Inner Classifier-Free Guidance and Its Taylor Expansion for Diffusion Models

Shikun Sun[1], Longhui Wei, Zhicai Wang, Zixuan Wang, Junliang Xing, Jia Jia, Qi Tian

[1]*Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China*

Email: [1]ssk52839916@gmail.com

Abstract

Classifier-free guidance (CFG) is a pivotal technique for balancing the diversity and fidelity of samples in conditional diffusion models. This approach involves utilizing a single model to jointly optimize the conditional score predictor and unconditional score predictor, eliminating the need for additional classifiers. It delivers impressive results and can be employed for continuous and discrete condition representations. However, when the condition is continuous, it prompts the question of whether the trade-off can be further enhanced. Our proposed inner classifier-free guidance (ICFG) provides an alternative perspective on the CFG method when the condition has a specific structure, demonstrating that CFG represents a first-order case of ICFG. Additionally, we offer a second-order implementation, highlighting that even without altering the training policy, our second-order approach can introduce new valuable information and achieve an improved balance between fidelity and diversity for Stable Diffusion.

Contributations

- We introduce ICFG and analyze the convergence of its Taylor expansion under specific conditions.
- We demonstrate that CFG can be regarded as a first-order ICFG and propose a second-order Taylor expansion for our ICFG.
- We apply the second-order ICFG to the Stable Diffusion model and observe that, remarkably, our new formulation yields valuable information and enhances the trade-off between fidelity and diversity, even without modifying the training policy.

Preliminary

We assume that this diffusion process follows a SDE:

$$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}. \tag{1}$$

The score function is defined as follows:

$$\mathbf{s}(\mathbf{x}, t) = \nabla_{\mathbf{x}_t} \log q(\mathbf{x}_t). \tag{2}$$

Then, the reverse-time SDE is:

$$\mathrm{d}\mathbf{x} = [\mathbf{f}(\mathbf{x}, t) - g(t)^2\mathbf{s}(\mathbf{x}, t)]\mathrm{d}t + g(t)\mathrm{d}\overline{\mathbf{w}}. \tag{3}$$

For the unconditional diffusion score $\epsilon^\theta(\mathbf{x}, t)$, using the same set of classifiers, the modified diffusion score is given by:

$$\begin{aligned}\bar{\epsilon}^\theta(\mathbf{x}_t, \mathbf{c}, t) &= \epsilon^\theta(\mathbf{x}_t, t) - (w+1)\beta_t\nabla_{\mathbf{x}_t}\log p_t^\theta(\mathbf{c}|\mathbf{x}_t) \\ &= -\beta_t\nabla_{\mathbf{x}_t}\left[\log q^\theta(\mathbf{x}_t) + (w+1)\log p_t^\theta(\mathbf{c}|\mathbf{x}_t)\right].\end{aligned} \tag{4}$$

The main idea behind CFG is to use a single model to simultaneously fit both the conditional score predictor and the unconditional score predictor. This is achieved by randomly replacing the condition $\mathbf{c}$ with $\varnothing$ (an empty value). By doing so, one can obtain the conditional score predictor $\epsilon^\theta(\mathbf{x}, \mathbf{c}, t)$ and the unconditional score predictor $\epsilon^\theta(\mathbf{x}, t)$, which is equivalent to $\epsilon^\theta(\mathbf{x}, \varnothing, t)$. Then, because

$$\begin{aligned}\nabla_{\mathbf{x}_t}\left[\log p_t(\mathbf{c}|\mathbf{x}_t)\right] &= \nabla_{\mathbf{x}_t}\left[\log q(\mathbf{x}_t|\mathbf{c}) - \log q(\mathbf{x}_t) + \log p(\mathbf{c})\right] \\ &= \nabla_{\mathbf{x}_t}\left[\log q(\mathbf{x}_t|\mathbf{c}) - \log q(\mathbf{x}_t)\right],\end{aligned} \tag{5}$$

which indicates that after applying the operator $\nabla_{\mathbf{x}_t}$, we can replace the last term of Equation (4) with $\log q^\theta(\mathbf{x}_t|\mathbf{c}) - \log q^\theta(\mathbf{x}_t)$ to achieve a similar effect. Then we get the enhanced diffusion score:

$$\begin{aligned}\hat{\epsilon}^\theta(\mathbf{x}_t, \mathbf{c}, t) &= (w+1)\epsilon^\theta(\mathbf{x}_t, \mathbf{c}, t) - w\epsilon^\theta(\mathbf{x}_t, t) \\ &= -\beta_t\nabla_{\mathbf{x}_t}\left[\log q^\theta(\mathbf{x}_t|\mathbf{c}) + w(\log q^\theta(\mathbf{x}_t|\mathbf{c}) - \log q^\theta(\mathbf{x}_t))\right] \\ &= -\beta_t\nabla_{\mathbf{x}_t}\left[\log q^\theta(\mathbf{x}_t) + (w+1)(\log q^\theta(\mathbf{x}_t|\mathbf{c}) - \log q^\theta(\mathbf{x}_t))\right],\end{aligned} \tag{6}$$

whose enhanced intermediate distribution is:

$$\overline{q}^\theta(\mathbf{x}_t|\mathbf{c}) \propto q^\theta(\mathbf{x}_t)\left[\frac{q^\theta(\mathbf{x}_t|\mathbf{c})}{q^\theta(\mathbf{x}_t)}\right]^{w+1}. \tag{7}$$

Methodology

Theorem 0.1. *Given condition $\mathbf{c}$, the enhanced transition kernel $\overline{q}_{0t}^\theta(\mathbf{x}_t|\mathbf{x}_0, \mathbf{c})$ by Eq. (7) equals to the original transition kernel $q_{0t}^\theta(\mathbf{x}_t|\mathbf{x}_0, \mathbf{c}) = q_{0t}^\theta(\mathbf{x}_t|\mathbf{x}_0)$ does not hold trivially. Specifically, when $w = 0$, the equation holds.*

The question arises: Can we always ensure that $\beta = 1$?

Assumption 0.1.

- *$\mathcal{C}$ is a cone, which means $\forall\beta \in \mathbb{R}^+, \forall\mathbf{c} \in \mathcal{C}, \beta\mathbf{c} \in \mathcal{C}$.*
- *For each $\mathbf{c} \in \mathcal{C}$, $\|\mathbf{c}\|$ represents the guidance strength and $\frac{\mathbf{c}}{\|\mathbf{c}\|}$ represents the guidance direction.*

Under Assumption 0.1, we define $\bar{q}^\theta(x_t|c) = q^\theta(\mathbf{x}_t|\mathbf{c}, \beta) \triangleq q^\theta(\mathbf{x}_t|\beta\mathbf{c})$. Based on this definition, we can state the following Corollary 0.1.1:

Corollary 0.1.1. *Given condition $\mathbf{c}$ and the guidance strength $\beta = w + 1$, we have:*

$$q_{0t}^\theta(\mathbf{x}_t|\mathbf{x}_0, \mathbf{c}, \beta) = q_{0t}^\theta(\mathbf{x}_t|\mathbf{x}_0).$$

The following algorithm offers a practical solution and can be effectively applied to mitigate the aforementioned problem.

Algorithm 3 Non-strict sample algorithm for second-order ICFG

Require: m: middle point for estimate second-order term
Require: w: first-order guidance strength on conditional score predictor
Require: v: second-order guidance strength on conditional score predictor
Require: $\mathbf{c}$: condition for sampling
Require: Require $\{t_1, t_2, ..., t_N\}$ increasing timestep sequence of sampling
Require: $Sample(\mathbf{z}_t, \epsilon_t)$: sample algorithm for diffusion models given $\mathbf{z}_t$ and ϵ_t
1: $\mathbf{z}_N \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $i = N, ..., 1$ **do**
3: $\quad\bar{\epsilon}_t = \epsilon^\theta(\mathbf{z}_i, \mathbf{c}) + w(\epsilon^\theta(\mathbf{z}_i, \mathbf{c}) - \epsilon^\theta(\mathbf{z}_i))$
$\qquad + v\frac{1}{m(1-m)}\left((1-m)\epsilon^\theta(\mathbf{z}_i) + m\epsilon^\theta(\mathbf{z}_i, \mathbf{c}) - \epsilon^\theta(\mathbf{z}_i, m\mathbf{c})\right)$
4: $\quad\mathbf{z}_{i-1} = Sample(\mathbf{z}_i, \bar{\epsilon}_t)$
5: **end for**
6: **return** $\mathbf{z}_0$

Experiment

Evaluation Metrics. We evaluate the widely-used Frechet Inception Score (FID) between the generated images and the target domain images, and CLIP Score between generated images and captions on the MS-COCO validation set.

Results.

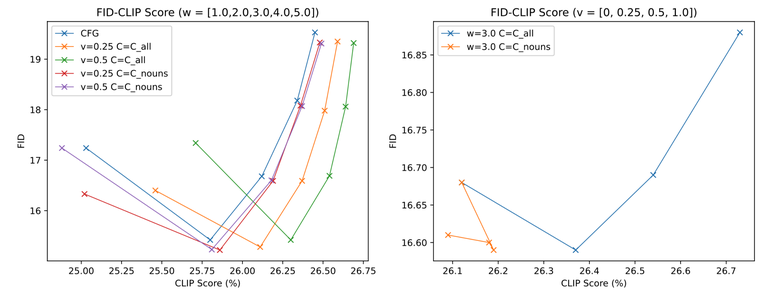

Figure 2: The FID-CLIP Score of varying w, v and $\mathcal{C}$.