Elucidating the SNR-t Bias of Diffusion Probabilistic Models

Supplementary Material

A. Difference from Prior Works

In this section, we outline the differences between the second experiment (Fig. 1c) in Sec 4 of this paper and prior work [37, 38]. We emphasize that ADM-ES [38] only provides a phenomenological conclusion and does not delve into the underlying causes of the phenomenon. In contrast, the SNR-t bias discovered in this paper, along with the sliding window experiments on neural networks based on Fig. 1b, provide in-depth explanations and evidence for this phenomenon. Additionally, this section offers more robust experimental analyses for the phenomenon.

(1) The SNR-t bias is the underlying cause of exposure bias proposed by ADM-IP [37] and ADM-ES [38]. ADM-IP and ADM-ES define the exposure bias as an intuitively inter-sample bias between the perturbed sample $\boldsymbol{x}_t$ and the predicted sample $\hat{\boldsymbol{x}}_t$. Meanwhile, ADM-ES also claims that exposure bias leads to the accumulation of errors, yet it fails to provide fundamental evidence for such error accumulation. In contrast, we explicitly demonstrate when the SNR of the input sample mismatches the timestep, the network's predictive output exhibits significant errors, as shown in the Key Finding 1 (Fig. 1b). Furthermore, since the SNR of reverse-process samples is consistently lower than the ideal level, as shown in the Key Finding 2 (Fig. 1c), the network's predictions during the reverse process are invariably erroneous, specifically manifesting as overestimated outputs. In summary, the SNR-t bias stems primarily from the forced coupling of sample SNR and timestep during training.

(2) Unlike ADM-ES, this paper focuses on drawing deeper conclusions and uncovering the underlying patterns. Specifically, Figure 2 in ADM-ES concludes that the L_2-norm of $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}_t, t)$ in the reverse process is always larger than that of $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$ in the forward process. However, ADM-ES does not explore the deep-seated reasons for this overestimation phenomenon. In this paper, we derive Finding 1 through the sliding window experiments in Sec. 4: for the fixed timestep s, when handling the sample $\boldsymbol{x}_t$ with a lower SNR, where $t > s$, the network tends to overestimate the predicted output. Conversely, when dealing with the sample $\boldsymbol{x}_t$ with a higher SNR, the predicted output is typically underestimated. Therefore, combining the findings of ADM-ES and Finding 1 of this paper, we arrive at Finding 2: Reverse denoising samples often exhibit lower SNR compared to their corresponding forward samples at the same timestep.

(3) Unlike exposure bias, an inter-sample bias, the SNR-t bias is a more specific SNR-timestep bias. Meanwhile, our method based on the SNR-t bias can be naturally integrated into state-of-the-art models for correcting exposure bias, such as ADM-IP, ADM-ES, and DPM-FR, further improving the generation quality of these correction models as shown in Sec. 6.2. Additionally, our method can significantly enhance the generation quality in the latest text-to-image models, as shown in Appendix E. Thus, these experiments further illustrate the differences between SNR-t bias and exposure bias, as well as the necessity of researching SNR-t bias.

Furthermore, we also provide more robust experimental evidence for Fig. 1c to eliminate interference caused by random seeds and sampling batch sizes. Specifically, we fix the sampling batch size at 2000 and then select different random number seeds (16, 42, and 99) to obtain distinct sampling trajectories, as illustrated in Figs. 5a, 5b, and 5c, respectively. Subsequently, we fix the random number seed and vary the sampling batch sizes (10, 100, and 1000), as shown in Figs. 5d, 5e, and 5f, respectively. Fig. 5 clearly demonstrates that regardless of the random number seed and sampling batch size, the network output of the reverse process is consistently larger than that of the forward process, which provides more robust evidence for our analysis.

B. Theoretical evidence of Assumption 5.1

Assumption 5.1. *During both the forward and reverse processes, the reconstruction sample $\boldsymbol{x}_{\boldsymbol{\theta}}^0(\boldsymbol{x}_t, t)$ can be expressed in terms of the original data $\boldsymbol{x}_0$ as follows:*

$$\boldsymbol{x}_{\boldsymbol{\theta}}^0(\hat{\boldsymbol{x}}_t, t) = \gamma_t \boldsymbol{x}_0 + \phi_t \boldsymbol{\epsilon}_t, \tag{22}$$

where $0 < \gamma_t \leqslant 1$, $\phi_t < M$, and M denotes a uniform upper bound constant across all timesteps.

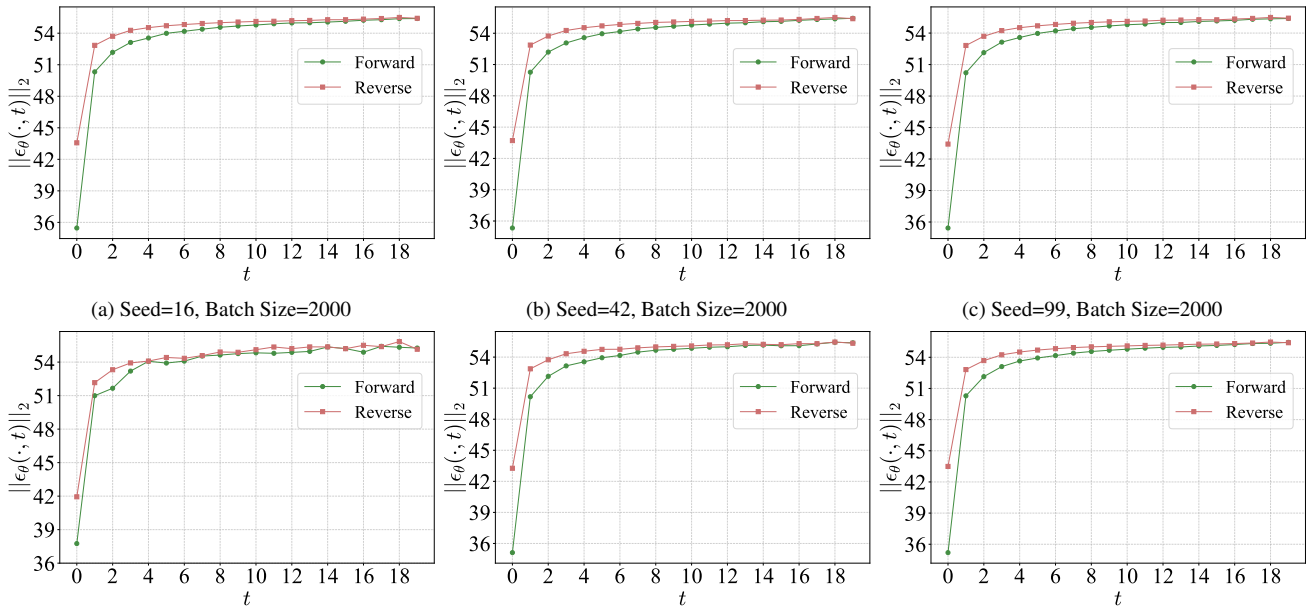
Specifically, we emphasize that both the forward reconstructed sample $\boldsymbol{x}_{\boldsymbol{\theta}}^0(\boldsymbol{x}_t, t)$ and the reverse reconstructed sample $\boldsymbol{x}_{\boldsymbol{\theta}}^0(\hat{\boldsymbol{x}}_t, t)$ adhere to the form specified in Eq. 22.

In this section, we present the detailed proof of Assumption 5.1. As stated in the main text, previous work proposed two distinct linear assumptions but lacked supporting evidence. However, we provide both experimental evidence and theoretical proofs to support our findings. Under Gaussian perturbation $\boldsymbol{q}_\sigma(\boldsymbol{y}|\boldsymbol{x})$, the Tweedie's formula is

$$\mathbb{E}[\boldsymbol{x}|\boldsymbol{y}] = \boldsymbol{y} + \sigma^2 \nabla_{\boldsymbol{y}} \log \boldsymbol{q}_\sigma(\boldsymbol{y}), \tag{23}$$

where $\boldsymbol{q}_\sigma(\boldsymbol{y}) := \int \boldsymbol{q}(\boldsymbol{y}|\boldsymbol{x})\boldsymbol{q}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$. Now, by substituting the forward perturbation distribution $\boldsymbol{q}(\boldsymbol{x}_t|\boldsymbol{x}_0)$ of DPMs into Eq. 23, we can obtain:

$$\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t] = \frac{\boldsymbol{x}_t + (1-\bar{\alpha}_t)\nabla_{\boldsymbol{x}_t} \log \boldsymbol{q}(\boldsymbol{x}_t)}{\sqrt{\bar{\alpha}_t}}. \tag{24}$$



(a) Seed=16, Batch Size=2000 (b) Seed=42, Batch Size=2000 (c) Seed=99, Batch Size=2000

(d) Seed=42, Batch Size=10 (e) Seed=42, Batch Size=100 (f) Seed=99, Batch Size=1000

Figure 5. Robust experimental results for Fig. 1c with varied random number seeds and sampling batch sizes. These figures show the network output $||\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\cdot,t)||_2$ using forward samples $\boldsymbol{x}_t$ via Eq. 2 and reverse predicted samples $\hat{\boldsymbol{x}}_t$ via Eq. 8, respectively. $||\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}_t,t)||_2$ is always larger than $||\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)||_2$ in every figure.

Based on the relationship between the score and the noise $\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t) = -\frac{\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)}{\sqrt{1-\bar{\alpha}_t}}$, we further derive:

$$\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t] = \frac{\boldsymbol{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)}{\sqrt{\bar{\alpha}_t}} = \boldsymbol{x}^0_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t), \quad (25)$$

which clearly demonstrates that the reconstructed sample $\boldsymbol{x}^0_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)$ is essentially the posterior mean based on the Tweedie formula. Furthermore, the score network trained with the L2 norm-MSE loss function always have a theoretical analytical solution [53], which is also the posterior mean:

$$\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t) = \mathbb{E}_{q(\boldsymbol{x}_0|\boldsymbol{x}_t)}\left[\nabla_{\boldsymbol{x}_t}\log q(\boldsymbol{x}_t \mid \boldsymbol{x}_0)\right]. \quad (26)$$

Based on the equivalence between the score and noise, the optimal solution for noise prediction is also the same posterior mean. Therefore, based on the mean tendency of denoising operations and network predictions, we can regard $\boldsymbol{x}^0_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)$ as the mean estimate $\bar{\boldsymbol{x}}_0$ of $\boldsymbol{x}_0$.

The variance formula is expressed as:

$$\mathbb{E}[\|\boldsymbol{x}_0\|^2] = \|\bar{\boldsymbol{x}}_0\|^2 + \mathrm{Var}(\|\boldsymbol{x}_0\|). \quad (27)$$

Based on the non-negativity of the variance, we obtain:

$$\|\bar{\boldsymbol{x}}_0\|^2 \leqslant \mathbb{E}[\|\boldsymbol{x}_0\|^2].$$

We substitute $\boldsymbol{x}^0_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)$ for $\bar{\boldsymbol{x}}_0$, then given that the expectation of a constant is the constant itself, we can take the expectation of both sides of the above equation to obtain:

$$\mathbb{E}[\|\boldsymbol{x}^0_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)\|^2] \leqslant \mathbb{E}[\|\boldsymbol{x}_0\|^2]. \quad (28)$$

Eq. 28 clearly demonstrates that the L2 norm of reconstructed samples is always smaller than that of real samples, which indicates that the reconstruction operation is always accompanied by information loss.

However, previous work [24, 38] argues that reconstructed samples should be modeled as:

$$\boldsymbol{x}^0_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t) = \boldsymbol{x}_0 + \phi_t\boldsymbol{\epsilon}_t, \quad (29)$$

which is clearly inconsistent with Eq. 28. Thus, We use the form in Eq. 22, consistent with the assumption of LA-DPM [64] and DPM-FR [63].

In addition, we also provide experimental evidence for the above proof. Following the experimental setup described in Sec. 4, we perform the following operations sequentially: (1) We generate perturbed samples $\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_T\}$ via Eq. 2, and feed $\boldsymbol{x}_t$ and timestep t into the network to obtain $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)$ to compute $\boldsymbol{x}^0_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)$ via Eq. 5. (2) Then, we initialize 2,000 standard Gaussian noise and iteratively denoise operation via Eq. 8 to obtain samples $\{\hat{\boldsymbol{x}}_1,\hat{\boldsymbol{x}}_2,\ldots,\hat{\boldsymbol{x}}_T\}$ and corresponding network outputs $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}_t,t)$ to compute $\boldsymbol{x}^0_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}_t,t)$ via Eq. 5. (3) Finally, we compute and plot the expectation of $||\boldsymbol{x}^0_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)||_2^2$, $||\boldsymbol{x}^0_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}_t,t)||_2^2$, and $||\boldsymbol{x}_0||_2^2$.

Fig. 6 clearly demonstrates that DPMs fail to fully reconstruct real data $\boldsymbol{x}_0$, both in the forward and reverse processes.

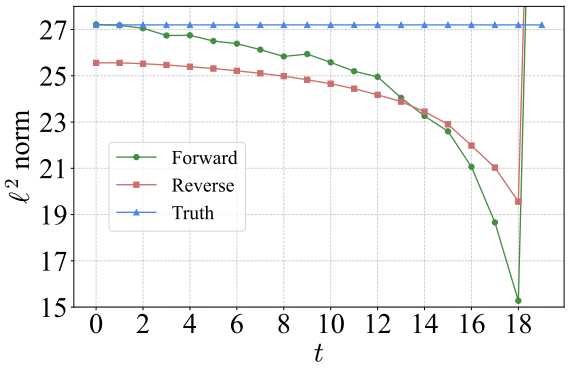


Figure 6. The expectation of $||\boldsymbol{x}_{\boldsymbol{\theta}}^0(\boldsymbol{x}_t, t)||_2^2$, $||\boldsymbol{x}_{\boldsymbol{\theta}}^0(\hat{\boldsymbol{x}}_t, t)||_2^2$, together with the ground-truth norm of $\boldsymbol{x}_0$.

This further indicates that the reconstruction operation incurs information loss. Notably, similar experiments are also reported in DPM-FR [63]. However, it focuses on the differences between the forward and reverse processes, whereas our work places greater emphasis on whether DPMs can fully reconstruct real data. Additionally, we argue that conducting experiments in the data space is more persuasive. This experiment further demonstrates that $\boldsymbol{x}_{\boldsymbol{\theta}}^0(\boldsymbol{x}_t, t)$ and $\boldsymbol{x}_{\boldsymbol{\theta}}^0(\hat{\boldsymbol{x}}_t, t)$ adhere to the form specified in Eq. 22.

C. Proofs of Theorem 5.1 and Eq. 15

In this section, we present the detailed proofs of Theorem 5.1 and Eq. 15. Our derivation process is mainly based on DPM-FR [63]. However, we provide a more rigorous derivation process, particularly for γ_t and $\hat{\gamma}_t$. Specifically, we focus on SNR, the core theme of this work.

Theorem 5.1. *For a specific timestep t in the reverse denoising process of DPMs, the SNR of the biased denoising sample $\hat{\boldsymbol{x}}_t$ is given by:*

$$\mathrm{SNR}(t) = \hat{\gamma}_t^2 \bar{\alpha}_t / \big(1 - \bar{\alpha}_t + (\frac{\sqrt{\bar{\alpha}_t}\beta_{t+1}}{1-\bar{\alpha}_{t+1}}\phi_{t+1})^2\big), \tag{30}$$

where $0 < \hat{\gamma}_t \leqslant 1$ and ϕ_{t+1} is derived from the reconstruction model $\boldsymbol{x}_{\boldsymbol{\theta}}^0(\hat{\boldsymbol{x}}_{t+1}, t+1)$ in Eq. 10.

Firstly, we emphasize that all subsequent noise terms $\boldsymbol{\epsilon}$ follow the standard Gaussian distribution. We rewrite the fundamental formula of DPMs and the forward noising process is expressed as:

$$\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_0. \tag{31}$$

We assume the current predicted sample is ideal. Thus, the reverse denoising process is expressed as:

$$\hat{\boldsymbol{x}}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)\right) + \sigma_t\boldsymbol{\epsilon}_1. \tag{32}$$

Then, substituting Eq. 25 into Eq. 32, we can obtain an equivalent form of the reverse denoising process:

$$\hat{\boldsymbol{x}}_{t-1} = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\boldsymbol{x}_{\boldsymbol{\theta}}^0(\boldsymbol{x}_t, t) + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{x}_t + \sqrt{\tilde{\beta}_t}\boldsymbol{\epsilon}_1, \tag{33}$$

By substituting Eqs. 31 and 22 into Eq. 33 to replace $\boldsymbol{x}_{\boldsymbol{\theta}}^0(\boldsymbol{x}_t, t)$ and $\boldsymbol{x}_t$, we can obtain:

$$\begin{aligned}
\hat{\boldsymbol{x}}_{t-1} &= \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}(\gamma_t\boldsymbol{x}_0 + \phi_t\boldsymbol{\epsilon}_t) + \\
&\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_0) + \sqrt{\tilde{\beta}_t}\boldsymbol{\epsilon}_1 \\
&= \left(\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t\gamma_t}{1-\bar{\alpha}_t} + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\sqrt{\bar{\alpha}_t}}{1-\bar{\alpha}_t}\right)\boldsymbol{x}_0 + \sqrt{\tilde{\beta}_t}\boldsymbol{\epsilon}_1 \\
&+ \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t\phi_t}{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\sqrt{1-\bar{\alpha}_t}}{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_0
\end{aligned} \tag{34}$$

For Eq. 34, we first focus on the coefficient of $\boldsymbol{x}_0$:

$$\begin{aligned}
&\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t\gamma_t}{1-\bar{\alpha}_t} + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\sqrt{\bar{\alpha}_t}}{1-\bar{\alpha}_t} \\
&= \frac{\sqrt{\bar{\alpha}_{t-1}}\big((1-\alpha_t)\gamma_t + \alpha_t(1-\bar{\alpha}_{t-1})\big)}{1-\bar{\alpha}_t}.
\end{aligned} \tag{35}$$

Given that $\gamma_t \leqslant 1$, we use the scaling method to amplify it to 1, yielding the following inequality:

$$\begin{aligned}
&\frac{\sqrt{\bar{\alpha}_{t-1}}\big((1-\alpha_t)\gamma_t + \alpha_t(1-\bar{\alpha}_{t-1})\big)}{1-\bar{\alpha}_t} \\
&\leqslant \frac{\sqrt{\bar{\alpha}_{t-1}}\big((1-\alpha_t) + \alpha_t(1-\bar{\alpha}_{t-1})\big)}{1-\bar{\alpha}_t} \\
&= \sqrt{\bar{\alpha}_{t-1}}
\end{aligned} \tag{36}$$

Given that $1-\alpha_t > 0, \gamma_t \leqslant 1$, We may rigorously define a novel coefficient $\hat{\gamma}_{t-1} \leqslant 1$ for $\hat{\boldsymbol{x}}_{t-1}$ where

$$\hat{\gamma}_{t-1}\sqrt{\bar{\alpha}_{t-1}} = \frac{\sqrt{\bar{\alpha}_{t-1}}\big((1-\alpha_t)\gamma_t + \alpha_t(1-\bar{\alpha}_{t-1})\big)}{1-\bar{\alpha}_t}. \tag{37}$$

For the standard Gaussian noise component in Eq. 34, based on the properties of the Gaussian distribution, we define a new coefficient $\hat{\psi}_{t-1}$ such that:

$$\begin{aligned}
\hat{\psi}_{t-1} &= (\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\phi_t)^2 + (\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\sqrt{1-\bar{\alpha}_{t-1}})^2 + \tilde{\beta}_t \\
&= (\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\phi_t)^2 + \frac{\alpha_t(1-\bar{\alpha}_{t-1})^2}{1-\bar{\alpha}_t} + \frac{(1-\bar{\alpha}_{t-1})(1-\alpha_t)}{1-\bar{\alpha}_t} \\
&= (\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\phi_t)^2 + \frac{\alpha_t(1-\bar{\alpha}_{t-1})^2 + (1-\bar{\alpha}_{t-1})(1-\alpha_t)}{1-\bar{\alpha}_t} \\
&= (\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\phi_t)^2 + 1 - \bar{\alpha}_{t-1}.
\end{aligned} \tag{38}$$

Table 8. FID and Recall (Rec) on DiT.

Model	Dataset	$T=20$		$T=50$	
		FID↓	Rec↑	FID↓	Rec↑
DiT	ImageNet 256	12.83	0.54	3.78	0.58
DiT-ES	ImageNet 256	10.00	-	3.30	-
DiT+Ours	**ImageNet** 256	**7.99**	**0.51**	**3.09**	**0.56**

Based on Eqs. 37 and (38), we can obtain

$$\hat{\boldsymbol{x}}_{t-1} = \hat{\gamma}_{t-1}\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_{t-1} + (\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\phi_t)^2}\boldsymbol{\epsilon}_{t-1} \tag{39}$$

Ultimately, based on Eq. 39, we obtain the SNR of $\boldsymbol{x}_{t-1}$ as:

$$\text{SNR}(t-1) = \hat{\gamma}_{t-1}^2\bar{\alpha}_{t-1}/\big(1-\bar{\alpha}_{t-1} + (\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\phi_t)^2\big) \tag{40}$$

By replacing the timestep in Eq. 40, we ultimately obtain the actual SNR of $\hat{\boldsymbol{x}}_t$ to complet the proof.

To obtain a more concise and intuitive form, we use the piecing-together method to derive:

$$\begin{aligned}\hat{\psi}_{t-1}^2 &= (\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\phi_t)^2 + (1-\hat{\gamma}_{t-1}^2)(1-\bar{\alpha}_{t-1}) \\ &\quad + \hat{\gamma}_{t-1}^2(1-\bar{\alpha}_{t-1})\end{aligned}$$

In conclusion, we have obtained the biased mean and variance of the reverse process:

$$\begin{aligned}&\hat{\boldsymbol{x}}_{t-1} = \hat{\gamma}_{t-1}\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0 + +\hat{\gamma}_{t-1}\sqrt{(1-\bar{\alpha}_{t-1})}\hat{\boldsymbol{\epsilon}}_3 \\ &+ \sqrt{(\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\phi_t)^2 + (1-\hat{\gamma}_{t-1}^2)(1-\bar{\alpha}_{t-1})}\tilde{\boldsymbol{\epsilon}}_3 \\ &= \hat{\gamma}_{t-1}\boldsymbol{x}_t + \psi_{t-1}\boldsymbol{\epsilon}_3,\end{aligned} \tag{41}$$

where $\psi_{t-1} = \sqrt{(\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\phi_t)^2 + (1-\hat{\gamma}_{t-1}^2)(1-\bar{\alpha}_{t-1})}$. Thus, we have completed the proof of Eq. 15. Finally, we emphasize again that γ_t is the coefficient of the reconstruction sample $\boldsymbol{x}_{\boldsymbol{\theta}}^0(\boldsymbol{x}_t, t)$ in Eq. 22, and $\hat{\gamma}_{t-1}$ is the coefficient of the predicted sample $\hat{\boldsymbol{x}}_{t-1}$ in Eqs. 39 and 41.

D. Additional Results

Given the extensive influence of transformer-based diffusion models, we select DiT [40] as the baseline model, ADM-ES [38] as the comparative model, and adopt Fréchet Inception Distance (FID) [15], Recall, and precision [15] as evaluation metrics to conduct the experiments. Tab. 8 clearly demonstrates that our method achieves a comprehensive reduction in the FID scores of DiT and outperforms the comparative models significantly.

Table 9. The search process of λ_l and λ_h on CIFAR-10 (CS) using A-DPM-DCW with 25 sampling steps.

Value	0.02	0.03	0.04	0.05	0.06	0.07	0.08
FID	7.64	7.37	7.24	7.18	7.19	7.35	7.66

E. Qualitative Comparison

To show the improvement effect of DCW on the generation quality of DPMs, we select two **state-of-the-art** text-to-image models, namely **Qwen-Image**, which demonstrates strong instruction following and text rendering ability, and **FLUX**, which is known for its high visual fidelity, to conduct extensive experiments. Given that our study focuses on the SNR-t bias, we conduct tests with a small number of steps to amplify the sampling errors of the baseline models as much as possible, thereby verifying how effectively DCW corrects such bias. As shown in Figs. 7, 8, 9, 10, 11, 12, and 13, DCW can significantly enhance the aesthetic quality.

Specifically, as shown in Figs. 7 and 10, our method consistently improves the visual quality of the generated images under a small number of sampling steps. Compared with the original models, our method produces results with more coherent scene structure, better semantic fidelity, and clearer details. It also alleviates common artifacts caused by sampling bias, leading to images that are more natural and visually appealing. These results demonstrate that DCW is effective across different baseline models and can reliably enhance generation quality in low-step sampling settings. Moreover, the improvements are consistently observed across diverse scenes and content types, further highlighting the robustness and generality of our method.

F. Parameter sensitivity

To demonstrate the insensitivity of DCW to hyperparameters λ_l and λ_h, we first apply DCW to A-DPM to obtain the optimal parameter λ_l on CIFAR-10 (CS). Then, based on the optimal parameter λ_l, we apply DCW to obtain the optimal parameter λ_h. Fig. 4 clearly shows that DCW can achieve performance gains over a wide range of λ_l and λ_h, indicating the insensitivity of DCW to hyperparameters. Due to the robustness of the method to parameters, the parameter search process is fast via the two-stage search. Firstly, a coarse search with a step size of 0.01 was performed. After identifying a turning point in the FID curve around 0.05, we conducted a fine-grained search with a step size of 0.001 and quickly determined the optimal value to be 0.052, as shown in Tab 9. Then, after fixing the optimal λ_l^* at 0.052, quickly derive the optimal parameter $\lambda_h^* = 0.010$ using the same method. In summary, the above experimental process further demonstrates the robustness and practicality of our method with respect to hyperparameters.

Figure 7. Qualitative comparison between **Qwen-Image** (first row) and **Qwen-Image-DCW** (second row) using **10 steps**, where the prompt is “A woman is walking on the beach by the sea”.

Figure 8. Qualitative comparison between **Qwen-Image** (first row) and **Qwen-Image-DCW** (second row) using **10 steps**, where the prompt is “There is a house and a path on a snowy mountain”.

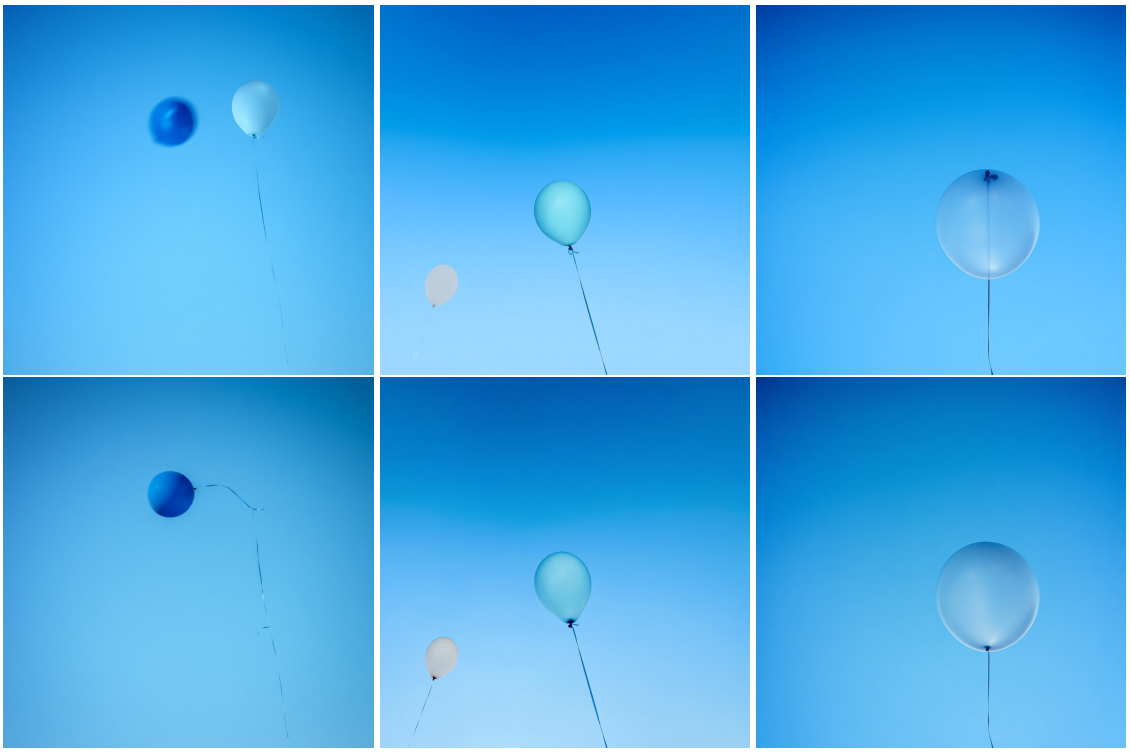


Figure 9. Qualitative comparison between **Qwen-Image** (first row) and **Qwen-Image-DCW** (second row) using **10 steps**, where the prompt is "**A balloon** gently climbs into a serene blue sky".

Figure 10. Qualitative comparison between **FLUX** (first row) and **FLUX-DCW** (second row) using **10 steps**, where the prompt is "There is a house and a path on a snowy mountain".

Figure 11. Qualitative comparison between **FLUX** (first row) and **FLUX-DCW** (second row) using **10 steps**, where the prompt is “A woman is walking on the beach by the sea”.

Figure 12. Qualitative comparison between **FLUX** (first row) and **FLUX-DCW** (second row) using **10 steps**, where the prompt is “A balloon gently climbs into a serene blue sky”.

Figure 13. Qualitative comparison between **Qwen-Image** (first row) and **Qwen-Image-DCW** (second row) using **20 steps**, where the prompt is "A woman is walking on the beach by the sea".

Figure 14. Qualitative comparison between **FLUX** (first row) and **FLUX-DCW** (second row) using **20 steps**, where the prompt is "There is a house and a path on a snowy mountain".