Self-Guidance: Boosting Flow and Diffusion Generation on Their Own

Tiancheng Li, Weijian Luo, Zhiyang Chen, Liyuan Ma, and Guo-Jun Qi, *Fellow, IEEE*

***Abstract*—Proper guidance strategies are essential to achieve high-quality generation results without retraining diffusion and flow-based text-to-image models. Existing guidance either requires specific training or strong inductive biases of diffusion model networks, which potentially limits their ability and application scope. Motivated by the observation that artifact outliers can be detected by a significant decline in the density from a noisier to a cleaner noise level, we propose Self-Guidance (SG), which can significantly improve the quality of the generated image by suppressing the generation of low-quality samples. The biggest difference from existing guidance is that SG only relies on the sampling score function of the original diffusion or flow model at different noise levels, with no need for any tricky and expensive guidance-specific training. This makes SG highly flexible to be used in a plug-and-play manner by any diffusion or flow models. We also introduce an efficient variant of SG, named SG-prev, which reuses the output from the immediately previous diffusion step to avoid additional forward passes of the diffusion network. We conduct extensive experiments on text-to-image and text-to-video generation with different architectures, including UNet and transformer models. With open-sourced diffusion models such as Stable Diffusion 3.5 and FLUX, SG exceeds existing algorithms on multiple metrics, including both FID and Human Preference Score. SG-prev also achieves strong results over both the baseline and the SG, with 50 percent more efficiency. Moreover, we find that SG and SG-prev both have a surprisingly positive effect on the generation of physiologically correct human body structures such as hands, faces, and arms, showing their ability to eliminate human body artifacts with minimal efforts.**

***Index Terms*—Diffusion models, flow-based generative models, self-guidance.**

Received 5 July 2025; revised 30 August 2025; accepted 9 September 2025. Date of publication 18 September 2025; date of current version 3 December 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 9246710 and in part by Zhejiang Leading Innovative and Entrepreneur Team Introduction Program under Grant 2024R01007. Recommended for acceptance by W.- H. Cheng. *(Corresponding authors: Weijian Luo; Guo-Jun Qi.)*

Tiancheng Li is with Zhejiang University, Hangzhou 310058, China, and also with MAPLE Lab, Westlake University, Hangzhou 310030, China (e-mail: litiancheng@westlake.edu.cn).

Weijian Luo is with MAPLE Lab, Westlake University, Hangzhou 310030, China, and also with Peking University, Beijing 100871, China (e-mail: luoweijian@stu.pku.edu.cn).

Zhiyang Chen, Liyuan Ma, and Guo-Jun Qi are with MAPLE Lab, Westlake University, Hangzhou 310030, China, and also with the Institute of Advanced Technology, Westlake Institute for Advanced Study, Zhejiang 310030, China (e-mail: chenzhiyang@westlake.edu.cn; maliyuan@westlake.edu.cn; guojunq@gmail.com).

We have released our code at https://github.com/maple-research-lab/Self-Guidance.

This article has supplementary downloadable material available at https://doi.org/10.1109/TPAMI.2025.3611831, provided by the authors.

Digital Object Identifier 10.1109/TPAMI.2025.3611831

I. Introduction

OVER the past decade, deep generative models have achieved remarkable advancements across various applications [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17]. Among them, diffusion models [18], [19], [20] and flow-based models [21], [22], [23] have notably excelled in producing high-resolution, text-driven data such as images [24], [25], [26], videos [5], [27], [28], and others [4], [29], [30], [31], [32], [33], [34], [35], pushing the boundaries of Artificial Intelligence Generated Contents.

In simple terms, diffusion models learn a multi-step transition from a prior distribution $p_T(\boldsymbol{x}_T)$ to a real data distribution $p_0(\boldsymbol{x}_0)$. However, default sampling methods for diffusion and flow-based models often lead to unsatisfactory generation quality, such as broken human hands and faces, and images with bad foreground and background. To address these issues, various guidance strategies have emerged as cheap yet effective ways to guide the generation process for better generation quality. For instance, classifier-free guidance (CFG) [36] modifies the velocity of diffusion and flow-based generative models by adding a delta term between class-conditional and unconditional velocities, which pushes generated samples to have high class probabilities.

Though these existing guidance have shown impressive performance improvements, they have various individual restrictions. For instance, the CFG relies on computing an additional unconditional velocity, which requires training the diffusion model under both conditional and unconditional settings, therefore, harms the modeling performances [37]. Auto-Guidance (AG) pays a significant price that requires training an additional *bad-version* model, which is tricky as well as requiring more memory costs. Other guidance, such as Perturbed-attention Guidance (PAG [37]), and self-attention guidance (SAG [38]), do not rely on additional training. However, as the PAG paper described, the effectiveness of PAG is highly sensitive to the selection of perturbed attention layers inside the neural network, making it less flexible to enhance models in general applications. We notice that *all these guidance variants focus on diffusion sampling strategies at a single timestep, while neglecting how diffusion sampling at various timesteps could be explored to improve the generation quality*. Furthermore, we find that without explicitly retraining or perturbing an existing diffusion model as in CFG and PAG, using the diffusion samplings from different timesteps yields a Self-Guidance (SG) approach.

Specifically, for a diffusion model, we find that the reason for generating unnatural images is due to the inadequate denoising

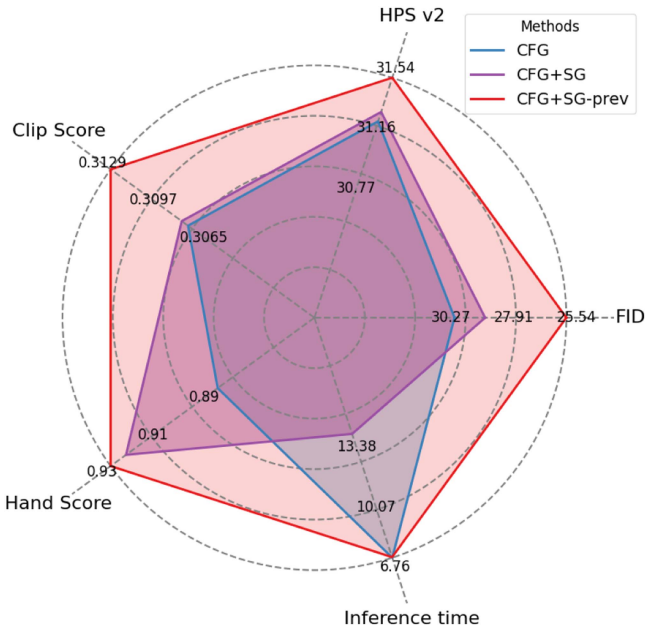


Fig. 1. Radar chart of 5 image benchmarks on Flux.1. Background numbers indicate reference values corresponding to normalized radii (0.5, 0.75, 1.0).

that could yield artifacts in the reverse diffusion process. As Fig. 2 shows, let us consider a simple diffusion model trained on a 1-dimensional toy dataset with two separate modes at ± 1 as the ground truth. At the noise level $t = 0.38$ for example, the probability $p_t(x_t)$ (i.e., the blue curve) is likely to draw artifact samples around the origin as it has a large density at $x_t = 0$.

To relieve it, an intuitive idea is to take a look at the diffused probability $p_{t+\delta(t)}(x_t)$ at a noisier level $t + \delta(t)$ (i.e., the green curve). We know the samples at this level could be noisier and thus contain more severe artifacts. Therefore, by comparing between these two levels, one can assume that if the density at the cleaner level t (although it is still high) declines significantly from that at the noisier level $t + \delta(t)$, the corresponding samples are likely to be artifacts, since the belief of sampling them as cleaner outputs has greatly dropped as reflected by the declining density. This is what we have seen in Fig. 2(a) and (b), where the current sampling probability $p_t(x_t)$ of the blue curve at $t = 0.38$ declines from that of the green curve at the noisier level $t + 0.2$ at the origin. This reveals that the samples around the origin could be artifacts, which is true. This decline can be measured by the ratio of these two probabilities, resulting in a new sampling strategy

$$p_t^{SG}(x_t) \propto p_t(x_t) \left\{ \frac{p_t(x_t)}{p_{t+\delta(t)}(x_t)} \right\}^{\omega}, \tag{1}$$

to guide the reverse diffusion process. Here, ω is the guidance scale. The higher the value of ω, the more sharply the artifact samples will be suppressed as shown in Fig. 2(b).

In this paper, we present such an inference-time sampling strategy called *Self-Guidance (SG)* since it only involves its own diffusion model at different noise levels. For image generation tasks, as shown in Fig. 3, compared with the benchmark diffusion sampling algorithm, applying SG successfully removes unwanted artifacts in generated images, fixing errors on human fingers and other generation errors, eliminating irrelevant objects, and improving text-image consistency.

This guidance can further be approximated using the output from the immediately previous diffusion step – i.e., $p_{t+1}(x_{t+1})$. We denote this approximation as *SG-prev*. SG-prev only needs one forward pass, thereby incurring no additional inference cost and keeping the overall inference time almost unchanged.

In addition, SG is architecture-independent and can therefore be integrated into almost all existing diffusion and flow-based models regardless of varying model architectures. Unlike CFG [36] and AG [39], SG does not require additional guidance-specific training. SG is orthogonal to many other guidance approaches [36], [37], [39], seamlessly works together with them in a plug-and-play manner for better performances.

In Section V-B, we apply SG and SG-prev on Flux.1-dev [23] and Stable-Diffusion 3.5 [21], [40], as well as a representative text-to-video model, CogVideoX [28]. We show that when using SG only, all models achieved improved FID [41] and higher human preference scores on the HPS v2.1 [42] benchmark. Combining SG and SG-prev with other established guidance methods, including CFG and PAG, leading to new state-of-the-art diffusion generation results, achieving 20.74 on FID, 31.56 on HPSv2, and 5.7845 on Aesthetic Score. Experiments on text-to-video generation also confirm the solid performance improvements brought by SG.

The main contributions of this work are as follows.

- We observe that the ratio of probabilities at different noise levels serves as a good indicator of artifacts during the reverse diffusion process.
- Leveraging this insight, we propose *Self-Guidance*—a plug-and-play guidance method that improves generation quality without requiring extra training, supervision, or external models. Moreover, such guidance can be implemented without incurring additional inference cost, using a variant we refer to as SG-prev.
- We show that SG and SG-prev are compatible with other sampling methods like CFG and PAG, leading to state-of-the-art image generation performances.

II. Related Works

There are roughly two kinds of diffusion guidance. The first kind mixes the outputs of multiple models to get the generative velocity. The classifier-free guidance (CFG) [36] uses a mix of conditional and unconditional score functions as the velocity. Most recently, the Auto-guidance (AG) [39] proposed to mix the current score output with a *bad-version* model to enhance persistence of velocity, resulting in improved performances. Both CFG and AG use two models for inference explicitly or implicitly. Some other studies have also elaborately studied guidance through the lens of diffusion solvers [43], [44], [45].

The second line of guidance defines the velocity by mixing the output of the score functions of a perturbed and the original sample. These methods usually need one model for inference. The self-attention guidance (SAG) [38] proposed to assign Gaussian perturbations to samples for guidance. Recently,

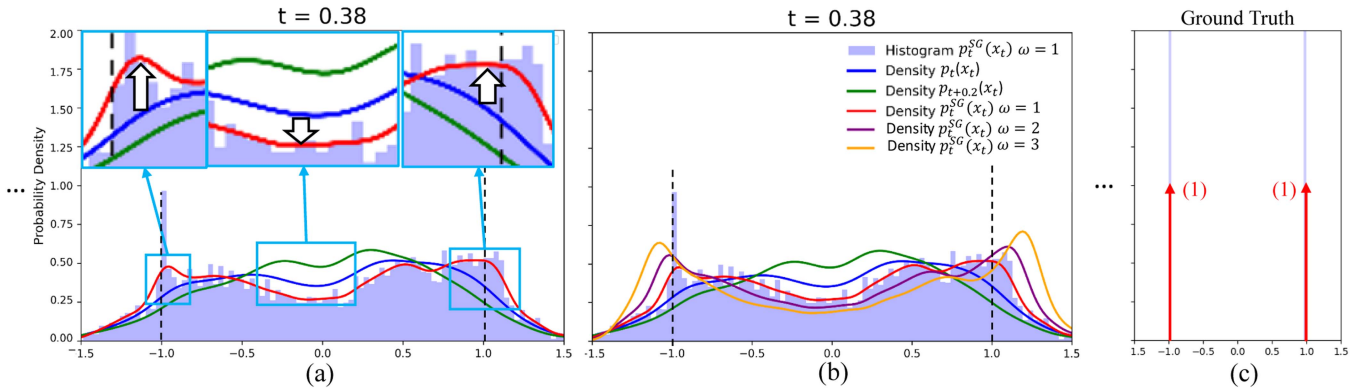


Fig. 2. We train a diffusion model on a 1-dimensional toy example with data drawn from two separate modes at ± 1. At different noise levels t, we fit and plot the distribution of generated samples in the reverse diffusion process. (a) The results on diffusion time $t = 0.38$. The blue curve plots the distribution of generated samples at the current noise level t, while the green curve plots the distribution of samples generated at the noisier level $t + \delta(t)$. The red curve shows the obtained SG sampling distribution after applying the ratio of these two probabilities with $\omega = 1$. (b) With various ω, we show that artifact samples around the origin are suppressed more sharply with a larger value of ω, while the density of samples around two groundtruth modes is boosted. (c) The ground truth two-mode distribution at ± 1.

Fig. 3. Qualitative comparisons between Flux.1 (baseline) and Self-Guidance (SG) diffusion samples. We compare the generated images of Flux and our SG from four parts. The red box in the figure represents the bad Flux generation and the better SG generation, and the enlarged image is shown in the lower right corner.

Perturbed-attention guidance (PAG) [37] found that modifying the self-attention map of the diffusion model network for guidance results in strong performances. PAG-like guidance has shown steady improvements. Another recent work [46] proposed the TimeStep Guidance (TSG), which perturbed the timestep embedding of diffusion models, sharing a similar spirit as PAG. For perturbation-based guidance, how to choose proper perturbations is mostly decided in an ad-hoc manner, which may vary significantly across different neural network architectures. Other works also study perturbation-based guidance paradigms from different perspectives [47], [48], [49].

Despite improvements in overall image structure and class-label alignment, existing guidelines have a common problem: they still generate images with artifacts, like six-finger hands or twisted human bodies, which brings significant concerns in AI safety and user experiences. As a comparison, our proposed Self-Guidance significantly improves fine-grained details of generated images, completing weaknesses of existing guidance in an orthogonal way. We will introduce details of SG in Section IV.

III. Preliminary

a) Diffusion Models: In this section, we introduce preliminary knowledge and notations about diffusion models [18], [50], [51]. The flow models [52], [53] share similar concepts that we put in Appendix B. Assume we observe data from the underlying distribution $q_d(\boldsymbol{x})$. The goal of generative modeling is to train models to generate new samples $\boldsymbol{x} \sim q_d(\boldsymbol{x})$. The forward diffusion process of DM transforms $q_0 = q_d$ towards some simple noise distribution,

$$\mathrm{d}\boldsymbol{x}_t = \boldsymbol{F}(\boldsymbol{x}_t, t)\mathrm{d}t + G(t)\mathrm{d}\boldsymbol{w}_t, \tag{2}$$

where $\boldsymbol{F}$ is a pre-defined drift function, $G(t)$ is a pre-defined scalar-value diffusion coefficient, and $\boldsymbol{w}_t$ denotes an independent Wiener process. A continuous-indexed score network $\boldsymbol{s}_\varphi(\boldsymbol{x}, t)$ is employed to approximate marginal score functions of the forward diffusion process (2). The learning of score networks is achieved by minimizing a weighted denoising score matching objective [50], [54],

$$\mathcal{L}_{DSM}(\varphi) = \int_{t=0}^{T} \lambda(t)\mathbb{E}_{\boldsymbol{x}_0\sim q_0, \boldsymbol{x}_t|\boldsymbol{x}_0\sim q_t(\boldsymbol{x}_t|\boldsymbol{x}_0)} \|\boldsymbol{s}_\varphi(\boldsymbol{x}_t, t) - \nabla_{\boldsymbol{x}_t}\log q_t(\boldsymbol{x}_t|\boldsymbol{x}_0)\|_2^2 \mathrm{d}t. \tag{3}$$

Here, the weighting function $\lambda(t)$ controls the importance of the learning at different time levels and $q_t(\boldsymbol{x}_t|\boldsymbol{x}_0)$ denotes the conditional transition of the forward diffusion (2). After training, the score network $\boldsymbol{s}_\varphi(\boldsymbol{x}_t, t) \approx \nabla_{\boldsymbol{x}_t}\log q_t(\boldsymbol{x}_t)$ is a good approximation of the marginal score function of the diffused data distribution. High-quality samples from a DM can be drawn by simulating generative SDE (4) which is implemented by replacing the score function $\nabla_{\boldsymbol{x}_t}\log q_t(\boldsymbol{x}_t)$ with the learned score network [50].

$$\mathrm{d}\boldsymbol{x}_t = \left\{\boldsymbol{F}(\boldsymbol{x}_t, t) - \frac{1+\tau(t)^2}{2}G^2(t)\boldsymbol{s}_\varphi(\boldsymbol{x}_t, t)\right\}\mathrm{d}t + \tau(t)G(t)\mathrm{d}\bar{\boldsymbol{w}}_t,\ t \in [0, T],\ \boldsymbol{x}_T \sim p_T. \tag{4}$$

b) Classifier-free Guidance: Though the diffusion model has a solid theoretical interpretation, directly using its score functions to simulate the generative SDE often leads to sub-optimal performances, especially for class-conditioned generation. Assume $\boldsymbol{c}$ is a class label, and $\boldsymbol{s}_\varphi(\boldsymbol{x}_t, t|\boldsymbol{c})$ is a conditional score network. The pioneering work [36] introduces the classifier-free guidance (CFG), which replaces the vanilla probability $q_t(\boldsymbol{x}_t|\boldsymbol{c})$ with a new one $\tilde{q}_t(\boldsymbol{x}_t|\boldsymbol{c}) := q_t(\boldsymbol{x}_t)(\frac{q_t(\boldsymbol{x}_t|\boldsymbol{c})}{q_t(\boldsymbol{x}_t)})^\omega$. $q_t(\boldsymbol{x}_t)$ represents the unconditional distribution which can be implemented by inputting an empty label $\varnothing$ as $q_t(\boldsymbol{x}_t|\varnothing)$. With this, the new score function turns to

$$\tilde{\boldsymbol{s}}_\varphi(\boldsymbol{x}_t, t|\boldsymbol{c}) := \boldsymbol{s}_\varphi(\boldsymbol{x}_t, t|\varnothing) + \omega\left\{\boldsymbol{s}_\varphi(\boldsymbol{x}_t, t|\boldsymbol{c}) - \boldsymbol{s}_\varphi(\boldsymbol{x}_t, t|\varnothing)\right\} \tag{5}$$

Such a guidance strategy has become a default setting for diffusion models, such as the Stable Diffusion series [40]. However, the CFG strategy has its limitations. First, CFG requires both a conditional and an unconditional score network, which either requires two separate models or is challenging to train with a single model. Second, the CFG is not available for tasks such as purely unconditional generation.

IV. Self-Guidance

Fig. 2 demonstrates that using the ratio $\frac{p_t(\boldsymbol{x}_t)}{p_{t+\delta(t)}(\boldsymbol{x}_t)}$ helps suppress artifacts arising from the original diffused probability $p_t(\boldsymbol{x}_t)$ at each noise level t.

Formally, when generating the image with a condition c, the self-guided generative distribution at noise level t can be formulated as discussed in Section I,

$$p_t^{SG}(\boldsymbol{x}_t|\boldsymbol{c}, t) \propto p_t(\boldsymbol{x}_t|\boldsymbol{c})\left\{\frac{p_t(\boldsymbol{x}_t|\boldsymbol{c})}{p_{t+\delta(t)}(\boldsymbol{x}_t|\boldsymbol{c})}\right\}^\omega. \tag{6}$$

This probability contains two parts. One is the original probability $p_t(x_t|c)$ that contains both the desired and artifact samples. The other is the ratio between probabilities of two diffusion times, and it is used to suppress the artifacts. The guidance scale ω acts as the combination weight.

Fig. 2 has already illustrated the influence of ω on the Self-Guidance sampling. Here we provide a more complex 2-dimensional example in Fig. 6 to show the effects of Self-Guidance with various values of ω. The ground truth data are sampled from a double-swirl pattern. By comparing the diffusion samplings with and without Self-Guidance, it demonstrates the ability of SG to remove the artifact outliers that do not reside on the swirls. With an increasing value of ω, the outlier artifacts are suppressed more sharply, which eventually leads to the high-quality sampling of the ground truth distribution. However, it is worth noting that when ω becomes too large (e.g., $\omega = 7$), the distribution of generated samples could start drifting away from the swirls. This is not surprising because too large ω could underweight the original density $p_t(\boldsymbol{x}_t|c)$ that aims to cover the samplings of the swirl data. This shows that a suitable guidance scale ω is necessary to balance between the sampling of true data density and the suppression of artifact outliers. In experiments, we will study the effect of its choices in the ablation.

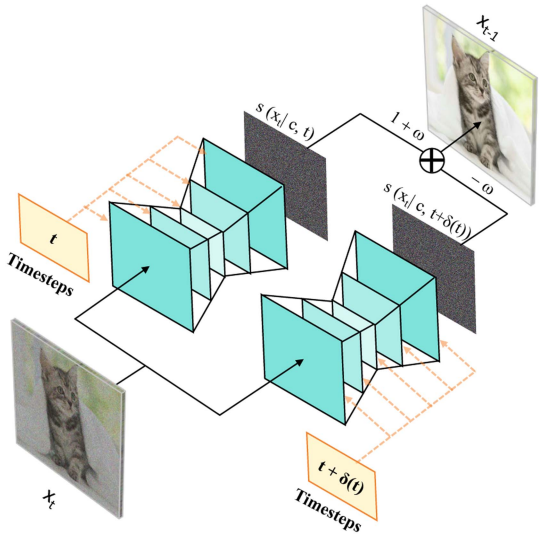


Fig. 4. An illustration of Self-Guidance of one iteration step during generation.

Since diffusion models output the score instead of the diffused probability, one can transform (7) into the score function by the relation of $\boldsymbol{s}(\boldsymbol{x}_t|\boldsymbol{c},t) = \nabla_{x_t} \log p(\boldsymbol{x}_t|c)$. Formally, we start with the log-density ratio between two nearby timesteps:

$$\log \frac{p_t(x_t|c)}{p_{t+\delta(t)}(x_t|c)} = \log p_t(x_t|c) - \log p_{t+\delta(t)}(x_t|c), \quad (7)$$

and take the gradient to x to obtain the score difference:

$$\begin{aligned}\nabla_x \log \frac{p_t(x_t|c)}{p_{t+\delta(t)}(x_t|c)} &= \nabla_x \log p_t(x_t|c) - \nabla_x \log p_{t+\delta(t)}(x_t|c) \\ &= \boldsymbol{s}(x_t|c,t) - \boldsymbol{s}(x_t|c,t+\delta(t)). \quad (8)\end{aligned}$$

Finally, the score with scale ω can be written as:

$$\begin{aligned}\boldsymbol{s}^{SG}(\boldsymbol{x}_t|\boldsymbol{c},t) := \; & \boldsymbol{s}(\boldsymbol{x}_t|\boldsymbol{c},t) \\ & + \omega\left\{\boldsymbol{s}(\boldsymbol{x}_t|\boldsymbol{c},t) - \boldsymbol{s}(\boldsymbol{x}_t|\boldsymbol{c},t+\delta(t))\right\}, \quad (9)\end{aligned}$$

Fig. 4 illustrates how Self-Guidance is applied in practice at each denoising step during the reverse diffusion process. For a noisy latent $\boldsymbol{x}_t$ at timestep t, the model simultaneously predicts the noise at t and at a noisier timestep $t+\delta(t)$. These two predictions are then linearly combined as in (9), and the combined prediction is subsequently used to denoise $\boldsymbol{x}_t$ and obtain the latent $\boldsymbol{x}_{t-1}$ for the next denoising step. This can be easily implemented by inference with the same model with two different timesteps within a batch. This means SG does not require an additional network or specific training tricks as in other guidance algorithms, but only uses the trained model itself. That is why we called the method Self-Guidance.

There are various choices of the shifted noise scale $\delta(t)$. In this paper, we consider a constant shift scale that is fixed independent of diffusion time t, and a dynamic one $\delta(t) = t/\sigma$ that becomes smaller as t approaches 0. The latter allows a stronger Self-Guidance at the beginning of the diffusion sampling process. Moreover, we can also simply reuse the output

Algorithm 1: Diffusion Model Inference With Self-Guidance and Other Guidance.

Input: Conditional DM $d_{cod(\theta)}$, Unconditional DM $d_{uncod(\theta)}$, Perturbed DM $\widehat{d}_{cod(\theta)}$, guidance scales $\omega_{CFG}, \omega_{PAG}, \omega_{SG}$, shift scales $\delta(t)$.

for t *in* `timesteps` **do**
- $\boldsymbol{s}(\boldsymbol{x}_t|\boldsymbol{c},t) = d_{cod(\theta)}(\boldsymbol{x}_t,\boldsymbol{c},t)$
- **if** *Classifier-free guidance* **then**
  - $\boldsymbol{s}(\boldsymbol{x}_t|t) = d_{uncod(\theta)}(\boldsymbol{x}_t,t)$
- **end**
- **if** *Perturbed-attention guidance* **then**
  - $\widehat{\boldsymbol{s}}(\boldsymbol{x}_t|\boldsymbol{c},t) = \widehat{d}_{cod(\theta)}(\boldsymbol{x}_t,\boldsymbol{c},t)$
- **end**
- **if** *Self-Guidance* **then**
  - $\boldsymbol{s}(\boldsymbol{x}_t|\boldsymbol{c},t+\delta(t))$
    - $= d_{\mathrm{cod}(\theta)}(\boldsymbol{x}_t,\boldsymbol{c},t+\delta(t))$ if SG
    - $\approx d_{\mathrm{cod}(\theta)}(\boldsymbol{x}_{t+1},\boldsymbol{c},t+1)$ if SG-prev
- **end**
- $\boldsymbol{s}^*(\boldsymbol{x}_t|\boldsymbol{c},t) = \boldsymbol{s}(\boldsymbol{x}_t|t)$
  - $+\omega_{CFG} * (\boldsymbol{s}(\boldsymbol{x}_t|\boldsymbol{c},t) - \boldsymbol{s}(\boldsymbol{x}_t|t))$
  - $+\omega_{PAG} * (\boldsymbol{s}(\boldsymbol{x}_t|\boldsymbol{c},t) - \widehat{\boldsymbol{s}}(\boldsymbol{x}_t|\boldsymbol{c},t))$
  - $+\omega_{SG} * (\boldsymbol{s}(\boldsymbol{x}_t|\boldsymbol{c},t) - \boldsymbol{s}(\boldsymbol{x}_t|\boldsymbol{c},t+\delta(t)))$
- $\boldsymbol{x}_{t-1}$ = `scheduler.step`$(\boldsymbol{s}^*(\boldsymbol{x}_t|\boldsymbol{c},t),t,\boldsymbol{x}_t)$

end

return $\boldsymbol{x}_0$

distribution $\boldsymbol{s}(\boldsymbol{x}_{t+1}|\boldsymbol{c},t+1)$ from the previous diffusion step to approximate the SG guidance term, as in (10)

$$\begin{aligned}\boldsymbol{s}^{SG\_prev}(\boldsymbol{x}_t|\boldsymbol{c},t) := \; & \boldsymbol{s}(\boldsymbol{x}_t|\boldsymbol{c},t) \\ & + \omega\left\{\boldsymbol{s}(\boldsymbol{x}_t|\boldsymbol{c},t) - \boldsymbol{s}(\boldsymbol{x}_{t+1}|\boldsymbol{c},t+1)\right\}. \quad (10)\end{aligned}$$

This avoids doubling inference time as it does not require an additional forward pass at each diffusion step. We call this approximation method SG-prev.

In Fig. 5, we present a comparison between SG and SG-prev on the same one-dimensional toy example as in Fig. 2. We find that the so-called SG-prev method is more efficient while bringing additional positive effects.

During image denoising, the early steps are mostly noise, and the image gradually becomes clear in the final steps. This is reflected by the probability density $p_t(x_t)$ in Fig. 5: at high-noise levels, the distribution is nearly Gaussian, so changes between steps are small; at low-noise levels, the distribution varies significantly, making the previous diffusion step probability $p_{t+1}(x_{t+1})$ (gray curve) near $x = 0$ higher than $p_{t+\delta(t)}(x_t)$ at the slightly noisier level $t+\delta(t)$ (green curve), which leads to stronger suppression of artifacts.

We detail how SG and SG-prev are used in inference in Algorithm 1, as well as if combined with other diffusion guidance methods, such as CFG and PAG. For a single step at the noise level t, the diffusion model applies different guidance individually by forwarding the model with the corresponding settings: unconditional prediction in CFG, perturbing the attention in PAG, a shifted timestep $t+\delta(t)$ in SG, and the previous timesteps' output in SG-prev. These correct terms provide orthogonal improvements. Combined with their corresponding

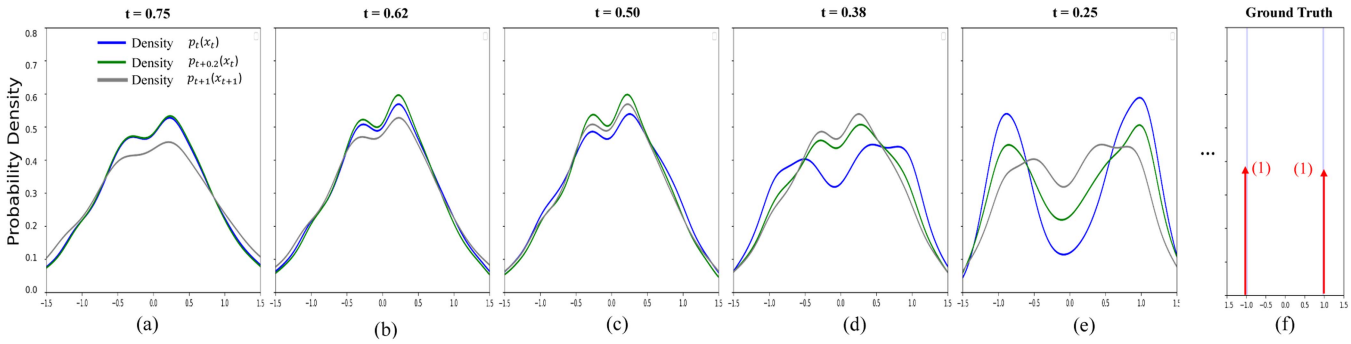


Fig. 5. Comparison of SG and SG-prev under the two-mode distribution. (a) - (e) are the results on different diffusion time from $t = 0.75$ to $t = 0.25$. The blue curve plots the distribution of generated samples at the current noise level t, the green curve plots the distribution of samples generated at the noisier level $t + \delta(t)$, while the gray curve plots the distribution of samples immediately generated at the immediately previous diffusion step. (f) The ground truth two-mode distribution at ± 1.

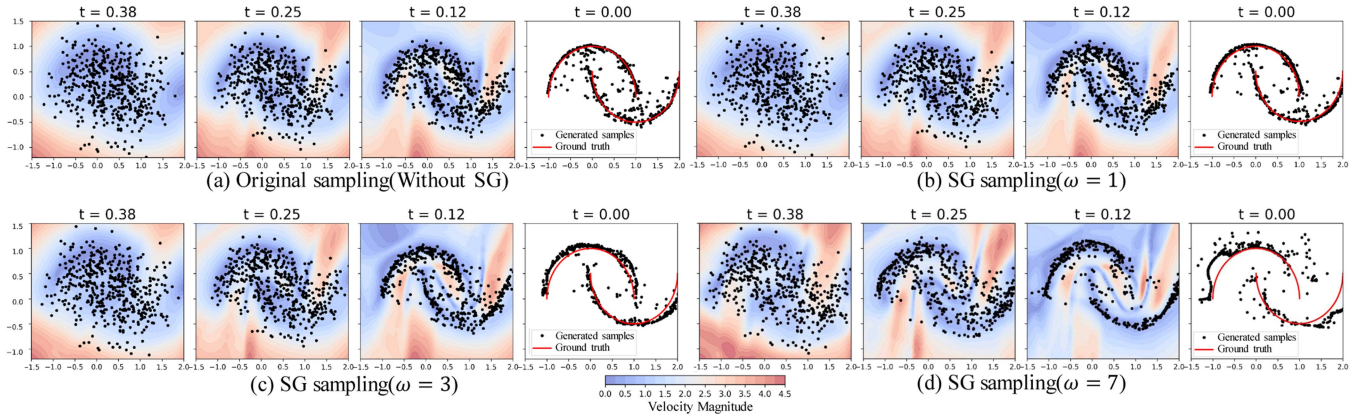


Fig. 6. We train a flow-based diffusion model on a two-dimensional example with data sampled from a double-swirl pattern. We plot the distribution of the generated samples at the last four noise levels. The heat map in (a) shows the magnitude of predicted velocity without self-guidance, while the subfigures (b), (c), and (d) show those with Self-Guidance with various ω and noise levels. With a larger ω, more artifact outliers are successfully removed. Meanwhile, a too large value of ω (e.g., $\omega = 7$) could underweight (or, in other words, oversuppress) the sampling from the original density $p_t(\boldsymbol{x}_t|c)$ that covers the swirl distribution. Thus, a suitable value of the guidance scale is necessary to balance the sampling coverage and the artifact suppression.

weight factors $\omega_{CFG}, \omega_{PAG}, \omega_{SG}$, the overall prediction is optimized to achieve higher consistency and control in the quality of the generation.

a) Some discussions about the motivation of SG: Besides the empirical motivation, Self-guidance also shares solid theoretical backgrounds. In this part, we show that suppressing artifacts with $\frac{p_t(\boldsymbol{x}_t|\boldsymbol{c})}{p_{t+\delta(t)}(\boldsymbol{x}_t|\boldsymbol{c})}$ can be motivated from the heat equation. Specifically, under Gaussian smoothing with kernel $\mathcal{N}(0, \sigma^2)$, the diffused density evolves as:

$$p_{t+\delta(t)}(x) = \int p_t(y)\,\mathcal{N}(x; y, \sigma^2\delta(t))\,dy$$
$$= (p_t * \mathcal{N}(0, \sigma^2))(x). \tag{11}$$

Differentiating with respect to t yields the classical heat equation:

$$\partial_t p_t(x) = \frac{1}{2} g(t)^2 \Delta_x p_t(x), \tag{12}$$

where $g(t)$ denotes the noise scaling function and Δ_x is the Laplacian operator. Now, assume that an artifact resides at a location x_{artifact} in a low-density region of $p_t(x)$, i.e., a "valley." Under the multivariate second-derivative test, the Hessian at this point is positive definite, implying

$$\Delta_x p_t(x_{\text{artifact}}) > 0. \tag{13}$$

Thus, the heat equation implies:

$$\partial_t p_t(x_{\text{artifact}}) > 0 \quad \Rightarrow \quad p_{t+\delta(t)}(x_{\text{artifact}}) > p_t(x_{\text{artifact}}) \tag{14}$$

when $\delta(t)$ is sufficiently small.

In other words, forward diffusion causes density to flow into low-density artifact regions, increasing their mass. Therefore, comparing p_t to $p_{t+\delta(t)}$ (via their ratio) effectively downweights these artifact regions and promotes sampling from the true data manifold.

V. Experiments

A. Implementation Details

a) Models: We apply our Self-Guidance on current state-of-the-art text-to-image and text-to-video models. For text-to-image, we consider both a classic diffusion model, Stable Diffusion 1.4 [40], and current state-of-the-art flow-based models, Stable-Diffusion 3 series [21], [22] and Flux.1-dev [23]. For text-to-video, we consider CogVideoX-5B [28], an open-source video generation model trained with flow matching.

TABLE I
QUANTITATIVE RESULTS OF DIFFERENT STABLE DIFFUSION MODELS AND FLUX WITH VARIOUS GUIDANCE. FID, CLIP SCORE, AND AESTHETIC SCORE ARE EVALUATED ON THE MS-COCO 2017 VALIDATION SUBSET. BOLDED VALUES HIGHLIGHT THE BEST PERFORMANCE.

Model	Approch	FID↓	HPS v2.1					Clip Score↑	Aesthetic Score↑
			concept-art↑	photo↑	anime↑	paintings↑	average↑		
SD 1.4[51]	baseline	50.76	21.22	21.24	21.37	20.52	21.09	0.2695	5.0025
	SG	42.73	22.13	22.23	21.84	21.73	21.98	0.2781	5.0766
	CFG[11]	27.07	23.61	24.91	23.60	24.83	24.24	0.3095	5.4427
	CFG+PAG[1]	27.37	23.61	24.87	24.86	23.73	24.26	0.3093	**5.4448**
	CFG+SG	26.67	**24.95**	25.17	24.13	24.12	24.59	0.3094	5.4201
	CFG+SG-prev	23.70	24.90	23.73	23.87	**25.10**	24.39	0.3110	5.2954
	CFG+PAG+SG	26.75	24.12	**25.28**	**24.99**	24.20	**24.65**	0.3094	5.4351
	CFG+PAG+SG-prev	**23.58**	23.94	25.17	24.91	23.75	24.44	**0.3115**	5.3444
SD 3[8]	baseline	51.47	21.17	18.91	21.11	21.62	20.70	0.2762	4.9628
	SG	48.82	22.01	19.49	20.79	21.47	20.83	0.2964	4.8888
	CFG[11]	27.00	30.26	28.29	30.81	30.62	30.00	0.3169	5.3915
	CFG+PAG[1]	27.34	30.26	28.45	30.78	30.73	30.05	0.3164	5.3978
	CFG+SG	25.21	**30.67**	28.49	**31.10**	**31.03**	30.32	0.3163	**5.4431**
	CFG+SG-prev	25.15	29.94	27.57	30.33	30.17	29.50	0.3217	5.3260
	CFG+PAG+SG	25.28	30.65	**28.72**	31.09	31.02	**30.37**	0.3162	5.4424
	CFG+PAG+SG-prev	**24.83**	29.92	27.79	30.34	30.26	29.58	**0.3219**	5.3238
SD 3.5[2]	baseline	39.69	22.06	22.98	23.87	22.03	22.74	0.3020	5.3068
	SG	38.40	23.74	24.74	25.08	23.39	24.24	0.3018	5.3986
	CFG[11]	23.86	31.67	29.48	**32.65**	31.43	31.31	0.3129	5.6152
	CFG+SG	22.70	**31.74**	**29.50**	32.57	**31.52**	**31.33**	0.3132	**5.6532**
	CFG+SG-prev	**20.74**	31.41	28.93	32.07	31.24	30.91	**0.3162**	5.6013
Flux.1[29]	CFG[11]	29.74	31.26	29.87	32.80	31.78	31.43	0.3080	5.7527
	CFG+SG	28.59	31.32	**30.02**	32.89	31.81	31.51	0.3084	**5.7845**
	CFG+SG-prev	**25.54**	**31.54**	29.71	**32.92**	**32.06**	**31.56**	**0.3129**	5.7391

To facilitate a fair comparison, we consistently evaluate the model with the same number of steps in inference: DDIM Solver with 50 steps for SD 1.4, Euler with 28 steps for SD 3 series, and FLUX. The guidance scales for CFG and PAG are kept the same as the original settings in the corresponding papers. We will elaborate in Appendix C3, available online.

b) Evaluation Metrics: We employ Fréchet Inception Distance (FID) [41], Human Preference Score (HPS v2.1) [42], CLIP Score [55], and Aesthetic Score [56] to comprehensively assess the quality of the image generation. For video generation, we employ VBench [57] to decompose video generation quality into multiple well-defined dimensions.

In addition, to address and analyze the performance on some long-standing problems in image generation, we introduce some specific evaluation metrics, including Hand-Conf [41], [58], Hand-FID (FID-H) [41] in hand generation, FaceScore [59], Face-FID (FID-F) [41] in face generation, Pick Score [60], ImageReward [61] in high-quality generation. We will elaborate on them in detail when used later.

B. Quantitative Comparisons

1) Text-to-Image Generation: Table I shows the effect of using SG itself on the three SD models. First of all, by incorporating SG itself on SD 1.4, we achieve an impressive 8-point (50.7602 →42.7362) improvement in the FID. Similar improvements appear in other metrics and models as well. This indicates that SG itself is able to improve the quality of generated images.

SG is specifically designed to address the artifact issues in CFG and PAG. As shown in Table I, incorporating SG with CFG leads to consistent improvements in HPS across all models, demonstrating that SG better aligns with human preferences. Moreover, our approach remains competitive with PAG in both CLIP and Aesthetic Score.

Then, we find that combining SG with CFG and PAG will have the best performance over using them alone. Fig. 7 shows the comparison between CFG, CFG+PAG, and CFG + PAG+SG. The results show that adding SG not only enhances fine-grained details, prompt alignment, and error correction but also achieves new SoTA better performance across multiple metrics (24.65 HPS v2.1 for SD 1.4, 25.28 FID for SD 3, 0.3132 CLIP Score in SD 3.5, and 5.7845 Aesthetic Score in Flux).

In addition, as shown in Fig. 1, SG-prev outperformed both compared baselines and the SG model, particularly in terms of FID and CLIP Score. This is achieved by applying SG-prev after the diffusion time (not step) t decreases below 500, according to Fig. 5. This demonstrates that the SG-prev successfully boosts the diffusion models without incurring any additional cost.

2) Text-to-Video Generation: Table III provides quantitative evaluation results across four key dimensions in VBench [57]. The results indicate that adding SG effectively improves the metrics in human action (0.8320 →0.8450), and multiple objects (0.4619 →0.4703). SG-prev also outperforms the baseline across all four metrics under the same sampling time, but slightly lags behind SG that doubles the sampling time. Additionally, We provide more video examples in Appendix A2. After having a close look at the generated examples in Fig. 8, we find that

Fig. 7. Qualitative comparison between CFG, CFG+PAG and CFG + PAG+SG. The red box in the figure indicates where the baseline (SD1.4 and SD3) is poorly generated, and the zoomed-in image is shown in the lower left corner.

TABLE II
QUANTITATIVE RESULTS OF SG PERFORMANCES IN SPECIFIC GENERATIVE DOMAINS COMPARED WITH CFG BASELINE FLUX.1 [23]

Methods	Hand Generation		Face Generation		Text-Image Alignment	Generation Quality		
	FID-H ↓	Hand-Conf ↑	FID-F ↓	FaceScore ↑	CLIP Score ↑	FID ↓	Pick Score↑	ImageReward↑
Flux.1[29]	67.0163	0.8902	34.1976	4.5448	0.3080	29.74	22.9802	1.0971
+SG	**66.4360**	**0.9283**	**34.0561**	**4.6634**	**0.3084**	**28.59**	**22.9960**	**1.1046**

TABLE III
QUANTITATIVE RESULTS OF CFG BASED TEXT-TO-VIDEO GENERATION MODEL COGVIDEOX WITH OUR SG AND SG-PREV. ALL SCORES ARE EVALUATED ON THE STANDARD PROMPT SUITE OF VBENCH [57]. BOLDED VALUES HIGHLIGHT THE BEST PERFORMANCE FOR EACH METRIC.

Model	Human Action↑	Overall Consistency↑	Multiple Objects↑	Appearance Style↑
CogVideoX-5B[66]	0.8320	0.2526	0.4619	0.2331
+ SG	**0.8450**	**0.2553**	0.4703	**0.2339**
+SG-prev	0.8376	0.2530	**0.4774**	0.2333

SG also effectively eliminates artifact problems (broken arms and legs or misplaced limbs) in video generation tasks and enhances text-video consistency. Due to this comparison, we may conclude that SG is also effective in improving video generation quality.

C. Guidance Performances in Specific Domains

Through our expriments and analysis (e.g., Fig. 3), we observed that SG is particularly effective when handling some long-standing problems in image generation. More specifically, SG can effectively generate hands with an exact correct number of fingers with natural shapes (Part 1), and remove artifacts like redundant arms or hands (Part 2). It also enhances Flux's comprehension of textual input (Part 4), effectively eliminating extraneous objects. This allows the main subject of the image to stand out while the background is appropriately blurred(Part 3).

To quantitatively reveal the effectiveness of SG in these aspects, we propose some domain-specific metrics and conduct a comparison with Flux.1 [23], the state-of-the-art text-to-image model. Here we introduce the metrics in detail.

a) Hand Generation: To evaluate the quality of the hands in the generated images, we propose two metrics, the Fréchet Inception Distance for hands (FID-H) and the HAND-CONF.

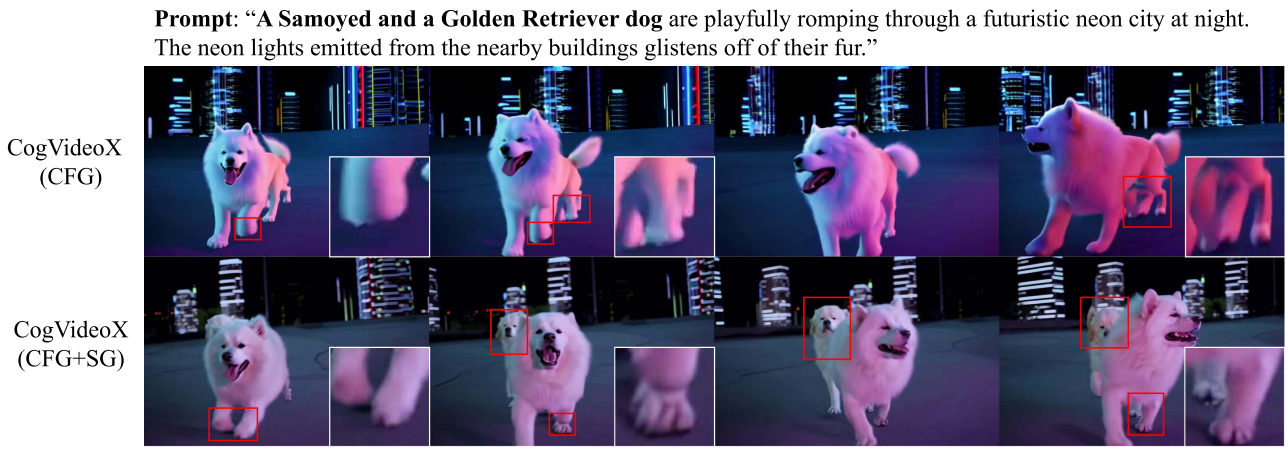


Fig. 8. Qualitative comparison between CFG and CFG+SG in text-to-video model CogVideoX-5B [28]. The red box in the figure indicates where the baseline(CFG) is poorly generated, and the zoomed-in image is shown in the lower left corner. Adding SG to CFG effectively eliminates artifacts (Broken arms and legs and misplaced limbs) in video generation. SG also enhances the consistency between prompt and video(Number of dogs in the video).

FID-H selects 5,000 human-hand-related prompts from the coco validation set, and calculates the distance between the generated image and the real images in COCO. To compute HAND-CONF, we detect hands on the generated images with a pretrained hand detector (i.e. Mediapip [62], and average the confidence scores on 5,000 prompts in the HandCaption-58 k dataset [63].

b) Face Generation: Fréchet Inception Distance for Faces (FID-F) is calculated with 5,000 face-related prompts and images in COCO val. Regarding FaceScore, we use a face quality-focused reward model [59] to score 10,000 images generated with prompts from the HumanCaption-10 M dataset [64].

Moreover, to assess text-image consistency, we measured the CLIP Score. FID, PickScore [60] and ImageReward [61] are also listed here for general image quality assessments.

As shown in Table II, SG achieves significant results in fine-grained generation tasks, such as hand and face generation, with improvements in FID-H (67.0163 to 66.4360) and FID-F (34.1976 to 34.0561). Additionally, the increase in PickScore and ImageReward indicates that our SG is more adept at producing higher-quality samples, thus minimizing generation errors, as illustrated in Part 4 of Fig. 3.

In summary, our SG improves diffusion and flow-based models in multiple aspects, generating high-quality images.

D. Ablation Studies

1) Guidance Scale: We conducted an experimental investigation into the effect of the guidance scale on the performance of Self-Guidance. Using Stable Diffusion 1.4 [40], we sampled 5000 images with guidance scales ranging from 0.0 to 5.0 in 1.0 intervals. We measure the FID [41] and HPS v2.1 [42] for these images. The results, as shown in Fig. 9, indicate that Self-Guidance achieves the best FID (42.23) at a guidance scale of 2.0, and the highest HPS (21.98) at 3.0.

2) Shift Scale

a) Constant Shift Scale: The shift scale $\delta(t)$ represents the extent to which the noise level deviates from the current level. A simple way is to set it to a constant independent of t. The larger the shift scale is, the more it can contrast with the current noise level to suppress the artifacts. However, too large a shift may damage the desired patterns that should be preserved

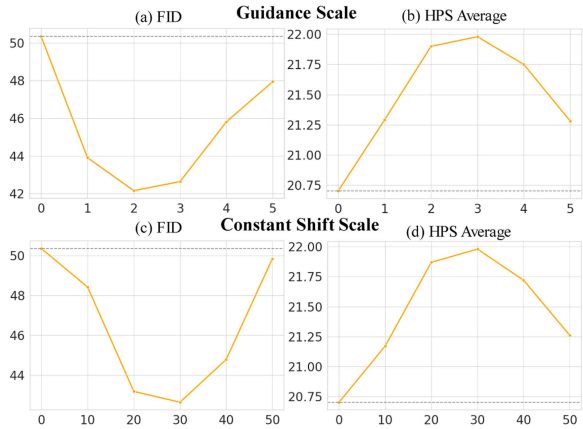


Fig. 9. Ablation study of the guidance scale and shift scale.

in the current sampling density, and thus a suitable shift scale should be adopted. Fig. 9 shows how the FID and HPS score change as the shift scale increases from 0 to 50 (within 1,000 total training timesteps). According to the results, SG achieves the lowest FID and highest HPS score with the shift scale set to 30.

b) Dynamic Shift Scale: One can also dynamically adjust $\delta(t)$ during the reverse diffusion process. At higher noise levels, a large shift of $\delta(t)$ can be adopted, as noisy artifacts can be rapidly removed at the early stage of diffusion sampling. Thus, we introduce $\delta(t)$ as a linear function of time ($\delta(t) = t/\sigma$), allowing stronger guidance at higher noise levels while reducing it at lower levels. More experiments are shown in Appendix A4. We find that dynamic shift scaling improves quality, avoiding blurred images or noise issues caused by excessive guidance as the noise level decreases to zero.

3) Hyperparameter Selection Strategy: Based on our experiments, setting the guidance scale (ω) to 3 generally yields the best results across different models. However, slight tuning may still be necessary depending on the specific model. If the generated images appear overly blurry after setting $\omega = 3$, increasing ω moderately (e.g., to 3.5–4) can help enhance

sharpness. Conversely, if excessive noise is observed, decreasing ω (e.g., to 1–2) is recommended.

For the shift scale ($\delta(t)$), we use a default value equal to 1% of the model's total diffusion time (e.g., $\delta(t) = 10$ if diffusion time $= 1000$). The tuning strategy for $\delta(t)$ follows the same principle as for ω: increase $\delta(t)$ (e.g., to 10-20) if images are too blurry, or decrease it (e.g., to 1–10) if observe excessive noise.

That said, we recommend using SG-prev as a more practical alternative, since it can be viewed as a form of dynamic shift scale. With SG-prev, only the guidance scale needs to be tuned, simplifying hyperparameter selection.

VI. Conclusion

This paper suggests that the reason for artifacts in image generation is incompetent denoising at each step, and proposes Self-Guidance (SG)—an inference-time strategy that modulates the output distribution using the difference between distributions at the current step t and a noisier step $t + \delta(t)$. SG is architecture-agnostic, training-free, and compatible with existing guidance methods. Extensive experiments show that SG consistently improves text-to-image and text-to-video generation, especially in challenging cases like generating realistic human hands and bodies. To reduce SG's two forward passes per sampling step, we introduce SG-prev, which approximates SG by directly reusing the output from the immediately previous timestep. As a result, SG-prev introduces no additional computational cost.

References

[1] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8110–8119.

[2] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," in *Proc. 36th Int. Conf. Neural Inf. Process. Syst.*, 2022, pp. 26565–26577.

[3] A. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *Int. Conf. Mach. Learn.*, 2021, pp. 8162–8171.

[4] A. V. D. Oord et al., "WaveNet: A generative model for raw audio," in *Proc. SSW*, 2016, pp. 125–125.

[5] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video diffusion models," *Neural Inf. Process. Syst.*, 2022, pp. 8633–8646.

[6] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "DreamFusion: Text-to-3D using 2D diffusion," in *Proc. 11th Int. Conf. Learn. Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=FjNys5c7VyY

[7] E. Hoogeboom, V. G. Satorras, C. Vignac, and M. Welling, "Equivariant diffusion for molecule generation in 3D," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 8867–8887.

[8] H. Kim, S. Kim, and S. Yoon, "Guided-TTS: A diffusion model for text-to-speech via classifier guidance," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 11119–11133.

[9] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 10236–10245.

[10] R. T. Chen, J. Behrmann, D. K. Duvenaud, and J.-H. Jacobsen, "Residual flows for invertible generative modeling," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 9916–9926.

[11] C. Meng, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon, "SDEdit: Image synthesis and editing with stochastic differential equations," 2021, *arXiv:2108.01073*.

[12] G. Couairon, J. Verbeek, H. Schwenk, and M. Cord, "DIFFEdit: Diffusion-based semantic image editing with mask guidance," in *Proc. 11th Int. Conf. Learn. Representations*, 2022. [Online] Available: https://openreview.net/forum?id=3lge0p5o-M-

[13] W. Luo, C. Zhang, D. Zhang, and Z. Geng, "Diff-Instruct∗: Towards human-preferred one-step text-to-image generative models," 2024, *arXiv:2410.20898*.

[14] W. Luo, T. Hu, S. Zhang, J. Sun, Z. Li, and Z. Zhang, "Diff-instruct: A universal approach for transferring knowledge from pre-trained diffusion models," *Neural Inf. Process. Syst.*, 2024, pp. 76525–76546.

[15] W. Luo, "Diff-instruct: Training one-step text-to-image generator model to align with human preferences," 2024, *arXiv:2410.18881*.

[16] Z. Geng, A. Pokle, W. Luo, J. Lin, and J. Z. Kolter, "Consistency models made easy," 2024, *arXiv:2406.14548*.

[17] W. Luo, Z. Huang, Z. Geng, J. Z. Kolter, and G.-j. Qi, "One-step diffusion distillation through score implicit matching," *Neural Inf. Process. Syst.*, 2024, pp. 115377–115408.

[18] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 6840–6851.

[19] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," 2020, *arXiv:2010.02502*.

[20] Y. Song, C. Durkan, I. Murray, and S. Ermon, "Maximum likelihood training of score-based diffusion models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 1415–1428.

[21] P. Esser et al., "Scaling rectified flow transformers for high-resolution image synthesis," in *Proc. Int. Conf. Mach. Learn.*, 2024, pp. 12606–12633.

[22] S. AI, "Introducing stable diffusion 3.5," 2023. Accessed: Oct. 30, 2024. [Online]. Available: https://stability.ai/news/introducing-stable-diffusion-3-5

[23] B. F. Labs, "Flux.1," 2023. [Online]. Available: https://blackforestlabs.ai/announcing-black-forest-labs/

[24] C. Saharia et al., "Photorealistic text-to-image diffusion models with deep language understanding," *Neural Inf. Process. Syst.*, 2022, pp. 36479–36494.

[25] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with CLIP latents," 2022, *arXiv:2204.06125*.

[26] A. Ramesh et al., "Zero-shot text-to-image generation," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8821–8831.

[27] A. Blattmann et al., "Stable video diffusion: Scaling latent video diffusion models to large datasets," 2023, *arXiv:2311.15127*.

[28] Z. Yang et al., "CogVideoX: Text-to-video diffusion models with an expert transformer," in *Proc. 13th Int. Conf. Learn. Representations*, 2024. [Online] Available: https://openreview.net/forum?id=LQzN6TRFg9

[29] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "DiffWave: A versatile diffusion model for audio synthesis," in *Proc. Int. Conf. Learn. Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=a-xFK8Ymz5J

[30] S. Shen et al., "DiffTalk: Crafting diffusion models for generalized audio-driven portraits animation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 1982–1991.

[31] Z. Huang, Z. Geng, W. Luo, and G.-j. Qi, "Flow generator matching," 2024, *arXiv:2410.19310*.

[32] W. Luo, B. Zhang, and Z. Zhang, "Entropy-based training methods for scalable neural implicit samplers," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, pp. 7137–7157.

[33] B. Zhang, W. Luo, and Z. Zhang, "Enhancing adversarial robustness via score-based optimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 51810–51829.

[34] Y. Wang, W. Bai, W. Luo, W. Chen, and H. Sun, "Integrating amortized inference with diffusion models for learning clean distribution from corrupted images," 2024, *arXiv:2407.11162*.

[35] W. Deng et al., ""Variational Schrödinger diffusion models," 2024, *arXiv:2405.04795*.

[36] J. Ho and T. Salimans, "Classifier-free diffusion guidance," in *Proc. NeurIPS Workshop Deep Generative Models Downstream Appl.*, 2022.

[37] D. Ahn et al., "Self-rectifying diffusion sampling with perturbed-attention guidance," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 1–17.

[38] S. Hong, G. Lee, W. Jang, and S. Kim, "Improving sample quality of diffusion models using self-attention guidance," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 7462–7471.

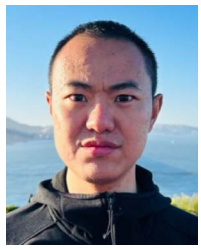
[39] T. Karras, M. Aittala, T. Kynkäänniemi, J. Lehtinen, T. Aila, and S. Laine, "Guiding a diffusion model with a bad version of itself," *Neural Inf. Process. Syst.*, 2024, pp. 52996–53021.

[40] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10684–10695.

[41] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 6629–6640.

[42] X. Wu et al., "Human preference score V2: A solid benchmark for evaluating human preferences of text-to-image synthesis," 2023, *arXiv:2306.09341*.
[43] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "DPM-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps," *Neural Inf. Process. Syst.*, 2022, pp. 5775–5787.
[44] L. Liu, Y. Ren, Z. Lin, and Z. Zhao, "Pseudo numerical methods for diffusion models on manifolds," in *Proc. Int. Conf. Learn. Representations*, 2022. [Online] Available: https://openreview.net/forum?id=PlKWVd2yBkY
[45] S. Xue et al., "Sa-solver: Stochastic adams solver for fast sampling of diffusion models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, pp. 77632–77674.
[46] S. Sadat, M. Kansy, O. Hilliges, and R. M. Weber, "No training, no problem: Rethinking classifier-free guidance for diffusion models," in *Proc. 13th Int. Conf. Learn. Representations*, 2024.
[47] S. Alemohammad, A. I. Humayun, S. Agarwal, J. Collomosse, and R. Baraniuk, "Self-improving diffusion models with synthetic data," 2024, *arXiv:2408.16333*.
[48] S. Hong, "Smoothed energy guidance: Guiding diffusion models with reduced energy curvature of attention," *Neural Inf. Process. Syst.*, 2024, pp. 66743–66772.
[49] Y. Luo et al., "FreeEnhance: Tuning-free image enhancement via content-consistent noising-and-denoising process," in *Proc. 32nd ACM Int. Conf. Multimedia*, New York, NY, USA, 2024, pp. 7075–7084, doi: 10.1145/3664647.3681506.
[50] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *Proc. Int. Conf. Learn. Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=PxTIG12RRHS
[51] J. Sohl-Dickstein, P. Battaglino, and M. R. DeWeese, "Minimum probability flow learning," 2009, *arXiv:0906.4779*.
[52] X. Liu, C. Gong, and Q. Liu, "Flow straight and fast: Learning to generate and transfer data with rectified flow," in *Proc. 11th Int. Conf. Learn. Representations*, 2022. [Online] Available: https://openreview.net/forum?id=XVjTT1nw5z
[53] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," in *Proc. 11th Int. Conf. Learn. Representations*, 2023. [Online] Available: https://openreview.net/forum?id=PqvMRDCJT9t
[54] P. Vincent, "A connection between score matching and denoising autoencoders," *Neural Comput.*, vol. 23, no. 7, pp. 1661–1674, 2011.
[55] S. Zhengwentai, "clip-score: CLIP score for PyTorch," 2023. [Online]. Available: https://github.com/taited/clip-score
[56] C. Schuhmann, "Improved aesthetic predictor," 2023. [Online]. Available: https://github.com/christophschuhmann/improved-aesthetic-predictor
[57] Z. Huang et al., "VBench: Comprehensive benchmark suite for video generative models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 21807–21818.
[58] G. Parmar, R. Zhang, and J.-Y. Zhu, "On aliased resizing and surprising subtleties in GAN evaluation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11410–11420.
[59] Z. Liao, Q. Xie, C. Chen, H. Lu, and Z. Deng, "FaceScore: Benchmarking and enhancing face quality in human generation," in *Proc. AAAI Conf. Artif. Intell.*, 2024, vol. 38, no. 13, pp. 14838–14846.
[60] Y. Kirstain, A. Polyak, U. Singer, S. Matiana, J. Penna, and O. Levy, "Pick-a-pic: An open dataset of user preferences for text-to-image generation," *Neural Inf. Process. Syst.*, 2023, pp. 36652–36663.
[61] J. Xu et al., "ImageReward: Learning and evaluating human preferences for text-to-image generation," *Neural Inf. Process. Syst.*, 2023, pp. 15903–15935.
[62] F. Zhang et al., "MediaPipe hands: On-device real-time hand tracking," 2020, arXiv:2006.10214.
[63] Nagolinc, "Hand captions," 2023. Accessed: Oct. 31, 2024. [Online]. Available: https://huggingface.co/datasets/nagolinc/hand_captions
[64] OpenFace and CQUPT, "HumanCaption-10M," 2023. Accessed: Oct. 31, 2024. [Online]. Available: https://huggingface.co/datasets/OpenFace-CQUPT/HumanCaption-10M
[65] T. Karras, M. Aittala, J. Lehtinen, J. Hellsten, T. Aila, and S. Laine, "Analyzing and improving the training dynamics of diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 24174–24184.
[66] K. Li et al., "Unmasked teacher: Towards training-efficient video foundation models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 19948–19960.
[67] X. Huang et al., "Tag2text: Guiding vision-language model via image tagging," 2024, *arXiv:2303.05657*.
[68] Y. Wang et al., "InternVid: A large-scale video-text dataset for multimodal understanding and generation," 2024, *arXiv:2307.06942*.

Tiancheng Li is currently working toward the CS PhD degree with the joint PhD Program of Zhejiang University, Hangzhou, China, and Westlake University, Hangzhou, supervised by professor Guo-jun Qi. His research interests lie in generative modeling, multimodal learning, with a focus on diffusion models and reinforcement learning.



Weijian Luo received the BS degree from the University of Science and Technology of China, Hefei, China, and the MS and PhD degrees in statistics and generative modeling from Peking University, Beijing, China. He is currently a RedStar senior research scientist with Humane Intelligence (HI) Laboratory, Xiaohongshu Inc. His research interests include large-scale generative models, including one-step text-to-image/video synthesis and reasoning in multimodal systems.

Zhiyang Chen received the BE degree in engineering automation from Xi'an Jiaotong University, Xi'an, China, in 2019, and the PhD degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2024. He is currently a postdoc with Westlake University, Hangzhou, China. His research interests include large multimodal models and reasoning models.

Liyuan Ma received the BS degree in electrical engineering from Southwest Jiaotong University, Chengdu, China, in 2016, and the PhD degree in information science & electronic engineering, Zhejiang University, Hangzhou, China, in 2023. Since 2023, he has been with the School of Engineering, The Westlake University, Hangzhou, China, where he is currently a research assistant professor. His research interests include generative model and embodied intelligence.

Guo-Jun Qi (Fellow, IEEE) is currently a Professor of Artificial Intelligence at Westlake University, Hangzhou, China. From August 2014 to August 2018, he was a faculty member with the Department of Computer Science, University of Central Florida. After that, he was a Technical VP and the chief scientist leading and overseeing the International Research and Development Team for multiple artificial intelligence services on the Huawei Cloud. His research interests include machine learning and knowledge discovery from multi-modal data to build smart and reliable information and decision-making systems.