Rethinking Peculiar Images by Diffusion Models: Revealing Local Minima's Role

Jinhyeok Jang, Chan-Hyun Youn*, Minsu Jeon, Changha Lee

KAIST
{jjh6297, chyoun, msjeon, changha.lee}@kaist.ac.kr

Abstract

Recent significant advancements in diffusion models have revolutionized image generation, enabling the synthesis of highly realistic images with text-based guidance. These breakthroughs have paved the way for constructing datasets via generative artificial intelligence (AI), offering immense potential for various applications. However, two critical challenges hinder the widespread adoption of synthesized data: computational cost and the generation of peculiar images. While computational costs have improved through various approaches, the issue of peculiar image generation remains relatively unexplored. Existing solutions rely on heuristics, extra training, or AI-based post-processing to mitigate this problem. In this paper, we present a novel approach to address both issues simultaneously. We establish that both gradient descent and diffusion sampling are specific cases of the generalized expectation-maximization algorithm. We hypothesize and empirically demonstrate that peculiar image generation is akin to the local minima problem in optimization. Inspired by optimization techniques, we apply naive momentum and positive-negative momentum to diffusion sampling. Last, we propose new metrics to evaluate the peculiarity. Experimental results show momentum effectively prevents peculiar image generation without extra computation.

Introduction

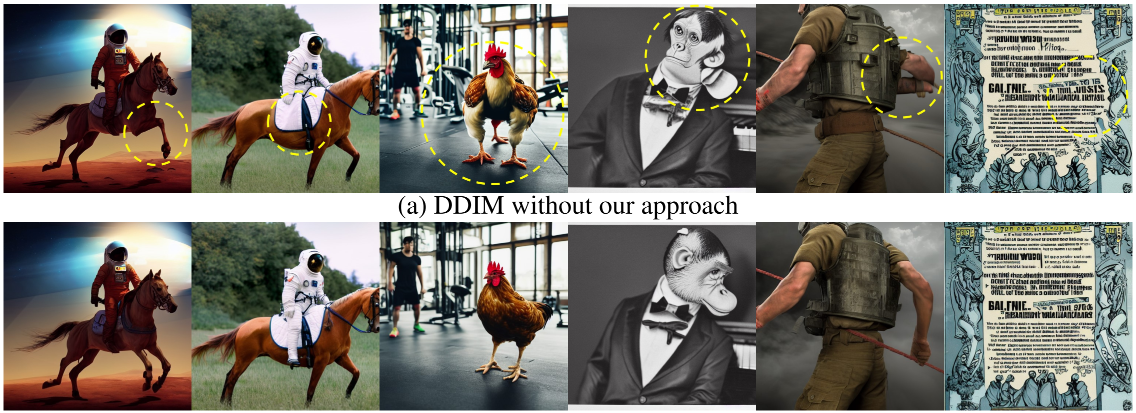
In recent years, significant advancements have been made in the field of generative artificial intelligence (AI), resulting in the remarkable ability to generate highly realistic data. Also, modern generative AI learn a wide range of knowledge and can easily produce images based on text prompts, enabling intentional control over the outputs. As a result, they can generate images, from realistic to imaginative, effectively alleviating the burden of dataset construction. Previously, data acquisition was challenging due to high costs and various limitations, such as risks, dangers, and privacy concerns. However, generative AI has introduced a novel approach to data collection, wherein the generated data becomes employed as training data for other AI models.

Nevertheless, current generative AI has some drawbacks and limitations. Firstly, generating an image demands significant time and computational costs. For instance, denoising diffusion probabilistic model (DDPM) (Ho, Jain, and Abbeel 2020) suggests that a substantial number of sequential steps (approximately 1,000) are required to generate data. Secondly, the generated images are not always entirely reliable, often displaying peculiar and imperfect characteristics. Notably, as highlighted in (Perez et al. 2023), diffusion models often return images of humans with extra fingers, animals with more or less legs. To address the challenge of generation cost, several research efforts such as denoising diffusion implicit model (DDIM) (Song, Meng, and Ermon 2021), PNDM(Liu et al. 2021), or DPM solver++ (Lu et al. 2022b) have been proposed. These methods reshape the denoising process or utilize ordinary differential equation (ODE) solvers in diffusion sampling. This reduces the number of steps needed from 1,000 to tens, with only a minor image quality decline. Conversely, these methods do not directly tackle the second issue of peculiar images, which continues to persist. Also, the peculiar images are another kind of degradation, so addressing this problem without extra costs must help mitigate the computation cost issue.

This paper tackles these issues by combining techniques that evade local minima in SGD. We hypothesize that SGD and diffusion sampling are categorized as a kind of generalized expectation-maximization (GEM) (Dempster, Laird, and Rubin 1977), and derive the two methods into a general form. Then, we compare the diffusion process with SGD-based optimization in terms of equations and visualization. Our analysis reveals that the generation of peculiar images is equivalent to local minima in SGD optimization. To surmount this challenge, we incorporate momentum—a commonly used strategy to escape local minima—along with its variant, positive-negative momentum, into diffusion sampling. Our experiments demonstrate the effectiveness of both momentum strategies in alleviating the production of peculiar artifacts and efficiently generating reasonable images. Our contribution can be summarized as follows:

1. We hypothesize and validate that both SGD and diffusion sampling are special classes of GEM.
2. We establish an equivalence between the peculiar image generation and local minima problems.
3. We propose metrics to measure the qualitative failure.
4. We implement momentum and verify its capacity to address qualitative failures without extra computations.

[*] Corresponding author



(a) DDIM without our approach

(b) DDIM with our approach

Figure 1: Comparison of results generated by stable diffusion with and without our approach.

Related Works

Efficient Diffusion Sampling

Diffusion model generates novel images by iterative denoising from random noise. The concept was originally proposed by (Sohl-Dickstein et al. 2015), but it did not gain much attention compared to generative adversarial network (GAN) (Goodfellow et al. 2014; Radford, Metz, and Chintala 2015; Karras et al. 2018; Brock, Donahue, and Simonyan 2019). Subsequently, DDPM (Ho, Jain, and Abbeel 2020) presented improvements in training the diffusion model and introduced the application of Langevin dynamics to diffusion sampling like (Welling and Teh 2011), leading to significant advancements. It formulated diffusion sampling as:

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{a_t}}\mathbf{x}_t + \frac{\sqrt{1-a_t}}{\sqrt{a_t}}\epsilon_\theta(\mathbf{x}_t, t) + \sigma_t\mathbf{z} \tag{1}$$

where $\mathbf{z} \sim \mathcal{N}(0, I)$. In both equations, $\mathbf{x}_t$ indicates the generated image of timestep $t = 0, 1, ..., T$. Then, ϵ_θ means a model for noise estimation. The noise level of the next step can be controlled by a_t. However, the DDPM demands extensive computation and time costs. To reduce the computation cost, DDIM reformulated the denoising process as:

$$\mathbf{x}_{t-1} = \sqrt{a_{t-1}}\underbrace{\left(\frac{\mathbf{x}_t - \sqrt{1-a_t}\epsilon_\theta(\mathbf{x}_t, t)}{\sqrt{a_t}}\right)}_{o(\mathbf{x}_t, t)\text{: expected } \mathbf{x}_0 \text{ from } \mathbf{x}_t \text{ at } t} + \underbrace{\sqrt{1-a_{t-1}-\sigma_t^2}\epsilon_\theta(\mathbf{x}_t, t)}_{\vec{\mathbf{x}}_t\text{: direction pointing to } \mathbf{x}_t} + \sigma_t\mathbf{z}, \tag{2}$$

It accomplished this by predicting the noise-free image and linearly recombining it with the current noisy image, resulting in less struggling. This approach alleviated some of the computational challenges. Subsequently, researchers have explored alternative approaches to further accelerate diffusion sampling by adopting ODE solvers such as PLMS (Liu et al. 2021), DEIS (Zhang and Chen 2022) and DPM solvers (Lu et al. 2022a,b). These methods successfully achieved similar results to DDPM but with a reduced number of steps by utilizing high-order approximation.

Escaping Qualitative Failure

Generating peculiar images is a widely known challenge in generative AI. Studies like (Borji 2023; Ma et al. 2023; Perez et al. 2023; Wang et al. 2023; Karras et al. 2023) emphasize that AI models often struggle with rendering body parts accurately, such as human or animal faces, limbs, hands, and fingers. Backgrounds in images also suffer from issues, as seen in (Ma et al. 2023; Borji 2023).

In the generative AI community, practitioners have shared various guidelines to address these issues. However, most solutions available are based on know-how or heuristic engineering, including hyper-parameter tuning (e.g., image resolution and aspect ratio), using detailed prompts, and negative prompts. These solutions often involve manual intervention. Few studies, such as (Perez et al. 2023), have explored this problem as a research topic and proposed automatic prompt adjusting or refinement techniques. Advanced diffusion models, as in (Ma et al. 2023; Nichol et al. 2021), have also been explored, but these require complex training and resources. Alternatively, post-processing methods such as image editing and restoration, suggested in (Wang et al. 2023), are commonly recommended for improving generated results. However, they require additional computations.

Problem Definition

The advancements in diffusion models and the integration of text encoder grounding have propelled image synthesis to remarkable levels, allowing for the creation of highly realistic images based on text prompts. This has positioned image synthesis as a promising way for dataset generation. However, we emphasize a critical issue within diffusion models, specifically regarding the accurate depiction of content details in the generated images. For instance, as evident in Figure 1, some images generated by stable diffusion (Rombach et al. 2022) exhibit unnatural features such as animals with three legs or people lacking legs. Then, these phenomena have been widely known issues (Borji 2023; Perez et al. 2023). Incorporating such unnatural images into training data mandates models to accommodate the anomalies, potentially introducing complexities in real-world applications. Moreover, resolving this peculiarity issue without adding extra computational burden serves to improve the computational efficiency, which is another challenge faced by diffusion models. Given the growing significance of image synthesis in dataset construction, the rapid generation of high-quality and trustworthy images is crucial for maintaining the reliability and resilience of AI systems.

Diffusion Sampling with Momentum

To address the two problems, we formulate a hypothesis: ***The qualitative failure in image generation is equivalent to local minima in optimization.*** Based on this, we apply techniques for escaping local minima to diffusion sampling. To validate, there are some questions that have to be addressed.

Diffusion Sampling as Optimization

According to (Ho, Jain, and Abbeel 2020), the noise estimation network ϵ_θ is a **learned gradient** of the clean dataset density. Also, DDIM predicts $o(\mathbf{x}_t, t)$ and updates slightly toward the it for every step. The $o(\mathbf{x}_t, t)$ indicates an *"expected noise-free image ($||o(\boldsymbol{x}_t, t) - \boldsymbol{x}_0||_2 \simeq 0$)"* at t, and the noise-free image is the goal of diffusion sampling (the expected minima). In other words, $\vec{\mathbf{x}}_t$ indicates gradient toward $o(\mathbf{x}_t, t)$. This process shares the idea of GEM algorithm (Dempster, Laird, and Rubin 1977) which is repetition of expectation-step (E-step) and partial-maximization-step (PM-step). In diffusion sampling, ϵ_θ works as an expectation of loss, and the update is equivalent to the PM-step.

Peculiar Image as Local Minima

The GEM algorithm is recognized for its ability to converge to points of zero gradient, yet it's also prone to becoming trapped in local minima or saddle points. A point satisfying two criteria is called a local minimum or saddle point: 1) the presence of alternative minima that are superior in value, and 2) the inability to make further updates. When applying these criteria to the generation of peculiar images, a clear parallel is observed. For instance, a naturally structured four-legged horse is undeniably a more accurate representation than one with three legs, highlighting the existence of superior alternatives. Moreover, when peculiar images are generated without any remaining noise to remove, the process reaches a convergence point beyond which no further denoising is possible. This suggests that despite achieving convergence, there are still better possible outcomes, akin to the challenges faced with local minima in optimization. Therefore, the generation of peculiar images in diffusion models can be likened to encountering local minima or saddle points in optimization processes.

Similarity between SGD and Diffusion Sampling

The alignment of diffusion sampling with the GEM algorithm has been demonstrated. However, an essential aspect remains: drawing a comparable level of similarity between Stochastic Gradient Descent (SGD) and diffusion sampling. This step is vital for facilitating the transfer of specific techniques from SGD to diffusion sampling.

Comparison in formula. Interestingly, SGD is also known as a class of GEM (Audhkhasi, Osoba, and Kosko 2016). SGD with backpropagation can be represented as:

$$\Theta_{t+1} = \Theta_t - \eta \frac{1}{B} \sum_{b=1}^{B} \frac{d\mathcal{L}(\mathbf{s}_b; \Theta_t)}{d\Theta_t}, \qquad \mathbf{s} \in_R \mathcal{D}, \tag{3}$$

where Θ_t means model weights at t-th iteration, η means learning rate. B indicates batch size, $\mathcal{L}(\mathbf{s}; \Theta)$ means the loss function on sample $\mathbf{s}$ with weights Θ. The $\mathbf{s}$ is randomly chosen from training dataset $\mathcal{D}$. Then, the sampling process of DDIM (Eq. (2)) can be written into SGD-like form as:

$$\begin{aligned} \mathbf{x}_{t-1} &= \sqrt{a_{t-1}}\, o(\mathbf{x}_t, t) + \vec{\mathbf{x}}_t + \sigma_t \mathbf{z} \\ &= \mathbf{x}_t - \left(\mathbf{x}_t + \sqrt{a_{t-1}}\, o(\mathbf{x}_t, t) + \vec{\mathbf{x}}_t\right) + \sigma_t \mathbf{z} \end{aligned} \tag{4}$$

The two concepts, SGD and diffusion sampling can be formulated into form of GEM with stochastic manner as:

Algorithm 1: Formulation of GEM for a variable $\mathbf{u}$

```
Input: $\mathbf{u}_T$
for $t$=$T$,$T-1$,...,1 do
    E-step: Compute $Q(\mathbf{u}_t)$
    PM-step: $\mathbf{u}_{t-1} = \mathbf{u}_t - \Delta\mathbf{u}_t + noise$,
end for
return $\mathbf{u}_0$
```

where $Q(\mathbf{u}_t)$ indicates cost function for $\mathbf{u}_t$, $\Delta\mathbf{u}_t$ means an update term, and $noise$ indicates stochastic noise. Then, the $Q(\mathbf{u}_t)$, $\Delta\mathbf{u}_t$ and $noise$ are summarized in Table. **??** and **??**.

Method	$\mathbf{u}_t$	$Q(\mathbf{u}_t)$
SGD	Θ_{T-t}	$\eta\frac{1}{B}\sum_{b=1}^{B}\mathcal{L}(\mathbf{s}_b; \Theta_{T-t})$
DDPM/DDIM	$\mathbf{x}_t$	$\epsilon_\theta(\mathbf{x}_t, t)$

Table 1: E-step Parts of SGD and diffusion sampling

Method	$\Delta\mathbf{u}_t$	$noise$
SGD	$\eta\frac{1}{B}\sum_{b=1}^{B}\frac{d\mathcal{L}(\mathbf{s}_b;\Theta_{T-t})}{d\Theta_{T-t}}$	0
DDPM	$(1-\frac{1}{\sqrt{a_t}})\mathbf{x}_t + \frac{\sqrt{1-a_t}}{\sqrt{a_t}}\epsilon_\theta(\mathbf{x}_t, t)$	$\sigma_t\mathbf{z}$
DDIM	$(1-\frac{\sqrt{a_{t-1}}}{\sqrt{a_t}})\mathbf{x}_t$ $+(\frac{\sqrt{a_{t-1}}\sqrt{1-a_t}}{\sqrt{a_t}} - \sqrt{1-a_{t-1}-\sigma_t^2})\epsilon_\theta(\mathbf{x}_t, t)$	$\sigma_t\mathbf{z}$

Table 2: PM-step Parts of SGD and diffusion sampling

In this context, previous methods like PNDM and DPM solver can be viewed as high-order optimization techniques such as Newton-Raphson or Gauss-Newton methods.

Comparison in visualization. We found similarity between SGD and diffusion sampling in the formulation of GEM. In addition, we visually compared them.

First, we generated CIFAR10(Krizhevsky, Hinton et al. 2009)-like images using a pretrained DDIM and saved every image ($\mathbf{x}_t$) from an initial random noise ($\mathbf{x}_T \sim \mathcal{N}(0, I)$,

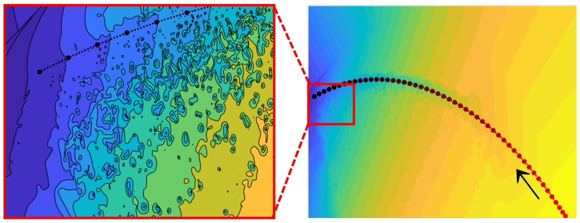


Figure 2: Landscape visualization of $\mathbf{x}_t$ for every step $t = 1, 2, ..., 50$. The color indicates $log(||o(\mathbf{x}_t+n, t)-\mathbf{x}_{\text{DDPM}}||_2)$. (The $\mathbf{x}_{\text{DDPM}}$ is obtained by DDPM sampling with $T = 1,000$ from the same $\mathbf{x}_T$), and $n \sim \mathcal{N}(0, I)$.

where T means the number of steps) to the converged $\mathbf{x}_0$. We then gathered the surroundings of $\mathbf{x}_t$ trajectory by injecting various random noise as $\mathbf{x}_t + n$, where $n \sim \mathcal{N}(0, I)$. Using PCA, we reduced dimension of the surrounding and trajectory images for landscape visualization. Figure 2 presents the trajectory and landscape. The dot color progresses from red to black, indicating t. Landscape color corresponds to $log(||o(\mathbf{x}_t + n, t) - \mathbf{x}_{\text{DDPM}}||_2)$. Since image generation lacks ground truth, we employed $\mathbf{x}_{\text{DDPM}}$ as an alternative. The $\mathbf{x}_{\text{DDPM}}$ was obtained by DDPM with $T = 1,000$ from the same $\mathbf{x}_T$. SGD features a continuous loss landscape (search space), but is characterized by numerous local minima (Li et al. 2018). A similar pattern is observed in diffusion sampling. This landscape illustrates not only the resemblance between the search spaces of SGD and diffusion sampling but also the presence of local minima in diffusion sampling.

Second, we conducted an analysis of the evolution of generated images concerning the total number of steps (T). Similar to the acknowledged impact of learning rates on the local minima phenomenon, we conjectured a comparable influence of T on the occurrence of peculiar images. Figure 3 visually represents changes in generated images as T varies, utilizing stable diffusion. As shown, diverse patterns emerge. The first row shows a discernible trend emerging where the quality of generation appears to be directly proportional to T. The second row shows poor generation from $T = 20$ to $T = 100$, with only the instance at $T = 50$ yielding an image of a normal appearance. This observation lends support to the presence of local minima. The last row shows normal images at $T = 20$ and $T \geq 180$ cases. These observations indicate that the search space of diffusion sampling is quite complex, so there are many local minima.

In summary, our key findings include: 1) SGD and diffusion sampling both align with GEM-based optimization, 2) Their smooth search spaces are similar, and 3) We observed complex patterns indicating local minima. Based on these, we focused on mitigating image peculiarity by applying optimization techniques to escape local minima.

Integrating Momentum into Diffusion Sampling

Now that we have identified the qualitative failure in image generation as local minima and established the similarity between SGD and diffusion sampling, we searched for methods to escape these local minima. Several studies have explored how to avoid local minima in SGD-based optimization. Among them, we adopted two techniques, 1) Momentum and 2) Stochastic gradient noise (SGN).

Momentum. Momentum is a well-known technique in optimization that can aid in escaping local minima (Sutskever et al. 2013; Zavriev and Kostyuk 1993; Jelassi and Li 2022). It introduces a moving average of past gradients, which helps the weights have non-zero gradients at local minima, facilitating escape from such points. Also, the momentum is similar to using a larger η. Studies, like (Zavriev and Kostyuk 1993; Gitman et al. 2019; Leclerc and Madry 2020), have mentioned the benefits of momentum. The SGD (Sutskever et al. 2013) with momentum can be written as:

$$m_t = \beta m_{t-1} + (1-\beta)(\eta_t \Delta\Theta_t), \tag{5}$$

$$\Theta_{t+1} = \Theta_t - m_t, \tag{6}$$

where β means momentum parameter about forgetting the previous states that satisfying $\beta \in [0, 1)$. When it is imported into DDIM, diffusion sampling can be formulated as:

Algorithm 2: Momentum for diffusion sampling

```
$\mathbf{x}_T \sim \mathcal{N}(0, I)$
$m_t = 0$
for $t$=$T$,...,1 do
    $\mathbf{x}'_{t-1}$ = DDIM($\mathbf{x}_t, t$)
    $m_t = \beta m_{t+1} + (1-\beta)(\mathbf{x}_t - \mathbf{x}'_{t-1})$
    $\mathbf{x}_{t-1} = \mathbf{x}_t - m_t$
end for
return $\mathbf{x}_0$
```

Momentum with Stochastic Gradient Noise. Next, we focused on the noise injection term in Eq. (1) (4). As described in (Ho, Jain, and Abbeel 2020), they adapted Langevin dynamics to the diffusion model, where the equation inherently incorporates the noise injection term to account for nondeterministic movements in molecular dynamics. A comparable concept, **SGN**, can also be found in optimization techniques. It is a kind of random noise on the update term during gradient descent. Stochastic noise is widely recognized as an effective approach for escaping saddle points or sharp minima during optimization. Consequently, introducing SGN has been observed to aid in locating flat minima. Numerous studies (Hochreiter and Schmidhuber 1994; Keskar et al. 2016; Ge et al. 2015; Jin et al. 2017; Zhu et al. 2019; Xie, Sato, and Sugiyama 2020) have explored the benefits of artificially injecting random noise into SGD and demonstrated improved generalization performance. Nevertheless, this approach to artificial noise injection does come with certain limitations (Wu et al. 2020).

In the context of diffusion sampling, the discrepancy between real noise and estimated noise can serve as a form of SGN. However, the naive momentum method is unable to effectively account for SGN (Xie et al. 2021), there is no change in SGN with or without naive momentum.

Incorporating SGN within a momentum-based framework can be achieved through the use of Positive-Negative momentum (Xie et al. 2021), as proposed in the literature. According to (Xie et al. 2021), this approach allows for the

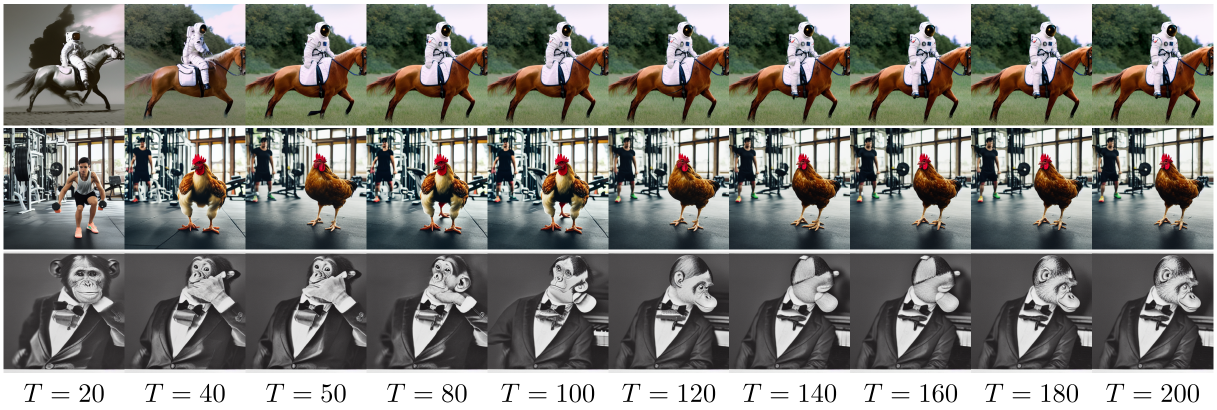


Figure 3: Generated images with varying number of steps (T). Note that the images of first row have to contain cock, not person. This figure illustrates the presence of various patterns in the generation process, indicating the existence of local minima.

control of the SGN magnitude. Positive-Negative momentum is defined as a momentum value that satisfies $\beta < 0$ instead of the conventional range of $\beta \in [0, 1)$. When t is an odd number, it can be represented as follows:

$$\begin{cases} m_t^{(even)} = \sum_{\tau=0,2,4...,t-1}(1+|\beta|)|\beta|^{t-\tau}(\mathbf{x}_\tau - \mathbf{x}'_{\tau-1}) \\ m_t^{(odd)} = \sum_{\tau=1,3,5...,t}(1+|\beta|)|\beta|^{t-\tau}(\mathbf{x}_\tau - \mathbf{x}'_{\tau-1}) \end{cases} \tag{7}$$

$$m_t = (1+|\beta|)m_t^{(odd)} - |\beta|m_t^{(even)}. \tag{8}$$

By using this positive-negative momentum, the magnitude of SGN becomes $[(1+|\beta|)^2 + |\beta|^2]$-times larger than one without momentum (Xie et al. 2021).

Conceptual Comparison with Previous Approaches. Prior works rooted in ODE solvers, such as PLMS and DPM solver++, as well as our integration of momentum, reflect a common concept: *incorporating historical information into updates*. However, momentum is applied across the entire update term ($\mathcal{A}(\mathbf{x}_t)$), whereas previous works were confined to $\epsilon\theta(\mathbf{x}_t, t)$ only. Furthermore, our approach aggregates the entirety of historical data, whereas previous works limited this consideration to only the past few steps, like 2 or 3 steps.

The concept of positive-negative momentum reveals a common thread in previous works as well. In specific, PLMS employed Adams–Bashforth methods, a numerical approximation technique, to integrate short-term momentum. It utilized specific weighting coefficients $[\frac{55}{24}, \frac{-59}{24}, \frac{37}{24}, \frac{-9}{24}]$ to account for the estimated noise across the current four steps. Also, DPM solver++ adopted [1+$\frac{1}{2r}$, -$\frac{1}{2r}$] weighting to accommodate the estimated noise within the current two steps, where r signifies the ratio of log-SNR between the steps.

Measuring Peculiarity in Generated Images

To the best of our knowledge, there isn't a defined evaluation metric to measure the peculiarities of image synthesis. While our work aims to enhance both the computational efficiency and the peculiarity of generated images, the assessment of peculiarity poses a significant challenge. To make it simple, we assessed only human images. Numerous analyses have indicated that diffusion models exhibit anomalies in limb and finger. For this, we introduced two metrics utilizing the OpenPose (Cao et al. 2017), a pose estimation method considering the interconnectivity between human joints. The first metric centers on the average confidence scores of all joints. For an image $\mathbf{x}$, OpenPose ($\mathcal{H}(\cdot)$) not only estimates 25 joints but also quantifies their confidence scores as:

$$\{(\mathbf{j}_i^p, c_i^p)|\, i = 1, 2, .., 25, \text{and } p = 1, 2, .., N_\mathbf{x}\} = \mathcal{H}(\mathbf{x}), \tag{9}$$

where $(\mathbf{j}_i^p, c_i^p) \in (\mathbf{J}^p, \mathbf{C}^p)$ means a joint and its confidence score of p-th person among $N_\mathbf{x}$ persons in $\mathbf{x}$. Our expectation is that peculiar images would yield lower confidence scores or no detected joints (zero confidence).

For our second metric, we initially considered a Fréchet Inception Distance (FID)-like metric, but recognized its potential bias towards frequently observed poses. Since pose naturalness should be evaluated without considering frequency – understanding that rare poses can be natural and common ones peculiar – we focused on measuring peculiarity independently of frequency. Drawing inspiration from Hwang et al. (Hwang, Yang, and Kwak 2020), we developed our metric using clustering applied to MS COCO dataset (Lin et al. 2014), which is rich in real human images. For each of N persons in MS COCO dataset, we performed sample-wise min-max normalization on their poses as:

$$\{\Xi^1, \Xi^2, ..., \Xi^N\} = \{norm(\mathbf{J}^p_{coco})\,|\, p = 1, 2, ..., N\}. \tag{10}$$

Then, we applied K-means clustering to the set of Ξ as:

$$\{\Psi^1, \Psi^2, ..., \Psi^K\} = kmeans(\{\Xi^1, \Xi^2, ..., \Xi^N\}, K), \tag{11}$$

where K indicates the number of clusters. Each cluster center Ψ means the representative normalized poses of real persons. Then, we measured the quality based on the distance for p-th person (d^p) to the nearest cluster center as:

$$d^p = min\{\, ||norm(\mathbf{J}^p) - \Psi^k||_2 \,|\, k = 1, 2, ..., K\}. \tag{12}$$

This metric measures the distance to the nearest representative pose, identified through the analysis of numerous real images. Utilizing these two metrics enables the assessment of peculiarity in human images.

Experiments

In this section, we evaluate how integrating momentum can improve diffusion sampling. Our source code is available[1].

Quantitative Analysis about Momentum

We conducted an analysis using pre-trained diffusion models[2] on the CIFAR10 and CelebA datasets (Liu et al. 2015), without any further training. Our aim was to assess the FID for the number of score function evaluations (NFE). To achieve this, we explored different sampling methods, including DDPM, DDIM, PLMS, and DPM solver. Additionally, we considered two distinct schedules: linear and quadratic skipping.

FID Analysis on Pretrained Diffusion Models. The objective of this analysis is to evaluate the efficacy of sampling methods. For PNDM, it incorporates the Runge-Kutta method for the initial steps, necessitating a 4× NFE due to high-order approximations. Then, it uses part of linear multi-step (PLMS) for the remainder steps. To ensure a fair comparison, we substituted the initial steps with those derived from DDIM. For DPM solver++, it introduces two techniques: one is employing multistep algorithm that reuses previously estimated noises. The other one is its own noise scheduling. To isolate the impact of the update term while negating the influence of other components, we adopted the noise scheduling from DDIM instead of the new scheduling. By implementing these modifications, we aimed to evaluate only sampling methods, focusing on their update terms without the confounding effects of additional components.

The FID scores for 50,000 synthetic images per method are presented in Table **??**. The results demonstrate that incorporating momentum into diffusion models leads to improved performance without requiring additional training. As shown, the naive or positive-negative momentum improves for all methods. Particularly noteworthy is the performance of DDIM with momentum, outperforming even the more recent works. These results support the hypothesis that local minima is a factor in peculiar image, and employing techniques to escape local minima can alleviate the issue. Additionally, Table **??** details the computation costs of generating a CelebA-like image using PLMS and linear scheduling, both with and without momentum. These results demonstrate that momentum can improve the previous methods without extra computations.

Ablation Study about β and T

Numerous studies have indicated a proportional relationship between the magnitude of SGN and the temperature ($\frac{\eta}{B}$). For diffusion models, while setting B to 1 is straightforward, defining η is challenging. Instead, a comparable parameter emerges in the form of the step size (T), which is similar to $\frac{1}{\eta}$ within the diffusion sampling framework.

Dataset	Schedule	Method	β	# NFE			
				50	100	250	1000
CIFAR10	Linear	DDPM	0.0	N/A	N/A	N/A	**3.19**
		DDIM	0.0	**7.19**	5.58	4.62	N/A
			0.2	11.09	4.91	4.02	N/A
			-0.8	13.24	**4.00**	**3.57**	N/A
		PLMS	0.0	4.15	**3.73**	**3.62**	N/A
			0.2	**3.63**	3.91	3.71	N/A
		DPM++	0.0	5.69	4.56	4.00	N/A
			0.2	**3.84**	**3.78**	**3.80**	N/A
	Quadratic Skip	DDIM	0.0	4.53	4.06	3.89	N/A
			0.2	6.01	4.77	4.12	N/A
			-0.8	**4.25**	**3.67**	**3.69**	N/A
		PLMS	0.0	**3.82**	3.78	3.83	N/A
			0.2	4.80	**3.66**	**3.79**	N/A
		DPM++	0.0	3.99	3.77	3.84	N/A
			0.2	5.33	4.40	4.04	N/A
			-0.3	**3.45**	**3.50**	**3.64**	N/A
CelebA	Linear	DDPM	0.0	N/A	N/A	N/A	**2.99**
		DDIM	0.0	5.99	4.13	2.88	N/A
			0.2	**3.91**	**3.30**	**2.67**	N/A
			-0.8	13.75	7.16	3.49	N/A
		PLMS	0.0	4.52	2.78	2.45	N/A
			0.2	**2.20**	**2.33**	**2.35**	N/A
		DPM++	0.0	3.81	2.77	2.39	N/A
			0.2	**2.97**	**2.58**	**2.39**	N/A
	Quadratic Skip	DDIM	0.0	**6.29**	6.05	6.15	N/A
			0.2	7.36	6.50	6.20	N/A
			-0.8	9.37	**5.93**	**5.96**	N/A
		PLMS	0.0	7.20	6.59	**6.26**	N/A
			0.2	**7.07**	**6.56**	6.28	N/A
		DPM++	0.0	**6.14**	**6.05**	6.18	N/A
			0.2	7.12	6.49	6.26	N/A
			-0.3	6.89	6.15	**6.09**	N/A

Table 3: Quality of generated images measured in FID.

Setting	GFLOPS/img	FID(↓)
$T = 250$ w/o momentum	250×46.74	2.45
$T = 50$ w/o momentum	50×46.74	4.52
$T = 50$ w/ momentum	50×46.74	**2.20**

Table 4: Computation cost for generating a 64×64 image.

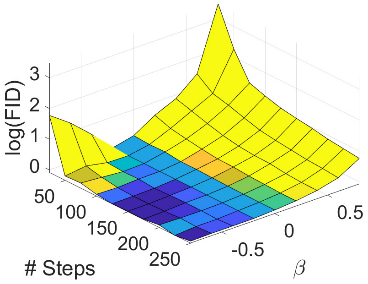
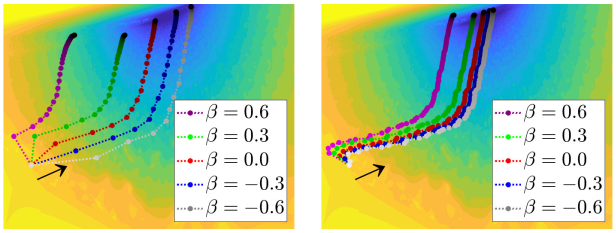


Figure 4: FID analysis for varying β and number of steps.

Analysis of FID over β. We thoroughly investigated FID enhancements by varying β and T, utilizing the pretrained CIFAR10 model with quadratic skipping as established in the preceding section. The results are depicted in Figure 4. Significantly, the results unveil that for small step sizes ($T \leq 25$), a negative β worsens FID. However, negative β effectively improves FID for all cases where $T > 25$. Moreover, a distinct pattern emerges: the optimal β is proportional to $\frac{1}{T}$. This observation means that the magnitude of SGN is already excessive for smaller T, negating the need for fur-

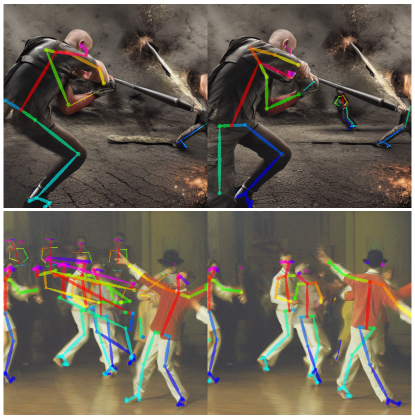
[1] https://github.com/jjh6297/momentum-diffusion-sampling

[2] https://github.com/tqch/ddpm-torch



$T = 50$ $T = 250$

Figure 5: Visualization of $o(\mathbf{x}_t, t)$ in DDIM for every step. Note that the color indicates $log(||o(\mathbf{x}_t + n, t) - \mathbf{x}_{\text{DDPM}}||_2)$ (The $\mathbf{x}_{\text{DDPM}}$ was obtained by DDPM samling with $T = 1,000$ from the same $\mathbf{x}_T$, and used as an alternative of ground truth, and $n \sim \mathcal{N}(0, I)$).



w/o momentum w/ momentum

Figure 6: Examples of human images with pose estimation

ther enlargement. Conversely, as T increases and is associated with a smaller η, the SGN's magnitude diminishes. In this context, emphasizing the magnitude becomes imperative, hence the efficacy of larger $|\beta|$ values. These results not only establish the equivalence between SGD and diffusion sampling but also validate our hypothesis.

Landscape and Trajectory Visualization. In addition, we analyzed and compared the effect of β over the trajectory of $o(\mathbf{x}_t, t)$ with $T = 50$ as illustrated in Figure 5. The procedure outlined in section Q3, involving the variation of β and T values, was replicated for $o(\mathbf{x}_t, t)$ in place of $\mathbf{x}_t$. As shown, the positive β suffers from velocity of the first few steps because the momentum is initialized as zero. Then, the more negative β is beneficial for the higher T. This fact is related to the previous ablation study shown in Figure 4. In contrast, $\beta = -0.3$ is the best for the $T = 50$ case.

Application to Stable Diffusion

At last, we applied the momentum ($\beta = -0.3$) to a pretrained stable diffusion model with DDIM sampling and $T = 100$. To ensure a fair evaluation, we conducted two separate tests using the same random seed: one with positive-negative momentum and the other without it.

Qualitative Evaluation. Figure 1 and 6 present the results obtained from the same initial random noise ($\mathbf{x}_T$), contrasted between the scenarios of employing and not employing positive-negative momentum. The comparison shows significant improvements when momentum is incorporated into the sampling. As depicted, the generated images are more completed and visually natural with positive-negative momentum. Animals are represented with their natural leg count, while figures of astronauts regain their legs. These demonstrate the efficacy of momentum-based optimization in refining the image generation process. Also, they highlight how momentum can tackle peculiar image generation in diffusion models, even when guided by text prompts.

β	Single Person	Groups
0.0	0.4351	0.1880
-0.3	**0.4383**	**0.1991**

Table 5: Average joint confidence(↑) of human images

# Persons	β	# Clusters				
		5	10	15	25	50
Single Person	0.0	0.2823	0.2589	0.2461	0.2407	0.2221
	-0.3	**0.2782**	**0.2515**	**0.2416**	**0.2378**	**0.2219**
Group	0.0	0.3086	0.2799	0.2627	0.2447	**0.2255**
	-0.3	**0.3067**	**0.2796**	**0.2625**	**0.2442**	0.2271

Table 6: Average $RMSE$(↓) to the nearest cluster center

Quantitative Evaluation of Human Image Peculiarity. Our proposed metrics were used to evaluate the quality of human images. Using the same random seed, we generated two types of full-body human images: one featuring a single individual and another depicting a group, as shown in Figure 6. The results are summarized in Table **??** and Table **??**, affirming the effectiveness of the momentum to the quality of generated images. Table **??** signifies that the utilization of momentum contributes to generating images with increased pose estimation certainty. Meanwhile, Table **??** demonstrates that images generated with momentum exhibit more natural poses compared to those generated without it.

Conclusion

In this paper, we address the issue of qualitative failure in image generation by tackling local minima. Based on the mention that noise estimation in diffusion models is equivalent to learned gradients, we conjectured this issue resembles a local minima problem in optimization. To counter this, we introduced momentum – a way for escaping local minima – into diffusion sampling. Our experimental section entails a comparative analysis of the diffusion sampling process and SGD, encompassing trend analysis and qualitative visualizations of outputs. By integrating two kinds of momentum into diffusion sampling, the generated results exhibited improved completeness and reliability. Lastly, we incorporated the positive-negative momentum into stable diffusion models, yielding successful generation results and quantitative performance. These experiments effectively validate the equivalency between the problem of peculiar image generation and the local minima issue. Moreover, they demonstrate the potential of momentum in alleviating this problem.

Acknowledgements

This work was partially supported by the Korea Institute of Marine Science & Technology Promotion (KIMST) under grant [RS-2023-00238652, Integrated Satellite-based Applications Development for Korea Coast Guard]. This work was also partially supported by the Korea Aerospace Industries (KAI) LTD.

References

Audhkhasi, K.; Osoba, O.; and Kosko, B. 2016. Noise-enhanced convolutional neural networks. *Neural Networks*, 78: 15–23.

Borji, A. 2023. Qualitative failures of image generation models and their application in detecting deepfakes. *Image and Vision Computing*, 137: 104771.

Brock, A.; Donahue, J.; and Simonyan, K. 2019. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *ICLR*.

Cao, Z.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 7291–7299.

Dempster, A. P.; Laird, N. M.; and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1): 1–22.

Ge, R.; Huang, F.; Jin, C.; and Yuan, Y. 2015. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on learning theory*, 797–842. PMLR.

Gitman, I.; Lang, H.; Zhang, P.; and Xiao, L. 2019. Understanding the role of momentum in stochastic gradient methods. *NeurIPS*, 32.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *NeurIPS*, 27.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *NeurIPS*, 33: 6840–6851.

Hochreiter, S.; and Schmidhuber, J. 1994. Simplifying neural nets by discovering flat minima. *NeurIPS*, 7.

Hwang, J.; Yang, J.; and Kwak, N. 2020. Exploring rare pose in human pose estimation. *IEEE Access*, 8: 194964–194977.

Jelassi, S.; and Li, Y. 2022. Towards understanding how momentum improves generalization in deep learning. In *ICML*, 9965–10040. PMLR.

Jin, C.; Ge, R.; Netrapalli, P.; Kakade, S. M.; and Jordan, M. I. 2017. How to escape saddle points efficiently. In *ICML*, 1724–1732. PMLR.

Karras, J.; Holynski, A.; Wang, T.-C.; and Kemelmacher-Shlizerman, I. 2023. Dreampose: Fashion image-to-video synthesis via stable diffusion. *arXiv preprint arXiv:2304.06025*.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *ICLR*.

Keskar, N. S.; Mudigere, D.; Nocedal, J.; Smelyanskiy, M.; and Tang, P. T. P. 2016. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. In *ICLR*.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images. *Technical report*, University of Toronto.

Leclerc, G.; and Madry, A. 2020. The two regimes of deep network training. *arXiv preprint arXiv:2002.10376*.

Li, H.; Xu, Z.; Taylor, G.; Studer, C.; and Goldstein, T. 2018. Visualizing the loss landscape of neural nets. *NeurIPS*, 31.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*, 740–755. Springer.

Liu, L.; Ren, Y.; Lin, Z.; and Zhao, Z. 2021. Pseudo Numerical Methods for Diffusion Models on Manifolds. In *ICLR*.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *ICCV*, 3730–3738.

Lu, C.; Zhou, Y.; Bao, F.; Chen, J.; Li, C.; and Zhu, J. 2022a. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *NeurIPS*, 35: 5775–5787.

Lu, C.; Zhou, Y.; Bao, F.; Chen, J.; Li, C.; and Zhu, J. 2022b. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*.

Ma, W.-D. K.; Lewis, J.; Kleijn, W. B.; and Leung, T. 2023. Directed diffusion: Direct control of object placement through attention guidance. *arXiv preprint arXiv:2302.13153*.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

Perez, A.; Elistratov, I.; Schmitt-Ulms, F.; Demir, E.; Lolla, S.; Ahmadi, E.; and Amini, A. 2023. Risk-Aware Image Generation by Estimating and Propagating Uncertainty. *ICML Workshops*.

Radford, A.; Metz, L.; and Chintala, S. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*, 10684–10695.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2256–2265. PMLR.

Song, J.; Meng, C.; and Ermon, S. 2021. Denoising diffusion implicit models. *ICLR*.

Sutskever, I.; Martens, J.; Dahl, G.; and Hinton, G. 2013. On the importance of initialization and momentum in deep learning. In *ICML*, 1139–1147. PMLR.

Wang, Y.; Yu, J.; Yu, R.; and Zhang, J. 2023. Unlimited-size diffusion restoration. In *CVPR*, 1160–1167.

Welling, M.; and Teh, Y. W. 2011. Bayesian learning via stochastic gradient Langevin dynamics. In *ICML*, 681–688.

Wu, J.; Hu, W.; Xiong, H.; Huan, J.; Braverman, V.; and Zhu, Z. 2020. On the noisy gradient descent that generalizes as sgd. In *ICML*, 10367–10376. PMLR.

Xie, Z.; Sato, I.; and Sugiyama, M. 2020. A Diffusion Theory For Deep Learning Dynamics: Stochastic Gradient Descent Exponentially Favors Flat Minima. In *ICLR*.

Xie, Z.; Yuan, L.; Zhu, Z.; and Sugiyama, M. 2021. Positive-negative momentum: Manipulating stochastic gradient noise to improve generalization. In *ICML*, 11448–11458. PMLR.

Zavriev, S.; and Kostyuk, F. 1993. Heavy-ball method in nonconvex optimization problems. *Computational Mathematics and Modeling*, 4(4): 336–341.

Zhang, Q.; and Chen, Y. 2022. Fast Sampling of Diffusion Models with Exponential Integrator. In *ICLR*.

Zhu, Z.; Wu, J.; Yu, B.; Wu, L.; and Ma, J. 2019. The Anisotropic Noise in Stochastic Gradient Descent: Its Behavior of Escaping from Sharp Minima and Regularization Effects. In *ICML*, 7654–7663. PMLR.