C^{2}FG: Control Classifier-Free Guidance via Score Discrepancy Analysis Supplementary

Jiayang Gao[1,*], Tianyi Zheng[2,*], Jiayang Zou[1], Fengxiang Yang[2], Shice Liu[2], Luyao Fan[1], Zheyu Zhang[2], Hao Zhang[2], Jinwei Chen[2], Peng-Tao Jiang[2], Bo Li[2,†], Jia Wang[1,†]
[1] Shanghai Jiao Tong University, Shanghai, China
[2]vivo BlueImage Lab, vivo Mobile Communication Co., Ltd., China
{gjy0515, jiawang}@sjtu.edu.cn, {zhengtianyi, libra}@vivo.com

Appendix Overview

This appendix provides additional details and supplementary results to support the main paper. In Section 1, we review related literature to place our work in a broader context. Section 2 presents the detailed proofs of the theoretical results introduced in the main text. In Section 3, we further explore the connection between our MSE bound and Harnack-type inequalities, highlighting their theoretical implications. In Section 4, we show the difference between standard CFG and our method. Finally, Section 5 reports additional experimental results and visualizations.

1. Related Work

A scaling factor for conditional diffusion models was first introduced in CG [7], which controls the trade-off between fidelity and diversity:

$$\hat{\mu} = \mu_{\theta,\text{uncond}} + \gamma\, \Sigma_\theta(x_t, t) \nabla \log p_t(y \mid x_t), \tag{1}$$

where $\mu_{\theta,\text{uncond}}$ denotes the predicted mean of the unconditional denoiser, $\Sigma_\theta(x_t, t)$ is the predicted covariance (or noise scale) at step t, and $\nabla \log p_t(y \mid x_t)$ represents the conditional score function with respect to the label y. The hyperparameter γ is the classifier-free guidance scale: $\gamma > 1$ strengthens conditioning at the cost of diversity, while $\gamma < 1$ weakens conditioning but increases sample diversity. This scaling modifies the reverse sampling distribution as:

$$\tilde{p}(x_{t-1} \mid x_t, y) = \frac{p(x_{t-1} \mid x_t)\, p^\gamma(y \mid x_{t-1})}{Z(x_t, y)}, \quad Z(x_t, y) = \sum_{x_{t-1}} p(x_{t-1} \mid x_t)\, p^\gamma(y \mid x_{t-1}). \tag{2}$$

Then CFG [11] eliminates the need for an external classifier by jointly training the network for both conditional and unconditional predictions:

$$\epsilon_\theta(x_t, t, y) = \mu_{\theta,\text{uncond}} + \Sigma_\theta(x_t, t) \nabla \log p_t(y \mid x_t), \quad \epsilon_\theta(x_t, t, \phi) = \mu_{\theta,\text{uncond}}. \tag{3}$$

where ϵ_θ is the neural network's output for noise prediction. Substituting these into equation 1 and setting $\gamma = w$ yields the CFG formulation:

$$\hat{\epsilon}(x_t, t, y) = w\, [\epsilon_\theta(x_t, t, y) - \epsilon_\theta(x_t, t, \phi)] + \epsilon_\theta(x_t, t, \varnothing), \tag{4}$$

We see that CFG and CG are using the same scaling factor. And for now CFG with this scaling technique that has been widely adopted in mainstream diffusion models, typically with a fixed CFG-scale.

However, recent studies have pointed out that using a constant guidance weight is not necessarily optimal and may lead to limitations in balancing fidelity and diversity. Specifically, several works have proposed various forms of dynamic or time-dependent scaling strategies to improve generation quality. [23] proposes Frequency-Decoupled Guidance (FDG), an improved version of classifier-free guidance that operates in the frequency domain, which chooses a low cfg-scale for low

frequencies and a high cfg-scale for high frequencies. [15] observe that applying a constant classifier-free guidance (CFG) weight across all noise levels is suboptimal: guidance harms diversity in the high-noise regime, has little effect in the low-noise regime, and is only beneficial in the middle. They propose restricting guidance to a limited interval of noise levels, which both improves sample fidelity and diversity while reducing computational cost. [20] propose a geometric guidance method for CG to address the vanishing gradient issue in late denoising stages of probabilistic approaches. Its core innovation enforces fixed-length gradient updates ($\|\nabla p\|$-normalized) proportional to data dimension ($\sqrt{D}/T$), maintaining consistent guidance strength throughout sampling. [17] rescale classifier-free guidance to prevent over-exposure. [18] propose β-adaptive scaling to address the trade-off between image quality and prompt alignment in standard CFG. It dynamically adjusts guidance strength via a time-dependent β-distribution $\beta(t)$, enforcing weak guidance at initial/final steps and strong guidance during critical mid-denoising phases. [26] investigate different time-dependent schedulers for the guidance weight. Their analysis and experiments confirm that dynamic weighting strategies outperform fixed weights, with high weights being beneficial in the mid-noise regime but detrimental at the extremes. [6] and [4] improve diffusion model performance by constraining CFG to the data manifold, enabling higher-quality generation, better inversion, and smoother interpolation at lower guidance scales. [24] mitigate spatial inconsistency in classifier-free guidance by introducing Semantic-aware CFG, which segments latent images into semantic regions via attention maps and adaptively assigns region-specific guidance scales, leading to more balanced semantics and higher-quality generations. [25] propose Diffusion-NPO, which incorporates non-parametric optimization into diffusion sampling via nearest-neighbor matching, improving sample diversity and quality without retraining and working across different models and datasets. [3] investigate PCG, a theoretical foundations of CFG, and reveal the differrence and relationship between DDPM and DDIM, embedding CFG in a broader design space of principled sampling methods. [27] introduce TFG, a framework which encompasses existing methods as special cases. In their framework, they defined a hyper-parameter space for their algorithm and analyze the underlying theoretical motivation of each hyper-parameter.

As for recent work, **RAAG** [29] recompute ω at every reverse step via a lightweight exponential map of the current RATIO: $\omega(\rho) = 1 + (\omega_{\max} - 1)\exp(-\alpha\rho)$, which is similar to the form of our C^{2}FG. However, RAAG is primarily designed for text-to-image generation under strong conditioning, whereas our analysis highlights intrinsic properties of diffusion dynamics, making the applicability of our framework broader and not restricted to text-to-image tasks. *Besides, their exponential design is motivated by empirical intuition, while ours is supported by formal theorems, providing a stronger theoretical grounding.* **Stage-wise Dynamic of CFG** [12] analyze CFG under multimodal conditionals and show that the sampling process can be seen as three successive stages: early *direction drift*, mid *mode separation* and late *constration*. And then propose a stage-wise guidance schedule. However, their dynamic guidance schedule can actually be seen as a piece-wise CFG guidance $\omega(t)$ just like [15]. Moreover, their method seems relies on task-specific properties such as large guidance scales $\omega(t)$, suffering from a lack of generality. **S^2-Guidance** [5] propose a novel method in which network parameters are stochastically masked to form a latent subnetwork during each forward pass, guiding the model away from potential low-quality predictions and toward higher-quality outputs. However, their approach also relies on task-specific settings, such as text-to-image and text-to-video generation. [22] introduce an extension of CFG called **TSG**, whose motivation is based on the structure of diffusion network. They compute the model outputs for the clean time-step embedding and a perturbed embedding and use their difference to guide the sampling. Importantly, the motivation behind TSG arises from the interaction between conditional and timestep embeddings within the network architecture, rather than from a theoretical analysis of the diffusion framework itself.

While these approaches have shown promising improvements, they are still largely heuristic in nature and often lack rigorous theoretical justification, leaving the principles of adaptive weight design not fully understood. To address this gap, our work provides a theoretical foundation for adaptive guidance. By establishing a sequence of results (Theorems 1–4), we uncover structural properties of diffusion processes under different initial distributions. These insights naturally motivate the design of adaptive, theoretically grounded scaling functions. In this way, our framework offers a more robust and general basis for conditional generation.

2. Proof of Theorems

In this section we give the proof of theorems below:

Theorem 1 (VP-SDE Score MSE Bound)**.** *Assume that the sample space is bounded and closed. Then we consider the VP-SDE*

$$\mathrm{d}x_t = -\frac{1}{2}\beta(t)x_t\,\mathrm{d}t + \sqrt{\beta(t)}\,\mathrm{d}w_t, \tag{5}$$

let $p(x,t)$ and $\tilde{p}(x,t)$ denote the probability densities at time t, induced by initial distributions $p(x_0)$ and $\tilde{p}(x_0)$, respectively.

Then, the mean-square error (MSE) between the scores satisfies the uniform bound

$$\|\nabla \log p(x,t) - \nabla \log \tilde{p}(x,t)\| \leq \frac{\alpha(t)}{\sigma^2(t)} C, \quad \forall x \in \textit{supp},\ t \geq 0, \tag{6}$$

where C is a constant, $\alpha(t) = \exp\left(-\frac{1}{2}\int_0^t \beta_s \mathrm{d}s\right)$, and $\sigma(t) = \alpha(t)\sqrt{\int_0^t \frac{\beta_s}{\alpha^2(s)}\mathrm{d}s}$.

Theorem 2 (VE-SDE Score MSE Bound)**.** *Assume that the sample space is bounded and closed. Then we consider the VE-SDE*

$$\mathrm{d}x_t = \sqrt{\frac{\mathrm{d}\sigma_t^2}{\mathrm{d}t}}\mathrm{d}w_t, \tag{7}$$

let $p(x,t)$ and $\tilde{p}(x,t)$ denote the probability densities induced by initial distributions $p(x_0)$ and $\tilde{p}(x_0)$, respectively. Assume that the sample space is bounded and closed.

Then, the mean-square error (MSE) between the conditional and unconditional scores satisfies the uniform bound

$$\|\nabla \log p(x,t) - \nabla \log \tilde{p}(x,t)\| \leq \frac{C}{\sigma^2(t)}, \quad \forall\, x \in \textit{supp},\ t \geq 0, \tag{8}$$

where C is a constant.

Theorem 3 (Harnack-type Inequality of VP-SDE)**.** *Let $p(x_t,t) \in C^{2,1}(\mathbb{R}^n \times [0,+\infty))$ denote the probability density function of the VP-SDE equation 5, and define*

$$s(t) = \frac{1}{2}\int_0^t \beta_r dr, \ \ t(s) = s^{-1}(t).$$

Then for any $\alpha > 1, x_1, x_2 \in \mathbb{R}^n, 0 < s_1 < s_2 < +\infty$, the following inequality holds:

$$p(x_1, t(s_1)) \leq p(x_2, t(s_2)) \left(\frac{s_2}{s_1}\right)^{\frac{m\alpha}{2}} \exp\left(\frac{\alpha^2\|x_1 - x_2\|^2}{4(s_2 - s_1)} + \frac{\|x_2\|^2 - \|x_1\|^2}{2}\right), \tag{9}$$

where $m \geq n$ and $\|\cdot\|$ denotes the Euclidean distance.

Theorem 4 (Harnack-type Inequality of VE-SDE)**.** *Similarly, let $p(x_t,t) \in C^{2,1}(\mathbb{R}^n \times [0,+\infty))$ denote the probability density function of the VE-SDE equation 7, and define*

$$s(t) = \sigma_t^2, \ \ t(s) = s^{-1}(t).$$

Then for any $\alpha > 1, x_1, x_2 \in \mathbb{R}^n, 0 < s_1 < s_2 < +\infty$, the following inequality holds for p:

$$p(x_1, t(s_1))) \leq p(x_2, t(s_2)) \left(\frac{s_2}{s_1}\right)^{\frac{n\alpha}{2}} \exp\left(\frac{\alpha^2\|x_1 - x_2\|^2}{4(s_2 - s_1)}\right). \tag{10}$$

2.1. Proof of Theorem 1

Proof of Theorem 1. For VP-SDE

$$\mathrm{d}x_t = -\frac{1}{2}\beta_t x_t \mathrm{d}t + \sqrt{\beta_t}\mathrm{d}w_t, \tag{11}$$

we can represent x_t with x_0:

$$x_t = \alpha(t)x_0 + \sigma(t)\xi_t, \tag{12}$$

where $\alpha(t) = \exp\left(-\frac{1}{2}\int_0^t \beta_s \mathrm{d}s\right), \sigma(t) = \alpha(t)\sqrt{\int_0^t \frac{\beta_s}{\alpha^2(s)}\mathrm{d}s}$, and $\xi_t \sim \mathcal{N}(0, I)$.

Hence we can get the $p(x_t, t|x_0)$:

$$p(x, t|x_0) = \frac{1}{(2\pi\sigma^2(t))^{n/2}} \exp\Big(- \frac{||x - \alpha(t)x_0||^2}{2\sigma^2(t)}\Big), \tag{13}$$

by using Bayes formula, we can get theprobablity density function :

$$p(x, t) = \int_{\mathbb{R}^n} \frac{1}{(2\pi\sigma^2(t))^{n/2}} \exp\Big(- \frac{||x - \alpha(t)x_0||^2}{2\sigma^2(t)}\Big) p(x_0)\mathrm{d}x_0, \tag{14}$$

then we can get the score:

$$\begin{aligned}
\nabla \log p(x, t) &= \frac{\nabla p(x_t, t)}{p(x_t, t)} && (15)\\
&= \frac{\int_{\mathbb{R}^n} \frac{\alpha(t)x_0 - x}{\sigma^2(t)} \exp\Big(- \frac{||x-\alpha(t)x_0||^2}{2\sigma^2(t)}\Big) p(x_0)\mathrm{d}x_0}{\int_{\mathbb{R}^n} \exp\Big(- \frac{||x-\alpha(t)x_0||^2}{2\sigma^2(t)}\Big) p(x_0)\mathrm{d}x_0} && (16)\\
&= \frac{1}{\sigma^2(t)}\Big(\alpha(t)\mathbb{E}[x_0|x_t = x] - x\Big). && (17)
\end{aligned}$$

Denote that $p(x_0|y) = \tilde{p}(x_0)$, consider the MSE:

$$\|\nabla \log p(x, t) - \nabla \log \tilde{p}(x, t)\| = \frac{\alpha(t)}{\sigma^2(t)} \|\mathbb{E}_{x_0\sim p}[x_0|x_t = x] - \mathbb{E}_{x_0'\sim \tilde{p}}[x_0'|x_t = x]\|, \tag{18}$$

then we try bounding $f(t, x) = \|\mathbb{E}_{x_t\sim p_t}[x_0|x_t = x] - \mathbb{E}_{x_t'\sim \tilde{p}_t}[x_0'|x_t' = x]\|$ term. Assume that $f(t, x)$ is a smooth function on $\mathbb{R}^n \times [0, +\infty)$, it's easy to find that

$$f(0, x) = 0, f(+\infty, x) = \|\mathbb{E}_{x_0\sim p}[x_0] - \mathbb{E}_{x_0'\sim \tilde{p}}[x_0']\|,$$

hence $f(t, x)$ is a bounded function on t, and we denote its bound by $C(x)$.Note that we cannot say that when $t \to 0$, $\frac{\alpha(t)}{\sigma^2(t)} \|\mathbb{E}_{x_0\sim p}[x_0|x_t = x] - \mathbb{E}_{x_0'\sim \tilde{p}}[x_0'|x_t = x]\| \to 0$, because $\sigma(t) \to 0$, too.

In practical engineering applications of diffusion models, the sample space is often assumed to be compact, reflecting the fact that physical quantities are naturally limited and numerical simulations are performed on finite domains. So $C(x)$ can be bounded by C without loss of convince.Assume that we talk about x_0 on any bounded domain K with $\sup_{z\in K} |z| \le R$. Let total variation distance be $\mathrm{TV}(\mu, \nu) = \frac{1}{2} \int |\mu(\mathrm{d}x) - \nu(\mathrm{d}x)|$

$$\begin{aligned}
f(t, x) &= \big\|\mathbb{E}_{x_t\sim p_t}\big[X_0 \mid x_t = x\big] - \mathbb{E}_{x_t\sim \tilde{p}_t}\big[X_0 \mid x_t = x\big]\big\| \\
&= \Big\| \int x_0 \big(p(x_0 \mid x_t = x) - \tilde{p}(x_0 \mid x_t = x)\big)\, \mathrm{d}x_0 \Big\| \\
&\le 2M \cdot \mathrm{TV}\big(p(\cdot \mid x_t = x),\, \tilde{p}(\cdot \mid x_t = x)\big) \\
&\le 2R = C.
\end{aligned}$$

Then we can rewrite equation 18

$$\|\nabla \log p(x, t) - \nabla \log \tilde{p}(x, t)\| \le \frac{2\alpha(t)}{\sigma^2(t)} R. \tag{19}$$

□

2.2. Proof of Theorem 2

Proof of Theorem 2. For VE-SDE

$$\mathrm{d}x_t = \sqrt{\frac{\mathrm{d}\sigma_t^2}{\mathrm{d}t}}\mathrm{d}w_t. \tag{20}$$

we can represent x_t with x_0:

$$x_t = x_0 + \sigma(t)\xi_t, \tag{21}$$

where $\xi_t \sim \mathcal{N}(0, I)$.

Like the proof of Theorem 1, we have

$$\|\nabla \log p(x,t) - \nabla \log \tilde{p}(x,t)\| \leq \frac{1}{\sigma^2(t)}C. \tag{22}$$

□

2.3. Proof of Theorem 3

First we give Lemma 1 and Lemma 2 without proof as below:

Lemma 1 (Cut-off Function [10])**.** *There exists a cut-off function $\eta \in C_c^\infty(B_R)$ with $0 \leq \eta \leq 1$, such that $\eta \equiv 1$ on $B_{\frac{R}{2}}$, and for any $x \in \mathbb{R}^n$,*

$$|\nabla\eta|(x) \leq \frac{C}{R}\eta^{\frac{1}{2}}, \quad \Delta\eta(x) \geq -\frac{C}{R^2} \tag{23}$$

where $C > 0$ depends only on the dimension n.

Lemma 2 (Bochner formula and Bakry–Émery Inequality of Heat equation with Witten Laplacian [2])**.** *Define linear operator $\mathrm{L} = \Delta - \nabla\phi \cdot \nabla$, and $\nabla^2\phi$ is positive semi-definite, then for any $g \in C^3$, we have*

$$\frac{1}{2}\mathrm{L}|\nabla g|^2 = |\nabla^2 g|^2 + \langle \nabla g, \nabla \mathrm{L} g\rangle + \nabla g^T \nabla^2 \phi \nabla g, \tag{24}$$

and furthermore

$$\frac{1}{2}\mathrm{L}|\nabla g|^2 \geq \frac{|\mathrm{L}g|^2}{m} + \langle \nabla g, \nabla \mathrm{L} g\rangle + \nabla g^T \nabla^2 \phi \nabla g \tag{25}$$

where $|\nabla^2 g|^2 = \Sigma_{i,j=1}^n (\partial_{ij} g)^2$ and $m \geq n$ denotes the virtual dimension.

Then we first prove such lemma:

Lemma 3 (Cut-off Function for Heat Equation with Witten Laplacian)**.** *There exists a cut-off function $\eta \in C_c^\infty(B_R)$ with $0 \leq \eta \leq 1$, such that $\eta \equiv 1$ on $B_{\frac{R}{2}}$, and for any $x \in \mathbb{R}^n$, $\phi = k(|x|)x, k \geq 0$ on B_R,*

$$|\nabla\eta|(x) \leq \frac{C}{R}\eta^{\frac{1}{2}}, \quad \Delta\eta(x) \geq -\frac{C}{R^2}, \quad \nabla\phi \cdot \nabla\eta(x) \leq 0, \tag{26}$$

$C > 0$ depends only on the dimension n.

Proof. **Step 1. Construction of the cutoff.** We construct a radial cutoff function by setting

$$\eta(x) = \psi\left(\frac{|x|}{R}\right),$$

where $\psi \in C_c^\infty([0,\infty))$ satisfies:

$$\psi \equiv 1 \text{ on } [0, 1/2], \qquad \psi \equiv 0 \text{ on } [1, \infty), \qquad \psi' \leq 0,$$

together with the standard cutoff estimates

$$|\psi'| \le C\sqrt{\psi}, \qquad |\psi''| \le C.$$

Step 2. Gradient estimate. Writing $r = |x|$, we compute

$$\nabla\eta(x) = \frac{1}{R}\psi'\big(\tfrac{r}{R}\big)\,\frac{x}{r}.$$

Hence

$$|\nabla\eta(x)| \le \frac{1}{R}\big|\psi'\big(\tfrac{r}{R}\big)\,\big| \le \frac{C}{R}\sqrt{\eta(x)}.$$

Step 3. Laplacian estimate. Using the radial Laplacian formula, we have

$$\Delta\eta(x) = \frac{1}{R^2}\psi''\big(\tfrac{r}{R}\big) + \frac{n-1}{rR}\psi'\big(\tfrac{r}{R}\big)\,.$$

The first term is bounded by C/R^2 since $|\psi''| \le C$. For the second term, note that $\psi' = 0$ when $r \le R/2$, and for $r \in [R/2, R]$, we have

$$\left|\frac{n-1}{rR}\psi'\big(\tfrac{r}{R}\big)\right| \le \frac{C}{R^2}.$$

Therefore

$$|\Delta\eta(x)| \le \frac{C}{R^2}, \qquad \Delta\eta(x) \ge -\frac{C}{R^2}.$$

Step 4. Witten Laplacian estimate. Finally,

$$\mathrm{L}\eta(x) = \Delta\eta(x) - kx\cdot\nabla\eta(x).$$

Since

$$x\cdot\nabla\eta(x) = \frac{r}{R}\psi'\big(\tfrac{r}{R}\big)\,,$$

and $\psi' \le 0$, the term $k(x)x\cdot\nabla\eta(x) \le 0$. □

Based on this we give proof of Theorem 5:

Theorem 5 (Gradient Estimate of Heat Equation with Witten Laplacian)**.** *Let u be a positive solution to the heat equation*

$$\partial_t u = (\Delta - \nabla\phi\cdot\nabla)u, \tag{27}$$

on $(0,T]\times B_R$. Assume that $\nabla^2\phi$ is positive semi-definite, $\phi = k(|x|)x, k \ge 0$ on B_R, then for any $(t,x) \in (0,T]\times B_{\frac{R}{2}}$, the following inequality holds:

$$\frac{|\nabla u|^2}{u^2} - \alpha\frac{\partial_t u}{u} \le \frac{m\alpha^2}{2t} + \frac{C\alpha^2}{R^2}\Big(1 + \frac{\alpha^2}{\alpha-1}\Big), \tag{28}$$

where $m \ge n$ denotes the virtual dimension, $C(m,n)$ is a constant depends on (m,n).

Proof. We define linear operator $\mathrm{L} = \Delta - \nabla\phi\cdot\nabla$, and function $f = \log u, F = t(|\nabla f|^2 - \alpha\partial_t f)$, then applying it into equation 27, we have

$$\partial_t f = \mathrm{L}f + |\nabla f|^2, \tag{29}$$

$$\mathrm{L}f = -|\nabla f|^2 + \partial_t f = -\frac{F}{\alpha t} - \frac{\alpha-1}{\alpha}|\nabla f|^2, \tag{30}$$

$$\Delta f = \mathrm{L}f + \langle\nabla\phi, \nabla f\rangle = -\frac{F}{\alpha t} + \left\langle \nabla\phi - \frac{\alpha-1}{\alpha}\nabla f, \nabla f\right\rangle. \tag{31}$$

Based on Lemma 2 and equation 29,equation 30, equation 31, we can get

$$\begin{aligned}
\mathrm{L}F &= t((\Delta - \nabla\phi\cdot\nabla)|\nabla f|^2 - \alpha\partial_t((\Delta-\nabla\phi\cdot\nabla)f)) \\
&= t\left(2|\nabla^2 f|^2 + 2\langle\nabla f, \nabla \mathrm{L}f\rangle - \alpha\partial_t(\mathrm{L}f) + 2\nabla f^T\nabla^2\phi\nabla f\right) \\
&\geq t\left(\frac{2}{m}|\mathrm{L}f|^2 + 2\langle\nabla f,\nabla\mathrm{L}f\rangle - \alpha\partial_t(\mathrm{L}f) + 2\nabla f^T\nabla^2\phi\nabla f\right) \\
&\geq t\frac{2|-\frac{F}{\alpha t} + \left\langle -\frac{\alpha-1}{\alpha}\nabla f,\nabla f\right\rangle|^2}{m} \\
&\quad + t\left(2\left\langle\nabla f,\nabla\left(-\frac{F}{\alpha t} - \frac{\alpha-1}{\alpha}|\nabla f|^2\right)\right\rangle - \alpha\partial_t(-\frac{F}{\alpha t} - \frac{\alpha-1}{\alpha}|\nabla f|^2)\right) \\
&= \left(\frac{2}{m\alpha^2}\left(\frac{F^2}{t} + 2(\alpha-1)F|\nabla f|^2 + (\alpha-1)^2 t|\nabla f|^4\right)\right) - \frac{2}{\alpha}\langle\nabla f,\nabla F\rangle \\
&\quad - \frac{2(\alpha-1)}{t}\alpha\left\langle\nabla f,\nabla|\nabla f|^2\right\rangle - \frac{F}{t} + \partial_t F + 2(\alpha-1)t\langle\nabla f,\partial_t\nabla f\rangle \\
&\geq \left(\frac{2}{m\alpha^2}\left(\frac{F^2}{t} + 2(\alpha-1)F|\nabla f|^2\right)\right) - \frac{2}{\alpha}\langle\nabla f,\nabla F\rangle - \frac{F}{t} + \partial_t F \\
&\quad + \frac{2(\alpha-1)}{t}\alpha\left\langle\nabla f,\nabla(-|\nabla f|^2 + \alpha\partial_t\nabla f)\right\rangle \\
&= \left(\frac{2}{m\alpha^2}\left(\frac{F^2}{t} + 2(\alpha-1)F|\nabla f|^2\right)\right) - 2\langle\nabla f,\nabla F\rangle - \frac{F}{t} + \partial_t F,
\end{aligned}$$

hence we have

$$(\partial_t - \mathrm{L})F \leq -\left(\frac{2}{m\alpha^2}\left(\frac{F^2}{t} + 2(\alpha-1)F|\nabla f|^2\right)\right) + \frac{F}{t} + 2\langle\nabla f,\nabla F\rangle. \tag{32}$$

Let us consider the cut-off function η which satisfies $\langle\nabla\phi,\nabla\eta\rangle \geq 0$(Lemma 3). We use the Bochner technique to estimate its upper bound, $\forall T' \in (0,T]$, suppose ηF attains its maximum over $(0,T']\times\bar{B}_R$ at (t_0,x_0). Without loss of generality, assume $(\eta F)(t_0,x_0) > 0$; otherwise, the conclusion of the theorem holds trivially. Consequently, we have $\eta(x_0), F(t_0,x_0) > 0$, which implies $x_0 \notin \partial B_R, t_0 > 0$. Thus, (t_0,x_0) lies in the interior of $(B_R)_T$. Then we consider

$$\begin{aligned}
(\partial_t - \mathrm{L})(\eta F) &= -F\cdot\mathrm{L}\eta - 2\langle\nabla\eta,\nabla F\rangle + \eta(\partial_t - \mathrm{L})F \\
&= -F\cdot\Delta\eta + F\cdot\langle\nabla\phi,\nabla\eta\rangle - 2\langle\nabla\eta,\nabla F\rangle + \eta(\partial_t-\mathrm{L})F \\
&\leq \frac{C}{R^2}F - 2\langle\nabla\eta,\nabla F\rangle + F\cdot\langle\nabla\phi,\nabla\eta\rangle \\
&\quad + \eta\left(-\left(\frac{2}{m\alpha^2}\left(\frac{F^2}{t} + 2(\alpha-1)F|\nabla f|^2\right)\right) + \frac{F}{t} + 2\langle\nabla f,\nabla F\rangle\right).
\end{aligned}$$

Applying $\nabla F = \frac{\nabla(\eta F)}{\eta} - \frac{\nabla\eta}{\eta}F$, we have

$$\begin{aligned}
(\partial_t - \mathrm{L})(\eta F)(t_0,x_0) &\leq \frac{C}{R^2}F - \frac{2}{\eta}\langle\nabla\eta,\nabla(\eta F))\rangle + 2\frac{|\nabla\eta|^2}{\eta}F + F\cdot\langle\nabla\phi,\nabla\eta\rangle \\
&\quad + \eta\left(-\left(\frac{2}{m\alpha^2}\left(\frac{F^2}{t_0} + 2(\alpha-1)F|\nabla f|^2\right)\right) + \frac{F}{t_0} + 2\langle\nabla f,\nabla F\rangle\right)
\end{aligned}$$

Using the properties of maximum

$$\nabla(\eta F)(t_0,x_0) = 0, \Delta(\eta F)(t_0,x_0) \leq 0, \partial_t(\eta F)(t_0,x_0) = 0,$$

and applying Lemma 1 so that

$$0 \leq \frac{C+2C^2}{R^2}F - \frac{2}{m\alpha^2}\frac{\eta F^2}{t_0} - \frac{4(\alpha-1)}{m\alpha^2}\eta F|\nabla f|^2 + \frac{\eta F}{t_0} \tag{33}$$

$$+ 2\eta\left\langle \nabla f, \nabla F\right\rangle + F\cdot\left\langle \nabla\phi, \nabla\eta\right\rangle \tag{34}$$

$$= \frac{C+2C^2}{R^2}F - \frac{2}{m\alpha^2}\frac{\eta F^2}{t_0} - \frac{4(\alpha-1)}{m\alpha^2}\eta F|\nabla f|^2 + \frac{\eta F}{t_0} \tag{35}$$

$$+ 2\left\langle \nabla f, \nabla(\eta F))\right\rangle - 2F\left\langle \nabla f, \nabla\eta\right\rangle + F\cdot\left\langle \nabla\phi, \nabla\eta\right\rangle \tag{36}$$

$$= \frac{C+2C^2}{R^2}F - \frac{2}{m\alpha^2}\frac{\eta F^2}{t_0} - \frac{4(\alpha-1)}{m\alpha^2}\eta F|\nabla f|^2 + \frac{\eta F}{t_0} \tag{37}$$

$$- 2F\left\langle \nabla f, \nabla\eta\right\rangle + F\cdot\left\langle \nabla\phi, \nabla\eta\right\rangle, \tag{38}$$

then let us consider two of the terms $\frac{4(\alpha-1)}{m\alpha^2}\eta F|\nabla f|^2 + 2F\left\langle \nabla f, \nabla\eta\right\rangle$,

$$\begin{aligned}\frac{4(\alpha-1)}{m\alpha^2}\eta F|\nabla f|^2 + 2F\left\langle \nabla f, \nabla\eta\right\rangle &\geq \frac{4(\alpha-1)}{m\alpha^2}\eta F|\nabla f|^2 - 2F|\nabla f||\nabla\eta| \\ &\geq \frac{4(\alpha-1)}{m\alpha^2}\eta F|\nabla f|^2 \\ &\quad - F(\frac{4(\alpha-1)R^2}{m\alpha^2C^2}|\nabla f|^2|\nabla\eta|^2 + \frac{m\alpha^2C^2}{4(\alpha-1)R^2}) \\ &\geq -\frac{m\alpha^2C^2}{4(\alpha-1)R^2}F\end{aligned}$$

then inequality 38 can be turn into

$$0 \leq \left(\frac{m\alpha^2C^2}{4(\alpha-1)R^2} + \frac{C+2C^2}{R^2}\right)F - \frac{2}{m\alpha^2}\frac{\eta F^2}{t_0} + \frac{\eta F}{t_0} + F\cdot\left\langle \nabla\phi, \nabla\eta\right\rangle,$$

then we divide F and then get

$$\begin{aligned}\eta F(t_0, x_0) &\leq \frac{m\alpha^2}{2}t_0\left(\frac{m\alpha^2C^2}{4(\alpha-1)R^2} + \frac{C+2C^2}{R^2} + \frac{\eta}{t_0} + \left\langle \nabla\phi, \nabla\eta\right\rangle\right) \\ &\leq \frac{m\alpha^2}{2}t_0\left(\frac{m\alpha^2C^2}{4(\alpha-1)R^2} + \frac{C+2C^2}{R^2} + \frac{1}{t_0}\right) \\ &\leq \frac{m\alpha^2}{2} + \frac{m\alpha^2}{2}\left(\frac{m\alpha^2C^2}{4(\alpha-1)R^2} + \frac{C+2C^2}{R^2}\right)t_0 \\ &\leq \frac{m\alpha^2}{2} + \frac{C_1\alpha^2}{R^2}\left(\frac{\alpha^2}{\alpha-1} + 1\right)T', \\ &(C_1 = \max\{m^2C^2/8, C^2 + C/2\}),\end{aligned}$$

On $B_{\frac{R}{2}}$, $\eta = 1, \nabla\eta = 0$, so for all $(t, x) \in (0, T'] \times B_{\frac{R}{2}}$

$$\begin{aligned}t(|\nabla f|^2 - \alpha\partial_t f)|_{t=T'} &= F(T', x) = \eta F(T', x) \leq \eta F(t_0, x_0) \\ &\leq \frac{m\alpha^2}{2} + \frac{C_1\alpha^2}{R^2}\Big(\frac{\alpha^2}{\alpha-1} + 1\Big)T',\end{aligned}$$

T' is arbitrary, so

$$(|\nabla f|^2 - \alpha\partial_t f) \leq \frac{m\alpha^2}{2t} + \frac{C_1\alpha^2}{R^2}\Big(\frac{\alpha^2}{\alpha-1} + 1\Big). \tag{39}$$

□

From Theorem 5 we can conclude Theorem 6

Theorem 6 (Harnack-type Inequality of Heat Equation with Witten Laplacian). *Let u be a positive solution of the heat equation $\partial_t u = \mathrm{L}u$ in $(0,T] \times B_R$, where $\alpha > 1$. For any $x_1, x_2 \in B_{\frac{R}{2}}$ and $0 < t_1 < t_2 \leq T$, the following inequality holds:*

$$u(x_1,t_1) \leq u(x_2,t_2)\left(\frac{t_2}{t_1}\right)^{\frac{m\alpha}{2}} \exp\left(\frac{\alpha^2\|x_1-x_2\|^2}{4(t_2-t_1)} + \frac{C\alpha}{R^2}\left(1+\frac{\alpha^2}{\alpha-1}\right)(t_2-t_1)\right), \tag{40}$$

where $C = C(m,n)$.

Proof. Let $f = \log u$. Consider the line segment

$$L(s) = (1-s)(t_2,x_2) + s(t_1,x_1).$$

We have

$$\begin{aligned}\log\frac{u(x_1,t_1)}{u(x_2,t_2)} &= \int_0^1 \frac{d}{ds} f(L(s))\,ds \\ &= \int_0^1 \left[\nabla f(L(s))\cdot(x_1-x_2) + \partial_t f(L(s))(t_1-t_2)\right] ds.\end{aligned}$$

Moreover, using the inequality

$$-\partial_t f \leq -\frac{1}{\alpha}|\nabla f|^2 + \frac{m\alpha}{2t} + \left[\frac{C\alpha}{R^2}\left(\frac{\alpha^2}{\alpha-1}+1\right)\right],$$

we get

$$\begin{aligned}\log\frac{u(x_1,t_1)}{u(x_2,t_2)} \leq \int_0^1 \Big[&|\nabla f(L(s))|\,|x_1-x_2| \\ &+ \Big(-\frac{1}{\alpha}|\nabla f|^2(L(s)) + \frac{m\alpha}{2[(1-s)t_2+st_1]} \\ &+ \frac{C\alpha}{R^2}\left(\frac{\alpha^2}{\alpha-1}+1\right)\Big)(t_2-t_1)\Big] ds.\end{aligned}$$

Using the inequality

$$|\nabla f(L(s))|\,|x_1-x_2| - \frac{t_2-t_1}{\alpha}|\nabla f|^2(L(s)) \leq \frac{\alpha\, d^2(x_1,x_2)}{4(t_2-t_1)},$$

we obtain

$$\log\frac{u(x_1,t_1)}{u(x_2,t_2)} \leq \frac{\alpha\, d^2(x_1,x_2)}{4(t_2-t_1)} + \frac{m\alpha}{2}\ln\frac{t_2}{t_1} + \frac{C\alpha}{R^2}\left(\frac{\alpha^2}{\alpha-1}+1\right)(t_2-t_1).$$

□

Finally, we can prove Theorem 3:

Proof of Theorem 3. The VP-SDE is given by

$$\mathrm{d}x_t = -\frac{1}{2}\beta_t x_t \mathrm{d}t + \sqrt{\beta_t}\mathrm{d}W_t, \tag{41}$$

and its corresponding Fokker-Planck equation (FPE) is

$$\frac{\partial p_t(x)}{\partial t} = \frac{1}{2}\beta_t\left(\nabla_x\cdot[xp_t(x)] + \Delta_x p_t(x)\right). \tag{42}$$

We can reparameterize t by letting $ds = \frac{1}{2}\beta_t dt$. Then,

$$s(t) = \frac{1}{2}\int_0^t \beta_r dr, \tag{43}$$

$$\frac{\mathrm{d}}{\mathrm{d}t} = \frac{1}{2}\beta_t \frac{\mathrm{d}}{\mathrm{d}s}. \tag{44}$$

Thus,

$$\frac{\partial p_{t(s)}(x)}{\partial s} = \frac{\partial p_t}{\partial t}\frac{\mathrm{d}t}{\mathrm{d}s} = \frac{\partial p_t(x)}{\partial t}\cdot\frac{1}{\frac{1}{2}\beta_t} = \nabla_x\cdot[xp_t(x)] + \Delta_x p_t(x). \tag{45}$$

For this new FPE

$$\frac{\partial p_{t(s)}(x)}{\partial s} = \nabla_x\cdot[xp_{t(s)}(x)] + \Delta_x p_{t(s)}(x), \tag{46}$$

the corresponding SDE is

$$\mathrm{d}x_{t(s)} = -x_{t(s)}\mathrm{d}s + \sqrt{2}\mathrm{d}W_s. \tag{47}$$

Assume $p(x,t)$ is a positive solution to this FPE, and let $u(x,t) = p(x,t)e^{|x|^2/2}$. Computing the right-hand side:

$$\nabla(xp) = x\nabla p + np = (nu + x\nabla u - |x|^2 u)e^{-|x|^2/2}, \tag{48}$$

$$\Delta p = \nabla\cdot[(\nabla u - xu)e^{-x^2/2}] = [\Delta u - nu - 2x\nabla u + |x|^2 u]e^{-|x|^2/2}, \tag{49}$$

$$\nabla(xp) + \Delta p = [\Delta u - x\nabla u]e^{-|x|^2/2}. \tag{50}$$

Thus, the FPE for u is

$$\frac{\partial u_{t(s)}(x)}{\partial s} = \Delta u - x\cdot\nabla u = \Delta u - \nabla\phi\cdot\nabla u, \quad \phi = \frac{|x|^2}{2}, \tag{51}$$

which satisfies the equation in **Theorem 6**, and we can easily figure out that $k(|x|) = 1 > 0$.

Therefore, for any $\alpha > 1, x_1, x_2 \in M, 0 < s_1 < s_2 < +\infty$, and let $R \to \infty$, the following inequality holds:

$$u(x_1, t(s_1)) \le u(x_2, t(s_2))\left(\frac{s_2}{s_1}\right)^{\frac{m\alpha}{2}} \exp\left(\frac{\alpha^2\|x_1 - x_2\|^2}{4(s_2 - s_1)}\right). \tag{52}$$

Rewriting it in terms of p, we obtain

$$p(x_1, t(s_1)) \le p(x_2, t(s_2))\left(\frac{s_2}{s_1}\right)^{\frac{m\alpha}{2}} \exp\left(\frac{\alpha^2\|x_1 - x_2\|^2}{4(s_2 - s_1)} + \frac{\|x_2\|^2 - \|x_1\|^2}{2}\right). \tag{53}$$

□

2.4. Proof of Theorem 4

First we give Lemma 4 without proof as below:

Lemma 4 (Bochner Formula and Bakry–Émery Inequality [2]). *For any $g \in C^3$, we have*

$$\frac{1}{2}\Delta|\nabla g|^2 = |\nabla^2 g|^2 + \langle\nabla g, \nabla\Delta g\rangle, \tag{54}$$

and furthermore

$$\frac{1}{2}\Delta|\nabla g|^2 \ge \frac{|\Delta g|^2}{n} + \langle\nabla g, \nabla\Delta g\rangle \tag{55}$$

where $|\nabla^2 g|^2 = \Sigma_{i,j=1}^n (\partial_{ij} g)^2$.

Based on this we give proof of Theorem 7:

Theorem 7 (Gradient Estimate of Heat equation)**.** *Let u be a positive solution to the heat equation*

$$\partial_t u = \Delta u, \tag{56}$$

on $(0,T] \times B_R$. Then for any $(t,x) \in (0,T] \times B_{\frac{R}{2}}$, the following inequality holds:

$$\frac{|\nabla u|^2}{u^2} - \alpha \frac{\partial_t u}{u} \leq \frac{n\alpha^2}{2t} + \frac{C\alpha^2}{R^2}\Big(1 + \frac{\alpha^2}{\alpha - 1}\Big), \tag{57}$$

where $C(n)$ is a constant depends on n.

Proof. Like the proof of Theorem 5, just turn L into Δ and then we can get the conclusion. □

From Theorem 7 we can conclude Theorem 8:

Theorem 8 (Harnack-type Inequality of Heat Equation)**.** *Let u be a positive solution of the heat equation $\partial_t u = \Delta u$ in $(0,T] \times B_R$, where $\alpha > 1$. For any $x_1, x_2 \in B_{\frac{R}{2}}$ and $0 < t_1 < t_2 \leq T$, the following inequality holds:*

$$u(x_1,t_1) \leq u(x_2,t_2)\left(\frac{t_2}{t_1}\right)^{\frac{m\alpha}{2}} \exp\left(\frac{\alpha^2\|x_1 - x_2\|^2}{4(t_2 - t_1)} + \frac{C\alpha}{R^2}\left(1 + \frac{\alpha^2}{\alpha - 1}\right)(t_2 - t_1)\right), \tag{58}$$

where $C = C(n)$.

Proof. Like the proof of 6. □

Finally, we can prove Theorem 4:

Proof of Theorem 4. The VE-SDE form is given by $\mathrm{d}x_t = \sqrt{\frac{\mathrm{d}\sigma_t^2}{dt}}\mathrm{d}W_t$, and its corresponding FPE form is

$$\frac{\partial p_t(x)}{\partial t} = \frac{1}{2}\frac{\mathrm{d}\sigma_t^2}{dt}\Delta_x(p_t(x)).$$

We can reparameterize t by letting $s = \frac{1}{2}\sigma_t^2$, which gives $\frac{\mathrm{d}s}{\mathrm{d}t} = \frac{1}{2}\frac{\mathrm{d}\sigma^2}{\mathrm{d}t}$. Therefore,

$$\frac{\partial p_{t(s)}(x)}{\partial s} = \frac{\partial p_t}{\partial t}\frac{\mathrm{d}t}{\mathrm{d}s} = \frac{\partial p_t(x)}{\partial t}\frac{1}{\frac{1}{2}\frac{\mathrm{d}\sigma^2}{\mathrm{d}t}} = \Delta_x(p_t(x)).$$

For this new FPE $\frac{\partial p_{t(s)}(x)}{\partial s} = \Delta_x(p_t(x))$, its corresponding SDE form is:

$$\mathrm{d}x_{t(s)} = \sqrt{2}\mathrm{d}W_s.$$

Assume $p(x,t)$ is the fundamental solution of this FPE, satisfying Theorem 8.

Thus, for any $\alpha > 1$, $x_1, x_2 \in M$, and $0 < s_1 < s_2 < +\infty$, let $R \to \infty$, the following inequality holds:

$$u(x_1, t(s_1))) \leq u(x_2, t(s_2))\left(\frac{s_2}{s_1}\right)^{\frac{n\alpha}{2}} \exp\left(\frac{\alpha^2\|x_1 - x_2\|^2}{4(s_2 - s_1)}\right). \tag{59}$$

□

3. Relationship Between MSE Bound and Harnack-type Inequality

In this section we provide a deeper insight into the connection between Theorems 1 and 3: they respectively lead to Theorems 12 and 10. In essence, these two results offer complementary perspectives on the evolution of the KL divergence.

3.1. Harnack-type inequality to KL-divergence

Starting from Harnack-type inequality, we can arrive at log-Harnack inequality. Consider SDE

$$\mathrm{d}X_t = -X_t\mathrm{d}t + \sqrt{2}\mathrm{d}W_t,$$

we obtain Theorem 9:

Theorem 9 (log-Harnack inequality)**.** *Let* $u(t,x) = P_t f(x) = \int \varphi_t(x,y)f(y)\mathrm{d}y$ *with the OU Mehler kernel* $\varphi_t(x,y) = (2\pi s_t)^{-n/2}\exp\big(-\frac{|y-e^{-t}x|^2}{2s_t}\big)$, $s_t = 1-e^{-2t}$. *Assume* $\mathrm{supp}(f) \subset B(0,R)$. *Then for every* $t>0$ *and every* $x,y\in\mathbb{R}^n$,

$$P_t\log f(y) \le \log P_t f(x) + \mid x-y \mid \sup_{z\in[x,y]} \sqrt{\frac{m\alpha^2}{2t} + \alpha\left(\frac{e^{-t}}{s_t}S'(z,t)\right)},$$

where $S'(x,t) = ((R^2 + e^{-2t}\mid x\mid^2 + 2e^{-t}R\mid x\mid)^2 + \mid x\mid R + e^{-t}\mid x\mid^2 - ne^{-t}), [x,y] := \{x+\theta(y-x) : \theta\in[0,1]\}$. *In particular, on any bounded domain* K *with* $\sup_{z\in K}|z| \le M$ *one has*

$$P_t\log f(y) \le \log P_t f(x) + \mid x-y\mid \sqrt{\frac{m\alpha^2}{2t} + \alpha\left(\frac{e^{-t}}{s_t}S'(\mid x\mid = M,t)\right)} \tag{60}$$

$$= \log P_t f(x) + S_K(t)\mid x-y\mid. \tag{61}$$

Proof. From Theorem 5 we conclude that a Gradient estimate holds on $\mathbb{R}^n$:

$$\frac{|\nabla u|^2}{u^2} - \alpha\frac{\partial_t u}{u} \le \frac{m\alpha^2}{2t},$$

where $\alpha>1, m>n$. For φ_t, we have

$$\nabla_x\log\varphi_t(x,y) = \frac{e^{-t}}{s_t}(y-e^{-t}x), \tag{62}$$

$$\Delta_x\log\varphi_t(x,y) = -\frac{ne^{-2t}}{s_t}. \tag{63}$$

Thus

$$\partial_t\log u = \frac{\mathrm{L}_x u}{u} \tag{64}$$

$$= \frac{\int(\Delta_x\varphi_t(x,y) - x\cdot\nabla_x\varphi_t(x,y))f(y)\mathrm{d}y}{\int\varphi_t(x,y)f(y)\mathrm{d}y} \tag{65}$$

$$= \frac{\int(\Delta_x\log\varphi_t(x,y) + \|\nabla_x\log\varphi_t(x,y)\|^2 - x\cdot\nabla_x\log\varphi_t(x,y))\varphi_t(x,y)f(y)\mathrm{d}y}{\int\varphi_t(x,y)f(y)\mathrm{d}y} \tag{66}$$

$$= \mathbb{E}_{Y\sim\pi_{t,x}}\left[\Delta_x\log\varphi_t(x,Y) + \|\nabla_x\log\varphi_t(x,Y)\|^2 - x\cdot\nabla_x\log\varphi_t(x,Y)\right], \tag{67}$$

$$= \mathbb{E}_{Y\sim\pi_{t,x}}\left[-\frac{ne^{-2t}}{s_t} + \frac{e^{-2t}}{s_t^2}\|Y-e^{-t}x\|^2 - \frac{e^{-t}}{s_t}(x\cdot Y - e^{-t}x^2)\right], \tag{68}$$

where $\pi_{t,x} = \frac{\varphi_t(x,y)f(y)}{\int\varphi_t(x,y)f(y)\mathrm{d}y}$. As $\mathrm{supp}(f)\subset B(0,R)$,

$$\begin{aligned}
\partial_t\log u &= \mathbb{E}_{Y\sim\pi_{t,x}}\left[-\frac{ne^{-2t}}{s_t} + \frac{e^{-2t}}{s_t^2}\|Y-e^{-t}x\|^2 - \frac{e^{-t}}{s_t}(x\cdot Y - e^{-t}\mid x\mid^2)\right] \\
&\le -\frac{ne^{-2t}}{s_t} + \frac{e^{-2t}}{s_t^2}(R^2 + e^{-2t}\mid x\mid^2 + 2e^{-t}R\mid x\mid)^2 \\
&\quad + \frac{e^{-t}}{s_t}(\mid x\mid R + e^{-t}\mid x\mid^2) \\
&\le \frac{e^{-t}}{s_t}((R^2 + e^{-2t}\mid x\mid^2 + 2e^{-t}R\mid x\mid)^2 + \mid x\mid R + e^{-t}\mid x\mid^2 - ne^{-t}) \\
&= \frac{e^{-t}}{s_t}S'(x,t).
\end{aligned}$$

Thus

$$\|\nabla \log u\|^2 \le \frac{m\alpha^2}{2t} + \alpha\left(\frac{e^{-t}}{s_t} S'_K(x,t)\right), \tag{69}$$

$$\|\nabla \log u\| \le \sqrt{\frac{m\alpha^2}{2t} + \alpha\left(\frac{e^{-t}}{s_t} S'(x,t)\right)}, \tag{70}$$

we can easily get that

$$\log u(t,y) - \log u(t,x) \le \mid x-y \mid \sup_{z\in[x,y]} \sqrt{\frac{m\alpha^2}{2t} + \alpha\left(\frac{e^{-t}}{s_t} S'(z,t)\right)}, \tag{71}$$

by Jesen's inequality, we have

$$P_t \log f(y) \le \log P_t f(x) + \mid x-y \mid \sup_{z\in[x,y]} \sqrt{\frac{m\alpha^2}{2t} + \alpha\left(\frac{e^{-2t}}{s_t} S'(z,t)\right)}, \tag{72}$$

as desired. □

Thus we obtain theorem below.

Theorem 10 (entropy–cost inequality)**.** *Let $K \subset \mathbb{R}^n$ be compact, and assume the transition kernels $P_t(x,\cdot) = \varphi_t(x,\cdot)\mathrm{d}y$ satisfy the pointwise log-Harnack inequality 60 above for all $x, y \in K$. Then for any two probability measures μ, ν supported in K and any coupling $\pi \in \Pi(\mu,\nu)$,*

$$\mathrm{KL}(P_t\nu \parallel P_t\mu) \le \iint |x-y|\, S_K(t)\, \pi(dx,dy) = S_K(t)\, \mathbb{E}_\pi[|X-Y|].$$

Taking the infimum over couplings,

$$\mathrm{KL}(P_t\nu \parallel P_t\mu) \le S_K(t)\, W_1(\mu,\nu) \le S_K(t)\, W_2(\mu,\nu), \tag{73}$$

so in particular the KL at time t is bounded by a compact-set constant $S_K(t)$ times the initial Wasserstein distance.

Proof. Recall the variational (Donsker–Varadhan) formula for relative entropy of two probability densities ρ, μ [8]:

$$\mathrm{KL}(P_t\nu \parallel P_t\mu) = \sup_{\phi \in B_b} \Big\{ \int \phi(z)\, P_t\nu(dz) - \log \int e^{\phi(z)}\, P_t\mu(dz) \Big\},$$

where B_b denotes bounded measurable functions, $P_t\nu(dz) = \int_y \varphi_t(y,z)\, \nu(dy)\, dz$, $P_t\mu(dz) = \int_x \varphi_t(x,z)\, \mu(dx)\, dz$.

For an arbitrary bounded ϕ set $f = e^\phi \ge 1$. Then

$$\int \phi(z)\, \varphi_t(y,z)\, dz \;\le\; \log \int e^{\phi(z)}\, \varphi_t(x,z)\, dz + |x-y|\, S_K(t).$$

Taking the supremum over all bounded ϕ yields exactly

$$\mathrm{KL}\Big(\varphi_t(y,\cdot) \Big\| \varphi_t(x,\cdot)\Big) \;\le\; |x-y|\, S_K(t).$$

Now fix any coupling $\pi \in \Pi(\mu,\nu)$. By convexity of KL under mixtures (or the standard coupling inequality),

$$\begin{aligned}\mathrm{KL}(P_t\nu \parallel P_t\mu) &= \mathrm{KL}\Big(\int \varphi_t(y,\cdot)\, \nu(dy) \;\Big\|\; \int \varphi_t(x,\cdot)\, \mu(dx)\Big) \\ &\le \iint \mathrm{KL}(\varphi_t(y,\cdot) \parallel \varphi_t(x,\cdot))\; \pi(dx,dy).\end{aligned}$$

Using the kernel bound and factoring $S_K(t)$ yields

$$\mathrm{KL}(P_t\nu \parallel P_t\mu) \le \iint |x-y| S_K(t)\, \pi(dx,dy) = S_K(t)\mathbb{E}_\pi[|X-Y|].$$

Taking infimum over π gives the W_1 form. Finally the monotonicity $W_1 \le W_2$ yields the stated W_2-bound. □

3.2. Score MSE bound to KL-divergence

Definition 1 (Relative Fisher Information)**.** *Let ν and μ be two probability measures on $\mathbb{R}^n$ such that ν is absolutely continuous with respect to μ. The* relative Fisher information *of ν with respect to μ is defined by*

$$I(\nu \,\|\, \mu) := \int_{\mathbb{R}^n} \left\| \nabla \log \frac{d\nu}{d\mu}(x) \right\|^2 d\nu(x),$$

where $\frac{d\nu}{d\mu}$ denotes the Radon–Nikodym derivative of ν with respect to μ, and $\nabla \log \frac{d\nu}{d\mu}$ is the score function *of ν relative to μ.*

Intuitively, $I(\nu \,\|\, \mu)$ measures the squared $L^2(\nu)$-distance between the score functions of ν and μ.

Theorem 11. *Let $X_t \in \mathbb{R}^n$ be the output of the SDE*

$$\mathrm{d}X_t = a(X_t, t)\mathrm{d}t + g(t)\mathrm{d}W_t. \tag{74}$$

Then for the above KL-divergence, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(P_t\nu \parallel P_t\mu) = -\frac{1}{2}g^2(t)I(P_t\nu \parallel P_t\mu). \tag{75}$$

For OU process

$$\mathrm{d}X_t = -X_t\mathrm{d}t + \sqrt{2}W_t,$$

we have the form:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(P_t\nu \parallel P_t\mu) = -I(P_t\nu \parallel P_t\mu). \tag{76}$$

Proof. We note $P_t\nu = p_t, P_t\mu = q_t$ for convenience. FPE of equation 74 is

$$\partial_t p_t = -\nabla \cdot (a p_t) + \frac{1}{2}\sigma^2(t)\Delta p_t,$$

as for differential entropy $H(X_t) = -\int p_t \log p_t \mathrm{d}x$, we obtain

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}H(X_t) &= -\int \partial_t p_t \log p_t \mathrm{d}x - \int \partial_t p_t \mathrm{d}x \\
&= -\int \partial_t p_t \log p_t \mathrm{d}x \\
&= \int \nabla \cdot (a p_t) \log p_t \mathrm{d}x - \int \frac{1}{2}g^2(t)\Delta p_t \log p_t \mathrm{d}x,
\end{aligned}$$

then we calculate the terms in the above equation,

$$\begin{aligned}
\int \nabla \cdot (a p_t) \log p_t \mathrm{d}x &= (-1)\int \left\langle a_t p_t, \frac{\nabla p_t}{p_t} \right\rangle \\
&= -\int \langle a_t, \nabla p_t \rangle \\
&= \mathbb{E}_{p_t}[\nabla \cdot a_t],
\end{aligned}$$

using $\Delta \log p = \Delta p/p - (\nabla \log p)^2$,

$$\begin{aligned}
\int \frac{1}{2}g^2(t)\Delta p_t \log p_t \mathrm{d}x &= \frac{1}{2}g^2(t)\int p_t \Delta \log p_t \\
&= \frac{1}{2}g^2(t)\int p_t(\Delta p_t/p_t - (\nabla \log p_t)^2) \\
&= -\frac{1}{2}g^2(t)\int p_t(\nabla \log p_t)^2,
\end{aligned}$$

so we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}H(X_t) = \frac{1}{2}g^2(t)\int p_t(\nabla \log p_t)^2 + \mathbb{E}_{p_t}[\nabla \cdot a_t].$$

Then we consider the term $S(p_t, q_t) = -\int p_t \log q_t \mathrm{d}x$,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}S(p_t, q_t) &= -\int \partial_t p_t \log q_t \mathrm{d}x - \int \frac{p_t}{q_t}\partial_t q_t \mathrm{d}x \\ &= \int \nabla \cdot (ap_t)\log q_t \mathrm{d}x - \int \frac{1}{2}g^2(t)\Delta p_t \log q_t \mathrm{d}x \\ &\quad + \int \nabla \cdot (aq_t)\frac{p_t}{q_t}\mathrm{d}x - \int \frac{1}{2}g^2(t)\Delta q_t \frac{p_t}{q_t}\mathrm{d}x,\end{aligned}$$

then we calculate the terms in the above equation,

$$\begin{aligned}\int \nabla \cdot (ap_t)\log q_t \mathrm{d}x &= (-1)\int \left\langle a_t p_t, \frac{\nabla q_t}{q_t}\right\rangle \\ &= -\int \langle a_t, \nabla \log q_t\rangle p_t,\end{aligned}$$

$$\begin{aligned}\int \nabla \cdot (aq_t)\frac{p_t}{q_t}\mathrm{d}x &= (-1)\int \left\langle a_t q_t, \frac{q_t\nabla p_t - p_t \nabla q_t}{q_t^2}\right\rangle \\ &= -\int \left\langle a_t, \nabla p_t - \frac{p_t \nabla q_t}{q_t}\right\rangle \\ &= \mathbb{E}_{p_t}[\nabla \cdot a_t] + \int \langle a_t, \nabla \log q_t\rangle p_t,\end{aligned}$$

$$\begin{aligned}\int \frac{1}{2}g^2(t)\Delta q_t \frac{p_t}{q_t}\mathrm{d}x &= -\frac{1}{2}g^2(t)\int \left\langle \nabla q_t, \frac{q_t\nabla p_t - p_t\nabla q_t}{q_t^2}\right\rangle \\ &= -\frac{1}{2}g^2(t)\int \langle \nabla \log q_t, \nabla \log p_t - \nabla \log q_t\rangle p_t,\end{aligned}$$

$$\int \frac{1}{2}g^2(t)\Delta p_t \log q_t \mathrm{d}x = -\frac{1}{2}g^2(t)\int\int \langle \nabla \log p_t, \nabla \log q_t\rangle p_t,$$

so we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}S(p_t, q_t) = -\frac{1}{2}g^2(t)\int p_t[(\nabla \log q_t)^2 - 2\langle \nabla \log p_t, \nabla \log q_t\rangle] + \mathbb{E}_{p_t}[\nabla \cdot a_t].$$

Then we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(p_t\|q_t) = -\frac{1}{2}g^2(t)I(p_t \,\|\, q_t).$$

□

Still, we consider SDE

$$\mathrm{d}X_t = -X_t \mathrm{d}t + \sqrt{2}W_t,$$

then we obtain conclusion below via Theorem 11 and 1:

Theorem 12 (KL Bound for Ornstein–Uhlenbeck SDE)**.** *Consider the Ornstein–Uhlenbeck SDE*

$$\mathrm{d}X_t = -X_t\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t,$$

by Theorem 1 we obtain

$$I(p_t \,\|\, q_t) \le 4R^2 \frac{e^{-2t}}{(1-e^{-2t})^2},$$

and let p_t and q_t be the distributions of two solutions with different initial conditions. Then, there exists a constant $C > 0$ such that for all $t \ge 0$,

$$\mathrm{KL}(p_t \,\|\, q_t) = \int_t^\infty I(p_s \,\|\, q_s)\,\mathrm{d}s \le \int_t^\infty 4R^2 \frac{e^{-2s}}{(1-e^{-2s})^2}\,\mathrm{d}s \le 2R^2 \frac{e^{-2t}}{1-e^{-2t}},$$

where $I(p_s \,\|\, q_s)$ denotes the relative Fisher information (or score MSE) of p_s with respect to q_s.

In particular, this provides an explicit upper bound for the KL divergence between p_t and q_t in terms of t.

3.3. Conclusion

Via Theorem 1 and 3, we can get Theorem 12 and 10, which both bound the KL-divergence $\mathrm{KL}(p_t \,\|\, q_t)$.

We can observe that these two approaches are closely related in spirit:

- The MSE-bound approach (Theorem 12) directly controls the relative Fisher information

$$I(p_t \,\|\, q_t) = \mathbb{E}_{p_t}[\,|s_{p_t} - s_{q_t}|^2\,],$$

  and then integrates it over time to obtain an explicit upper bound for the KL-divergence.
- The Harnack inequality approach (Theorem 10) instead provides a pointwise control on the semigroup, which, via coupling and Wasserstein distances, leads to a KL upper bound of the form

$$\mathrm{KL}(P_t\nu \parallel P_t\mu) \le S_K(t)\, W_1(\mu,\nu) \le S_K(t)\, W_2(\mu,\nu).$$

- In essence, both methods link the KL divergence at time t to some notion of discrepancy at the initial time: MSE-bound does it via the score difference (relative Fisher information), while Harnack-bound does it via transport distances (W_1 or W_2). The MSE bound can be seen as a "local-in-space" version of the Harnack control: if the pointwise kernel control from Harnack implies a bound on $\nabla \log p_t$, then integrating it yields a Fisher-information-type bound. Thus, the two approaches are complementary perspectives on how initial differences propagate under the dynamics of the SDE.

This observation highlights that controlling either the score differences or the pointwise semigroup can provide rigorous quantitative bounds on the evolution of KL divergence in diffusion processes.

4. Algorithm Comparison

Figure 1 compares standard CFG and C^{2}FG. At each timestep t during generation, the C^{2}FG update replaces the standard CFG as follows:

$$\hat{\boldsymbol{\epsilon}}_{\boldsymbol{c}}^{\omega}(\boldsymbol{x}_t) = \hat{\boldsymbol{\epsilon}}_{\varnothing}(\boldsymbol{x}_t) + \omega(t)\big[\hat{\boldsymbol{\epsilon}}_{\boldsymbol{c}}(\boldsymbol{x}_t) - \hat{\boldsymbol{\epsilon}}_{\varnothing}(\boldsymbol{x}_t)\big].$$

Algorithm 1 Reverse Diffusion with CFG

Require: $\boldsymbol{x}_T \sim \mathcal{N}(0, \mathbf{I}_d), 0 \le \omega \in \mathbb{R}$
for $i = T$ **to** 1 **do**
  $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{c}}^{\omega}(\boldsymbol{x}_t) = \hat{\boldsymbol{\epsilon}}_{\varnothing}(\boldsymbol{x}_t) + \omega[\hat{\boldsymbol{\epsilon}}_{\boldsymbol{c}}(\boldsymbol{x}_t) - \hat{\boldsymbol{\epsilon}}_{\varnothing}(\boldsymbol{x}_t)]$
  $\hat{\boldsymbol{x}}_{\boldsymbol{c}}^{\omega}(\boldsymbol{x}_t) \leftarrow (\boldsymbol{x}_t - \sqrt{1-\bar{\alpha}_t}\hat{\boldsymbol{\epsilon}}_{\boldsymbol{c}}^{\omega}(\boldsymbol{x}_t))/\sqrt{\bar{\alpha}_t}$
  $\boldsymbol{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\hat{\boldsymbol{x}}_{\boldsymbol{c}}^{\omega}(\boldsymbol{x}_t) + \sqrt{1-\bar{\alpha}_{t-1}}\hat{\boldsymbol{\epsilon}}_{\boldsymbol{c}}^{\omega}(\boldsymbol{x}_t)$
end for
return $\boldsymbol{x}_0$

Algorithm 2 Reverse Diffusion with Our Method

Require: $\boldsymbol{x}_T \sim \mathcal{N}(0, \mathbf{I}_d), \omega(t) \in \mathrm{C}[0, +\infty)$
for $i = T$ **to** 1 **do**
  $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{c}}^{\omega}(\boldsymbol{x}_t) = \hat{\boldsymbol{\epsilon}}_{\varnothing}(\boldsymbol{x}_t) + \omega(t)[\hat{\boldsymbol{\epsilon}}_{\boldsymbol{c}}(\boldsymbol{x}_t) - \hat{\boldsymbol{\epsilon}}_{\varnothing}(\boldsymbol{x}_t)]$
  $\hat{\boldsymbol{x}}_{\boldsymbol{c}}^{\omega}(\boldsymbol{x}_t) \leftarrow (\boldsymbol{x}_t - \sqrt{1-\bar{\alpha}_t}\hat{\boldsymbol{\epsilon}}_{\boldsymbol{c}}^{\omega}(\boldsymbol{x}_t))/\sqrt{\bar{\alpha}_t}$
  $\boldsymbol{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\hat{\boldsymbol{x}}_{\boldsymbol{c}}^{\omega}(\boldsymbol{x}_t) + \sqrt{1-\bar{\alpha}_{t-1}}\hat{\boldsymbol{\epsilon}}_{\boldsymbol{c}}^{\omega}(\boldsymbol{x}_t)$
end for
return $\boldsymbol{x}_0$

Figure 1. Comparison between reverse diffusion process by CFG and C^{2}FG. Our C^{2}FG guidance weight $\omega(t)$ is a time-decay function.

5. Additional Experiments

More Visualized Analysis on Theorem 1. In Figure 2, each pixel in the heatmap corresponds to the logarithmic ratio of the conditional prediction to the unconditional prediction at a specific spatial location and channel. A value of zero (shown as white) indicates no difference (ratio=1). Positive values (red) indicate amplification of the conditional prediction relative to the unconditional one, while negative values (blue) indicate suppression. Importantly, the further a pixel's value deviates from zero—whether red or blue—the larger the discrepancy between the two predictions. Thus, both strong red and strong blue regions highlight locations where the conditional and unconditional outputs differ most significantly.

Building on Theorem 1, these heatmaps provide a visual representation of how the score discrepancy evolves over time and across spatial locations. In particular, the early timesteps (larger t indices in the backward diffusion process) show relatively mild color variations, consistent with the theoretical bound $\|\nabla \log p - \nabla \log \tilde{p}\| \propto \alpha(t)/\sigma^2(t)$, which predicts smaller score differences at well-mixed later times. Conversely, at timesteps closer to the end of the reverse diffusion (smaller t indices), the heatmaps exhibit more pronounced red and blue regions, indicating larger deviations between conditional and unconditional predictions. This aligns with the theoretical observation that the MSE between scores can be large near small diffusion times, where initial distribution differences are amplified. Therefore, the heatmaps not only highlight spatially localized discrepancies but also corroborate the temporal trend predicted by Theorem 1, illustrating that both strong positive (red) and negative (blue) regions correspond to locations and timesteps with significant score mismatch.

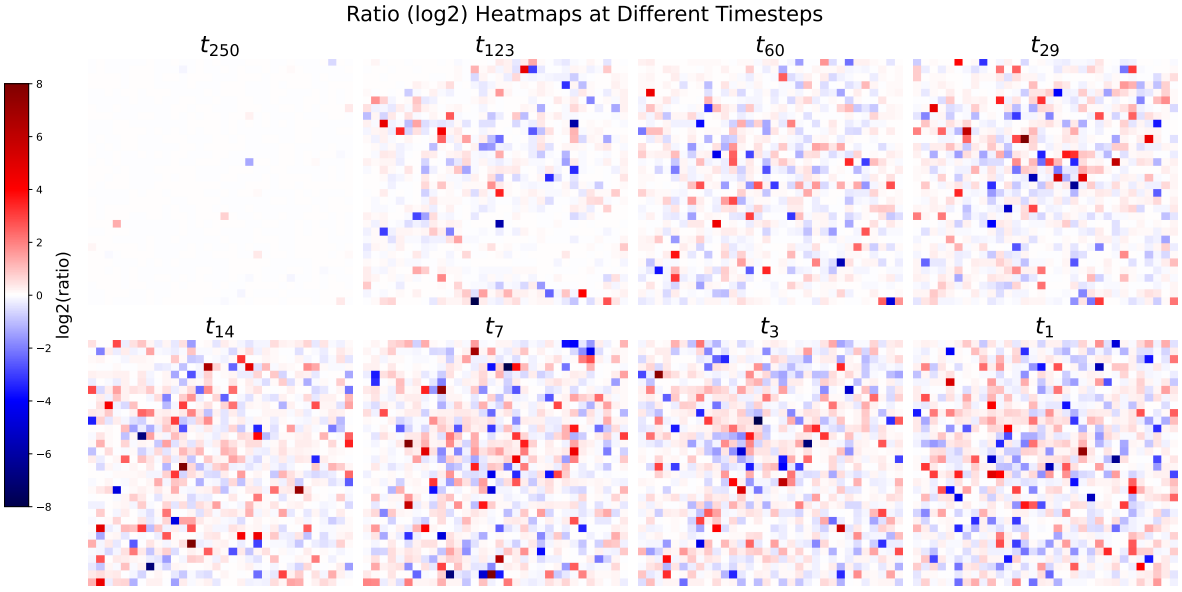


Figure 2. Heatmaps of the logarithmic ratio ($\log_2$) between conditional and unconditional predictions at selected timesteps. White indicates no difference (ratio=1), while red and blue highlight amplification and suppression, respectively. Stronger colors denote larger deviations between the two predictions.

Comparisons of various forms of $\omega(t)$. As shown in Figure 3, we compare the performance of our method with the DiT-XL/2 baseline [19] under a fixed parameter $\lambda = 1.0$. We observe that the curve corresponding to our method consistently lies below that of the baseline, indicating a strictly better IS–FID trade-off. In Figure 4b, we further evaluate several alternative choices of the scheduling function $\omega(t)$ in [26], including $\sin\big((t/t_m)\pi\big)$, t/t_m, $1 - t/t_m$, together with our proposed formulation($\exp(1 - t/t_m)$), whose trends are shown in Figure 4a. We observe that certain choices such as sine-based $\omega(t)$ perform even worse than the DiT baseline. Besides, although some of these functions share a broadly similar decreasing trend with our design, they are not aligned with the approximate exponential upper bound derived from our framework. Consequently, their empirical IS–FID trade-off performance is consistently inferior to ours.

These results highlight that the improvement does not merely come from tuning the scaling magnitude, but primarily from how the temporal modulation interacts with the diffusion dynamics. In particular, our schedule suppresses error amplification in early steps while preserving semantic consistency in later denoising stages, yielding more stable and efficient generation across the entire sampling trajectory.

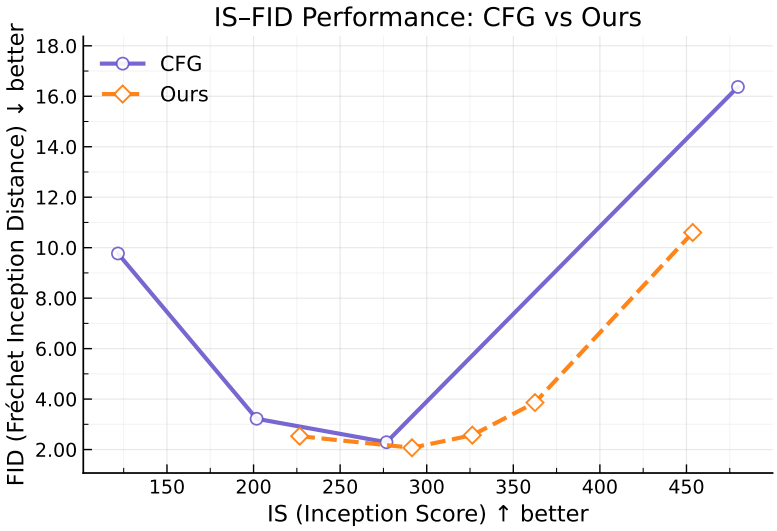
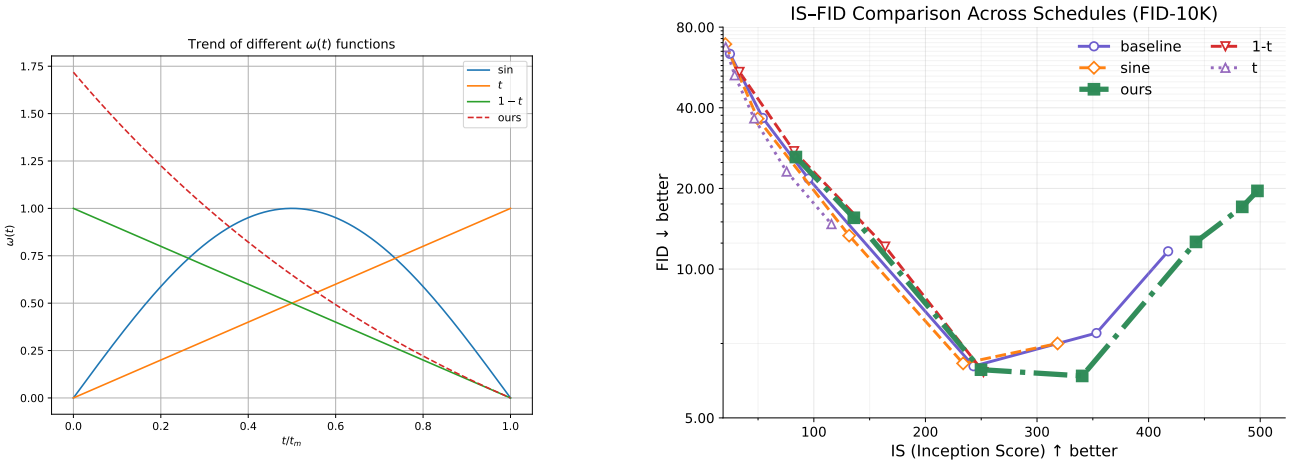


Figure 3. Impact of the initial schedule weight ω_0 on IS–FID performance (with fixed $\lambda = 1.0$, 250 inference steps).



(a) Trend of various $\omega(t)$

(b) Effect of guidance scale ω_0 on IS–FID curves (50 steps, 10K samples)

Figure 4. Comparison of IS–FID performance under different hyperparameter settings on DiT-XL/2 and ImageNet-256.

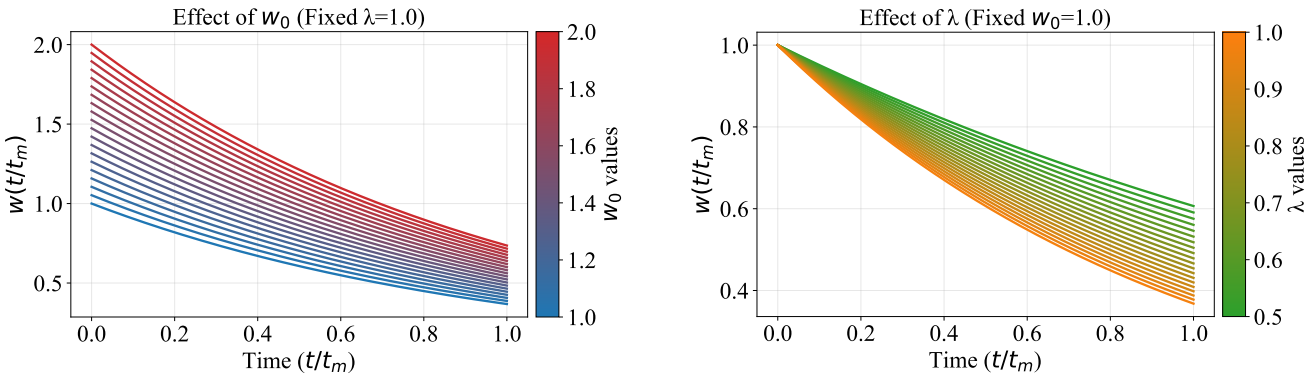
ImageNet(256×256)				
Model 25 inference timesteps	FID↓	IS ↑	Prec↑	Rec↑
DiT-XL/2 (baseline, $\omega = 1.5$)	11.88	192.13	0.7176	**0.6651**
DiT-XL/2+β-CFG ($\omega = 1.5, a = b = 2.0$)	8.42	254.42	0.8038	0.6026
DiT-XL/2 + RAAG ($\omega_{\max} = 18.0, \alpha = 12.0$)	21.57	**444.57**	**0.8360**	0.2636
DiT-XL/2 + Ours ($\omega_0 = 1.0, \lambda = 1.0$)	**7.70**	294.03	0.7994	0.6357

Table 1. Performance comparison between our method and existing adaptive CFG approaches on ImageNet-256.

To further validate our approach, we compare it against recent time-varying strategies, specifically RAAG [29] and β-CFG [18]. We follow the official hyperparameters from the original papers (β-CFG: $a = b = 2.0$; RAAG: $\omega_{\max} = 18.0, \alpha = 12.0$). These comparisons are conducted on the DiT-XL/2 model using ImageNet-256, with the results summarized in Table 1. The quantitative comparison reveals that, under identical settings, our approach achieves superior overall performance, particularly in term of FID (7.70). While β-CFG yields marginally higher Precision, it significantly lags in all other metrics.

Furthermore, we observe that though RAAG obtains the highest IS score (444.57) and the highest Precision, it suffers from severely degraded FID (21.57) and Recall (0.2636). In other word, its guidance mechanism emphasizes semantic alignment (IS) rather than accurate distribution fitting (FID), leading to degraded performance. We attribute this to its design focus on text-to-image generation, which appears to generalize poorly to class-conditional settings. In contrast, our method demonstrates superior generalization capabilities, proving robust across diverse tasks and model architectures.

Analysis of Parameters in C^{2}FG. As shown in Figure 5a and 5b, ω_0 sets the initial or maximum guidance strength, and λ controls the rate of exponential decay. Moreover, Table 2 presents an ablation study on the hyperparameter λ. While the results demonstrate that various λ values are effective for enhancing performance, the best outcome is achieved with $\lambda = \log e = 1.0$. The results indicate that this C^{2}FG design is effective.



(a) Effect of ω_0 (Fixed $\lambda = 1.0$)

(b) Effect of λ (Fixed $\omega_0 = 1.0$)

Figure 5. (a) demonstrates the impact of initial weight ω_0; (b) illustrates how different λ values affect the decay profile.

ImageNet(256×256), 50k samples, 250 SDE inference timesteps	
Model	FID↓
REPA (Fixed CFG = 1.35)	1.80
REPA ($\lambda = \log 2$)	1.68
REPA ($\lambda = 1(\log e)$)	**1.51**
REPA ($\lambda = \log 3$)	1.58

Table 2. Comparison between the different effect of λ, fixing $\omega_0 = 1.0$.

Results on More Framework. In Table 3, we show the results of our C^{2}FG on autoguidance introduced by [13] with the model of EDM2 [14]. Autoguidance involves two denoiser networks $D_0(x;\sigma,c)$ and $D_1(x;\sigma,c)$ and the guiding effect is achieved by extrapolating between the two denoising results by a factor ω:

$$D_\omega(x;\sigma_t,c) = \omega D_1(x;\sigma_t,c) + (1-\omega)D_0(x;\sigma_t,c),$$

based on their method, we make ω be a time-variance function $\omega(t)$ with the same formula of C^{2}FG: $\omega(t) = \omega_0 \exp(1-t/t_{\max})$. As shown in Table 3, our dynamic guidance $\omega(t)$ consistently improves over the static guidance baseline. On ImageNet-64, where the model operates directly in the pixel domain, our method achieves lower FID and FD-DINOv2 [1], indicating that dynamic weighting not only preserves fidelity but also enhances semantic alignment. On high-resolution ImageNet-512, which is considerably more challenging, we also observe clear gains under the same setting, confirming that the proposed C^{2}FG can robustly integrate with autoguidance across scales. These results highlight the generality of our approach: the time-dependent extrapolation scheme provides a more adaptive balance between fidelity and diversity than a fixed scalar weight.

Denoising Process. As shown in Figure 6, we provide a qualitative comparison of intermediate decoding results between our C^{2}FG and the baseline across the denoising trajectory. From step 250 down to 50, both methods generate visually similar results. However, in the final refinement stage (from step 50 to 0), the difference becomes more pronounced: our C^{2}FG produces sharper structures and more coherent details, highlighting the benefit of dynamically adjusting the guidance strength in the later denoising steps.

ImageNet(64×64)		
Model	FID↓	FD_{DINOv2} ↓
EDM2-S-autoguidance ($\omega = 1.7$)	1.044	56.3
EDM2-S-autoguidance+Ours($\omega_0 = 0.9, \lambda = 0.7$)	**1.028**	**52.7**
ImageNet (512×512), 10k samples		
EDM2-S-autoguidance ($\omega = 1.4$)	5.27	121.2
EDM2-S-autoguidance+Ours($\omega_0 = 0.9, \lambda = 0.5$)	**5.15**	**116.7**

Table 3. We evaluated conditional image generation on ImageNet with EDM2 and Autoguidance.

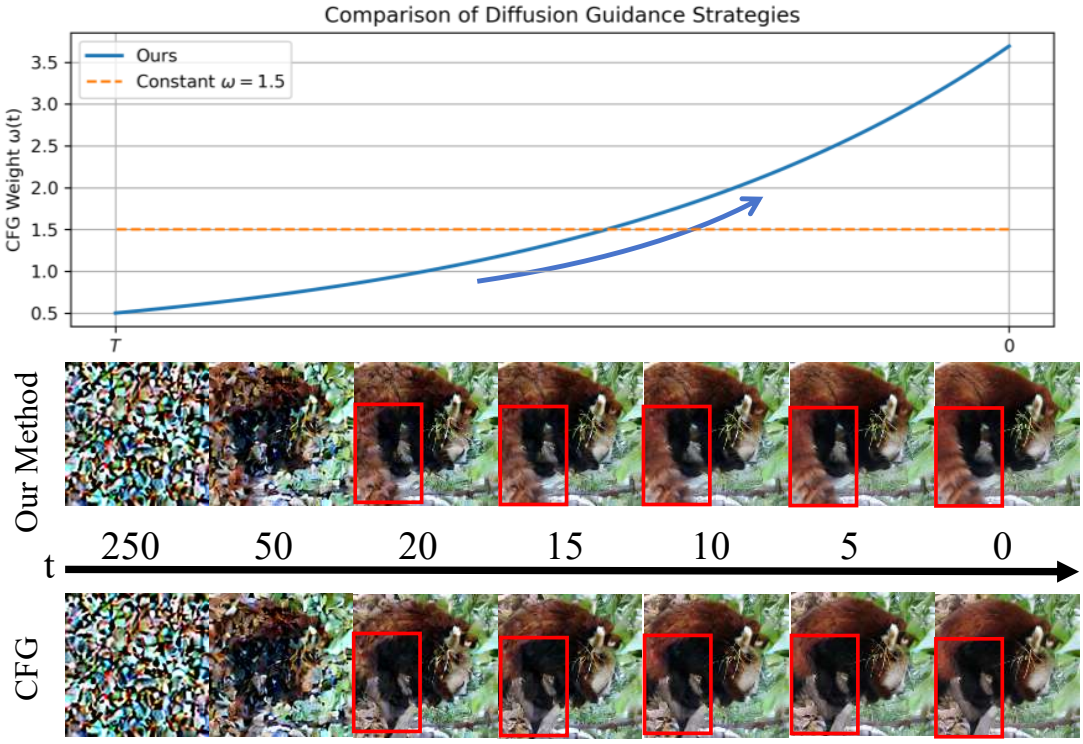


Figure 6. Comparison between results during the denoising process of C^2FG and Baseline.

Table 4. Additional Comparisons. **Left:** Comparisons with dynamic guidance methods on SD1.5 (MS-COCO) and SiT (ImageNet). **Right:** Results on modern T2I models (Flux, SD3).

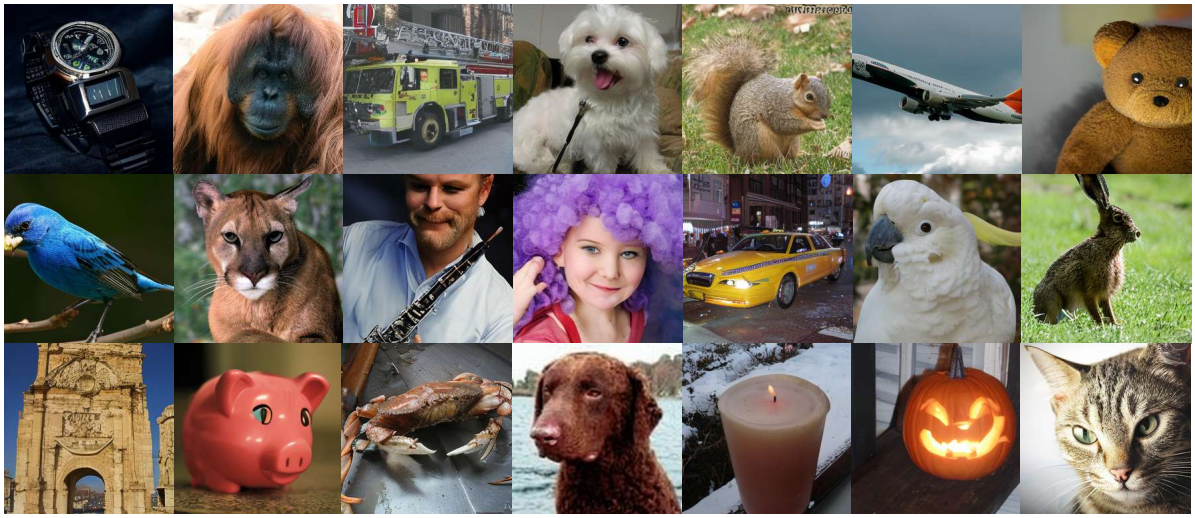
Compare	Fixed SD1.5,MS-COCO				Fixed SiT,ImageNet		Compare T2I models (CLIP↑)		
Method	CFG	CFG++	β-CFG	C^2FG **(4,1)**	FDG	C^2FG**(1.7,0.15)**	Models	Flux **(1.5,1)**	SD3 **(5,1)**
FID(10k) ↓	19.32	18.87	16.74	**16.71**	6.15	**3.20**	CFG	31.4	31.4
CLIP(10k) ↑	32.0	32.0	31.7	**32.0**	–	–	C^2FG	**31.5**	**31.5**

Additional Results. In Table 4 we compare our methods with other methods on different models. On SD1.5 [21], C^2FG achieves the **best FID&CLIP**, surpassing CFG++ [6] and β-CFG [18]. It also consistently improves Flux [16] and SD3 [9]. On SiT [28], C^2FG outperforms FDG [23]. Thus C^2FG consistently outperforms other methods across diverse tasks, verifying its strong generality. And Figure 7 shows our visualized results on T2I tasks. Additionally, Figure 8 shows additional results using our C^2FG method on DiT and SiT models.

Figure 7. **Visual Comparison.** C^{2}FG produces images that better align with the text prompt than standard CFG, yielding more faithful details, consistent with the quantitative gains in Table 4.

Prompt: *A rustic wooden signpost sticking out of the grass in a beautiful garden. The text 'Control CFG' is carved into the wood. Sunlight, lens flare, detailed textures.*

(a) Images generated by the DiT-XL/2 model with C^{2}FG on ImageNet-256.



(b) Images generated by the SiT-XL/2 (REPA) model with C^{2}FG on ImageNet-256.

Figure 8. Additional results for C^{2}FG.

References

[1] Junyeong Ahn. FD-DINOv2: FD Score via DINOv2. https://github.com/justin4ai/FD-DINOv2, 2024. Version 0.1.0. 19

[2] Dominique Bakry and Michel Émery. Diffusions hypercontractives. In *Séminaire de Probabilités XIX 1983/84: Proceedings*, pages 177–206. Springer, 2006. 5, 10

[3] Arwen Bradley and Preetum Nakkiran. Classifier-free guidance is a predictor-corrector. *Transactions on Machine Learning Research*, 2025. 2

[4] Ao Chen, Lihe Ding, and Tianfan Xue. Diffier: Optimizing diffusion models with iterative error reduction, 2025. 2

[5] Chubin Chen, Jiashu Zhu, Xiaokun Feng, Nisha Huang, Meiqi Wu, Fangyuan Mao, Jiahong Wu, Xiangxiang Chu, and Xiu Li. S^2-guidance: Stochastic self guidance for training-free enhancement of diffusion models, 2025. 2

[6] Hyungjin Chung, Jeongsol Kim, Geon Yeong Park, Hyelin Nam, and Jong Chul Ye. CFG++: Manifold-constrained classifier free guidance for diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025. 2, 20

[7] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *ArXiv*, abs/2105.05233, 2021. 1

[8] M. D. Donsker and S. R. S. Varadhan. Asymptotic evaluation of certain markov process expectations for large time, i. *Communications on Pure and Applied Mathematics*, 28(1):1–47, 1975. 13

[9] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. 20

[10] L.C. Evans. *Partial Differential Equations*. American Mathematical Society, 1998. 5

[11] Jonathan Ho. Classifier-free diffusion guidance. *ArXiv*, abs/2207.12598, 2022. 1

[12] Cheng Jin, Qitan Shi, and Yuantao Gu. Stage-wise dynamics of classifier-free guidance in diffusion models, 2025. 2

[13] Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. In *Proc. NeurIPS*, 2024. 19

[14] Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. In *Proc. CVPR*, 2024. 19

[15] Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. *Advances in Neural Information Processing Systems*, 37: 122458–122483, 2024. 2

[16] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. 20

[17] Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 5404–5411, 2024. 2

[18] Dawid Malarz, Artur Kasymov, Maciej Zięba, Jacek Tabor, and Przemysław Spurek. Classifier-free guidance with adaptive scaling. *arXiv preprint arXiv:2502.10574*, 2025. 2, 18, 20

[19] William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022. 17

[20] Mateusz Poleski, Jacek Tabor, and Przemysław Spurek. Geoguide: Geometric guidance of diffusion models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 297–305. IEEE, 2025. 2

[21] 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*, pages 10684–10695, 2022. 20

[22] Seyedmorteza Sadat, Manuel Kansy, Otmar Hilliges, and Romann M. Weber. No training, no problem: Rethinking classifier-free guidance for diffusion models, 2025. 2

[23] Seyedmorteza Sadat, Tobias Vontobel, Farnood Salehi, and Romann M. Weber. Guidance in the frequency domain enables high-fidelity sampling at low cfg scales. *ArXiv*, abs/2506.19713, 2025. 1, 20

[24] Dazhong Shen, Guanglu Song, Zeyue Xue, Fu-Yun Wang, and Yu Liu. Rethinking the spatial inconsistency in classifier-free diffusion guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9370–9379, 2024. 2

[25] Fu-Yun Wang, Yunhao Shui, Jingtan Piao, Keqiang Sun, and Hongsheng Li. Diffusion-npo: Negative preference optimization for better preference aligned generation of diffusion models. *arXiv preprint arXiv:2505.11245*, 2025. 2

[26] Xi Wang, Nicolas Dufour, Nefeli Andreou, Marie-Paule Cani, Victoria Fernández Abrevaya, David Picard, and Vicky Kalogeiton. Analysis of classifier-free guidance weight schedulers. *arXiv preprint arXiv:2404.13040*, 2024. 2, 17

[27] Haotian Ye, Haowei Lin, Jiaqi Han, Minkai Xu, Sheng Liu, Yitao Liang, Jianzhu Ma, James Zou, and Stefano Ermon. Tfg: Unified training-free guidance for diffusion models. 2024. 2

[28] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations*, 2025. 20

[29] Shangwen Zhu, Qianyu Peng, Yuting Hu, Zhantao Yang, Han Zhang, Zhao Pu, Ruili Feng, and Fan Cheng. Raag: Ratio aware adaptive guidance, 2025. 2, 18