Self-Guided Generation of Minority Samples Using Diffusion Models

Soobin Um and Jong Chul Ye

KAIST, Daejeon, Republic of Korea
{sum,jong.ye}@kaist.ac.kr

Abstract. We present a novel approach for generating minority samples that live on low-density regions of a data manifold. Our framework is built upon diffusion models, leveraging the principle of guided sampling that incorporates an arbitrary energy-based guidance during inference time. The key defining feature of our sampler lies in its *self-contained* nature, *i.e.*, implementable solely with a pretrained model. This distinguishes our sampler from existing techniques that require expensive additional components (like external classifiers) for minority generation. Specifically, we first estimate the likelihood of features within an intermediate latent sample by evaluating a reconstruction loss w.r.t. its posterior mean. The generation then proceeds with the minimization of the estimated likelihood, thereby encouraging the emergence of minority features in the latent samples of subsequent timesteps. To further improve the performance of our sampler, we provide several time-scheduling techniques that properly manage the influence of guidance over inference steps. Experiments on benchmark real datasets demonstrate that our approach can greatly improve the capability of creating realistic low-likelihood minority instances over the existing techniques without the reliance on costly additional elements. Code is available at `https://github.com/soobin-um/sg-minority`.

Keywords: Diffusion models · Image generation · Minority sampling

1 Introduction

Contemporary large-scale datasets often exhibit long-tailed distributions, containing *minority* samples that lie on low-density regions of the data manifold. The minority samples are less common and often possess unique characteristics rarely seen in the majority of the data. Generating these less probable data points are indispensable in a variety of applications like classification [31], anomaly detection [10,11], and medical diagnosis [43] where augmenting additional instances of rare attributes could enhance the predictive capabilities of the focused tasks. Such augmentation is also significant in promoting fairness, aligning with social vulnerabilities often associated with minority instances [34, 43]. Moreover, the unique features within these minority instances are of paramount importance in use-cases like creative AI applications [15, 35], where the ability to generate samples with exceptional creativity is crucial.

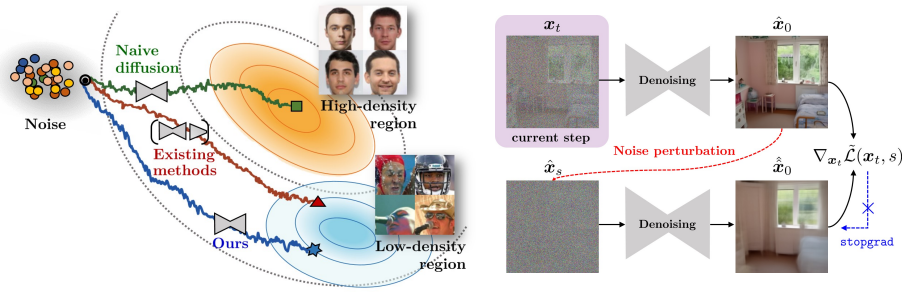

Fig. 1: (Left) Existing methods vs. our self-guided approach. Unlike previous methods that rely upon external components (*e.g.*, classifiers) to guide the generation process towards low-density regions [38, 43], our approach yields low-density guidance solely based on a pretrained diffusion model, thereby offering a self-contained minority generation achievable without any aids of expensive extra elements. **(Right) Overview of our self guidance for minority data.** Specifically to yield low-density guidance given a current latent instance $\boldsymbol{x}_t$ during inference, we first obtain its denoised version $\hat{\boldsymbol{x}}_0$ via the use of Tweedie's formula [7,33] implemented with a pretrained model. We then perturb $\hat{\boldsymbol{x}}_0$ into $\hat{\boldsymbol{x}}_s$ via the DDPM forward process and denoise $\hat{\boldsymbol{x}}_s$ to $\hat{\hat{\boldsymbol{x}}}_0$ via the pretrained model. A discrepancy between $\hat{\boldsymbol{x}}_0$ and $\hat{\hat{\boldsymbol{x}}}_0$ (denoted as $\tilde{\mathcal{L}}(\boldsymbol{x}_t, s)$ in the figure) is then computed, and we subsequently use its gradient as low-density guidance with the `stopgrad` technique [5] applied on $\hat{\hat{\boldsymbol{x}}}_0$; see Sec. 3 for details.

The problem is that under standard generative frameworks (*e.g.*, GANs [13], DDPM [18]), curating even a small number of low-probability samples requires substantial investments in terms of time and computational resources [16,38,43]. Several efforts were made to address this issue particularly under supervised settings [26, 31, 38, 46]. However, most of their methods bear an inherent reliance on annotations (*e.g.*, minority labels) that are difficult to obtain in many practically relevant scenarios [26, 31, 38, 46]. Furthermore, they often require specialized training procedures [26, 31, 46] which restrict the utilization of powerful pretrained models (like ADM [9] and Stable Diffusion [35]), further limiting their practical significance in real-world applications.

One recent study by [43] addressed the challenges using diffusion-based generative frameworks [18, 40]. Their approach leverages a pretrained diffusion model and unlabeled data to implement a classifier-guided sampler that targets low-density regions by optimizing their own minority metric. While this method demonstrates substantial enhancements even under unsupervised settings, it often introduces significant overhead in obtaining the classifier particularly for large-scale benchmarks. For instance, training the classifier for the ImageNet-64 experiments in [43] requires significant resource investments of **40** V100-days [9]. Also, the classifier construction necessitates access to a substantial number of real samples, which may be prohibitive in data-limited situations.

Contribution. In this work, we propose a *minimalist* approach that eliminates the necessity for such expensive external components. Leveraging diffusion-based

generative frameworks [18, 40], we develop a sampling technique that can be executed using only a pretrained model. One defining feature of our sampler lies in its *self-guided* nature: unlike prior arts that obtain guidance toward low-density regions with the help of costly outside-components (*e.g.*, separately trained classifiers [38,43]), our guidance for low-density regions is obtained in an *autonomous* fashion by leveraging the knowledge of a given pretrained model; see Fig. 1 for an overview (the left plot).

Specifically to identify the direction of a low-density region during inference time, we develop a metric that evaluates the uniqueness of features that are contained in an intermediate instance $\boldsymbol{x}_t$. More precisely, we employ a reconstruction loss w.r.t. the posterior mean $\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t]$, inspired by the low-likelihood measure proposed in [43]. We provide both theoretical and empirical evidence that demonstrates a close connection between ours and existing likelihood measures for validation.

We then establish a sampler that optimizes the proposed metric over the generation process of diffusion models to encourage evolution towards low-likelihood minority features captured by our metric. We highlight that our method could be highly effective even with intermittent usage (*e.g.*, incorporating our guidance every 5 sampling steps) thereby significantly reducing the computational costs associated with our approach. To improve the sample quality of generated minority samples, we further provide several scheduling methods that properly scale the strength of our guidance over the sampling timesteps.

We also conduct extensive experiments across a diverse range of real benchmarks. We demonstrate that the proposed sampler can significantly boost up the capability of generating high-quality minority samples, as reflected in high values of low-density metrics such as Average k-Nearest Neighbor and better quality metrics like Fréchet Inception Distance [17]. To highlight the practical significance of our work, we also delve into a downstream application, investigating the advantages of our sampler for data augmentation in classifier training. We emphasize that the benefits of our framework stem exclusively from a pretrained diffusion model, which is in stark contrast with existing techniques [26,31,38,43,46] that require additional expensive resources to improve the minority-generating capability.

Related work. In addition to [43], the capability of producing minority data has been explored under several distinct conditions and scenarios [21, 26, 31, 38, 46]. One close instance to our work is [38] where the authors develop a diffusion sampler that can encourage the sampling process of diffusion models to evolve toward low-density regions w.r.t. a specific class via an external classifier and a class-conditional model. The key distinction w.r.t. our sampler is that their method requires access to the class predictor which is often expensive to obtain.

The authors in [21, 26, 31, 46] investigate a slightly different goal, improving representations of minority instances to enhance data coverage close to the ground-truth data distribution. However, their methods rely upon label information that indicate minority samples, which is inherently distinct from our work. [37] develops a sampler for text-to-image (T2I) diffusion models [35] to

specifically enhance the quality of generated samples prompted with unique concepts (*e.g.*, shaking hands) rarely observed during training. However, their method requires access to a set of high-quality images that describe the target concepts and therefore not directly comparable to ours.

The exploration of diversity of diffusion models has received less attention compared to their quality aspects which have been scrutinized from various perspectives [19, 20]. One notable progress was recently made in [36] wherein the authors discovered that simply incorporating noise perturbations (yet properly annealed over time) to class embeddings could significantly improve the diversity of generated samples. A key difference from our approach lies in the fact that their method is designed to generate diverse samples that adhere to the ground-truth data distribution, rather than focusing on low-density regions of the distribution (as ours and [38, 43]).

2 Background

2.1 Diffusion-based generative models

Diffusion models [18, 40, 41] are latent variable models described by a forward diffusion process and the associated reverse process. The forward process is basically a Markov chain with a Gaussian transition, where data is gradually perturbed by Gaussian noise according to a variance schedule $\{\beta_t\}_{t=1}^T$: $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) \coloneqq \mathcal{N}(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \boldsymbol{I})$ where $\{\boldsymbol{x}_t\}_{t=1}^T$ are latent variables with the same dimensionality as data $\boldsymbol{x}_0 \sim q(\boldsymbol{x}_0)$. One important property of the forward process is that it admits *one-shot* sampling of $\boldsymbol{x}_t$ at any timestep $t \in \{1, \dots, T\}$:

$$q_{\alpha_t}(\boldsymbol{x}_t \mid \boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\alpha_t}\boldsymbol{x}_0, (1-\alpha_t)\boldsymbol{I}), \tag{1}$$

where $\alpha_t \coloneqq \prod_{s=1}^t (1-\beta_s)$. The variance schedule is designed to respect $\alpha_T \approx 0$ so that $\boldsymbol{x}_T \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. The reverse process is another Markov Chain with learnable Gaussian transition $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) \coloneqq \mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t), \boldsymbol{\Sigma}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t))$. The mean is expressible in terms of a noise-conditioned score network as $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) = \frac{1}{\sqrt{1-\beta_t}}(\boldsymbol{x}_t + \beta_t \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t))$, where the score network is parameterized to approximate the score function of the perturbed distribution: $\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) \coloneqq \nabla_{\boldsymbol{x}_t} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}_t) \approx \nabla_{\boldsymbol{x}_t} \log q_{\alpha_t}(\boldsymbol{x}_t)$. The variance of the reverse process is often fixed, *e.g.*, $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) = \beta_t \boldsymbol{I}$ [18]. One common way to construct the score network is through a denoising score matching [42, 44]:

$$\min_{\boldsymbol{\theta}} \sum_{t=1}^T w_t \mathbb{E}_{q(\boldsymbol{x})q_{\alpha_t}(\tilde{\boldsymbol{x}}|\boldsymbol{x})}[\|\boldsymbol{s}_{\boldsymbol{\theta}}(\tilde{\boldsymbol{x}}, t) - \nabla_{\tilde{\boldsymbol{x}}} \log q_{\alpha_t}(\tilde{\boldsymbol{x}} \mid \boldsymbol{x})\|_2^2],$$

where $w_t \coloneqq 1 - \alpha_t$. One notable point is that this procedure is equivalent to training a noise-prediction network $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)$ that predicts noise added on clean data $\boldsymbol{x}_0$ through the forward process in Eq. (1) [42, 44]. This establishes an intimate connection between the two networks: $\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) = -\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)/\sqrt{1-\alpha_t}$.

Given a pretrained score model, the generation can be done by starting from $\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and iteratively going through the reverse process down to $\boldsymbol{x}_0$:

$$\boldsymbol{x}_{t-1} = \boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) + \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{1/2}(\boldsymbol{x}_t, t)\boldsymbol{z}, \quad \boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}). \tag{2}$$

This process, often called *ancestral sampling* [42], is actually a discretized simulation of a stochastic differential equation that defines $\{p_{\boldsymbol{\theta}}(\boldsymbol{x}_t)\}_{t=0}^{T}$ [42], which guarantees to sample from $p_{\boldsymbol{\theta}}(\boldsymbol{x}_0) \approx q(\boldsymbol{x}_0)$.

2.2 Guided sampling with diffusion models

One instrumental feature of diffusion models is that their generative processes are often amenable to various optimization signals for conditioning generations in post-hoc fashions. Specifically at each time step t, one can incorporate an arbitrary energy-based *guidance* into the sampling process (*e.g.*, Eq. (2)) to encourage the evolution toward a desired direction [12]:

$$\boldsymbol{x}_{t-1} = \boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) + \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{1/2}(\boldsymbol{x}_t, t)\boldsymbol{z} + w_t \boldsymbol{g}(\boldsymbol{x}_t, t), \tag{3}$$

where $\boldsymbol{g}(\boldsymbol{x}_t, t)$ is a (sign-flipped) energy-based guidance function, and w_t corresponds to the strength of the guidance term possibly scheduled over time. The guidance function may incorporate a target condition $\boldsymbol{c}$, in which case the function becomes $\boldsymbol{g}(\boldsymbol{x}_t, t; \boldsymbol{c})$. Notice that plugging the gradient of a classifier log-likelihood (*e.g.* $\nabla_{\boldsymbol{x}_t} \log p_{\boldsymbol{\phi}}(y|\boldsymbol{x}_t)$) into Eq. (3) (alongside $w_t = w\boldsymbol{\Sigma}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$ where w is a fixed constant) recovers the famous classifier-guided sampler [9].

Guidance for minority data. The principles of existing minority samplers are centered around the classifier guidance [9]. Particularly for low-likelihood generation with a conditional diffusion model, [38] propose to leverage the classifier guidance in the *opposite* direction. Their guidance function is expressible as:

$$\boldsymbol{g}(\boldsymbol{x}_t, t; y) = -\nabla_{\boldsymbol{x}_t} \log p_{\boldsymbol{\phi}}(y|\boldsymbol{x}_t).$$

where y indicates a target class for the focused conditional generation. The descending gradient makes the sampling process get closer to low-likelihood regions (w.r.t. the target class y), thereby encouraging generation of low-probability instances of the focused class y. On the other hand, the guidance developed by [43] uses the same sign of the guidance as [9] while incorporating a distinct classifier, specifically trained to predict the degree of uniqueness of features within $\boldsymbol{x}_t$:

$$\boldsymbol{g}(\boldsymbol{x}_t, t; l) = \nabla_{\boldsymbol{x}_t} \log p_{\boldsymbol{\psi}}(l|\boldsymbol{x}_t), \tag{4}$$

where l indicates the uniqueness level w.r.t. noisy latent instance $\boldsymbol{x}_t$. Their focused uniqueness metric, which is called *minority score*, is shown as being inversely correlated with the likelihood (*i.e.*, higher minority score, lower the likelihood) [43], and therefore the gradient ascent w.r.t. the metric can serve to encourage generation of highly unique (*i.e.*, less-probable) instances.

While both techniques offer great improvements in the capability of producing minority instances [38, 43], their guidance functions bear inherent reliance on

external components (like classifiers) that are often challenging (or even impossible) to acquire in practical settings. Our primary contribution lies in untethering such intrinsic dependencies and developing a *self-contained* guidance function implementable solely through a pretrained diffusion model, thereby significantly enhancing the accessibility and practicality of minority instance generation.

3 Method

3.1 Towards an inference-time minority metric

Our approach starts by investigating a metric to be incorporated in the guidance function (*i.e.*, $\boldsymbol{g}$ in Eq. (3)). Specifically in the context of self-guided generation of minority data, the metric should satisfy two criteria: (i) the ability to assess the likelihood of features underlain in an intermediate latent sample $\boldsymbol{x}_t$; (ii) accessibility via a pretrained diffusion model.

One can naturally think of leveraging an ODE-based likelihood estimator [42] to compute $\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_t)$ and incorporating the estimate in the guidance function. However, despite its capability of providing highly-accurate estimates, ODE-based estimators are often computationally expensive [42], *e.g.*, requiring many Jacobian computations proportional to the number of diffusion timestep T. More importantly, the direct use of the log-likelihood in the guidance function (*e.g.*, $g(\boldsymbol{x}_t, t) = \nabla_{\boldsymbol{x}_t} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}_t)$) may drive the sampling process *out-of-manifold*. This is because a low-likelihood in a perturbed distribution may imply a noisy instance that does not belong to the data manifold. The downside is evident by poor performance of the high-temperatured sampler of diffusion models; see details in Sec. G in [9].

We take a distinct approach that sidesteps the above challenges. To this end, we first introduce *minority score*, a low-likelihood measure proposed in [43]. The metric quantifies the degree of uniqueness (*i.e.*, low-densitiness) of features contained in a given *clean* sample $\boldsymbol{x}_0$, mathematically written as:

$$\mathcal{L}(\boldsymbol{x}_0; t) \coloneqq \mathbb{E}_{q_{\alpha_t}(\boldsymbol{x}_t|\boldsymbol{x}_0)}[d(\boldsymbol{x}_0, \hat{\boldsymbol{x}}_0(\boldsymbol{x}_t))], \tag{5}$$

where t refers to the timestep used for perturbing $\boldsymbol{x}_0$, and $d(\cdot, \cdot)$ is a discrepancy measure (*e.g.*, LPIPS [47]). $\hat{\boldsymbol{x}}_0$ denotes the posterior mean obtained via Tweedie's formula [7, 33] implemented with a pretrained model $\boldsymbol{s}_{\boldsymbol{\theta}}$:

$$\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t) \coloneqq \mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t] = \frac{1}{\sqrt{\alpha_t}} \left(\boldsymbol{x}_t + (1 - \alpha_t)\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)\right). \tag{6}$$

Intuitively, minority score can be interpreted as a reconstruction loss of a clean sample, measured with the posterior mean of a noise-perturbed version. The key benefit of this metric is in its computational efficiency, requiring only a single forward pass of a diffusion model while serving as a good proxy for the log-likelihood [43]. The problem is that the metric is defined w.r.t. clean samples $\boldsymbol{x}_0$, making it impossible to be directly employed in the guidance function (that should work with $\boldsymbol{x}_t$). The authors in [43] circumvented this issue by introducing

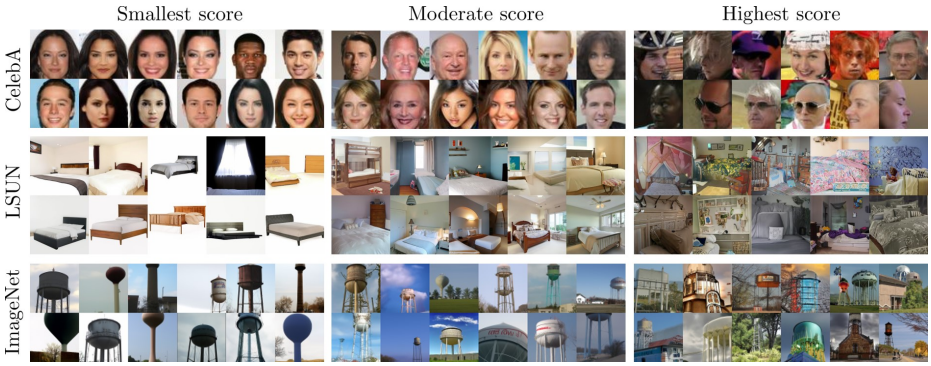

Fig. 2: Effectiveness of our metric (*i.e.*, Eq. (7)) for identifying low-likelihood minority features during inference time. Generated samples with the smallest (left column), moderate (middle column), and the highest (right column) values of the proposed metric are exhibited. We employed ADM [9] for the generations of all three benchmarks. The metric values were calculated during inference time via Eq. (7).

a separately-trained classifier (*e.g.*, in Eq. (4)), but as mentioned earlier, it is often expensive to obtain.

We make progress by providing a new metric that accommodates noisy intermediates $\boldsymbol{x}_t$. The key idea is to consider the posterior mean $\hat{\boldsymbol{x}}_0$ as a clean-surrogate of $\boldsymbol{x}_t$ and to measure the uniqueness of features therein. Specifically we employ a reconstruction loss of $\hat{\boldsymbol{x}}_0$:

$$\tilde{\mathcal{L}}(\boldsymbol{x}_t; s) \coloneqq \mathbb{E}_{q_{\alpha_s}(\hat{\boldsymbol{x}}_s|\hat{\boldsymbol{x}}_0)}[d(\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t), \hat{\hat{\boldsymbol{x}}}_0(\hat{\boldsymbol{x}}_s(\boldsymbol{x}_t)))], \tag{7}$$

where s refers to the timestep used for noise-corrupting $\hat{\boldsymbol{x}}_0$, and $\hat{\boldsymbol{x}}_s$ indicates the perturbed instance: $\hat{\boldsymbol{x}}_s(\boldsymbol{x}_t) \coloneqq \sqrt{\alpha_s}\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t) + \sqrt{1-\alpha_s}\boldsymbol{\epsilon}$. $\hat{\hat{\boldsymbol{x}}}_0(\hat{\boldsymbol{x}}_s)$ denotes the posterior mean of $\hat{\boldsymbol{x}}_s$ obtained by applying Eq. (6) on $\hat{\boldsymbol{x}}_s$. $d(\cdot,\cdot)$ is a distance measure that plays the same role as in Eq. (5). See the supplementary for details on our choices of s and $d(\cdot,\cdot)$. Note that our metric can be understood as minority score of $\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t)$, a variant that extends the metric to describe the uniqueness w.r.t. noisy instances $\boldsymbol{x}_t$.

Fig. 2 visualizes the effectiveness of the proposed metric on several real-world benchmarks. Observe that the instances in the left column, which are determined as high-likelihood samples by our metric, exhibit features that are commonly observed in the corresponding datasets (*e.g.*, frontal-view faces in CelebA [27]). While in the right column, we see samples containing uncommon visual attributes of the benchmarks. For instance, "Wearing_Hat" and "Eyeglasses" attributes observed in the CelebA samples are famously known as minority features [1,46]. Also, complicated visual attributes seen in the LSUN and ImageNet samples are actually key defining features of low-density instances in real-world benchmarks [2,39]. In the supplementary, we illustrate the correlations of the proposed metric and existing low-density metrics; see details therein.

Connection to log-likelihood. To further validate our metric in a theoretical manner, we establish a mathematical connection of the proposed metric with log-likelihood. To this end, we first prove that minority score, when integrated over timesteps with a properly chosen $d(\cdot,\cdot)$, becomes equivalent to the negative Evidence Lower BOund (ELBO) – a well-known proxy for log-likelihood. Then by leveraging the relation of our metric with minority score, we establish the connection with the ELBO. Below we provide formal statements of the claim. See the supplementary for the proofs.

Proposition 1. *Consider minority score in Eq.* (5) *with the squared-error distance loss* $\|\cdot\|_2^2$*. Its weighted sum over timesteps is equivalent (upto a constant factor) to the negative ELBO considered in [18]:*

$$\sum_{t=1}^{T} \bar{w}_t \mathcal{L}(\boldsymbol{x}_0; t) = \sum_{t=1}^{T} \mathbb{E}_{p(\boldsymbol{\epsilon})}[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\sqrt{\alpha_t}\boldsymbol{x}_0 + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}, t)\|_2^2] \gtrapprox -\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0),$$

where $\bar{w}_t \coloneqq \alpha_t/(1-\alpha_t)$ *and* $p(\boldsymbol{\epsilon}) \coloneqq \mathcal{N}(\boldsymbol{\epsilon}; \mathbf{0}, \boldsymbol{I})$*.*

Corollary 1. *The proposed metric in Eq.* (7) *with the squared-error loss is equivalent to the negative ELBO w.r.t.* $\log p_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t))$ *when integrated over timesteps with* $\bar{w}_s \coloneqq \alpha_s/(1-\alpha_s)$*.*

3.2 Self-guidance for low-density regions

Our next step is to develop the guidance function that incorporates our metric for minority generation. Since we are interested in encouraging $\boldsymbol{x}_t$ to evolve toward low-likelihood regions (that could yield high values of Eq. (7)), a natural choice for $\boldsymbol{g}$ would be to use the gradient of the proposed metric. Employing the gradient of our measure as $\boldsymbol{g}$ gives:

$$\boldsymbol{g}(\boldsymbol{x}_t, t; s) \coloneqq \nabla_{\boldsymbol{x}_t} \tilde{\mathcal{L}}(\boldsymbol{x}_t; s) = \nabla_{\boldsymbol{x}_t} \mathbb{E}_{q_{\alpha_s}(\hat{\boldsymbol{x}}_s|\hat{\boldsymbol{x}}_0)}[d(\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t), \hat{\hat{\boldsymbol{x}}}_0(\hat{\boldsymbol{x}}_s(\boldsymbol{x}_t)))]. \tag{8}$$

Notice that this guidance function does not require any external elements for computation, which is in stark contrast with the prior methods on low-density guidance [38,43]. We empirically found that simply adopting the above guidance function can yield great improvements in the capability to produce minority instances. However, the gradient computation in Eq. (8) requires two backward passes through the model $\boldsymbol{s}_{\boldsymbol{\theta}}$, which often comes with considerable computational overhead.

We handle the issue by leveraging the stop-gradient technique [5]. More specifically, we employ the stop-gradient on $\hat{\hat{\boldsymbol{x}}}_0$ that incurs the additional backward pass. Our modified guidance reflecting the stop-gradient reads:

$$\boldsymbol{g}^*(\boldsymbol{x}_t, t; s) \coloneqq \nabla_{\boldsymbol{x}_t} \mathbb{E}_{q_{\alpha_s}(\hat{\boldsymbol{x}}_s|\hat{\boldsymbol{x}}_0)}[d(\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t), \mathtt{sg}(\hat{\hat{\boldsymbol{x}}}_0(\hat{\boldsymbol{x}}_s(\boldsymbol{x}_t))))],$$

where $\mathtt{sg}(\cdot)$ indicates the stop-gradient operator. Notice that only a single backward pass now suffices for computing the gradient. Importantly, we found that it

often preserves the great performance benefit of low-likelihood guidance offered by the guidance function in Eq. (8). Conversely, incorporating the stop-gradient on $\hat{\boldsymbol{x}}_0$ (instead of $\hat{\hat{\boldsymbol{x}}}_0$) is not as effective, yielding little improvements over standard diffusion samplers. We conjecture that this trend is because the impact of concerning $\hat{\boldsymbol{x}}_0$ is much more significant in optimizing the proposed metric than the influence w.r.t. $\hat{\hat{\boldsymbol{x}}}_0$ which is less relevant to the current sample $\boldsymbol{x}_t$ due to the perturbation $\boldsymbol{\epsilon}$. See Tab. 2b for empirical validation regarding our stop-gradient approach. An overview of our guidance can be found in Fig. 1 (the right-plot). We found that our guidance could be robust to the off-manifold issue (mentioned in Sec. 3.1); see the supplementary for a detailed analysis on this point.

Intermittent guidance for reduced computations. We discovered that the computational costs of our sampler could be significantly reduced through *intermittent* usage, *i.e.*, incorporating the guidance once every n sampling steps. For instance, employing the guidance every 2 steps (*i.e.*, $n = 2$) could lead to a notable reduction of 37.7% in inference time with marginal impact on performance (see Tab. 1 for instance). Empirically, we observed that employing $n = 5$ yields satisfactory performance across diverse benchmarks. See the supplementary for detailed analysis on the impact of n.

3.3 Time-scheduling for improved sample quality

Now we move onto the story on w_t, a scaling factor that controls the strength of the guidance over time. A naive approach is to use a constant scale (*i.e.*, $w_t = w$), but we observed that it often leads to non-trivial degradation in sample quality for high values of w. We hypothesized that this comes from conflicting influences between the reverse process and our guidance, particularly occurring during the later timesteps. Specifically, the sampling process in these later steps often focuses on articulating fine details of images [6, 22, 24]. If our guidance remains consistently strong during these stages, it may impede the articulation process since our guidance could potentially encourage structural changes diverging from the refinement task. To avoid the conflict, we explored several time-scheduled scaling methods that employ decreasing w_t's over time. We found that they exhibit the same trend of yielding better sample quality over constant scales with a slight compromise in the low-densitiness.

Algorithm 1 Self-guided minority sampler

Require: T, n, s, w.
1: $\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
2: **for** $t \leftarrow T$ to 1 **do**
3: $\quad \boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ if $t > 1$, else $\boldsymbol{z} = \mathbf{0}$
4: $\quad \boldsymbol{x}_{t-1} \leftarrow \boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) + \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{1/2}(\boldsymbol{x}_t, t)\boldsymbol{z}$
5: $\quad$ **if** $t \bmod n = 0$ **then**
6: $\quad\quad \boldsymbol{x}_{t-1} \leftarrow \boldsymbol{x}_{t-1} + w\boldsymbol{\Sigma}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)\boldsymbol{g}^*(\boldsymbol{x}_t, t; s)$
7: $\quad$ **end if**
8: **end for**
return $\boldsymbol{x}_0$

We provide two distinct time schedules herein. The first one is a simple switch-off-type schedule that discontinues incorporating the guidance after a specific timestep:

$$w_t = w \cdot \mathbb{1}\{t \geq t_{\text{mid}}\},$$

(a) ADM [9] **(b)** Um and Ye [43] **(c)** Ours

Fig. 3: Sample comparison on LSUN-Bedrooms. We share the same random seed for all methods.

where t_{mid} is a pre-defined timestep that determines when to stop. The other proposal is one that leverages the noise variance of the reverse process (*i.e.*, the same choice as [9]):

$$w_t = w \cdot \boldsymbol{\Sigma}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t),$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$ can be either learned or fixed; see Sec. 2.1 for its formal definition. While the switch-off yields some improvements over the sampler with fixed scales, we empirically observed that the variance-based schedule generally yields better performance; see Tab. 2c for instance. The proposed minority sampler with the variance schedule is formulated as:

$$\boldsymbol{x}_{t-1} = \boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) + \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{1/2}(\boldsymbol{x}_t, t)\boldsymbol{z} + w\boldsymbol{\Sigma}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)\boldsymbol{g}^*(\boldsymbol{x}_t, t; s). \tag{9}$$

See Algorithm 1 for pseudocode of our sampler that incorporates the intermittent technique. The generation due to Eq. (9) can be interpreted as sampling from a modified density $\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{x}_t) \propto p_{\boldsymbol{\theta}}(\boldsymbol{x}_t)e^{\tilde{\mathcal{L}}(\boldsymbol{x}_t;s)}$. We note that instances with high values of $\tilde{\mathcal{L}}(\boldsymbol{x}_t; s)$ would have more chances to be generated compared to the original density $p_{\boldsymbol{\theta}}(\boldsymbol{x}_t)$, which aligns well with our focus to encourage the generation of minorities.

4 Experiments

4.1 Setup

Datasets and pretrained models. We employ four real-world benchmarks that include both unconditional and conditional data. Our unconditional benchmarks are CelebA 64^2 [27] and LSUN-Bedrooms 256^2 [45]. For the class-conditional datasets, we employ ImageNet 64^2 and 256^2 [8]. The pretrained model for CelebA was constructed by ourselves by following the settings in [43]. The models for LSUN-Bedrooms and ImageNet were taken from the checkpoints provided in [9]. In addition to our primary focus on these four benchmarks, we further explore

(a) ADM [9] **(b)** Sehwag *et al.* [38] **(c)** Ours

Fig. 4: Sample comparison on ImageNet-256. Generated samples from two classes are exhibited: Water tower (top row) and Bald eagle (bottom row). For each row, we share the same random seed across all three methods.

challenging scenarios of T2I generation and medical imaging to scrutinize the boundaries of our approach. See the supplementary for explicit details on these experimental tasks.

Baselines. We compare a variety of frameworks with our approach, which encompasses existing minority samplers as well as generic frameworks that are not specifically tailored for low-density generation. For the CelebA experiments, we employ BigGAN [4], ADM [9] with ancestral sampling, and Um and Ye [43]. As in [43], we incorporate an additional baseline on CelebA, which implements a conditioned generation of minority instances by using the classifier guidance [9] with minority annotations given in the dataset (*e.g.*, "Eyeglasses" [46]). We also compare [38] on CelebA by extending their sampler to admit unconditional data, where we used the same approach as [43] for the extension. For LSUN-Bedrooms, we consider four baselines: (i) StyleGAN [23]; (ii) ADM [9]; (iii) LDM [35]; (iv) Um and Ye [43]. We focus on four diffusion-based frameworks on ImageNet-64: (i) ADM [9]; (ii) EDM [22]; (iii) Sehwag *et al.* [38]; (iv) Um and Ye [43]. For the ImageNet-256 experiments, we consider: (i) ADM [9]; (ii) DiT [30]; (iii) CADS [36]; (iv) Sehwag *et al.* [38]; (v) Um and Ye [43].

Evaluation metrics. We respect the choices in the previous studies [38, 43] and employ three distinct measures for evaluating low-densitiness of the considered methods: (i) Average k-Nearest Neighbor (AvgkNN); (ii) Local Outlier Factor (LOF) [3]; (iii) Rarity Score [15]. For all three measures, a higher value implies that the instance is less likely than its neighborhood samples [15, 38, 43]. We also employ a range of quantitative metrics to assessing quality and diversity, including: (i) Clean Fréchet Inception Distance (cFID) [29]; (ii) Spatial FID (sFID) [28]; (iii) Improved Precision & Recall [25]. Specifically to evaluate the

Table 1: Comparison of sample quality and diversity. "ADM-ML" refers to a classifier-guided sampler implemented on ADM [9], which conditions on **M**inority **L**abels [43]. "+ intermittent" refers to our sampler that incorporates intermittent guidance. For baseline real data, we employ the most unique samples that yield the highest AvgkNN values. The best results are marked in **bold**, and the second bests are underlined.

Method	cFID	sFID	Prec	Rec
CelebA 64×64				
ADM [9]	75.41	17.11	**0.97**	0.23
BigGAN [4]	80.58	16.80	**0.97**	0.19
ADM-ML [43]	51.99	13.40	0.94	0.30
Sehwag *et al.* [38]	28.25	10.64	0.82	0.42
Um and Ye [43]	27.32	8.66	0.89	0.33
Ours	**18.57**	**8.20**	0.83	**0.48**
+ intermittent	19.34	8.85	0.82	0.47
ImageNet 64×64				
ADM [9]	18.37	5.39	0.79	0.53
EDM [22]	19.09	4.73	0.73	0.59
Sehwag *et al.* [38]	11.37	4.69	**0.80**	0.52
Um and Ye [43]	12.47	3.13	0.76	0.56
Ours	**11.08**	**3.09**	0.72	**0.63**
+ intermittent	11.24	3.17	0.73	0.62

Method	cFID	sFID	Prec	Rec
LSUN Bedrooms 256×256				
ADM [9]	63.30	8.00	0.89	0.15
LDM [35]	63.53	7.73	**0.90**	0.13
StyleGAN [23]	57.17	7.78	0.89	0.14
Um and Ye [43]	41.75	7.26	0.87	0.10
Ours	**35.79**	**4.94**	0.86	**0.16**
+ intermittent	36.94	5.13	0.87	0.15
ImageNet 256×256				
ADM [9]	13.22	7.66	**0.86**	0.39
DiT [30]	21.51	6.76	0.80	0.46
CADS [36]	15.95	6.18	0.81	**0.48**
Sehwag *et al.* [38]	10.93	6.66	0.85	0.39
Um and Ye [43]	11.44	4.63	0.85	0.42
Ours	**9.93**	**4.19**	0.83	0.46
+ intermittent	9.98	4.35	0.83	0.45

proximity to real minority data, we follow the same approach as in [43] and employ instances with the lowest likelihoods (*e.g.*, the ones yielding the highest AvgkNNs) as baseline real data for calculating our quality and diversity metrics.

4.2 Results

Qualitative comparisons. Fig. 3 compares generated samples on the LSUN-Bedrooms dataset. Notice that both minority generators (*i.e.*, Um and Ye [43] and ours) are more likely to yield low-likelihood features of the dataset (*e.g.*, complex visual attributes [2,39]) compared to a standard ancestral sampling implemented with ADM [9]. An important distinction herein is that our method yields this performance benefit solely with a pretrained model, in contrast to [43] that requires significant resources to train a separate classifier. Fig. 4 exhibits generated samples on another challenging benchmark, ImageNet 256×256. We see similar benefits of our method compared to the baselines, further demonstrating the effectiveness of our sampler on challenging large-scale benchmarks. See the supplementary for samples on CelebA.

Quantitative evaluation. Tab. 1 exhibits quality and diversity evaluations on our focused benchmarks. For the baseline real data, we employ the most unique samples that yield the highest AvgkNN values. Notice that our sampler yields better (or comparable) results than the baseline approaches in all datasets,

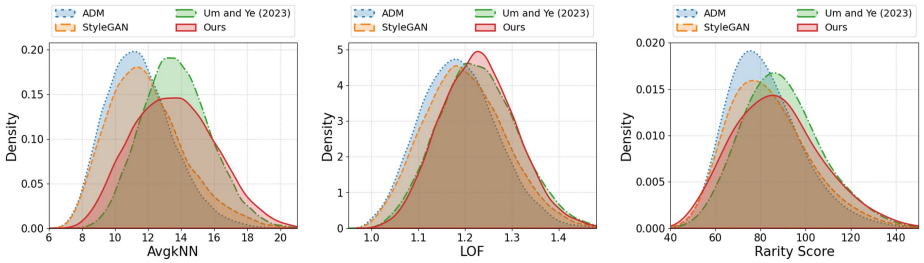

Fig. 5: Comparison of neighborhood density on LSUN-Bedrooms. "AvgkNN" refers to Average k-Nearest Neighbor, and "LOF" is Local Outlier Factor [3]. "Rarity Score" indicates a low-density metric proposed by [15]. The higher values, the less likely samples for all three measures.

Table 2: Ablation study investigating various design elements in our sampler. "Cost" denotes inference time measured in sec/sample. "✗" refers to ours not incorporating the stop gradient. "$\mathtt{sg}(\hat{\boldsymbol{x}}_0)$" and "$\mathtt{sg}(\hat{\tilde{\boldsymbol{x}}}_0)$" indicates the cases with the stop gradient either on $\hat{\boldsymbol{x}}_0$ and $\hat{\tilde{\boldsymbol{x}}}_0$, respectively. "S/O" denotes the case that uses the sudden switch-off: $w_t = w \cdot \mathbb{1}\{t \geq t_{\text{mid}}\}$. "Var" refers to ours using the variance schedule: $w_t = w \cdot \boldsymbol{\Sigma}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$.

(a) Influence of the scale w

w	cFID ↓	Rec ↑
0.0	84.51	0.18
2.0	57.98	0.56
6.0	**34.01**	**0.63**

(b) Impact of using $\mathtt{sg}(\cdot)$

Type	cFID ↓	Cost ↓
✗	**32.09**	2.65
$\mathtt{sg}(\hat{\boldsymbol{x}}_0)$	81.08	2.62
$\mathtt{sg}(\hat{\tilde{\boldsymbol{x}}}_0)$	34.01	**1.75**

(c) Effect of the schedule w_t

Type	cFID ↓	Prec ↑
Fixed	34.01	0.56
S/O	**31.79**	0.57
Var	37.78	**0.87**

demonstrating its ability to produce high-quality diverse samples close to the baseline real minorities. We emphasize that this superior performance persists even with the intermittent technique that can significantly reduce our inference time to a similar extent as existing samplers; see the supplementary for a detailed complexity analysis. We further highlight that the benefit of ours comes solely from a pretrained model, implying the practical importance of our approach.

Neighborhood density results. Fig. 5 compares our focused density measures on LSUN-Bedrooms. Observe that for all three metrics, both our method and [43] greatly improve the capability of generating minorities over ancestral sampling (implemented with ADM [9]), which corroborates the visual inspections made on Fig. 3. However, we highlight that our results were attained only with a pretrained model, which is in contrast to [43] helped by an external classifier to yield the benefit. See the supplementary for density results on other datasets.

Ablation studies. Tab. 2 exhibits ablations on our important design choices. Notice in Tab. 2a that increasing w yields improvements in proximity to real minority instances, validating the role of w as a knob to control the strength of our guidance. Tab. 2b shows the benefit of incorporating the stop-gradient on

$\hat{\tilde{\boldsymbol{x}}}_0$, where we see noticeable gain in the inference time with little compromise in performance. Tab. 2c investigates the impact of time-scheduling strategies. Note that the variance-based scheduling demonstrates significant improvement in sample quality while maintaining respectable cFID performance, providing justification for adopting the variance schedule as our final scheduling strategy. See the supplementary for further analyses and ablations on other parameters.

Downstream classification. To further emphasize the practical significance of our work, we explore a potential application of our sampler. Specifically, we investigate whether our minority-enhanced generated samples could enhance the performance of a classifier trained on a synthetically augmented dataset. We consider the classification of 40 distinct attributes of CelebA and train ResNet-18 models on three different datasets: (i) CelebA trainset; (ii) CelebA + 50K samples from ADM [9]; (iii) CelebA + 50K samples from ours. We incorporated an off-the-shelf classifier for labeling the generated samples. Tab. 3 compares the classification metrics of the three considered cases. Note that the ADM-augmented classifier fails to improve upon the non-augmented case, which we conjecture is due to the limited diversity of the ADM samples, a factor that has been observed to potentially hinder performance improvement [14, 32, 48]. Nonetheless, the classifier complemented with our samples enhances the performance across all metrics, highlighting the benefit of ours for downstream applications.

Table 3: Classification results for different training datasets. All settings were evaluated on the CelebA testset and averaged over three distinct runs.

Training data	F1	Prec	Rec
CelebA trainset	0.746	0.815	0.710
+ ADM gens (50K)	0.742	0.808	0.711
+ Ours gens (50K)	**0.757**	**0.822**	**0.724**

5 Conclusion

We develop a novel framework for generating minority data using diffusion models. Our self-guided sampler, based on our new minority metric, optimizes the generation process of diffusion models to evolve towards low-likelihood minority features. We additionally provide several techniques to further improve the complexity and fidelity due to our sampler. Extensive experiments across real data benchmarks demonstrate significant improvements over existing minority samplers. Importantly, the benefits stem solely from a pretrained diffusion model, distinguishing our approach from existing frameworks requiring additional components to improve the minority-generating capability over standard samplers.

Limitation and potential negative impact. One disadvantage is that the proposed sampler introduces additional inference costs compared to standard samplers. A potential concern is the misuse of our sampler to intentionally suppress the generation of minority-featured samples. This malicious use could be realized by employing negative values of w in Eq. (9), directing the focus towards producing instances dominated by high-likelihood majority features. It is crucial to acknowledge and address this risk, emphasizing the need for responsible usage of our framework to uphold fairness and inclusivity in generative modeling.

Acknowledgments

This work was partly supported by the National Research Foundation of Korea under Grant (No. RS-2024-00336454), by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT)) (No. RS-2022-II220984, Development of Artificial Intelligence Technology for Personalized Plug-and-Play Explanation and Verification of Explanation), by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No. RS-2019-II190075 Artificial Intelligence Graduate School Program (KAIST)), by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2023, and by Field-oriented Technology Development Project for Customs Administration funded by the Korea government (the Ministry of Science & ICT and the Korea Customs Service) through the National Research Foundation (NRF) of Korea under Grant NRF2021M3I1A1097910.

References

1. Amini, A., Soleimany, A.P., Schwarting, W., Bhatia, S.N., Rus, D.: Uncovering and mitigating algorithmic bias through learned latent structure. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. pp. 289–295 (2019)
2. Arvinte, M., Cornelius, C., Martin, J., Himayat, N.: Investigating the adversarial robustness of density estimation using the probability flow ode. arXiv preprint arXiv:2310.07084 (2023)
3. Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J.: Lof: identifying density-based local outliers. In: Proceedings of the 2000 ACM SIGMOD international conference on Management of data. pp. 93–104 (2000)
4. Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096 (2018)
5. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15750–15758 (2021)
6. Choi, J., Lee, J., Shin, C., Kim, S., Kim, H., Yoon, S.: Perception prioritized training of diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11472–11481 (2022)
7. Chung, H., Kim, J., Mccann, M.T., Klasky, M.L., Ye, J.C.: Diffusion posterior sampling for general noisy inverse problems. arXiv preprint arXiv:2209.14687 (2022)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
9. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
10. Du, X., Sun, Y., Zhu, X., Li, Y.: Dream the impossible: Outlier imagination with diffusion models. In: Advances in Neural Information Processing Systems (2023)
11. Du, X., Wang, Z., Cai, M., Li, Y.: Vos: Learning what you don't know by virtual outlier synthesis. arXiv preprint arXiv:2202.01197 (2022)

12. Epstein, D., Jabri, A., Poole, B., Efros, A.A., Holynski, A.: Diffusion self-guidance for controllable image generation. arXiv preprint arXiv:2306.00986 (2023)
13. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
14. Gowal, S., Rebuffi, S.A., Wiles, O., Stimberg, F., Calian, D.A., Mann, T.A.: Improving robustness using generated data. Advances in Neural Information Processing Systems **34**, 4218–4233 (2021)
15. Han, J., Choi, H., Choi, Y., Kim, J., Ha, J.W., Choi, J.: Rarity score: A new metric to evaluate the uncommonness of synthesized images. arXiv preprint arXiv:2206.08549 (2022)
16. Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., Song, D.: Natural adversarial examples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15262–15271 (2021)
17. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
18. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
19. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
20. Hong, S., Lee, G., Jang, W., Kim, S.: Improving sample quality of diffusion models using self-attention guidance. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7462–7471 (2023)
21. Huang, G., Jafari, A.H.: Enhanced balancing gan: Minority-class image generation. Neural computing and applications **35**(7), 5145–5154 (2023)
22. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. Advances in Neural Information Processing Systems **35**, 26565–26577 (2022)
23. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019)
24. Kim, D., Shin, S., Song, K., Kang, W., Moon, I.C.: Soft truncation: A universal training technique of score-based diffusion model for high precision score estimation. arXiv preprint arXiv:2106.05527 (2021)
25. Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., Aila, T.: Improved precision and recall metric for assessing generative models. Advances in Neural Information Processing Systems **32** (2019)
26. Lin, Z., Liang, H., Fanti, G., Sekar, V., Sharma, R.A., Soltanaghaei, E., Rowe, A., Namkung, H., Liu, Z., Kim, D., et al.: Raregan: Generating samples for rare classes. arXiv preprint arXiv:2203.10674 (2022)
27. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV) (December 2015)
28. Nash, C., Menick, J., Dieleman, S., Battaglia, P.W.: Generating images with sparse representations. arXiv preprint arXiv:2103.03841 (2021)
29. Parmar, G., Zhang, R., Zhu, J.Y.: On aliased resizing and surprising subtleties in gan evaluation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11410–11420 (2022)
30. Peebles, W., Xie, S.: Scalable diffusion models with transformers. arXiv preprint arXiv:2212.09748 (2022)

31. Qin, Y., Zheng, H., Yao, J., Zhou, M., Zhang, Y.: Class-balancing diffusion models. arXiv preprint arXiv:2305.00562 (2023)
32. Ravuri, S., Vinyals, O.: Classification accuracy score for conditional generative models. Advances in neural information processing systems **32** (2019)
33. Robbins, H.E.: An empirical bayes approach to statistics. In: Breakthroughs in statistics, pp. 388–394. Springer (1992)
34. Roh, Y., Nie, W., Huang, D.A., Whang, S.E., Vahdat, A., Anandkumar, A.: Dr-fairness: Dynamic data ratio adjustment for fair training on real and generated data. Transactions on Machine Learning Research (2023)
35. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
36. Sadat, S., Buhmann, J., Bradely, D., Hilliges, O., Weber, R.M.: Cads: Unleashing the diversity of diffusion models through condition-annealed sampling. arXiv preprint arXiv:2310.17347 (2023)
37. Samuel, D., Ben-Ari, R., Raviv, S., Darshan, N., Chechik, G.: It is all about where you start: Text-to-image generation with seed selection. arXiv preprint arXiv:2304.14530 (2023)
38. Sehwag, V., Hazirbas, C., Gordo, A., Ozgenel, F., Canton, C.: Generating high fidelity data from low-density regions using diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11492–11501 (2022)
39. Serrà, J., Álvarez, D., Gómez, V., Slizovskaia, O., Núñez, J.F., Luque, J.: Input complexity and out-of-distribution detection with likelihood-based generative models. arXiv preprint arXiv:1909.11480 (2019)
40. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning. pp. 2256–2265. PMLR (2015)
41. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. Advances in Neural Information Processing Systems **32** (2019)
42. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
43. Um, S., Ye, J.C.: Don't play favorites: Minority guidance for diffusion models. arXiv preprint arXiv:2301.12334 (2023)
44. Vincent, P.: A connection between score matching and denoising autoencoders. Neural computation **23**(7), 1661–1674 (2011)
45. Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., Xiao, J.: Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv preprint arXiv:1506.03365 (2015)
46. Yu, N., Li, K., Zhou, P., Malik, J., Davis, L., Fritz, M.: Inclusive gan: Improving data and minority coverage in generative models. In: European Conference on Computer Vision. pp. 377–393. Springer (2020)
47. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
48. Zhao, B., Bilen, H.: Synthesizing informative training samples with gan. arXiv preprint arXiv:2204.07513 (2022)