Rectified Diffusion Guidance for Conditional Generation

Mengfei Xia[1,2*] Nan Xue[2] Yujun Shen[2†] Ran Yi[3] Tieliang Gong[4] Yong-Jin Liu[1†]
[1]Tsinghua University [2]Ant Group
[3]Shanghai Jiao Tong University [4]Xi'an Jiao Tong University

Abstract

Classifier-Free Guidance (CFG), which combines the conditional and unconditional score functions with two coefficients summing to one, serves as a practical technique for diffusion model sampling. Theoretically, however, denoising with CFG cannot be expressed as a reciprocal diffusion process, which may consequently leave some hidden risks during use. In this work, we revisit the theory behind CFG and rigorously confirm that the improper configuration of the combination coefficients (i.e., the widely used summing-to-one version) brings about expectation shift of the generative distribution. To rectify this issue, we propose `ReCFG`[1] *with a relaxation on the guidance coefficients such that denoising with* `ReCFG` *strictly aligns with the diffusion theory. We further show that our approach enjoys a* ***closed-form*** *solution given the guidance strength. That way, the rectified coefficients can be readily pre-computed via traversing the observed data, leaving the sampling speed barely affected. Empirical evidence on real-world data demonstrate the compatibility of our post-hoc design with existing state-of-the-art diffusion models, including both class-conditioned ones (e.g., EDM2 on ImageNet) and text-conditioned ones (e.g., SD3 on CC12M), without any retraining. Code is available at https://github.com/thuxmf/recfg.*

1. Introduction

Diffusion probabilistic models (DPMs) [13, 27, 29], known simply as diffusion models, have achieved unprecedented capability improvement of high-resolution image generation. It is well recognized that, DPMs are the most prominent generative paradigm for a broad distribution (*i.e.*, text-to-image generation) [3, 9, 22]. Among DPM literature, Classifier-Free Guidance (CFG) [12] serves as an essential factor, enabling better conditional sampling in various fields [23, 25]. Vanilla conditional sampling via DPMs introduces the conditional score function $s_t(\mathbf{x}, c) = \nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t|c)$, resulting in poor performance in which synthesized samples appear to be visually incoherent and not faithful to the condition, even for large-scale models [25]. By drawing lessons from Bayesian theory, CFG employs an interpolation between conditional and unconditional score functions with a preset weight γ, *i.e.*,

$$s_{t,\gamma}(\mathbf{x}, c) = \gamma \nabla_{\mathbf{x}} \log q_t(\mathbf{x}|c) + (1-\gamma) \nabla_{\mathbf{x}} \log q_t(\mathbf{x}), \quad (1)$$

in which $\nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t)$ is the unconditional score function by annihilating the condition effect. By doing so, DPMs turn out to formulate the underlying distribution with a gamma-powered distribution [1], *i.e.*,

$$q_{t,\gamma}(\mathbf{x}|c) = q_t(\mathbf{x}|c)^{\gamma} q_t(\mathbf{x})^{1-\gamma}, \quad (2)$$

which is proportional to $q_t(\mathbf{x}) q_t(c|\mathbf{x})^{\gamma}$. Enlarging $\gamma > 1$ focuses more on the classifier effect $q_t(c|\mathbf{x})$, concentrating on better exemplars of given condition and thereby sharpening the gamma-powered distribution. In other words, CFG is designed to promote the influence of the condition.

However, inspired by seminal works [1], we argue that denoising with CFG cannot be expressed as a reciprocal of vanilla diffusion process by adding Gaussian noises, since the normally nonzero score function expectation of gamma-powered $q_{t,\gamma}(\mathbf{x}|c)$ violates the underlying theory of DPMs. Theoretically, score functions with zero expectation at all timesteps guarantee that the denoised $\tilde{\mathbf{x}}_0$ has expectation $\mathbb{E}[\tilde{\mathbf{x}}_0] = \frac{\alpha_0}{\alpha_T}\mathbb{E}[\mathbf{x}_T]$, thus $\mathbb{E}[\tilde{\mathbf{x}}_0] = \mathbb{E}[\mathbf{x}_0]$ and no bias on the conditional fidelity. Therefore, this theoretical flaw leaves some hidden risks during use, manifesting as a severe expectation shift phenomenon, *i.e.*, the expectation of the gamma-powered distribution will be shifted away from the ground-truth of the conditional distribution $q_t(\mathbf{x}|c)$. This is more conspicuous when applying larger γ. Fig. 1 clearly clarifies the expectation shift, in which the peak of induced distribution via CFG in red fails to coincide with that of ground-truth $q_0(\mathbf{x}_0|c)$. This theoretical flaw is known in theory [1, 8, 15], while being largely ignored in practice.

*Work finished during internship at Ant Group.
†Corresponding author.
[1]ReCFG, pronounced as "reconfigure", is the abbreviation for "rectified Classifier-Free Guidance".

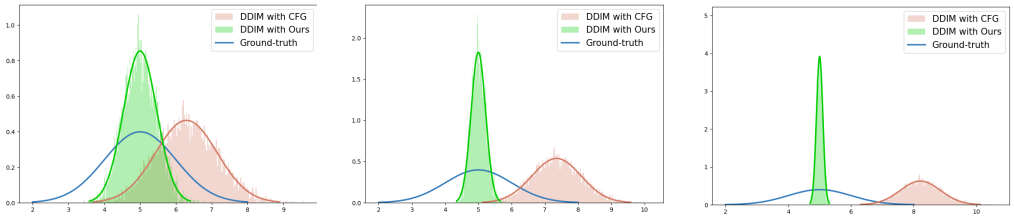


Figure 1. **Visualization** of expectation shift. The demonstrated toy data is simulated by $q_0(\mathbf{x}_0|c) \sim \mathcal{N}(c, 1)$, $q(c) \sim \mathcal{N}(0, 1)$, $q_0(\mathbf{x}_0) \sim \mathcal{N}(0, 2)$. Gamma-powered distribution $q_{0,\gamma}(\mathbf{x}_0|c)$ from CFG [12] fails to recover the same conditional expectation as ground-truth due to expectation shift (*i.e.*, probability density function and histogram by DDIM [28] sampler in **red**). To make a further step, larger γ suggests more severe expectation shift, *i.e.*, the peak of $q_{0,\gamma}(\mathbf{x}_0|c)$ tends further away from $q_0(\mathbf{x}_0|c)$ (*i.e.*, probability density function in **blue**) as γ goes from 1.5 to 2.5. As a comparison, our `ReCFG` successfully recovers the ground-truth expectation and smaller variance (*i.e.*, probability density function and histogram by DDIM [28] sampler in **green**), consistent with the motivation of guided sampling.

In this work, we first revisit the formulation of native CFG, theoretically confirming its flaw that we concluded above and summarizing as Theorem 1. Then, to quantitatively reveal the consequent expectation shift phenomenon by CFG, we employ a toy distribution, enjoying closed-form description of the behavior on the gamma-powered distribution. Under the toy settings, we analytically calculate the function of the precise value of expectation shift in correspondence with γ, as summarized in Theorem 2. Motivated by theoretical compatibility and canceling the expectation shift, we apply relaxation on the guidance coefficients in native CFG by circumventing the constraint that two coefficients sum to one, enabling a more flexible control on the induced distributions. To be more concrete, we propose to formulate the underlying distribution with *two* coefficients, *i.e.*,

$$q_{t,\gamma_1,\gamma_0}(\mathbf{x}|c) = q_t(\mathbf{x}|c)^{\gamma_1} q_t(\mathbf{x})^{\gamma_0}. \tag{3}$$

Aiming at consistency with the diffusion theory and thus better guidance efficacy, we specially design the constraints on γ_1 and γ_0, and theoretically confirm the feasibility. We further provide a closed-form solution to the constraints, and propose an algorithm to analytically determine γ_0 from a pre-computed lookup table in a post-hoc fashion. Thanks to the neat formulation, we can employ pixel-wise γ_0 according to the lookup table involving guidance strength γ_1, condition c and timestep t, as demonstrated in Fig. 2. We name the above process `ReCFG`. Compared with a global CFG weight applied on all denoising steps and pixels, `ReCFG` may achieve more flexible and accurate guidance. Experiments with state-of-the-art DPMs, including both class-conditioned ones (*e.g.*, EDM2 [16]) and text-conditioned ones (*e.g.*, SD3 [9]) under different NFEs and guidance strengths show that our `ReCFG` can achieve better guidance efficacy without retraining or extra time cost during inference stage. Hence, our work offers a new perspective on guided sampling of DPMs, encouraging more studies in the field of guided generation.

2. Related Work

DPMs and conditional generation. Diffusion probabilistic model (DPM) introduces a new scheme of generative modeling, formulated by forward diffusing and reverse denoising processes in a differential equation fashion [13, 27, 29]. Practically, it is trained by optimizing the variational lower bound. Benefiting from this breakthrough, DPM achieves high generation fidelity, and even beat GANs on image generation. By drawing lessons from conditional distribution, conditional generation [5, 14] takes better advantage of intrinsic intricate knowledge of data distribution, making DPM easier to scale up and the most promising option for generative modeling. Among the literature, text-to-image generation injects the embedding of text prompts to DPM, faithfully demonstrating the text content [3, 9, 22].

Classifier-Free Guidance. Classifier-Free Guidance (CFG) serves as the successor of Classifier Guidance (CG) [7], circumventing the usage of a classifier for noisy images. Both CFG and CG are based on Bayesian theory, and attempt to formulate the underlying distribution by concentrating more on condition influence, achieving better conditional fidelity. Despite great success in large-scale conditional generation, CFG faces a technical flaw that the guided distribution is not theoretically guaranteed to recover the ground-truth conditional distribution [1, 4, 8, 15]. To be more detailed, there exists a shifting issue that the expectation of guided distribution is drifted away from the correct one [1, 4]. This phenomenon may harm the condition faithfulness, especially for extremely broad distribution (*e.g.*, open-vocabulary synthesis).

3. Method

3.1. Background on conditional DPMs and CFG

Let $\mathbf{x}_0 \in \mathbb{R}^D$ be a D-dimensional random variable with an unknown distribution $q_0(\mathbf{x}_0|c)$, where $c \sim q(c)$ is the given condition. DPM [13, 27, 29] introduces a forward process

$\{\mathbf{x}_t\}_{t\in(0,T]}$ by gradually corrupting data signal of $\mathbf{x}_0$ with Gaussian noise, *i.e.*, the following transition distribution holds for any $t \in (0, T]$:

$$q_{0t}(\mathbf{x}_t|\mathbf{x}_0, c) = q_{0t}(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\alpha_t\mathbf{x}_0, \sigma_t^2\mathbf{I}), \quad (4)$$

in which $\alpha_t, \sigma_t \in \mathbb{R}^+$ are differentiable functions of t with bounded derivatives, referred to as the *noise schedule*. Let $q_t(\mathbf{x}_t|c)$ be the marginal distribution of $\mathbf{x}_t$ conditioned on c, DPM ensures that $q_T(\mathbf{x}_T|c) \approx \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ for some $\sigma > 0$, and the signal-to-noise-ratio (SNR) α_t^2/σ_t^2 is strictly decreasing with respect to timestep t [17].

Seminal works [17, 29] studied the underlying stochastic differential equation (SDE) and ordinary differential equation (ODE) theory of DPM. The forward and reverse processes are as below for any $t \in [0, T]$:

$$\mathrm{d}\mathbf{x}_t = f_t\mathbf{x}_t\mathrm{d}t + g_t\mathrm{d}\mathbf{w}_t, \quad \mathbf{x}_0 \sim q_0(\mathbf{x}_0|c), \quad (5)$$

$$\mathrm{d}\mathbf{x}_t = [f_t\mathbf{x}_t - g_t^2\nabla_{\mathbf{x}_t}\log q_t(\mathbf{x}_t|c)]\mathrm{d}t + g_t\mathrm{d}\bar{\mathbf{w}}_t, \quad (6)$$

where $\mathbf{w}_t, \bar{\mathbf{w}}_t$ are standard Wiener processes in forward and reverse time, respectively, and f_t, g_t have closed-form expressions with respect to α_t, σ_t. The unknown $\nabla_{\mathbf{x}_t}\log q_t(\mathbf{x}_t|c)$ is referred to as the conditional score function. Probability flow ODE (PF-ODE) from Fokker-Planck equation enjoys the identical marginal distribution at each t as that of the SDE in Eq. (6), *i.e.*,

$$\frac{\mathrm{d}\mathbf{x}_t}{\mathrm{d}t} = f_t\mathbf{x}_t - \frac{1}{2}g_t^2\nabla_{\mathbf{x}_t}\log q_t(\mathbf{x}_t|c). \quad (7)$$

Technically, DPM implements sampling by solving the reverse SDE or ODE from T to 0. To this end, it introduces a neural network $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, c, t)$, namely the noise prediction model, to approximate the conditional score function from the given $\mathbf{x}_t$ and c at timestep t, *i.e.*, $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, c, t) = -\sigma_t\nabla_{\mathbf{x}_t}\log q_t(\mathbf{x}_t|c)$, where the parameter θ can be optimized by the objective below:

$$\mathbb{E}_{\mathbf{x}_0,\boldsymbol{\epsilon},c,t}[\omega_t\|\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, c, t) - \boldsymbol{\epsilon}\|_2^2], \quad (8)$$

where ω_t is the weighting function, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $c \sim q(c)$, $\mathbf{x}_t = \alpha_t\mathbf{x}_0 + \sigma_t\boldsymbol{\epsilon}$, and $t \sim \mathcal{U}[0, T]$.

For better condition fidelity, during denoising stage, CFG [12] turns to use a linear interpolation between conditional and unconditional score functions, *i.e.*,

$$s_{t,\gamma}(\mathbf{x}, c) = \gamma\nabla_{\mathbf{x}}\log q_t(\mathbf{x}|c) + (1-\gamma)\nabla_{\mathbf{x}}\log q_t(\mathbf{x}). \quad (9)$$

Then PF-ODE can be rewritten as

$$\frac{\mathrm{d}\mathbf{x}_t}{\mathrm{d}t} = f_t\mathbf{x}_t - \frac{1}{2}g_t^2 s_{t,\gamma}(\mathbf{x}_t, c). \quad (10)$$

We further describe the CFG under the original DDIM theory. Recall that DDIM turns out to formulate discrete non-Markovian forward diffusing process such that the reverse denoising process obeys the distribution with parameters $\{\delta_t\}_{t=0}^T$ [28]:

$$q_\delta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0, c) = q_\delta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \quad (11)$$

$$\sim \mathcal{N}\left(\alpha_{t-1}\mathbf{x}_0 + \sqrt{\sigma_{t-1}^2 - \delta_t^2}\cdot\frac{\mathbf{x}_t - \alpha_t\mathbf{x}_0}{\sigma_t}, \delta_t^2\mathbf{I}\right). \quad (12)$$

Trainable generative process $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, c)$ is designed to leverage $q_\delta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0, c)$ with a further designed denoised observation $\mathbf{f}_\theta^t$ with noise prediction model $\boldsymbol{\epsilon}_\theta$, *i.e.*,

$$\mathbf{f}_\theta^t(\mathbf{x}_t, c) = \frac{1}{\alpha_t}(\mathbf{x}_t - \sigma_t\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, c, t)), \quad (13)$$

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, c) = \begin{cases} q_\delta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{f}_\theta^t(\mathbf{x}_t, c), c), & t > 1, \\ \mathcal{N}(\mathbf{f}_\theta^t(\mathbf{x}_1), \sigma_1^2\mathbf{I}), & t = 1. \end{cases} \quad (14)$$

DDIM proves that for any $\{\delta_t\}_t$, score matching of non-Markovian process above is equivalent to native DPM. With CFG weight γ, we generalize the theory as below:

$$\hat{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t, c, t) = \gamma\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, c, t) + (1-\gamma)\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t), \quad (15)$$

$$\hat{\mathbf{f}}_\theta^t(\mathbf{x}_t, c) = \frac{1}{\alpha_t}(\mathbf{x}_t - \sigma_t\hat{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t, c, t)), \quad (16)$$

$$\hat{p}_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, c) = \begin{cases} q_\delta(\mathbf{x}_{t-1}|\mathbf{x}_t, \hat{\mathbf{f}}_\theta^t(\mathbf{x}_t, c), c), & t > 1, \\ \mathcal{N}(\hat{\mathbf{f}}_\theta^t(\mathbf{x}_1, c), \sigma_1^2\mathbf{I}), & t = 1. \end{cases} \quad (17)$$

Native DDIM theory still holds since $q_\delta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0, c) = q_\delta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$, *i.e.*, with the definition

$$J_{\delta,\gamma}(\boldsymbol{\epsilon}_\theta) = \mathbb{E}_{q_\delta(\mathbf{x}_{0:T}|c)}\left[\log\frac{q_\delta(\mathbf{x}_{1:T}|\mathbf{x}_0, c)}{\hat{p}_\theta(\mathbf{x}_{0:T}|c)}\right], \quad (18)$$

we have the following theorem. Proof is in Appendix A.1.

Theorem 1. *For any $\{\delta_t\}_t$ and $\gamma > 1$, $J_{\delta,\gamma}$ is equivalent to native DPM under CFG up to a constant. However, denoising with CFG is not a reciprocal of the original diffusion process with Gaussian noise due to nonzero expectation of unconditional score function $\mathbb{E}_{q_t(\mathbf{x}_t|c)}[\nabla_{\mathbf{x}_t}\log q_t(\mathbf{x}_t)]$.*

Remark 1. *$\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, c, t)$ and $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$ are proportional to $\nabla_{\mathbf{x}_t}\log q_t(\mathbf{x}_t|c)$ and $\nabla_{\mathbf{x}_t}\log q_t(\mathbf{x}_t)$ with coefficients being each minus standard deviation respectively, and empirically we use the same fixed variance for both $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0, c)$ and $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$. Therefore, Theorem 1 is consistent with the original CFG using score functions in Eq. (9).*

3.2. Misconceptions on Expectation Shift

CFG is designed to concentrate on better exemplars for each denoising step by sharpening the gamma-powered distribution as below [1]:

$$q_{t,\gamma}(\mathbf{x}|c) = q_t(\mathbf{x}|c)^\gamma q_t(\mathbf{x})^{1-\gamma}. \quad (19)$$

We first generalize the counterexample in [1] to confirm the expectation shift phenomenon. For VE-SDE with deterministic sampling recipe, we consider the 1-dimensional distribution with $q_0(\mathbf{x}_0|c) \sim \mathcal{N}(c, 1)$, $q(c) \sim \mathcal{N}(0, 1)$, $q_0(\mathbf{x}_0) \sim \mathcal{N}(0, 2)$. Then we can formulate the forward process and score functions as below:

$$q_t(\mathbf{x}_t|c) \sim \mathcal{N}(c, 1+t),\ \nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t|c) = -\frac{\mathbf{x}_t - c}{1+t}, \quad (20)$$

$$q_t(\mathbf{x}_t) \sim \mathcal{N}(0, 2+t),\ \nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t) = -\frac{\mathbf{x}_t}{2+t}. \quad (21)$$

We state the theorem below describing the expectation shift. Proof is addressed in Appendix A.2.

Theorem 2. *Denote by $q^{\text{deter}}_{0,\gamma}(\mathbf{x}_0|c)$ the conditional distribution by solving PF-ODE in Eq. (10) with $\gamma > 1$. Then $q^{\text{deter}}_{0,\gamma}(\mathbf{x}_0|c)$ follows the closed-form expression as below.*

$$q^{\text{deter}}_{0,\gamma}(\mathbf{x}_0|c) \sim \mathcal{N}\left(c\phi(\gamma, T), 2^{1-\gamma}\psi(\gamma, T)\right), \quad (22)$$

in which

$$\phi(\gamma, T) = \frac{2^{\frac{1-\gamma}{2}}}{(T+1)^{\frac{\gamma}{2}}(T+2)^{\frac{1-\gamma}{2}}} \quad (23)$$

$$+ \frac{\gamma}{2^{\frac{\gamma+1}{2}}} \int_0^T \frac{(s+1)^{-\frac{\gamma+2}{2}}}{(s+2)^{\frac{1-\gamma}{2}}} \mathrm{d}s, \quad (24)$$

$$\psi(\gamma, T) = \frac{T+1}{(T+1)^{\gamma}(T+2)^{1-\gamma}}. \quad (25)$$

Specifically, when $T \to +\infty$, denote by $\phi(\gamma)$ with

$$\phi(\gamma) = \lim_{T \to +\infty} \phi(\gamma, T), \quad (26)$$

we have $\phi(\gamma) \geqslant \gamma \frac{7}{15} \left(\frac{10}{7}\right)^{\frac{5-\gamma}{2}}$ for $\gamma \in [1, 3]$, $\phi(1) = 1$, $\phi(3) = 2$, $\phi(\gamma) \geqslant 2$ for all $\gamma > 3$, and

$$q^{\text{deter}}_{0,\gamma}(\mathbf{x}_0|c) \sim \mathcal{N}(c\phi(\gamma), 2^{1-\gamma}). \quad (27)$$

However, note that the ground-truth conditional distribution $q_0(\mathbf{x}_0|c) \sim \mathcal{N}(c, 1)$, indicating that the ground-truth expectation is equal to c. That is to say, denoising with CFG achieves at least twice as large expectation as the ground-truth one. Fig. 1 clearly describes the phenomenon.

3.3. Rectified Classifier-Free Guidance

Recall that the constraint of the two coefficients with summation one disables the compatibility with diffusion theory and indicates expectation shift. Theorem 2 quantitatively describes the expectation shift, claiming that the two coefficients of conditional and unconditional score functions in Eq. (9) dominate both the expectation and variance of $q^{\text{deter}}_{0,\gamma}(\mathbf{x}_0|c)$. To this end, we propose to rectify CFG with relaxation on the guidance coefficients, *i.e.*,

$$s_{t,\gamma_1,\gamma_0}(\mathbf{x}, c) = \gamma_1 \otimes \nabla_{\mathbf{x}} \log q_t(\mathbf{x}|c) \quad (28)$$

$$+ \gamma_0 \otimes \nabla_{\mathbf{x}} \log q_t(\mathbf{x}), \quad (29)$$

in which γ_1, $\gamma_0 \in \mathbb{R}^D$ are functions with respect to condition c and timestep t, and $\otimes$ indicates element-wise product. Denote by $q^{\text{deter}}_{0,\gamma_1,\gamma_0}(\mathbf{x}_0|c)$ the attached conditional distribution following PF-ODE in Eq. (10) with $s_{t,\gamma_1,\gamma_0}(\mathbf{x}, c)$.

To make guided sampling compatible with the diffusion theory and annihilate expectation shift, it suffices to choose more appropriate γ_1 and γ_0 according to input condition c and timestep t. Intuitively, we need the constraint such that:

- Each component of γ is larger than one for strengthened conditional fidelity, *i.e.*, $\gamma_{1,i} > 1$,
- Denoising with PF-ODE and Eq. (28) is theoretically the reciprocal of forward process, thus $q^{\text{deter}}_{0,\gamma_1,\gamma_0}(\mathbf{x}_0|c)$ enjoys the same expectation as the ground-truth $q_0(\mathbf{x}_0|c)$,
- $q^{\text{deter}}_{0,\gamma_1,\gamma_0}(\mathbf{x}_0|c)$ enjoys smaller or the same variance as the ground-truth $q_0(\mathbf{x}_0|c)$ for sharper distribution and thus concentrated better exemplars.

In the sequel, we omit $\otimes$ for simplicity. We first focus on the compatibility with the diffusion theory. We have claimed in Theorem 1 that CFG cannot satisfy the diffusion theory due to nonzero $\mathbb{E}_{q_t(\mathbf{x}_t|c)}[\nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t)]$. To this end, it suffices to annihilate the expectation shift as below:

$$\mathbb{E}_{q_t(\mathbf{x}_t|c)}[s_{t,\gamma_1,\gamma_0}(\mathbf{x}, c)] = \mathbf{0}. \quad (30)$$

To confirm the feasibility and precisely describe the expectation of $q^{\text{deter}}_{0,\gamma_1,\gamma_0}(\mathbf{x}_0|c)$, resembling Eqs. (16) and (17) we write denoised observation and denoising process as below:

$$\hat{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t, c, t) = \gamma_1 \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, c, t) + \gamma_0 \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t), \quad (31)$$

$$\hat{\mathbf{f}}^t_\theta(\mathbf{x}_t, c) = \frac{1}{\alpha_t}(\mathbf{x}_t - \sigma_t \hat{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t, c, t)), \quad (32)$$

$$\hat{p}_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, c) = \begin{cases} q_\delta(\mathbf{x}_{t-1}|\mathbf{x}_t, \hat{\mathbf{f}}^t_\theta(\mathbf{x}_t, c), c), & t > 1, \\ \mathcal{N}(\hat{\mathbf{f}}^t_\theta(\mathbf{x}_1, c), \sigma_1^2 \mathbf{I}), & t = 1. \end{cases} \quad (33)$$

We have the theorem below, proof is in Appendix A.3.

Theorem 3. *Let $\mathbf{x}_t \sim q_t(\mathbf{x}_t|c)$, $\tilde{\mathbf{x}}_t \sim \hat{p}_\theta(\tilde{\mathbf{x}}_t|c)$ induced from DDIM sampler in Eq. (33). Assume that all $\delta_t = 0$, denote by Δ_t the difference between expectation of $\mathbf{x}_t$ and $\tilde{\mathbf{x}}_t$, by $\boldsymbol{\epsilon}^{c,t}_{\gamma_1,\gamma_0}$ the interpolation between score functions, i.e.,*

$$\Delta_t = \mathbb{E}_{q_t(\mathbf{x}_t|c)}[\mathbf{x}_t] - \mathbb{E}_{\hat{p}_\theta(\tilde{\mathbf{x}}_t|c)}[\tilde{\mathbf{x}}_t], \quad (34)$$

$$\boldsymbol{\epsilon}^{c,t}_{\gamma_1,\gamma_0}(\mathbf{x}) = (\gamma_1 - 1)\boldsymbol{\epsilon}_\theta(\mathbf{x}, c, t) + \gamma_0 \boldsymbol{\epsilon}_\theta(\mathbf{x}, t). \quad (35)$$

Then we have the following recursive equality:

$$\Delta_{t-1} = \frac{\sigma_{t-1}}{\sigma_t}\Delta_t - (\sigma_{t-1} - \frac{\alpha_{t-1}}{\alpha_t}\sigma_t)\mathbb{E}_{\tilde{\mathbf{x}}_t}[\boldsymbol{\epsilon}^{c,t}_{\gamma_1,\gamma_0}(\tilde{\mathbf{x}}_t)]. \quad (36)$$

Specifically, when $\Delta_t = 0$, we have:

$$\Delta_{t-1} = -(\sigma_{t-1} - \frac{\alpha_{t-1}}{\alpha_t}\sigma_t)\mathbb{E}_{\mathbf{x}_t}[\boldsymbol{\epsilon}^{c,t}_{\gamma_1,\gamma_0}(\mathbf{x}_t)]. \quad (37)$$

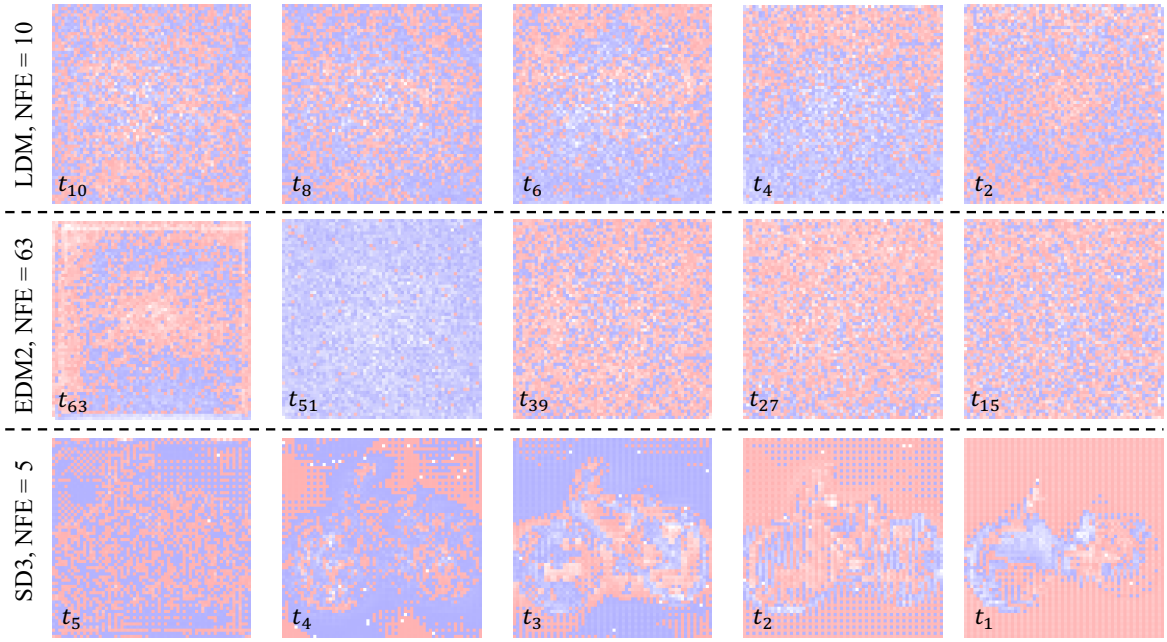


Figure 2. **Visualization** of the lookup table on LDM [25], EDM2 [16], and SD3 [9], each of which consists of the expectation ratio $\mathbb{E}_{\mathbf{x}_t}[\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, c, t)]/\mathbb{E}_{\mathbf{x}_t}[\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)]$. Each pixel represents the scale of the pixel-wise ratio, *i.e.*, color **red** implies that ratio is greater than one, while color **blue** stands for ratio smaller than one. The darker the color is, the farther the ratio appears away from one. We report in each row the expectation ratios on five timesteps uniformly sampled from the whole trajectory, under different DPMs and NFEs. It is noteworthy that expectation ratios at the same timestep vary largely by different pixels, and there is no general pattern along with timesteps or pixels.

Theorem 3 studies the difference between expectation of denoising with Eq. (28) and the ground-truth. Note that

$$\mathbb{E}_{\mathbf{x}_t}[\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, c, t)] = \mathbb{E}_{\mathbf{x}_t}[\mathbb{E}[\boldsymbol{\epsilon}|\mathbf{x}_t]] = \mathbb{E}_{\mathbf{x}_t}[\boldsymbol{\epsilon}] = \mathbf{0}, \tag{38}$$

therefore we have

$$\begin{aligned} &\mathbb{E}_{\mathbf{x}_t}[\boldsymbol{\epsilon}^{c,t}_{\gamma_1,\gamma_0}(\mathbf{x}_t)] && (39)\\ =\ &\mathbb{E}_{\mathbf{x}_t}[(\gamma_1 - 1)\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, c, t) + \gamma_0\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)] && (40)\\ =\ &\mathbb{E}_{\mathbf{x}_t}[\gamma_1\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, c, t) + \gamma_0\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)], && (41)\end{aligned}$$

which coincides with Eq. (30), indicating the feasibility and a closed-form solution given c and t as below:

$$\gamma_0 = (1 - \gamma_1)\mathbb{E}_{\mathbf{x}_t}[\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, c, t)]/\mathbb{E}_{\mathbf{x}_t}[\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)]. \tag{42}$$

As for variance, however, normally we cannot analytically calculate the variance of $\hat{p}_\theta(\mathbf{x}_t|c)$. Instead, we study the variance of toy data in Sec. 3.2 as an empirical evidence in the following theorem, where proof is in Appendix A.4.

Theorem 4. *Under settings in Theorem 2, denote by* $q^{\text{deter}}_{0,\gamma_1,\gamma_0}(\mathbf{x}_0|c)$ *the conditional distribution by PF-ODE with* γ_1 *and* γ_0 *as in Eq. (28). Then we have*

$$\text{var}_{q^{\text{deter}}_{0,\gamma_1,\gamma_0}(\mathbf{x}_0|c)}[\mathbf{x}_0] = 2^{\gamma_0}(T+1)^{1-\gamma_1}(T+2)^{-\gamma_0}. \tag{43}$$

According to Theorem 4, it is noteworthy that variance of $q^{\text{deter}}_{0,\gamma_1,\gamma_0}(\mathbf{x}_0|c)$ under toy setting is guaranteed to be smaller than the ground-truth $\text{var}_{q_0(\mathbf{x}_0|c)}[\mathbf{x}_0] = 1$ when each component satisfies that $\gamma_{0,i} \leqslant 0$ and $\gamma_{1,i} + \gamma_{0,i} \geqslant 1$, especially when $T \to +\infty$.

Now we formally propose the constraints. First, we need each component $\gamma_{1,i} > 1$ for strengthened conditional fidelity. Then for expectation, it is noteworthy that $\Delta_T = 0$ satisfies the assumption in Theorem 3. Therefore by induction, it is feasible to annihilate Δ_0 by annihilation of Eq. (30) at all intermediate timesteps t. Finally as for variance, we empirically set $\gamma_{0,i} \leqslant 0$ and $\gamma_{1,i} + \gamma_{0,i} \geqslant 0$.

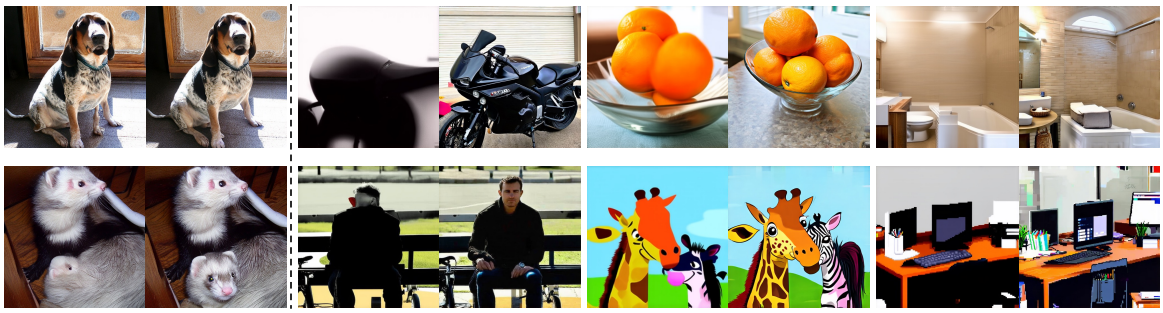
Practically, we can determine γ_0 according to the guidance strength γ_1, condition c, and timestep t, according to the closed-form solution in Eq. (42). Concretely, given condition c, it is feasible to pre-compute a collection of $\{(\epsilon_\theta(\mathbf{x}_t, c, t), \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t))\}_t$ by traversing $q_0(\mathbf{x}_0|c)$, and maintain a lookup table consisting of $\mathbb{E}_{\mathbf{x}_t}[\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, c, t)]/\mathbb{E}_{\mathbf{x}_t}[\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)]$. Then given any γ_1, we can directly achieve γ_0 by multiplying $-(\gamma_1 - 1)$ with the expectation ratio. Pseudo-code is addressed in Appendix B.

We make further discussion about `ReCFG`. By Cauchy-Schwarz inequality and Eq. (37) we have:

$$\|\Delta_{t-1}\|_2^2 \leqslant (\sigma_{t-1} - \frac{\alpha_{t-1}}{\alpha_t}\sigma_t)^2\mathbb{E}_{\mathbf{x}_t}[\|\boldsymbol{\epsilon}^{c,t}_{\gamma_1,\gamma_0}(\mathbf{x}_t)\|_2^2]. \tag{44}$$

Then we can define the objective resembling DPMs as below, optimizing reversely from $t = T$ to 0.

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}_t,t}[\|(\gamma_1 - 1)\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, c, t) + \gamma_0\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\|_2^2]. \tag{45}$$



(a) EDM2 (b) SD3

Figure 3. **Qualitative comparison** on EDM2 and SD3. Left and right in each cell suggest samples via CFG and ReCFG, respectively.

Resembling Theorem 1, with Eq. (45), we can also show the compatibility of ReCFG with DDIM, which is summarized as the theorem below. Proof is addressed in Appendix A.5

Theorem 5. *For any $\{\delta_t\}_t$, ReCFG with $\mathcal{L}$ is compatible with native DPM up to a constant.*

4. Experiments

4.1. Experimental Setups

Datasets and baselines. We apply ReCFG to seminal class-conditioned and text-conditioned DPMs, including LDM [25] and DiT [21] on ImageNet 256 [6], EDM2 [16] on ImageNet 512, and SD3 [9] on CC12M [2], respectively.

Evaluation metrics. As for class-conditioned LDM, DiT, and EDM2, we draw 50,000 samples for Fréchet Inception Distance (FID) [11] and $\text{FD}_{\text{DINOv2}}$ [30] to evaluate the fidelity and global coherency of the synthesized images, respectively. We further use Improved Precision (Prec.) and Recall (Rec.) [18] to separately measure sample fidelity (Precision) and diversity (Recall). As for text-conditioned SD3, following the official implementation, we use CLIP Score (CLIP-S) [10, 24], FID, and $\text{FD}_{\text{DINOv2}}$ on CLIP features [26] on 1,000 samples to evaluate conditional faithfulness and fidelity of the synthesized images, respectively. We also use MPS [31] to evaluate aesthetic scores. All four metrics are evaluated on the same MS-COCO validation split [20] as in official implementation [9].

Implementation details. We implement ReCFG with NVIDIA A100 GPUs, and employ pre-trained LDM[2], DiT[3], EDM2[4], and SD3[5] checkpoints provided in official implementation. We reproduce all the experiments with official and more other configurations including NFEs and guidance strengths.

[2]https://github.com/CompVis/latent-diffusion
[3]https://github.com/facebookresearch/DiT
[4]https://github.com/NVlabs/edm2
[5]https://huggingface.co/stabilityai/stable-diffusion-3-medium-diffusers

4.2. Results on Toy Example in Section 3.2

We first confirm the effectiveness of our method on toy data, as presented in Sec. 3.2. Given the closed-form expressions of score functions, we are able to precisely describe the distributions of both gamma-powered distribution $q_{0,\gamma}(\mathbf{x}_0|c)$ by native CFG and $q^{\text{deter}}_{0,\gamma_1,\gamma_0}(\mathbf{x}_0|c)$ by our ReCFG. The theoretical and numerical DDIM-based simulation value of probability density functions of both $q_{0,\gamma}(\mathbf{x}_0|c)$ and $q^{\text{deter}}_{0,\gamma_1,\gamma_0}(\mathbf{x}_0|c)$ are shown in Fig. 1. It is noteworthy that native CFG drifts the expectation of $q_{0,\gamma}(\mathbf{x}_0|c)$ further away from the peak of the ground-truth $q_0(\mathbf{x}_0|c)$ as γ becomes larger, consistent with Theorem 2. As a comparison, the peaks of $q^{\text{deter}}_{0,\gamma_1,\gamma_0}(\mathbf{x}_0|c)$ and $q_0(\mathbf{x}_0|c)$ coincide, while $q^{\text{deter}}_{0,\gamma_1,\gamma_0}(\mathbf{x}_0|c)$ is sharpened with smaller variance. Therefore, by adopting relaxation on coefficients γ_1 and γ_0 with specially proposed constraints, our ReCFG manages to annihilate expectation shift, enabling better guidance and thus better conditional fidelity.

4.3. Results on Real Datasets

We showcase some results in Fig. 3. One can see that ReCFG could fix artifacts on EDM2. It is also noteworthy that ReCFG significantly improves synthesis quality on SD3, especially detailed textures. Beyond the exhibited visualization, we conduct extensive quantitative experiments on state-of-the-art DPMs to further convey the efficacy of ReCFG. From Tabs. 1 and 2, we can tell that ReCFG is capable of better performance on both class-conditioned and text-conditioned DPMs under various guidance strengths and NFEs especially CLIP-S, indicating better conditional fidelity on open-vocabulary synthesis. Furthermore, correction for theoretical flaws of CFG enables strong compatibility of ReCFG with other empirical strategies such as RescaleCFG [19], achieving better performance.

4.4. Analyses

Variance of lookup table. Note that we need to precompute the lookup table consisting of expectation ratios for all conditions c, which is time-consuming and impractical for open-vocabulary distributions (*e.g.*, text-conditioned

Table 1. **Sample quality** on ImageNet [6].

ImageNet 256x256						
Model	Method	NFE (↓)	FD_{DINOv2} (↓)	FID (↓)	Prec. (↑)	Rec. (↑)
DiT-XL/2	CFG	250	120.07	2.27	0.83	0.57
$\gamma_1 = 1.50$	ReCFG	250	**118.71**	**2.13**	0.83	**0.58**
DiT-XL/2	CFG	250	162.68	3.22	0.76	0.62
$\gamma_1 = 1.25$	ReCFG	250	**145.79**	**3.01**	**0.77**	**0.63**
LDM	CFG	20	180.60	18.87	0.95	0.15
$\gamma_1 = 5.0$	ReCFG	20	**169.41**	**16.95**	**0.91**	**0.18**
LDM	CFG	20	149.79	11.46	**0.94**	0.27
$\gamma_1 = 3.0$	ReCFG	20	**142.54**	**9.78**	0.91	**0.32**
LDM	CFG	20	152.51	5.32	0.88	0.42
$\gamma_1 = 2.0$	ReCFG	20	**149.91**	**4.40**	0.88	**0.45**
LDM	CFG	20	203.17	5.36	0.80	0.51
$\gamma_1 = 1.5$	ReCFG	20	**198.44**	**4.78**	0.80	**0.53**
LDM	CFG	10	156.41	16.78	**0.94**	0.16
$\gamma_1 = 5.0$	ReCFG	10	**150.47**	**14.46**	0.89	**0.22**
LDM	CFG	10	153.97	10.13	0.91	0.28
$\gamma_1 = 3.0$	ReCFG	10	**142.04**	**8.26**	0.91	**0.33**
LDM	CFG	10	183.39	7.83	0.81	0.38
$\gamma_1 = 2.0$	ReCFG	10	**182.04**	**5.98**	**0.83**	**0.42**
LDM	CFG	10	251.07	13.19	0.69	0.46
$\gamma_1 = 1.5$	ReCFG	10	**248.23**	**11.27**	**0.72**	**0.49**
ImageNet 512x512						
Model	Method	NFE (↓)	FD_{DINOv2} (↓)	FID (↓)	Prec. (↑)	Rec. (↑)
EDM2-S	CFG	63	52.32	2.29	0.83	0.59
	ReCFG	63	**50.56**	**2.23**	0.83	0.59
EDM2-M	CFG	63	41.98	2.12	0.81	0.60
	ReCFG	63	**41.55**	**2.06**	0.81	0.61
EDM2-L	CFG	63	38.20	1.96	0.81	0.62
	ReCFG	63	**36.75**	**1.89**	0.81	0.62

Table 2. **Sample quality** on CC12M [2]

CC12M 512x512, SD3 [9]					
Method	γ_1	NFE (↓)	CLIP-S (↑)	FD_{DINOv2} (↓)	MPS (↑)
CFG	7.5	10	0.262	1105.51	9.828
ReCFG	7.5	10	0.263	1010.14	10.250
RescaleCFG [19]	7.5	10	0.267	1011.62	11.258
RescaleCFG + ReCFG	7.5	10	**0.268**	**979.87**	**11.336**
CFG	5.0	10	0.268	1053.44	10.883
ReCFG	5.0	10	**0.269**	999.48	11.031
RescaleCFG [19]	5.0	10	0.267	1009.67	11.242
RescaleCFG + ReCFG	5.0	10	**0.269**	**984.25**	**11.297**
CFG	2.5	10	0.265	1016.79	10.367
ReCFG	2.5	10	0.265	977.39	10.438
RescaleCFG [19]	2.5	10	0.265	1003.64	10.445
RescaleCFG + ReCFG	2.5	10	**0.266**	**963.21**	**10.477**
CFG	7.5	5	0.209	1466.91	3.189
ReCFG	7.5	5	0.229	1323.49	3.979
RescaleCFG [19]	7.5	5	**0.258**	1114.92	8.102
RescaleCFG + ReCFG	7.5	5	**0.258**	**1070.65**	**8.219**
CFG	5.0	5	0.248	1218.18	6.484
ReCFG	5.0	5	0.258	1074.60	7.398
RescaleCFG [19]	5.0	5	0.265	1087.98	8.719
RescaleCFG + ReCFG	5.0	5	**0.266**	**1040.01**	**8.813**
CFG	2.5	5	0.261	1119.06	7.902
ReCFG	2.5	5	**0.263**	1058.86	8.172
RescaleCFG [19]	2.5	5	0.262	1093.85	8.133
RescaleCFG + ReCFG	2.5	5	**0.263**	**1041.53**	**8.266**

Table 3. **Variance** of lookup table over condition c and timestep t. Note that we employ pixel-wise lookup table involving both c and t. We report the the mean and variance of lookup table over c and t, respectively, which are computed by averaging on all pixels.

Config.	LDM, NFE $= 10$	EDM2, NFE $= 63$	SD3, NFE $= 5$	SD3, NFE $= 10$
Variance over c	1.0050 ± 0.0012	1.0060 ± 0.0119	1.0250 ± 0.0369	1.0125 ± 0.0281
Variance over t	1.0050 ± 0.0013	1.0060 ± 0.1545	1.0250 ± 0.0359	1.0125 ± 0.0306

DPMs). In Tab. 3 we report the mean and variance of expectation ratios over condition c, which is averaged on all timesteps and pixels. One can observe that larger NFE suggests smaller ratio with also smaller variance. It is also noteworthy that the variance of text-conditioned DPMs is larger than that of class-conditioned ones due to far more complex open-vocabulary conditions, while both of which is insignificant compared to the mean. Therefore, it is feasible to prepare the lookup table for only part of all potential conditions and use the mean for all conditions, serving as a practically adequate strategy to improve time efficiency. Variance over timestep t averaged on all pixels and part of conditions is also reported in Tab. 3, where similar conclusion could be achieved.

Ablation studies. Recall that we pre-compute the lookup table by traversing $q_0(\mathbf{x}_0|c)$ for each condition c. Comprehensive ablation studies reported in Tabs. 4 and 5 convey a direct and clear picture of the efficacy of ReCFG under different numbers of traversals. We can conclude that larger number of traversals suggests better guidance performance, yet improvements from 100 to 500 traversals are relatively inconspicuous, especially on text-conditoned DPMs. In other words, employing 500 samples per condition is adequate in practice to serve as an empirical setting.

Time cost. Given the analyses on variance of expectation ratios over condition c and ablation on traversals, preparing the lookup table is quite efficient. In practice, we sample 500 images for a subset of 100 conditions, which takes ~ 3 hours using 1 NVIDIA A100 GPU. The time cost is very close to performing FID evaluation using 50,000 images.

Pixel-wise lookup table. ReCFG enables pixel-specific guidance coefficients γ_1 and γ_0 with the same shape as score functions, thanks to the closed-form solution in Eq. (42), *i.e.*, we can assign γ_0 for each pixel by maintaining the lookup table of pixel-wise expectation ratios. Fig. 2 demonstrates the ratios on LDM, EDM2, and SD3 at uniformly sampled timesteps under different NFEs. Both LDM and EDM2 show the generation of class 0 in ImageNet

Table 4. **Ablation study** of the number of traversals (the number after ReCFG) for lookup table on ImageNet [6]. For clearer demonstration, baselines of native CFG are highlighted in **gray**.

ImageNet 256x256, LDM [25]					
γ_1	γ_0	NFE (↓)	FID (↓)	Prec. (↑)	Rec. (↑)
3.0	-2.0	10	10.13	0.91	0.28
3.0	ReCFG-10	10	8.88	**0.92**	0.30
3.0	ReCFG-100	10	8.70	**0.92**	0.31
3.0	ReCFG-500	10	**8.26**	0.91	**0.33**
ImageNet 512x512, EDM2-S [16]					
γ_1	γ_0	NFE (↓)	FID (↓)	Prec. (↑)	Rec. (↑)
2.5	-1.5	63	5.87	**0.85**	0.46
2.5	ReCFG-10	63	5.06	0.84	0.47
2.5	ReCFG-100	63	4.99	0.84	0.45
2.5	ReCFG-500	63	**4.84**	0.84	**0.48**
2.0	-1.0	63	4.18	**0.85**	0.52
2.0	ReCFG-10	63	3.70	0.84	0.52
2.0	ReCFG-100	63	3.66	0.84	0.52
2.0	ReCFG-500	63	**3.61**	0.84	0.52

Table 5. **Ablation study** of the number of traversals (the number after ReCFG) for lookup table on CC12M [2]. For clearer demonstration, baselines of native CFG are highlighted in **gray**.

CC12M 512x512, SD3 [9]				
γ_1	γ_0	NFE (↓)	CLIP-S (↑)	FID (↓)
5.0	-4.0	25	0.267	72.37
5.0	ReCFG-10	25	0.267	72.15
5.0	ReCFG-100	25	**0.268**	72.03
5.0	ReCFG-500	25	**0.268**	**71.95**
5.0	-4.0	10	0.268	72.55
5.0	ReCFG-10	10	0.268	71.61
5.0	ReCFG-100	10	0.268	70.64
5.0	ReCFG-500	10	**0.269**	**70.31**
5.0	-4.0	5	0.248	115.51
5.0	ReCFG-10	5	0.252	107.09
5.0	ReCFG-100	5	0.256	103.25
5.0	ReCFG-500	5	**0.258**	**101.82**

(*i.e.*, "tench"), while SD3 adopts the prompt "A bicycle replica with a clock as the front wheel". One can observe that expectation ratios at the same timestep vary largely by different pixels, and there appears no general rules on the relation between γ_1 and γ_0. Therefore, it is indicated that trivially setting γ_1 and γ_0 to be scalars is less reasonable. As a comparison, our method makes it possible to employ more precise control on guided sampling in a simple and post-hoc fashion without further fine-tuning, enabling better performance. It is also noteworthy that expectation ratio of SD3 exhibits noticeable shapes, probably due to more informative text prompts than one-hot class labels and more powerful model thanks to the training scale.

4.5. Discussions

Classifier-Free Guidance is designed from Bayesian theory to facilitate conditional sampling, yet appears incompatible with original diffusion theory. Therefore, we believe ReCFG is attached to great importance on guided sampling by fixing the theoretical flaw of CFG. Despite the success on better conditional fidelity, our algorithm has several potential limitations. Theoretically, we need to pre-compute the lookup table by traversing the dataset to achieve rectified coefficients for each condition. Although we conduct extensive ablation studies on the number of traversals and variance over condition c, providing an adequate strategy especially for open-vocabulary datasets on text-conditioned synthesis, the optimal strategy is unexplored. Besides, we at present cannot provide precise control on variance of ReCFG due to incomputable variance in denoising process, and turn to employ empirical values. Therefore, how to further conquer these problems (*e.g.*, employing a predictor network $\boldsymbol{\omega}(c, t)$ for better γ_0 on open-vocabulary datasets according to Eq. (45)) will be an interesting avenue for future research. Although leaving the variance behavior unexplored, we hope that ReCFG will encourage the community to close the gap in the future.

5. Conclusion

In this paper, we analyze the theoretical flaws of native Classifier-Free Guidance technique and the induced expectation shift phenomenon. We theoretically claim the exact value of expectation shift on a toy distribution. Introducing a relaxation on coefficients of CFG and novel constraints, we manage to complete the theory of guided sampling by fixing the incompatibility between CFG and diffusion theory. Accordingly, thanks to the closed-form solution to the constraints, we propose ReCFG, a post-hoc algorithm aiming at more faithful guided sampling by determining the coefficients from a pre-computed lookup table. We further study the behavior of the lookup table, proposing an adequate strategy for better time efficiency in practice. Comprehensive experiments demonstrate the efficacy of our method on various state-of-the-art DPMs under different NFEs and guidance strengths.

Acknowledgements

This work was partially supported by the Natural Science Foundation of China (62461160309, 62302297), Beijing Science and Technology plan project (Z231100005923029), NSFC-RGC Joint Research Scheme (N_HKU705/24), Ant Group Research Intern Program, and Shanghai Sailing Program (22YF1420300).

References

[1] Arwen Bradley and Preetum Nakkiran. Classifier-free guidance is a predictor-corrector. *arXiv preprint arXiv:2408.09000*, 2024. 1, 2, 3, 4

[2] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 6, 7, 8

[3] Junsong Chen, Jincheng YU, Chongjian GE, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *Int. Conf. Learn. Represent.*, 2024. 1, 2

[4] Muthu Chidambaram, Khashayar Gatmiry, Sitan Chen, Holden Lee, and Jianfeng Lu. What does guidance do? a fine-grained analysis in a simple setting. In *Adv. Neural Inform. Process. Syst. Worksh.*, 2024. 2

[5] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. In *Int. Conf. Comput. Vis.*, 2021. 2

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2009. 6, 7, 8

[7] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In *Adv. Neural Inform. Process. Syst.*, 2021. 2

[8] Yilun Du, Conor Durkan, Robin Strudel, Joshua B. Tenenbaum, Sander Dieleman, Rob Fergus, Jascha Sohl-Dickstein, Arnaud Doucet, and Will Grathwohl. Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and mcmc. In *Int. Conf. Mach. Learn.*, 2023. 1, 2

[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. *arXiv preprint arXiv:2403.03206*, 2024. 1, 2, 5, 6, 7, 8

[10] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: a reference-free evaluation metric for image captioning. In *Empi. Met. Nat. Lang. Proc.*, 2021. 6

[11] 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. In *Adv. Neural Inform. Process. Syst.*, 2017. 6

[12] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *Adv. Neural Inform. Process. Syst. Worksh.*, 2021. 1, 2, 3

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Adv. Neural Inform. Process. Syst.*, 2020. 1, 2

[14] Lianghua Huang, Di Chen, Yu Liu, Shen Yujun, Deli Zhao, and Zhou Jingren. Composer: Creative and controllable image synthesis with composable conditions. In *ICML*, 2023. 2

[15] Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. *arXiv preprint arXiv:2406.02507*, 2024. 1, 2

[16] Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024. 2, 5, 6, 8

[17] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. In *Adv. Neural Inform. Process. Syst.*, 2021. 3

[18] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *arXiv preprint arXiv:1904.06991*, 2019. 6

[19] Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. In *IEEE Winter Conf. App. Comput. Vis.*, 2024. 6, 7

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context. *arXiv preprint arXiv:1405.0312*, 2015. 6

[21] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Int. Conf. Comput. Vis.*, 2023. 6

[22] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *Int. Conf. Learn. Represent.*, 2024. 1, 2

[23] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *Int. Conf. Learn. Represent.*, 2023. 1

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Int. Conf. Mach. Learn.*, 2021. 6

[25] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 1, 5, 6, 8

[26] Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. Projected gans converge faster. In *Adv. Neural Inform. Process. Syst.*, 2021. 6

[27] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Int. Conf. Mach. Learn.*, 2015. 1, 2

[28] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *Int. Conf. Learn. Represent.*, 2021. 2, 3

[29] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *Int. Conf. Learn. Represent.*, 2020. 1, 2, 3

[30] George Stein, Jesse C. Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Leigh Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L. Caterini, J. Eric T. Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. In *Adv. Neural Inform. Process. Syst.*, 2023. 6

[31] Sixian Zhang, Bohan Wang, Junqiang Wu, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang. Learning multi-dimensional human preference for text-to-image generation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024. 6