Because softmax amplifies local logit differences, the perturbations can create high-curvature attention patterns, attention collapse, or loss of salient token interactions, especially at low bit-widths. We therefore apply the LdG energy in (17) and its flow in (20) to the perturbed logit field S_q before the softmax. For small step size, the explicit LdG step can be viewed as a first-order approximation to the proximal correction

$$S_q^+ \approx \arg\min_S \left\{ \frac{1}{2}\|S - S_q\|_F^2 + \eta\mathcal{E}_{\mathrm{LdG}}[S] \right\}, \tag{25}$$

The corrected attention output is then computed as

$$P_q^+ = \mathrm{softmax}(S_q^+ + M), \qquad O_q^+ = P_q^+ V_q. \tag{26}$$

This relaxation is applied purely during inference without updating model weights.

6 Experimental evaluations

We first examine the behavior of the LdG update in isolation, then evaluate its impact on text-to-image generation through qualitative comparisons and quantitative benchmarks. Finally, we check whether the same training-free update can improve generation quality under 4-bit model quantization.

6.1 Evaluation of LdG-based update

The goal of our proposed LdG based update is not just to make attention maps smoother. Notice that a good guidance update should suppress unstable fluctuations while preserving the directional and spatial organization. So, we compare LdG with Gaussian smoothing on synthetic attention-like fields and evaluate four complementary properties: edge strength, peak value preservation, high-frequency retention, and structural similarity (SSIM). Additional details are in Appendix A.6.

Summary. Figure 4 shows that Gaussian smoothing reduces variation aggressively: it weakens edges, suppresses peak responses, removes high-frequency content, and lowers SSIM. In contrast, LdG preserves stronger edge responses and peak values, retains substantially more high-frequency

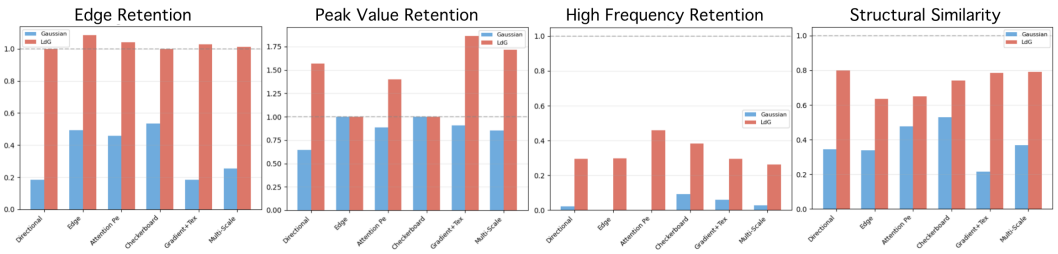

Figure 4: Comparison of Gaussian smoothing and the proposed Landau–de Gennes (LdG) update across multiple structural metrics on synthetic attention patterns. While Gaussian smoothing uniformly suppresses edges, peaks, and high-frequency content, LdG preserves directional structure, salient responses, and global spatial organization, as reflected by consistently higher edge retention, peak value retention, high-frequency retention, and SSIM. These results highlight the advantage of LdG as a structure-preserving, anisotropic regularizer for attention in text-to-image diffusion models.

energy, and maintains higher structural similarity across the tested patterns. The takeaway is that LdG behaves as a structure-preserving relaxation rather than a generic low-pass filter. It dissipates irregular fluctuations while retaining the local and global geometry of the attention field, supporting its use as an internal guidance mechanism in the following text-to-image experiments.

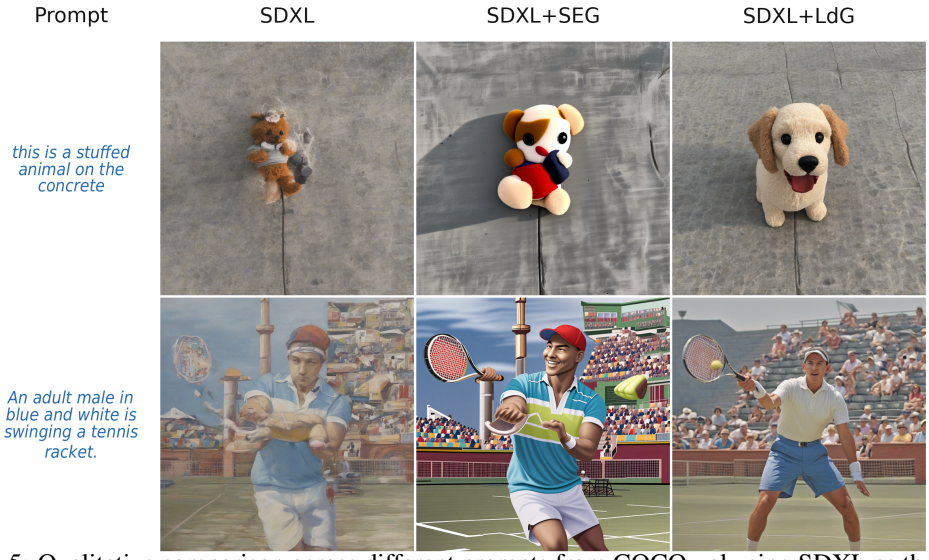

Figure 5: Qualitative comparison across different prompts from COCO-val using SDXL as the base model. For each prompt, we compare unguided generation with SEG and our LdG-based anisotropic relaxation. LdG consistently preserves structure while enhancing sharpness and reducing artifacts.

6.2 Qualitative comparison.

After checking that the proposed Landau–de Gennes (LdG) update preserves critical structural properties of attention maps, we now evaluate how these improvements translate to end-to-end image generation when the update rule is applied within a pre-trained diffusion model. We begin with a qualitative comparison to assess perceptual quality, structural fidelity, and semantic consistency of the generated samples, followed by a quantitative evaluation in the subsequent subsection.

Figure 5 compares samples produced by Stable Diffusion XL (SDXL) (Rombach et al., 2022), SEG (Hong, 2024), and the proposed LdG-based relaxation across four representative prompts covering still-life scenes, indoor environments, natural landscapes, and crowded multi-object settings.

Method	FID↓	CLIP↑	PickScore↑	Aesthetic↑
SDXL	50.59	0.2665	19.98	5.662
SEG	25.67	0.2889	21.32	**5.885**
LdG	**21.63**	**0.2915**	**21.34**	5.862

Table 1: Quantitative comparison on 30K MS-COCO validation set. SEG and LdG correspond to attention guidance methods.

Summary. Figure 5 shows two representative COCO-val prompts. Across both, the unguided SDXL samples exhibit diffuse geometry and weakened object structure, while SEG improves coherence but can introduce over-smoothed or stylized details. LdG better preserves object boundaries

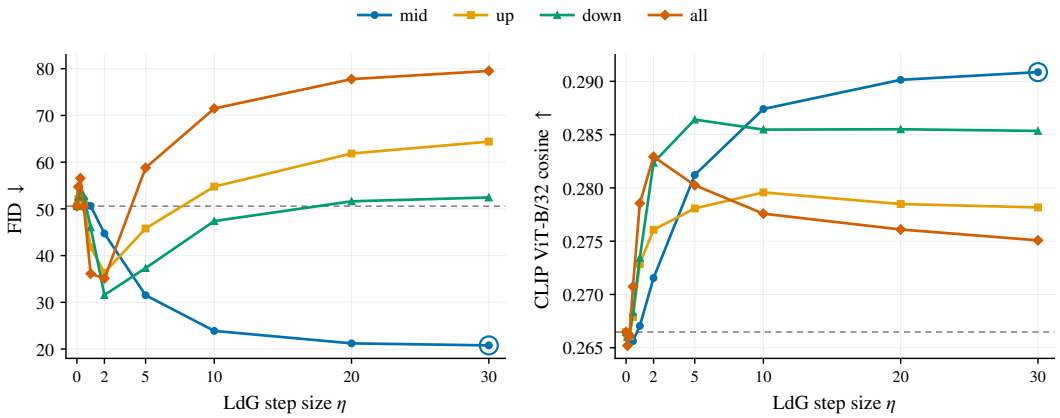

Figure 6: **Guidance strength interacts with layer placement.** FID (lower is better) and CLIP ViT-B/32 raw cosine similarity (higher is better) for four U-Net block placements with varying η. The x-axis is linear. Dashed lines mark the mean $\eta = 0$ control across placements. Both metrics are best at $\eta = 30$ with LdG applied to mid blocks. The full grid is reported in Table 3.

and spatial layout: the stuffed-animal example retains a clearer subject and surface structure, and the tennis scene maintains a more plausible player pose and court geometry. Additional qualitative examples are provided in Appendix A.5.

6.3 Quantitative Evaluation on MS-COCO

We evaluate our method on the MS-COCO 2014 (Lin et al., 2014) validation set following (Hong, 2024). We generate 30K samples for each configuration and report both the *Fréchet Inception Distance* (FID) (Heusel et al., 2017) and the *CLIP* score (Radford et al., 2021; Hessel et al., 2021). Our method is applied to the self-attention maps of SDXL, using the same sampling schedule, timestep configuration, and guidance scale as the baseline. The only modification is the inclusion of a small number of LdG relaxation steps at each denoising iteration. This enables selective dissipation of high-curvature anisotropic modes in the attention energy landscape.

Summary. We see from Table 1 that our proposed LdG method improves FID and CLIP scores relative to SDXL and SEG on 30K validation images from MS-COCO. These results demonstrate improvements in both distributional fidelity and text–image alignment. Since the proposed LdG update operates directly on the attention field, it preserves semantic directions while reducing noise, thereby avoiding the over-smoothing and saturation effects observed in isotropic relaxation.

Ablation. We demonstrate the interplay and effect of guidance strength, η and the U-Net blocks to which the LdG update is applied, using 30K images from MS-COCO validation set. From Figure 6 we see that applying LdG to up, down, or all blocks improves FID only up to moderate step sizes and then degrades at larger η, indicating that stronger perturbations are not uniformly beneficial. In contrast, mid-block guidance improves both metrics from $\eta = 0.5$ through 30, where it achieves the best tested FID and CLIP. This supports using mid-block attention as the primary intervention site in our experiments, while also showing that LdG guidance is sensitive to where and how strongly the attention flow is applied. Additional details and the complete results are in Appendix A.7.

6.4 Improving Generation Quality Under Quantization

After evaluating LdG guidance on full-precision SDXL, we next test whether the same attention-space relaxation remains effective under model compression. Quantization reduces memory requirements but can degrade generation quality, making 4-bit SDXL a useful setting for evaluating training-free correction without changing model weights. We quantized SDXL to 4 bits using Quanto (Corvoysier, 2024) and applied LdG guidance on the mid-layer attention modules. We generated 10K samples and measure FID and CLIP score along with four learned human-preference metrics: HPS-v2 (Wu et al., 2023), ImageReward (Xu et al., 2023), PickScore (Kirstain et al., 2023), and Aesthetic-v2 (Schuhmann, 2022). These evaluate generated images based on human visual preference, text-image semantic alignment, comparative user selection and aesthetics, serving as a measure of overall generation quality.

Precision	Guidance	FID↓	CLIP↑	ImageReward↑	Aesthetic-v2↑	PickScore↑	HPSv2↑
FP16	—	52.903	0.5071	-1.035	5.282	19.971	0.1761
INT4	—	80.127	0.4782	-1.422	4.904	19.368	0.1535
INT4	SEG	27.308	0.5814	-0.215	5.574	21.087	0.2404
INT4	**LdG**	**23.888**	**0.6050**	**0.176**	**5.719**	**21.721**	**0.2703**

Table 2: Generation quality of 4-bit quantized SDXL with and without guidance. LdG outperforms the unguided quantized model, quantized SEG, and the unguided FP16 baseline across all metrics.

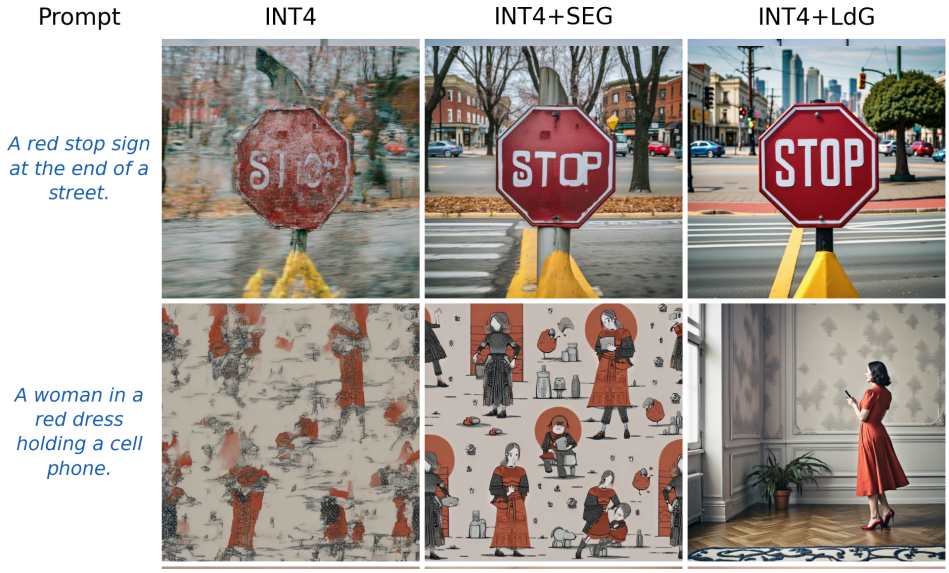

Figure 7: Generated images by the INT4 quantized SDXL model with different corrections applied. We can see that LdG preserves greater detail in the images while staying aligned to the prompt.

Summary. Table 2 compares LdG with SEG and the unquantized FP16 SDXL model. We apply SEG with blur $\sigma \to \infty$ (its best setting). For LdG, we use $\eta = 2$. Without additional guidance, 4-bit quantization worsens all six reported metrics relative to the FP16 baseline. LdG improves all six metrics relative to both the unguided quantized model and quantized SEG, and is the only evaluated configuration with a positive ImageReward score. The quantized model with LdG also outperforms the unguided FP16 baseline across all reported metrics. This comparison shows the benefit of guidance under quantization. Figure 7 shows the generated images for qualitative comparison. We further evaluated INT4-quantized SDXL variants; vanilla (no guidance), SEG and LdG on a uniformly sampled set of 340 shared prompts from MS-COCO validation dataset, using blind, position-randomized pairwise judgments from Claude Opus-4.8 across prompt fidelity, visual quality, artifacts, aesthetics, and overall preference. LdG was preferred over SEG on 72.8% of prompts and over the vanilla baseline on 97.6%, showing a strong improvement in perceived generation quality.

7 Conclusion

We introduced an anisotropic relaxation mechanism for diffusion models derived from the Landau–de Gennes (LdG) PDE. By interpreting attention as a spatially structured tensor field, we showed that a small number of PDE steps provides a principled way to dissipate high-curvature modes while preserving semantically meaningful directional structure. This contrasts with existing unconditional guidance strategies such as SAG, PAG, and isotropic smoothing approaches like SEG which often rely on heuristic manipulations of attention and may introduce oversmoothing, saturation, or loss of geometric detail. Our formulation yields an update motivated by energy minimization that requires no additional training or network modification. Through quantitative evaluation on MS-COCO and qualitative comparisons across a range of prompts, we demonstrate that LdG consistently improves sample quality over both the SDXL baseline and prior attention-based guidance methods. We further demonstrate that the same training-free update improves generation quality in 4-bit quantized SDXL, outperforming both SEG and the unguided FP16 baseline across all reported metrics. The method remains stable across a wide range of prompt types, including cases involving noticeable geometric structure or multiple objects. Code will be made available upon acceptance.

AI USE STATEMENT

AI tools were used as part of the LLM-as-judge experiment described in the paper. We took use of AI tools for integrating the core implementation to run larger experiments in optimized frameworks. Some of the images were polished using these. They were not used to generate research ideas, design experiments, perform analysis, or write the manuscript.

ETHICS STATEMENT

This work is intended to advance understanding of machine learning methods and does not involve human subjects, private personal data, or deployed decision-making systems. We do not anticipate direct societal harm from the proposed methodology, though, as with related machine learning techniques, misuse may be possible if applied in inappropriate or high-stakes settings without proper validation.

REPRODUCIBILITY STATEMENT

We have aimed to make the work reproducible by providing the main methodological details, experimental setup, implementation choices, and evaluation protocol in the paper and appendix. We will release code and configuration details upon acceptance.

REFERENCES

Donghoon Ahn, Hyoungwon Cho, Jaewon Min, Wooseok Jang, Jungwoo Kim, SeonHwa Kim, Hyun Hee Park, Kyong Hwan Jin, and Seungryong Kim. Self-rectifying diffusion sampling with perturbed-attention guidance. In *European Conference on Computer Vision*, pp. 1–17. Springer, 2024.

John M Ball. Mathematics and liquid crystals. *Molecular Crystals and Liquid Crystals*, 647(1):1–27, 2017.

Arwen Bradley and Preetum Nakkiran. Classifier-free guidance is a predictor-corrector. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=3W4F8jjeN6.

Huayu Chen, Kai Jiang, Kaiwen Zheng, Jianfei Chen, Hang Su, and Jun Zhu. Visual generation without guidance. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 8275–8295. PMLR, 2025. URL https://proceedings.mlr.press/v267/chen25ac.html.

Zhengdao Chen. On the interpolation effect of score smoothing in diffusion models. In *The Fourteenth International Conference on Learning Representations*, 2026.

Xiaoyuan Cheng, Xiaohang Tang, and Yiming Yang. Safe and stable control via lyapunov-guided diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

David Corvoysier. Optimum quanto: A pytorch quantization backend for optimum, 2024. URL https://github.com/huggingface/optimum-quanto. Apache-2.0 license.

Pierre-Gilles De Gennes and Jacques Prost. *The physics of liquid crystals*. Number 83. Oxford university press, 1993.

Tim Dockhorn, Arash Vahdat, and Karsten Kreis. Score-based generative modeling with critically-damped langevin diffusion. In *International Conference on Learning Representations*.

Wassim M Haddad and VijaySekhar Chellaboina. *Nonlinear dynamical systems and control: a Lyapunov-based approach*. Princeton university press, 2008.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7514–7528, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Susung Hong. Smoothed energy guidance: Guiding diffusion models with reduced energy curvature of attention. *Advances in Neural Information Processing Systems*, 37:66743–66772, 2024.

Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. Improving sample quality of diffusion models using self-attention guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7462–7471, 2023.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Jeeyung Kim, Erfan Esmaeili, and Qiang Qiu. Text embedding is not all you need: Attention control for text-to-image semantic alignment with text self-attention maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8031–8040, 2025.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation, 2023. URL https://arxiv.org/abs/2305.01569.

Xiuyu Li, Yijiang Liu, Long Lian, Huanrui Yang, Zhen Dong, Daniel Kang, Shanghang Zhang, and Kurt Keutzer. Q-diffusion: Quantizing diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17535–17545, 2023.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Chaoyi Pan, Zeji Yi, Guanya Shi, and Guannan Qu. Model-based diffusion for trajectory optimization. *Advances in Neural Information Processing Systems*, 37:57914–57943, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Javad Rajabi, Soroush Mehraban, Seyedmorteza Sadat, and Babak Taati. Token perturbation guidance for diffusion models. In *Advances in Neural Information Processing Systems*, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/60d166781779c460d8ad83d78f2ba03a-Abstract-Conference.html.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Seyedmorteza Sadat, Manuel Kansy, Otmar Hilliges, and Romann Weber. No training, no problem: Rethinking classifier-free guidance for diffusion models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Uqhlk77s0i.

Shankar Sastry. Lyapunov stability theory. In *Nonlinear systems: analysis, stability, and control*, pp. 182–234. Springer, 1999.

Christoph Schuhmann. LAION-Aesthetics. https://laion.ai/blog/laion-aesthetics/, August 2022. Accessed: 2026-08-21.

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*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Epifanio G Virga. *Variational theories for liquid crystals*. Chapman and Hall/CRC, 2018.

Max Welling, Sirui Lu, and Lars Holdijk. *Generative AI and Stochastic Thermodynamics: A Tale of Free Energies*. Cambridge University Press, 2026. doi: 10.1017/9781009709071.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis, 2023. URL https://arxiv.org/abs/2306.09341.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation, 2023. URL https://arxiv.org/abs/2304.05977.

Kentaro Yano. *The theory of Lie derivatives and its applications*. Courier Dover Publications, 2020.

Xinyu Zhou, Jiawei Zhang, and Stephen J. Wright. Smoothing the score function for generalization in diffusion models: An optimization-based explanation framework. *arXiv preprint arXiv:2601.19285*, 2026.

Haiming Zhu, Yangyang Xu, Chenshu Xu, Tingrui Shen, Wenxi Liu, Yong Du, Jun Yu, and Shengfeng He. Stable score distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16597–16606, 2025.

A Appendix

A.1 Connection to score smoothing

The tensor $Q(q)$ introduced in Section 4 measures anisotropy of the attended keys: it vanishes when their weighted covariance is isotropic and grows when their dispersion differs across directions. The same covariance structure also arises exactly in the empirical diffusion score. Let $\{y_j\}_{j=1}^n \subset \mathbb{R}^D$ denote the training samples and let t be the Gaussian noise variance. The posterior of y_j for a noisy point x is

$$w_j(x,t) = \mathrm{softmax}_j\left(\frac{x^\top y_j - \|y_j\|^2/2}{t}\right), m_t(x) = \sum_j w_j(x,t) y_j, \quad s_t^{(n)}(x) = \frac{m_t(x) - x}{t}. \tag{27}$$

Thus, Eq. (27) is precisely the attention map in Eq. (12), with x as query and y_j as both key and value. A single dominant responsibility retrieves one training sample, whereas non-degenerate responsibilities produce a barycenter Chen (2026, Eqs. (5)–(7)).

Applying the attention-covariance identity with $q = x$, $k_j = y_j$, and $\tau = t$ gives, for $C_t(x) := \mathrm{Cov}_{w(x,t)}(y)$,

$$\nabla_x^2 \log p_t^{(n)}(x) = \frac{C_t(x)}{t^2} - \frac{I_D}{t}, Q_t^{\mathrm{emp}}(x) := \mathrm{dev}(C_t(x)) = t^2 \,\mathrm{dev}\Big(\nabla_x^2 \log p_t^{(n)}(x)\Big). \tag{28}$$

Consequently, the induced order parameter is, in this model, exactly the anisotropic component of the score Hessian. This is the equal-variance specialization of the covariance-minus-isotropic decomposition in Zhou et al. (2026, Eq. (18)).

This identification also gives the quadratic term of the energy in Eq. (17) a direct diffusion interpretation. For a pre-softmax logit field S, the quadratic part of the update in Eq. (22) satisfies

$$S^+ = (1 - \eta a)S = \rho S, \quad \mathrm{softmax}(S^+) = \mathrm{softmax}(S/T_{\mathrm{eff}}), \quad T_{\mathrm{eff}} = \rho^{-1} > 1, \tag{29}$$

for $0 < \rho < 1$. Hence quadratic LdG relaxation is equivalent to temperature smoothing, which reduces single-sample dominance (Zhou et al., 2026, Eq. (6) and the temperature definition following Eq. (7)); the nonlinear bulk and elastic terms extend this scalar operation with state-dependent spectral control and spatial coherence. Geometrically, if all y_j lie in a linear subspace U, normalized reweighting changes only the component of the score within U; its normal component remains

$$P_{U^\perp} s_t(x) = -t^{-1} P_{U^\perp} x.$$

Thus, temperature smoothing can alter interpolation within U without weakening contraction toward it, consistent with Chen (2026, Eqs. (19) and (22)–(24)).

A.2 Interpreting as Implicit Reparameterization

The update acts on attention, but what does it imply for the underlying network? Let $s_\theta(x,t)$ denote the score and $A_\theta(x,t)$ the attention tensor. Since the score depends on attention through downstream transformation,

$$s_\theta(x,t) = \psi(A_\theta(x,t), x, t), \tag{30}$$

an attention perturbation $A \mapsto A + \delta A$ induces a first-order score shift

$$\delta s_A \approx \frac{\partial s_\theta}{\partial A}\,\delta A. \tag{31}$$

Meanwhile, parameter perturbations $\delta\theta$ induces attention variations:

$$\delta A \approx \frac{\partial A}{\partial \theta}\,\delta\theta. \tag{32}$$

We can ask: what parameter change would produce the same score shift as our attention perturbation? A parameter perturbation acting through attention gives

$$\delta s = J_\theta^{(\mathrm{attn})}\,\delta\theta, \qquad J_\theta^{(\mathrm{attn})} = \frac{\partial s_\theta}{\partial A}\frac{\partial A}{\partial \theta}. \tag{33}$$

Solving the associated least-squares problem produces the minimum-norm $\delta\theta$ that matches the attention-induced shift δs_A, which therefore is

$$\delta\theta = \left(J_\theta^{(\mathrm{attn})}\right)^+ \delta s_A. \tag{34}$$

Summary. LdG relaxation acts as an implicit, local reparameterization nudging the model toward a score field consistent with regularized attention, without explicitly modifying weights. This connects attention-guided sampling to classical gradient flows, but operating on internal representations rather than parameters.

A.3 Calculating the functional derivative

The gradient flow that minimizes $\mathcal{E}[A]$ requires calculating $\delta\mathcal{E}/\delta A$. We will treat each term separately. Then, we can write down and interpret the combined effect.

Quadratic/Magnitude term. For the quadratic penalty, let $A \to A + \epsilon\,\delta A$. We can expand,

$$\mathrm{Tr}\big[(A+\epsilon\delta A)^2\big] = \mathrm{Tr}(A^2) + 2\epsilon\,\mathrm{Tr}(A\,\delta A) + O(\epsilon^2). \tag{35}$$

Using $\mathrm{Tr}(XY) = \langle X^\top, Y\rangle_F$ gets us

$$\frac{\delta}{\delta A}\left(\tfrac{a}{2}\mathrm{Tr}(A^2)\right) = aA^\top. \tag{36}$$

This term produces a linear restoring force that uniformly damps all attention activations.

Anisotropy term. The quartic penalty requires the chain rule. Let $g(A) = \mathrm{Tr}(A^2)$. Then

$$\delta g(A) = 2\mathrm{Tr}(A\,\delta A), \qquad \delta[g(A)^2] = 2g(A)\,\delta g(A), \tag{37}$$

which gives

$$\frac{\delta}{\delta A}\left(\tfrac{c}{4}g(A)^2\right) = c\,\mathrm{Tr}(A^2)\,A^\top. \tag{38}$$

This contribution is *state-dependent*: the damping strength scales with $\mathrm{Tr}(A^2)$, so regions of high anisotropy experience stronger suppression. This allows selectively attenuating sharp, over-concentrated attention patterns.

Elastic term. The spatial penalty $\frac{\lambda}{2}\|\nabla A\|^2$ encourages neighboring attention values to remain similar. Integration by parts under standard boundary conditions gives

$$\frac{\delta}{\delta A}\left(\tfrac{\lambda}{2}\|\nabla A\|^2\right) = -\lambda\Delta A. \tag{39}$$

This is a diffusion operator: it smooths A by averaging over spatial neighbors, reducing high-frequency variation while preserving large-scale structure.

Combined dynamics. Assembling these contributions:

$$\frac{\delta\mathcal{E}}{\delta A} = \big(a + c\,\mathrm{Tr}(A^2)\big)A^\top - \lambda\Delta A. \tag{40}$$

The first term provides nonlinear damping involving uniform suppression from a, plus adaptive suppression from the anisotropy penalty. The second term provides spatial smoothing. Together, we get a relaxation that reduces attention magnitude and sharpness while encouraging spatial coherence.

A.4 Background

LdG interpretation. The Landau–de Gennes alignment tensor $Q(x)$ is typically symmetric and traceless: symmetry encodes unoriented local alignment, while tracelessness separates directional order from isotropic scale. In the classical liquid-crystal setting, the parameter a in Eq. (6) often controls proximity to the isotropic phase, b introduces the first-order isotropic-nematic transition in three-dimensional nematics, $c > 0$ ensures stability at large $\|Q\|$, and $\lambda > 0$ sets the strength of spatial smoothing. These physical roles motivate, but do not require, the attention-space interpretation used in the main text.

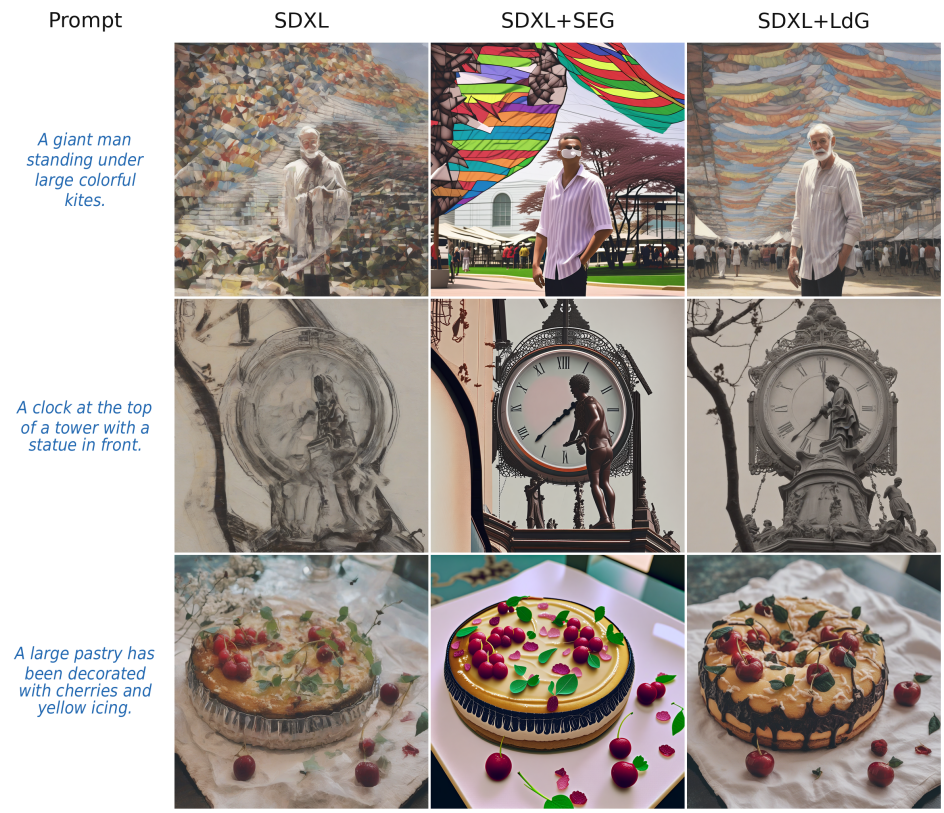

Figure 8: Additional qualitative comparison on COCO-val prompts using SDXL as the base model. These examples complement Figure 5.

The energy dissipation statement in Eq. (4) follows from the variational form of the molecular field. Under standard boundary conditions, the dynamics in Eq. (2) move the order parameter in the negative variational direction of F_{LdG}, after projection onto the traceless subspace through Eq. (3). Thus the free energy cannot increase along the flow. This monotone relaxation property is the specific aspect of LdG that we transfer to attention: rather than applying heuristic smoothing, we define a functional whose gradient flow explicitly dissipates an attention-field energy.

Lyapunov stability. For a dynamical system $\dot{x}(t) = f(x(t), t)$, an equilibrium or reference state $x^\star$ is stable if nearby trajectories that start close to $x^\star$ remain close for all future time, and asymptotically stable if they additionally converge to $x^\star$. The Lyapunov functional in Eq. (7) certifies this behavior because trajectories remain inside nested sublevel sets of V whenever $\dot{V} \leq 0$.

The controlled dynamics in Eq. (8) expose the same principle. Substituting the stabilizing feedback $u(t) = -B(t)^\top \nabla V(x(t))$ into the Lie derivative gives Eq. (9): the control contributes the non-positive term $-\|B(t)^\top \nabla V(x(t))\|^2$. Consequently, the controlled system dissipates energy at least as strongly as the uncontrolled system along the control directions. This is the basis for interpreting guided diffusion as stabilized dynamics and motivates acting on an internal representation, the attention tensor, through the energy step in Eq. (10).

A.5 Additional Qualitative Results

Figure 8 shows additional qualitative examples corresponding to the text-to-image comparison in Figure 5. These examples further illustrate the same trend observed in the main paper: LdG preserves object boundaries, local texture, and spatial structure more consistently than isotropic smoothing.



Figure 9: Additional qualitative examples for 4-bit quantized SDXL with different corrections applied. These examples complement Figure 7.

Figure 9 provides additional examples for the 4-bit quantized SDXL setting from Figure 7. Across these prompts, LdG guidance produces sharper and more semantically faithful generations than the unguided INT4 model and INT4 with SEG.

A.6 Attention-Field Structure Metrics

Section 6.1 evaluates the proposed update directly on synthetic attention-like fields before applying it inside the diffusion sampler. We use four complementary metrics, each capturing a different aspect of structure preservation.

Edge strength preservation. We measure edge strength as the average magnitude of spatial gradients in the attention field. This quantifies sharp spatial transitions, which in attention maps can correspond to abrupt changes in token interactions, object boundaries, or shifts in contextual alignment. Excessive attenuation of these gradients indicates over-smoothing.

Peak value preservation. Peak value preservation measures whether high-magnitude attention regions are retained after smoothing. These peaks correspond to salient token–token interactions and determine where semantic or spatial information is strongly propagated. Suppressing them can weaken conditioning signals and reduce local detail.

High-frequency retention. We measure high-frequency retention as the fraction of spectral energy in the high-frequency band after smoothing, normalized by the corresponding energy in the original attention map. This metric captures fine-scale structure such as localized alignments, texture-like variation, and detailed spatial correlations.

Structural similarity. Structural similarity (SSIM) measures global agreement between the original and smoothed attention fields in terms of local intensity, contrast, and spatial correlation. While edge strength, peak preservation, and frequency retention capture specific local properties, SSIM reflects whether the overall spatial organization of the field is preserved.

Together, these metrics distinguish structure-preserving relaxation from indiscriminate smoothing. A Gaussian blur can reduce noise, but it also suppresses boundaries, peaks, and high-frequency structure. The LdG update is designed to dissipate irregular fluctuations while retaining anisotropic organization, so improvement across all four metrics provides direct evidence for the mechanism used in the main experiments.

A.7 Ablation

We provide additional experimental details for the guidance-strength and layer-placement ablation summarized in Section 6.3. The ablation varies two choices in the implemented guidance: the LdG step size η and the U-Net blocks to which the update is applied. We sweep $\eta \in \{0, 0.1, 0.25, 0.5, 1, 2, 5, 10, 20, 30\}$ across four placements - mid, up, down, and all blocks - giving 40 configurations. Each configuration generates and scores $30{,}000$ images from the same MS-COCO val2014 examples under a fixed per-image seed assignment, for $1{,}200{,}000$ generations in total. We use SDXL base 1.0 at 1024×1024 resolution with 25 `EulerDiscreteScheduler` steps and FP16 weights at a pinned revision. The sweep uses the repository's `diffuse` update on the Query tensor, followed by SEG-style residual guidance with scale 1; classifier-free guidance is disabled. All other settings, including each image's hardware assignment, are held fixed.

FID is computed with `clean-fid` against the same $30{,}000$ real images, and CLIP is the mean raw ViT-B/32 cosine similarity. The $\eta = 0$ rows retain the two-branch pipeline with an identity update and serve as matched controls. Their FID ranges from 50.5786 to 50.5809, a spread of 2.3×10^{-3} that quantifies agreement among the controls, rather than uncertainty across independent runs.

Figure 6 shows a strong interaction between guidance strength and placement. The small positive steps $\eta = 0.1$ and 0.25 worsen FID relative to the matched control for all four placements. The up, down, and all placements achieve their best tested FID at $\eta = 2$ (36.34, 31.56, and 35.08, respectively), then deteriorate at every larger tested step. For all blocks, FID rises from 35.08 at $\eta = 2$ to 79.52 at $\eta = 30$, which is 57.2% higher than its $\eta = 0$ control. Stronger perturbations therefore do not necessarily improve image quality, especially when applied throughout the U-Net. The CLIP optima differ: up, down, and all peak at $\eta = 10$, 5, and 2, respectively.

Mid-block placement behaves differently. Its FID initially increases to 52.13 at $\eta = 0.5$, then decreases at every tested step through $\eta = 30$, while CLIP increases over that same interval. At $\eta = 30$, FID reaches 20.7813, a 58.9% reduction from the matched control, and CLIP rises from 0.2665 to 0.2909. Extending the sweep from $\eta = 10$ to 30 lowers mid-block FID from 23.88 to 20.78, a further 13.0% reduction. Mid, $\eta = 30$ is the best tested configuration for both metrics, supporting mid-block placement in this experimental setting. These results establish a placement-dependent response to guidance strength, but do not isolate its mechanism. The largest tested step is still the best for mid blocks, so an optimum beyond $\eta = 30$ remains undetermined. Each configuration uses one fixed subset and seed assignment without independent repeats; we therefore report descriptive comparisons and make no statistical significance claims.

Table 3 reports the complete grid shown in Figure 6: all ten step sizes and four placements, with FID and CLIP for each of the 40 configurations.

Table 3: **Full $\eta\times$ placement grid on MS-COCO 30K.** All 40 configurations in Figure 6 are reported. Best value per row is bold. All cells use $n = 30{,}000$; CLIP is a raw ViT-B/32 cosine. The $\eta = 0$ column is the matched identity-update control.

	LdG step size η									
Applied blocks	0	0.1	0.25	0.5	1	2	5	10	20	30
FID ↓										
mid	50.58	51.28	51.88	52.13	50.62	44.73	31.52	23.88	21.21	**20.78**
up	50.58	52.13	53.03	51.14	41.77	**36.34**	45.80	54.76	61.85	64.39
down	50.58	52.57	53.91	52.77	46.06	**31.56**	37.33	47.39	51.62	52.44
all	50.58	54.73	56.56	50.47	36.13	**35.08**	58.79	71.48	77.77	79.52
CLIP ↑										
mid	0.2665	0.2661	0.2657	0.2656	0.2671	0.2715	0.2812	0.2874	0.2901	**0.2909**
up	0.2665	0.2660	0.2661	0.2679	0.2729	0.2761	0.2781	**0.2796**	0.2785	0.2782
down	0.2665	0.2660	0.2664	0.2683	0.2734	0.2824	**0.2864**	0.2855	0.2855	0.2853
all	0.2665	0.2652	0.2661	0.2707	0.2786	**0.2829**	0.2803	0.2776	0.2761	0.2751