This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Minority-Focused Text-to-Image Generation via Prompt Optimization

Soobin Um Jong Chul Ye
Graduate School of AI
KAIST, Daejeon, Republic of Korea
{sum,jong.ye}@kaist.ac.kr

Figure 1. **Example results from our minority generation approach using SDXL-Lightning.** Our framework is designed to produce unique *minority* samples w.r.t. user-provided prompts, which are rarely generated by standard samplers like DDIM [46]. Due to its low-likelihood encouraging nature, our sampler often demonstrates counteracting results against demographic biases in text-to-image models [13]. See the samples in the last row for instance, where our sampler mitigates prevalent age and racial biases (*e.g.*, associating "man" with "young" and "woman" with "white") by modifying the demographic traits of the subjects.

Abstract

We investigate the generation of minority samples using pretrained text-to-image (T2I) latent diffusion models. Minority instances, in the context of T2I generation, can be defined as ones living on low-density regions of text-conditional data distributions. They are valuable for various applications of modern T2I generators, such as data augmentation and creative AI. Unfortunately, existing pretrained T2I diffusion models primarily focus on high-density regions, largely due to the influence of guided samplers (like CFG) that are essential for high-quality generation. To address this, we present a novel framework to counter the high-density-focus of T2I diffusion models. Specifically, we first develop an online prompt optimization framework that encourages emergence of desired properties during inference while preserving semantic contents of user-provided prompts. We subsequently tailor this generic prompt optimizer into a specialized solver that promotes generation of minority features by incorporating a carefully-crafted likelihood objective. Extensive experiments conducted across various types of T2I models demonstrate that our approach significantly enhances the capability to produce high-quality minority instances compared to existing samplers. Code is available at https://github.com/soobin-um/MinorityPrompt.

1. Introduction

Text-to-image (T2I) generative models [35, 40, 54] have recently attracted substantial interest for their capability to convert textual descriptions into visually striking images. At the forefront of the surge are diffusion models [20, 47], augmented by guidance techniques [11, 19] such as classifier-free guidance (CFG) [19]. The guided T2I samplers encourage generations from high-density regions of a data manifold [11], producing realistic images that faithfully respect the provided prompts.

A key challenge is that the inherent high density focus of modern T2I samplers makes it difficult to generate *minority* samples – instances that reside in low-density regions of the manifold. This limitation is particularly significant as T2I-generated data is increasingly incorporated in downstream applications [1, 49, 50] where the majority-focused bias within the data may be perpetuated. Furthermore, the unique attributes found in minority instances are crucial for applications like creative AI [16, 41], where generating novel and highly creative outputs is essential.

In this work, we present a novel approach dubbed as *MinorityPrompt* that counteracts the high-density bias of T2I samplers to improve their capability of minority generation. Our framework is built upon the concept of *prompt optimization*, an intuitive technique that exhibits strong performance in enhancing T2I diffusion models for various tasks [6, 14, 36]. Unlike existing T2I-based online prompt-tuning methods that modify the entire input prompts (*e.g.*, by optimizing their text-embeddings during inference), our approach updates the prompts in a *selective* fashion to preserve the intended semantics while encouraging generations of unique low-density features.

Specifically during inference, we incorporate learnable tokens into the input prompts, *e.g.*, by appending them to the end of the text. The embeddings of these tokens are iteratively refined across sampling timesteps to optimize the proposed objective for minority generation, which approximates the likelihood of noisy intermediate samples in T2I generation. See Fig. 2 for an overview. We highlight that our prompt optimization framework is versatile, *i.e.*, it can be applied to various tasks with distinct optimization objectives beyond minority generation.

Comprehensive experiments validate that our method can significantly improve the ability of creating minority instances of modern widely-adopted T2I models (including Stable Diffusion (SD) [41]) with minimal compromise in sample quality and text-image alignment. In addition, we emphasize that our framework can work on distilled backbones like SDXL-Lightning [29], which demonstrates its robustness and practical relevance. As an additional application, we explore the potential of our prompt optimization framework to improve the diversity of T2I models, further exhibiting its versatility as a general-purpose solver applicable across various tasks.

Our key contributions are summarized as follows:

- We propose a token-based online prompt optimization framework that iteratively updates learnable tokens during inference, achieving superior text-image alignment over existing online prompt tuners.
- We develop a novel objective for minority sampling in the T2I context, which mathematically approximates the target log-likelihood in T2I generation.
- We empirically demonstrate that our approach achieves state-of-the-art performance in generating minority samples for T2I generation.

2. Related Work

The generation of minority samples has been explored in a range of different scenarios and generative frameworks [22, 31, 38, 44, 51, 52, 56]. However, significant progress has been recently made with the introduction of diffusion models, due to their ability to faithfully capture data distributions [44, 51, 52]. As an initial effort, [44] incorporate separately-trained classifiers into the sampling process of diffusion models to yield guidance for low-density regions. The approach by [51] shares similar intuition of integrating an additional classifier into the reverse process for low-density guidance. A limitation is that their methods rely upon external classifiers that are often difficult to obtain, especially for large-scale datasets such as T2I benchmarks [43]. The challenge was recently addressed by [52] where the authors develop a self-contained minority sampler that works without expensive extra components (such as classifiers). However, their method is tailored for canonical image benchmarks (like LSUN [55] and ImageNet [9]) and exhibits limited performance gain in more challenging scenarios like T2I generation.

A related yet distinct objective is enhancing the diversity of diffusion models, an area that has been relatively overlooked compared to improving their quality. Significant progress was recently made in [42], where the authors demonstrated that adding noise perturbations, if gradually annealed over time, to conditional embeddings could greatly enhance the diversity of generated samples. However, unlike our approach, their method focuses on producing diverse samples that remain consistent with the ground-truth data distribution, rather than targeting the low-density regions of the distribution. Another notable contribution was done by [8]. Their idea is to repel intermediate latent samples that share the same condition, thereby encouraging the final generated samples to exhibit distinct features. A disadvantage is that it requires generating multiple instances for each prompt, which can be redundant in many practical scenarios.

Prompt optimization has been widely explored in the context of T2I diffusion models due to their strong depen-

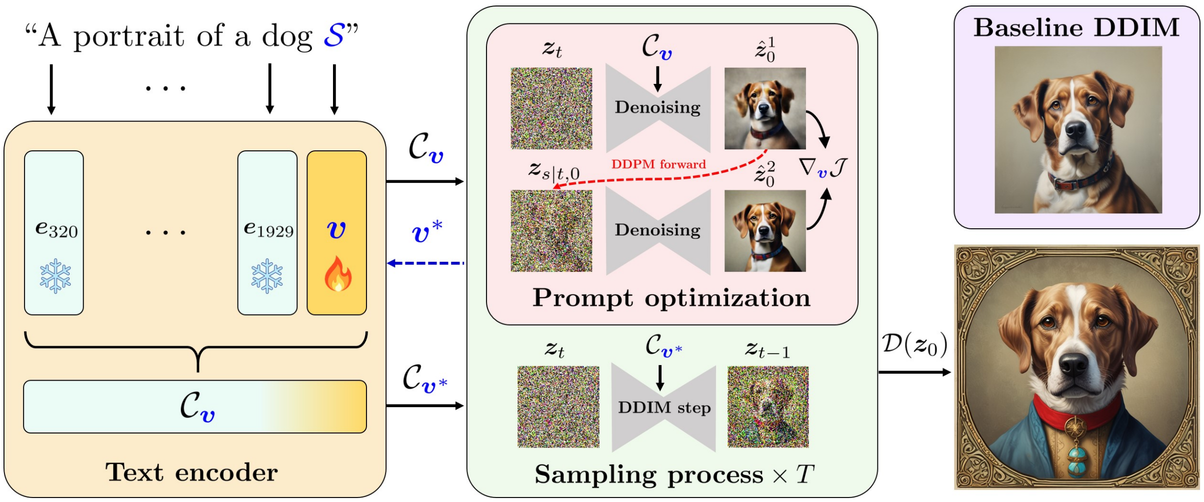

Figure 2. **Overview of MinorityPrompt.** Unlike existing online prompt tuning approaches that adjust the entire text-embedding (*e.g.*, the output of the text-encoder) during inference, our framework focuses on optimizing a dedicated *token-embedding* to better preserve the semantics within the prompt. Specifically given a user-prompt (*e.g.*, "A portrait of a dog"), we integrate a placeholder string (*e.g.*, $\mathcal{S}$ in the figure) into the prompt, marking the position of the learnable token embedding $\boldsymbol{v}$. With the text-embedding $\mathcal{C}_{\boldsymbol{v}}$ that incorporates the contents of $\boldsymbol{v}$, we update $\boldsymbol{v}$ *on-the-fly* during the inference process to maximize the reconstruction loss of the denoised version of $\boldsymbol{z}_t$ (i.e., $\hat{\boldsymbol{z}}_0^1$ in the figure). The optimized token $\boldsymbol{v}^*$ is subsequently used to progress the inference at the corresponding timestep; see Sec. 3 for details.

dence on language models. This approach has exhibited significant performance across various tasks, including inverse problems [6] and image editing [33, 36]. A key difference is that most existing methods in these lines tune the *entire* prompts to find the ones that best perform the focused tasks (*e.g.*, minimizing data consistency loss [6]). In contrast, our framework updates only the attached learnable tokens, thereby preserving the original prompt's semantics while encouraging the emergence of low-density features. Additional use cases of prompt tuning include personalization [14, 15] and object counting [57]. Similar to ours, their frameworks introduce variable tokens and tune their embeddings. However, their optimizations aim to learn visual concepts captured in user-provided images, whereas our focus is to invoke low-density features through optimized prompts. Also, their methods are not online, requiring separate training procedure which can be potentially expensive.

3. Method

Our focus is to generate high-quality minority instances using text-to-image (T2I) diffusion models, which faithfully reflect user-provided prompts while featuring unique visual attributes rarely produced via standard generation techniques[1]. To this end, we start with providing a brief overview on T2I diffusion frameworks and the essential background necessary to understand the core of our work. We subsequently present our proposed framework for minority generation based on the idea of prompt optimization.

3.1. Background and preliminaries

The task of T2I diffusion models is to generate an output image $\boldsymbol{x}_0 \in \mathbb{R}^d$ from a random noise vector $\boldsymbol{z}_T \in \mathbb{R}^k$ (where typically $k < d$), given a user-defined text prompt $\mathcal{P}$. Similar to standard (non-T2I) diffusion frameworks, the core of T2I diffusion sampling lies in an iterative denoising process that progressively removes noise from $\boldsymbol{z}_T$ until a clean version $\boldsymbol{z}_0$ is obtained. This denoising capability is learned through noise-prediction training [20, 47], mathematically written as:

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{z}_0, y, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), t \sim \text{Unif}\{1,\ldots,T\}}[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{z}_t, \mathcal{C})\|_2^2],$$

where $\boldsymbol{z}_0 := \mathcal{E}(\boldsymbol{x}_0)$, yielded by passing a training image $\boldsymbol{x}_0$ through a compressive model $\mathcal{E}$ (*e.g.*, the encoder of VQ-VAE [12, 41]). Here, $\boldsymbol{z}_t$ represents a noise-perturbed version of $\boldsymbol{z}_0$, given by $\boldsymbol{z}_t := \sqrt{\alpha_t}\boldsymbol{z}_0 + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}$, where $\{\alpha_t\}_{t=1}^T$ defines the noise-schedule. $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$ refers to a T2I diffusion model parameterized to predict the noise $\boldsymbol{\epsilon}$, and $\mathcal{C}$ represents the embedding of the text prompt $\mathcal{P}$. See below for details on how to obtain $\mathcal{C}$ from $\mathcal{P}$.

[1] More formally, this can be expressed as drawing instances from $\mathcal{S}_{\boldsymbol{c}} := \{\boldsymbol{z} \in \mathcal{M}_{\boldsymbol{c}} : p_{\boldsymbol{\theta}}(\boldsymbol{z}|\mathcal{C}) < \epsilon\}$, where $\mathcal{C}$ is the prompt, $\mathcal{M}_c$ represents the (latent) data manifold associated with $\mathcal{C}$, and $p_{\boldsymbol{\theta}}$ denotes the probability density captured by the T2I diffusion model. Here ϵ is a small positive constant.

Once trained, T2I generation can be done by starting from $\boldsymbol{z}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and implementing an iterative noise removal process guided by the text embedding $\mathcal{C}$. A common approach is to follow the deterministic DDIM sampling [5, 46]:

$$\boldsymbol{z}_{t-1} = \sqrt{\alpha_{t-1}}\hat{\boldsymbol{z}}_0(\boldsymbol{z}_t, \mathcal{C}) + \sqrt{1-\alpha_{t-1}}\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{z}_t, \mathcal{C}), \quad (1)$$

where $\hat{\boldsymbol{z}}_0(\boldsymbol{z}_t, \mathcal{C}) := (\boldsymbol{z}_t - \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{z}_t, \mathcal{C}))/\sqrt{\alpha_t}$. Here $\hat{\boldsymbol{z}}_0(\boldsymbol{z}_t, \mathcal{C})$ indicates a denoised estimate of $\boldsymbol{z}_t$ conditioned on the text embedding $\mathcal{C}$, implemented via Tweedie's formula [4].

To further strengthen the impact of text conditioning, classifier-free guidance (CFG) [19] is commonly integrated into the sampling process. In particular, one can obtain a high-density-focused noise estimation through extrapolation using an unconditional prediction:

$$\tilde{\boldsymbol{\epsilon}}^w_{\boldsymbol{\theta}}(\boldsymbol{z}_t, \mathcal{C}) := w\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{z}_t, \mathcal{C}) + (1-w)\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{z}_t), \quad (2)$$

where $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{z}_t)$ indicates an unconditional noise prediction, often implemented via null-text conditioning [19]. CFG refers to the technique that employs $\tilde{\boldsymbol{\epsilon}}^w_{\boldsymbol{\theta}}(\boldsymbol{z}_t, \mathcal{C})$ in place of $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{z}_t, \mathcal{C})$ (in Eq. (1)), which has been shown in various scenarios to significantly improve both sample quality and text alignment yet at the expense of diversity [42].

Text processing. A key distinction from non-T2I diffusion models is the incorporation of the text embedding $\mathcal{C}$, a continuous vector yielded by a text encoder $\mathcal{T}$ (such as BERT [10]) based on the user prompt $\mathcal{P}$. To obtain this embedding, each word (or sub-word) in $\mathcal{P}$ is first converted into a token – an index in a pre-defined vocabulary. Each token is then mapped to a unique embedding vector through an index-based lookup. These token-wise embedding vectors, often referred to as *token* embeddings, are typically learned as part of the text encoder. The token embeddings are then passed through a transformer model, yielding the final text embedding $\mathcal{C}$. For simplicity, we denote this text processing operation as the forward pass of the text encoder $\mathcal{T}$; thus, $\mathcal{C} = \mathcal{T}(\mathcal{P})$.

Prompt optimization. In the context of T2I diffusion models, prompt tuning is performed by intervening in the text-processing stage. A common approach is to adjust the text embedding $\mathcal{C}$ over inference time, which is widely adopted in existing online prompt optimizers [6, 36]. Specifically at sampling timestep t, existing online prompt tuners can be formulated as the following optimization problem:

$$\mathcal{C}^*_t := \arg\max_{\mathcal{C}} \mathcal{J}(\boldsymbol{z}_t, \mathcal{C}), \quad (3)$$

where $\boldsymbol{z}_t$ is a noisy latent at step t, and $\mathcal{J}$ represents a task-specific objective function, such as data consistency in inverse problems [6]. Once $\mathcal{C}^*_t$ is obtained, it is used as a drop-in replacement for $\mathcal{C}$ at time t (*e.g.*, in Eq. (1)), encouraging the desired property to manifest in subsequent timesteps.

A problem is that the optimization in Eq. (3) may lead to a loss of user-intended semantics in $\mathcal{P}$, due to the comprehensive updating of the entire text-embedding $\mathcal{C}$. This is critical, especially in the context of our focused T2I minority generation where preserving prompt semantics is essential; see the supplementary for our empirical results that support this. One can resort to tuning the null-text embedding while keeping $\mathcal{C}$ intact (as suggested by [33]). However, this method requires reserving the null-text dimension for this specific purpose, limiting its potential use for improving sample quality or serving other functions. In the following section, we present an online prompt optimization framework designed to better preserve semantics. Building on this foundation, we develop our T2I minority sampler, which promotes the generation of minority features while maintaining both sample quality and text-alignment performance.

3.2. Semantic-preserving prompt optimization

The key idea of our optimization approach is to incorporate learnable tokens into a given prompt $\mathcal{P}$ and update its embedding *on-the-fly* during inference. Specifically, we append a placeholder string[2] $\mathcal{S}$ to the prompt $\mathcal{P}$, which acts as a mark for the learnable tokens. For instance, the augmented prompt could be $\mathcal{P}_{\mathcal{S}} :=$ "A portrait of a dog $\mathcal{S}$". This additional string is treated as a new vocabulary item for the text-encoder $\mathcal{T}$. We assign a token embedding $\boldsymbol{v}$ to $\mathcal{S}$, and denote the text encoder incorporating it as $\mathcal{T}(\,\cdot\,; \boldsymbol{v})$.

We propose optimizing this embedding $\boldsymbol{v}$ rather than $\mathcal{C}$. The proposed online prompt optimization at sampling step t can then be formalized as follows:

$$\boldsymbol{v}^*_t := \arg\max_{\boldsymbol{v}} \mathcal{J}(\boldsymbol{z}_t, \mathcal{C}_{\boldsymbol{v}}), \quad (4)$$

where $\mathcal{C}_{\boldsymbol{v}} := \mathcal{T}(\mathcal{P}_{\mathcal{S}}; \boldsymbol{v})$. Afterward, the optimized text-embedding $\mathcal{C}_{\boldsymbol{v}^*_t}$ is obtained by text-processing $\mathcal{P}_{\mathcal{S}}$ with the updated token-embedding of $\mathcal{S}$, therefore $\mathcal{C}_{\boldsymbol{v}^*_t} := \mathcal{T}(\mathcal{P}_{\mathcal{S}}; \boldsymbol{v}^*_t)$.

Note that our optimization does not affect the embeddings of the tokens w.r.t. the original prompt $\mathcal{P}$. This is inherently more advantageous for preserving semantics compared to existing methods, which alter the entire text-embedding $\mathcal{C}$ and thereby effectively impact all token embeddings. We also highlight that unlike existing learnable-token-based approaches that share the same embedding throughout inference [14, 15, 57], our framework allows the token embedding $\boldsymbol{v}$ to change over timesteps t. This adaptive feature offers potential advantages, since the role of $\boldsymbol{v}$ in maximizing $\mathcal{J}$ can vary with $\boldsymbol{z}_t$ that changes over timesteps. This point is also implied in previous works that employ adaptive text-embeddings over time [6, 36].

[2]The placeholder string can be placed at any position in the prompt, but we empirically found that inserting it at the end of the prompt yields the best performance; see the supplementary for details.

Algorithm 1 MinorityPrompt

Require: $\epsilon_{\theta}, \mathcal{T}, \mathcal{D}, \boldsymbol{v}_T^{(0)}, \mathcal{P}_{\mathcal{S}}, \mathcal{C}, N, K, w, T, s, \lambda$.

$\boldsymbol{z}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
for $t \leftarrow T$ to 1 **do**
  $\mathcal{C}_{\boldsymbol{v}_t^*} \leftarrow \mathcal{C}$
  if $t \bmod N = 0$ **then**
    $\boldsymbol{v}_t^* \leftarrow \textsc{OptimizeEmb}(\boldsymbol{z}_t, \boldsymbol{v}_t^{(0)}, \epsilon_{\theta}, \mathcal{T}, K, s, \lambda)$
    $\mathcal{C}_{\boldsymbol{v}_t^*} \leftarrow \mathcal{T}(\mathcal{P}_{\mathcal{S}}; \boldsymbol{v}_t^*)$
  end if
  $\tilde{\epsilon}_{\theta}^w \leftarrow w\epsilon_{\theta}(\boldsymbol{z}_t, \mathcal{C}_{\boldsymbol{v}_t^*}) + (1-w)\epsilon_{\theta}(\boldsymbol{z}_t)$
  $\hat{\boldsymbol{z}}_0^w \leftarrow (\boldsymbol{z}_t - \sqrt{1-\alpha_t}\tilde{\epsilon}_{\theta}^w)/\sqrt{\alpha_t}$
  $\boldsymbol{z}_{t-1} \leftarrow \sqrt{\alpha_{t-1}}\hat{\boldsymbol{z}}_0^w + \sqrt{1-\alpha_{t-1}}\tilde{\epsilon}_{\theta}^w$
  $\boldsymbol{v}_{t-1}^{(0)} \leftarrow \boldsymbol{v}_t^*$
end for
return $\boldsymbol{x}_0 \leftarrow \mathcal{D}(\boldsymbol{z}_0)$

Algorithm 2 Prompt optimization

function $\textsc{OptimizeEmb}(\boldsymbol{z}_t, \boldsymbol{v}_t^{(0)}, \epsilon_{\theta}, \mathcal{T}, K, s, \lambda)$
  for $k \leftarrow 1$ to K **do**
    $\mathcal{C}_{\boldsymbol{v}} \leftarrow \mathcal{T}(\mathcal{P}_{\mathcal{S}}; \boldsymbol{v}_t^{(k-1)})$
    $\epsilon_{\theta}^1 \leftarrow \epsilon_{\theta}(\boldsymbol{z}_t, \mathcal{C}_{\boldsymbol{v}})$
    $\hat{\boldsymbol{z}}_0^1 \leftarrow (\boldsymbol{z}_t - \sqrt{1-\alpha_t}\epsilon_{\theta}^1)/\sqrt{\alpha_t}$
    $\epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
    $\boldsymbol{z}_{s|t,0} \leftarrow \sqrt{\alpha_s}\hat{\boldsymbol{z}}_0^1 + \sqrt{1-\alpha_s}\epsilon$
    $\epsilon_{\theta}^2 \leftarrow \epsilon_{\theta}(\boldsymbol{z}_{s|t,0}, \mathcal{C})$
    $\hat{\boldsymbol{z}}_0^2 \leftarrow (\boldsymbol{z}_{s|t,0} - \sqrt{1-\alpha_s}\epsilon_{\theta}^2)/\sqrt{\alpha_s}$
    $\mathcal{J}_t \leftarrow \|\hat{\boldsymbol{z}}_0^1 - \mathtt{sg}(\hat{\boldsymbol{z}}_0^2)\|_2^2 + \lambda\|\mathtt{sg}(\hat{\boldsymbol{z}}_0^1) - \hat{\boldsymbol{z}}_0^2\|_2^2$
    $\boldsymbol{v}_t^{(k)} \leftarrow \boldsymbol{v}_t^{(k-1)} + \mathtt{AdamGrad}(\mathcal{J}_t)$
  end for
  return $\boldsymbol{v}_t^* \leftarrow \boldsymbol{v}_t^{(K)}$
end function

Intuitively, our optimization can be understood as capturing a specific concept relevant to noisy latent $\boldsymbol{z}_t$ within the token $\boldsymbol{v}_t^*$, guided by the objective function $\mathcal{J}$. Thanks to its general design that accommodates any arbitrary objective function $\mathcal{J}$, this framework is versatile and can be employed in various contexts beyond minority generation. For instance, it can be used to diversify the outputs of T2I models. See details in Tab. 3.

3.3. MinorityPrompt: minority-focused prompt tuning

We now specialize the generic solver in Eq. (4) for the task of minority generation. The key question is how to formulate an appropriate objective function $\mathcal{J}$ for this purpose. To address this, we draw inspiration from Um and Ye [52], employing their likelihood metric as the starting point for developing our objective function.

Since the metric was originally defined in the pixel domain using non-T2I diffusion models (see the supplementary for details), we initially perform a naive adaptation to accommodate the latent space of interest, $\boldsymbol{z}_t \in \mathbb{R}^k$, and integrate text conditioning using CFG as is typical in the T2I context [23]. The adapted version of the metric reads:

$$\mathcal{J}(\boldsymbol{z}_t, \mathcal{C}) := \mathbb{E}_{\epsilon}\big[\|\hat{\boldsymbol{z}}_0^w(\boldsymbol{z}_t, \mathcal{C}) - \mathtt{sg}(\hat{\boldsymbol{z}}_0^w(\boldsymbol{z}_{s|t,0}^w, \mathcal{C}))\|_2^2\big], \quad (5)$$

where $\hat{\boldsymbol{z}}_0^w(\boldsymbol{z}_t, \mathcal{C})$ represents a clean estimate of $\boldsymbol{z}_t$ using the CFG noise term $\tilde{\epsilon}_{\theta}^w(\boldsymbol{z}_t, \mathcal{C})$ (in Eq. (2)). Here $\boldsymbol{z}_{s|t,0}^w$ indicates a noised version of $\hat{\boldsymbol{z}}_0^w(\boldsymbol{z}_t, \mathcal{C})$ w.r.t. timestep s: $\boldsymbol{z}_{s|t,0}^w := \sqrt{\alpha_s}\hat{\boldsymbol{z}}_0^w(\boldsymbol{z}_t, \mathcal{C}) + \sqrt{1-\alpha_s}\epsilon$, and $\hat{\boldsymbol{z}}_0^w(\boldsymbol{z}_{s|t,0}^w, \mathcal{C})$ is a clean version of $\boldsymbol{z}_{s|t,0}^w$ conditioned on $\mathcal{C}$. $\mathtt{sg}(\cdot)$ denotes the stop-gradient operator for reducing computational cost when used in guided sampling [52]. Notice that the squared L2 error is used as the discrepancy loss, rather than the originally used LPIPS [58], due to its incompatibility with our latent space. The quantity in Eq. (5) is interpretable as a reconstruction loss of $\hat{\boldsymbol{z}}_0^w(\boldsymbol{z}_t, \mathcal{C})$. As exhibited in Um and Ye [52], the loss may become large if $\boldsymbol{z}_t$ (represented by $\hat{\boldsymbol{z}}_0^w(\boldsymbol{z}_t, \mathcal{C})$) contains highly-unique minority features that often vanish during the reconstruction process. The comprehensive details regarding the original metric due to Um and Ye [52] are provided in the supplementary.

Considering Eq. (5) as the objective function, a natural approach for minority-focused prompt tuning would be to incorporate $\mathcal{C}_{\boldsymbol{v}}$ and optimize for the best $\boldsymbol{v}$:

$$\boldsymbol{v}_t^* := \arg\max_{\boldsymbol{v}} \mathcal{J}(\boldsymbol{z}_t, \mathcal{C}_{\boldsymbol{v}}). \quad (6)$$

However, we argue that this naively extended framework has theoretical issues that lead to limited performance gain over standard samplers. Specifically, three aspects of this objective weaken the desired connection to the target log-likelihood $\log p_{\theta}(\boldsymbol{z}_0|\mathcal{C})$ that we aim to capture: (i) the reliance on the CFG-based clean predictions $\hat{\boldsymbol{z}}_0^w$; (ii) obstructed gradient flow through the second term in the squared L2 loss due to $\mathtt{sg}(\cdot)$; and (iii) the incorporation of $\mathcal{C}_{\boldsymbol{v}}$ within the second term in the loss. See the supplementary on a detailed analysis on these points.

Hence, we propose the following optimization to address the theoretical issues:

$$\boldsymbol{v}_t^* := \arg\max_{\boldsymbol{v}} \mathcal{J}_{\mathcal{C}}(\boldsymbol{z}_t, \mathcal{C}_{\boldsymbol{v}})$$
$$\text{where } \mathcal{J}_{\mathcal{C}}(\boldsymbol{z}_t, \mathcal{C}_{\boldsymbol{v}}) := \mathbb{E}_{\epsilon}\big[\|\hat{\boldsymbol{z}}_0(\boldsymbol{z}_t, \mathcal{C}_{\boldsymbol{v}}) - \hat{\boldsymbol{z}}_0(\boldsymbol{z}_{s|t,0}, \mathcal{C})\|_2^2\big]. \quad (7)$$

Here $\hat{\boldsymbol{z}}_0(\boldsymbol{z}_t, \mathcal{C}_{\boldsymbol{v}}) := (\boldsymbol{z}_t - \sqrt{1-\alpha_t}\epsilon_{\theta}(\boldsymbol{z}_t, \mathcal{C}_{\boldsymbol{v}}))/\sqrt{\alpha_t}$, indicating a non-CFG clean estimate. $\boldsymbol{z}_{s|t,0}$ is a perturbed version of $\hat{\boldsymbol{z}}_0(\boldsymbol{z}_t, \mathcal{C}_{\boldsymbol{v}})$ w.r.t. timestep s: $\boldsymbol{z}_{s|t,0} := \sqrt{\alpha_s}\hat{\boldsymbol{z}}_0(\boldsymbol{z}_t, \mathcal{C}_{\boldsymbol{v}}) + \sqrt{1-\alpha_s}\epsilon$. Notice that this formulation eliminates the problematic components in Eq. (6): $\hat{\boldsymbol{z}}_0^w$, $\mathtt{sg}(\cdot)$ and $\mathcal{C}_{\boldsymbol{v}}$ in the second term. We found that the proposed optimization maintains a close connection to the focused log-likelihood. Below we provide a formal statement of our finding. See the supplementary for the proof.

Proposition 1. *The objective function in Eq. (7) is equivalent (upto a constant factor) to the negative ELBO w.r.t.* $\log p_{\boldsymbol{\theta}}(\hat{\boldsymbol{z}}_0(\boldsymbol{z}_t, \mathcal{C}_{\boldsymbol{v}}) \mid \mathcal{C})$ *when integrated over timesteps with* $\bar{w}_s := \alpha_s/(1-\alpha_s)$*:*

$$\sum_{s=1}^{T} \bar{w}_s \mathcal{J}_{\mathcal{C}}(\boldsymbol{z}_t, \mathcal{C}_{\boldsymbol{v}}) = \sum_{s=1}^{T} \mathbb{E}_{\boldsymbol{\epsilon}}[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{z}_{s|t,0}, \mathcal{C})\|_2^2] \\ \gtrapprox -\log p_{\boldsymbol{\theta}}(\hat{\boldsymbol{z}}_0(\boldsymbol{z}_t, \mathcal{C}_{\boldsymbol{v}}) \mid \mathcal{C}),$$

where $\boldsymbol{z}_{s|t,0} := \sqrt{\alpha_s}\hat{\boldsymbol{z}}_0(\boldsymbol{z}_t, \mathcal{C}_{\boldsymbol{v}}) + \sqrt{1-\alpha_s}\boldsymbol{\epsilon}$.

Intuitively, our optimization seeks to make the text-conditioned clean view $\hat{\boldsymbol{z}}_0(\boldsymbol{z}_t, \mathcal{C}_{\boldsymbol{v}})$ of the current sample $\boldsymbol{z}_t$ as unique as possible, from the perspective of the log-likelihood $\log p_{\boldsymbol{\theta}}(\hat{\boldsymbol{z}}_0(\boldsymbol{z}_t, \mathcal{C}_{\boldsymbol{v}})|\mathcal{C})$.

Techniques for improvement. In practice, we found that our optimization could be further stabilized by introducing a `sg`-related trick into the objective function:

$$\begin{aligned} \tilde{\mathcal{J}}_{\mathcal{C}} &:= \mathcal{J}_{\mathcal{C}}^1 + \lambda \mathcal{J}_{\mathcal{C}}^2, \quad \lambda > 0 \\ \text{where} \quad \mathcal{J}_{\mathcal{C}}^1 &:= \mathbb{E}_{\boldsymbol{\epsilon}}\left[\left\|\hat{\boldsymbol{z}}_0(\boldsymbol{z}_t, \mathcal{C}_{\boldsymbol{v}}) - \mathtt{sg}\left(\hat{\boldsymbol{z}}_0(\boldsymbol{z}_{s|t,0}, \mathcal{C})\right)\right\|_2^2\right] \\ \mathcal{J}_{\mathcal{C}}^2 &:= \mathbb{E}_{\boldsymbol{\epsilon}}\left[\left\|\mathtt{sg}\left(\hat{\boldsymbol{z}}_0(\boldsymbol{z}_t, \mathcal{C}_{\boldsymbol{v}})\right) - \hat{\boldsymbol{z}}_0(\boldsymbol{z}_{s|t,0}, \mathcal{C})\right\|_2^2\right]. \end{aligned} \tag{8}$$

In our empirical results, setting $\lambda = 1$ consistently produces the best performance across all considered T2I models. We note that this technique allows the gradient flow through the second term (contrary to the case of Eq. (6)), thereby sidestepping the gradient blocking issue that we mentioned earlier. Another significant improvement comes from the use of an annealed timestep s, which was originally adhered to a fixed value in Um and Ye [52]. We empirically found that employing an annealing schedule based on the inverse of the sampling step (*e.g.*, $s = T - t$) outperforms other fixed choices of s. Similar to Um and Ye [52], we conduct our prompt optimization intermittently (*i.e.*, once every N sampling steps) to reduce computational costs. We found that during non-optimizing steps, employing the base prompt $\mathcal{C}$ instead of $\mathcal{C}_{\boldsymbol{v}}$ (with the most recently updated token embedding) yields improvements in text-alignment and sample quality. See Algorithms 1 and 2 for the pseudocode of our approach.

Enhanced semantic controllability. A key benefit of our prompt optimization approach is its ability to provide an additional dimension of semantic control over the generated samples. Specifically, by selecting an appropriate initial point for $\boldsymbol{v}$ (*i.e.*, $\boldsymbol{v}_T^{(0)}$ in Algorithm 1), such as a word embedding with relevant semantics, one can impart the desired semantics to the generated output; see Fig. 3 for instance. Note that the controllability is not achievable with existing minority samplers that rely upon latent-space optimizations [44, 51, 52]. We found that properly choosing

Figure 3. **Improved semantic controllability by MinorityPrompt.** The samples in the first column are generations due to DDIM using the two base prompts (*e.g.*, "A chef in a white coat leans on a table" for the second row). The second and third columns exhibit generated samples from our framework, where we selected the corresponding word embeddings as the starting points of the prompt optimizations. In the the last column, we also present DDIM samples produced using attached prompts with the corresponding words for comparison. All samples were obtained using SDXL-Lightning [29].

initial words can yield improved minority generation performance compared to approaches that rely upon random starting points; see the supplementary for detailed results.

4. Experiments

4.1. Setup

T2I backbones and dataset. Our experiments were conducted using three distinct versions of Stable Diffusion (SD) [41], encompassing both standard and distilled versions to demonstrate the robustness of our approach. Specifically, we consider: (i) SDv1.5; (ii) SDv2.0; (iii) SDXL-Lightning (SDXL-LT) [29]. For all pretrained models, we employed the widely-adopted HuggingFace checkpoints trained on LAION [43] without any further modifications. As convention, we randomly selected 10K captions from the validation set of MS-COCO [30] for our main results (*e.g.*, in Tab. 1). While for analyses, smaller subsets of captions were used to enhance efficiency.

Baselines. The same four baselines were considered over all SD versions: (i) the standard DDIM [46]; (ii) a null-prompted DDIM; (iii) CADS [42]; (iv) SGMS [52]. The null-prompted DDIM serves as a baseline that leverages T2I models' capability to visualize specific concepts for minority generation by incorporating an appropriate null-text prompt (*e.g.*, "commonly-looking"). CADS [42] is the state-of-the-art diversity-focused sampler that may rival our approach in minority generation, while SGMS [52] is the state-of-the-art of minority generation outside the T2I domain. We adhered to standard sampling setups for all methods. Specifically, 50 DDIM steps (*i.e.*, $T = 50$) with

Figure 4. **Sample comparison on SDXL-Lightning.** Generated samples from three different approaches: (i) DDIM [46]; (ii) SGMS [52]; (iii) MinorityPrompt (ours). Six distinct prompts were used for this comparison, and random seeds were shared across all three methods.

$w = 7.5$ were used for SDv1.5 and SDv2.0, while $w = 1.0$ was employed for the 4-step SDXL-Lightning model.

Evaluations. For evaluating text-alignment, we consider three distinct quantities: (i) ClipScore [17]; (ii) PickScore [25]; (iii) Image-Reward [53]. We note that the latter two metrics also describe quality of generated samples, in the perspectives of user-preference [25, 53]. In addition, we employ two pairs of metrics for quality and diversity: (i) Precision and Recall [26]; (ii) Density and Coverage [34]. For the likelihood of generated samples, we rely upon the exact likelihood computation method based on PF-ODE as proposed by [48]. We also conduct user study to more investigate human preferences. Notably, we do not include Fréchet Inception Distance (FID) [18] as an evaluator, since FID measures closeness to baseline real data (*e.g.*, the MS-COCO validation set), which diverges from our focus on promoting generations in low-density regions.

4.2. Results

Qualitative comparisons. Fig. 4 presents a comparison of generated samples of our approach with two baselines. Notice that our MinorityPrompt tends to yield highly more distinct and complex features (*e.g.*, intricate visual elements [2, 45]) compared to the baseline samplers. A significant observation, also reflected in Fig. 1, is that MinorityPrompt often counters the inherent demographic biases of T2I models, *e.g.*, by adjusting age or skin color. See the samples in the second and third rows of the figure. A more extensive set of generated samples, including those from SDv1.5 and v2.0, can be found in the supplementary.

Quantitative evaluations. Tab. 1 exhibits performance comparisons across three distinct T2I models. Observe that our sampler outperforms all baselines in generating low-likelihood samples while maintaining reasonable performance in text-alignment and user preference; see the supplementary for the corresponding log-likelihood distributions. An important point is that MinorityPrompt significantly improves the previous state-of-the-art in minority generation (*i.e.*, SGMS [52]) in almost all cases, highlighting the effectiveness of our approach in the T2I context. As expected, our advantage often comes with some compromise in image quality, *e.g.*, evidenced by lower PickScore and Density values compared to the DDIM sampler. We leave the user study results in the supplementary.

Ablation studies. Tab. 2 exhibits the impacts of the three theoretical flaws in the naive framework in Eq. (6). Observe that incorporating any of these flaws into our framework results in immediate performance degradation, validating our claim made in Sec. 3.3. See the supplementary for theoretical evidence. A more comprehensive analysis and ablation study, including explorations of other design choices and applications to trending sampling techniques

Model	Method	CLIPScore ↑	PickScore ↑	ImageReward ↑	Precision ↑	Recall ↑	Density ↑	Coverage ↑	Likelihood ↓
	DDIM	31.4801	21.4830	0.2106	**0.5907**	0.6328	**0.6072**	**0.7492**	1.0367
	DDIM + null	31.1007	**21.5391**	**0.2422**	0.5660	0.6236	0.5362	0.7134	1.0339
SDv1.5	CADS [42]	31.4178	21.2836	0.1012	0.5696	**0.6346**	0.5562	0.7388	1.0127
	SGMS [52]	31.1665	21.2126	0.1230	0.4943	0.5960	0.4357	0.6470	0.9540
	MinorityPrompt	**31.5376**	21.3111	0.2352	0.5671	0.6228	0.5375	0.7328	**0.8971**
	DDIM	31.8490	21.6801	0.3821	0.5930	0.6292	**0.6592**	**0.7760**	1.1100
	DDIM + null	31.7223	**21.7190**	0.4024	0.5861	**0.6308**	0.5959	0.7378	1.0769
SDv2.0	CADS [42]	31.7687	21.5225	0.2981	0.5811	0.6194	0.5865	0.7388	1.0851
	SGMS [52]	31.4750	21.4457	0.2981	0.5166	0.6130	0.4713	0.6718	0.9898
	MinorityPrompt	**31.9586**	21.5958	**0.4249**	**0.6047**	0.6100	0.6192	0.7602	**0.9143**
	DDIM	31.5238	22.6733	0.7331	**0.5323**	0.6116	0.5206	0.6686	0.6082
	DDIM + null	**31.5259**	**22.6884**	**0.7368**	0.5256	0.6144	**0.5368**	**0.6700**	0.6077
SDXL-LT	CADS [42]	31.0418	22.3554	0.5017	0.5211	0.6176	0.5220	0.6560	0.6019
	SGMS [52]	31.2961	22.5784	0.6801	0.4823	**0.6616**	0.4018	0.5852	0.5462
	MinorityPrompt	31.3366	22.6050	0.7098	0.4777	0.6580	0.3856	0.5770	**0.5457**

Table 1. **Quantitative comparisons.** "SDXL-LT" denotes SDXL-Lightning (4-step version) [29]. "DDIM + null" indicates a baseline that leverages a properly-chosen null-prompt to encourage minority generations, where we used "commonly-looking" for the results herein. "CADS [42]" is the state-of-the-art in diverse sampling, while SGMS [52] denotes a minority sampler similar to ours, representing the state-of-the-art outside the T2I context. "Likelihood" represents log-likelihood values measured in bpd (bits per dimension).

Method	CLIPScore ↑	PickScore ↑	ImageReward ↑	Precision ↑	Recall ↑	Density ↑	Coverage ↑	Likelihood ↓
DDIM	31.4395	**21.4570**	0.1845	**0.6070**	0.7094	**0.6460**	**0.8410**	1.0465
Eq. (7) (proposed)	**31.7369**	21.3522	**0.2839**	0.5420	**0.7340**	0.5534	0.7860	**0.9230**
Eq. (7) + $\hat{z}_0^w$	30.5193	20.7307	-0.1468	0.4890	0.7182	0.4910	0.7450	0.9399
Eq. (7) + $\mathtt{sg}$	31.6597	21.3114	0.2738	0.5230	0.7284	0.4986	0.7470	0.9290
Eq. (7) + $\mathcal{C}_{\boldsymbol{v}}$	31.6676	21.3652	0.2808	0.5550	0.7262	0.5414	0.7500	0.9281
Eq. (7) + all (*i.e.*, Eq. (6))	30.2994	20.4840	-0.1944	0.4760	0.6864	0.4762	0.7220	0.9245

Table 2. **Ablation study results.** "+ $\hat{z}_0^w$" indicates the case that further incorporates the CFG clean predictions into Eq. (7). "+ $\mathtt{sg}$" refers to the one employing the stop-gradient on $\hat{\boldsymbol{z}}_0(\boldsymbol{z}_{s|t,0}, \mathcal{C})$. "+ $\mathcal{C}_{\boldsymbol{v}}$" represents the setting of feeding $\mathcal{C}_{\boldsymbol{v}}$ in the computations of $\hat{\boldsymbol{z}}_0(\boldsymbol{z}_{s|t,0}, \mathcal{C})$ in place of $\mathcal{C}$. "+ all" is the case that employs all the above three flawed choices, *i.e.*, Eq. (6). We observe clear performance benefits of our theory-driven design choices over the naive framework in Eq. (6). The results were obtained on SDv1.5.

Method	CS ↑	PS ↑	Prec ↑	Rec ↑	Den ↑	Cov ↑	IBS ↓
DDIM	**31.4393**	**21.2478**	**0.5860**	**0.6390**	0.7688	0.8220	0.6164
CADS [42]	31.2692	21.0262	0.5620	0.5980	**0.7964**	0.8180	0.5494
Ours	31.2724	21.0404	0.5480	0.6316	0.7672	**0.8460**	**0.5439**

Table 3. **Effectiveness of our diversity-focused prompt optimization framework in Eq. (9).** "IBS" refers to In-Batch Similarity, a diversity metric [8] that measures cosine similarity in the DINO feature space [3]. We employed SDv1.5 for the results.

(like CFG++ [7]), is presented in the supplementary.

Further application. Beyond our primary focus on minority generation, we explore a distinct realm of diverse generation with our optimizer in Eq. (4) to demonstrate the versatility of the proposed optimization framework for solving various tasks. Our specific goal herein is to encourage diversity in an inference batch that shares the same text prompt $\mathcal{P}$, similar to the focus in Corso et al. [8]. To achieve this, we introduce a new objective function that enforces repulsion between intermediate instances, formally written as:

$$\bar{\mathcal{J}} := \sum_{i=1}^{\mathcal{B}} \sum_{k \neq i} \|\hat{z}_0(z_t^{(i)}, \mathcal{C}_{\boldsymbol{v}}) - \hat{z}_0(z_t^{(j)}, \mathcal{C}_{\boldsymbol{v}})\|_2^2, \quad (9)$$

where $\mathcal{B}$ is the batch size, and $\{\boldsymbol{z}_t^{(i)}\}_{i=1}^{\mathcal{B}}$ denotes the intermediate samples in the batch at t. We found that incorporating Eq. (9) into Eq. (4) yields impressive results, even rivaling the state-of-the-art diverse sampler [42] (see Tab. 3). We leave generated samples in the supplementary.

5. Conclusion

We developed a novel framework for generating minority samples in the context of T2I generation. Built upon our prompt optimization framework that updates the embeddings of additional learnable tokens, our minority sampler offers significant performance improvements compared to existing approaches. To accomplish this, we meticulously tailor the objective function with theoretical justifications and implement several techniques for further enhancements. Beyond our focus of minority generation, we further demonstrated the potential of our framework in promoting diversity in generated samples. During this process, we also showed that the proposed optimization framework can serve as a general solution, with potential applicability to various optimization tasks associated with T2I generation.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) under Grants RS-2024-00336454 and RS-2023-00262527, and by Basic Science Research Program through the NRF funded by the Ministry of Education (RS-2024-00406726).

References

[1] Arash Afkanpour, Vahid Reza Khazaie, Sana Ayromlou, and Fereshteh Forghani. Can generative models improve self-supervised representation learning? *arXiv preprint arXiv:2403.05966*, 2024. 2

[2] Marius Arvinte, Cory Cornelius, Jason Martin, and Nageen Himayat. Investigating the adversarial robustness of density estimation using the probability flow ode. *arXiv preprint arXiv:2310.07084*, 2023. 7

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 8

[4] Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022. 4

[5] Hyungjin Chung, Suhyeon Lee, and Jong Chul Ye. Decomposed diffusion sampler for accelerating large-scale inverse problems. *arXiv preprint arXiv:2303.05754*, 2023. 4

[6] Hyungjin Chung, Jong Chul Ye, Peyman Milanfar, and Mauricio Delbracio. Prompt-tuning latent diffusion models for inverse problems. *arXiv preprint arXiv:2310.01110*, 2023. 2, 3, 4

[7] Hyungjin Chung, Jeongsol Kim, Geon Yeong Park, Hyelin Nam, and Jong Chul Ye. Cfg++: Manifold-constrained classifier free guidance for diffusion models. *arXiv preprint arXiv:2406.08070*, 2024. 8, 2, 4, 5

[8] Gabriele Corso, Yilun Xu, Valentin De Bortoli, Regina Barzilay, and Tommi Jaakkola. Particle guidance: non-iid diverse sampling with diffusion models. *arXiv preprint arXiv:2310.13102*, 2023. 2, 8, 3

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[10] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 4

[11] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

[12] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 3

[13] Felix Friedrich, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Patrick Schramowski, Sasha Luccioni, and Kristian Kersting. Fair diffusion: Instructing text-to-image generation models on fairness. *arXiv preprint arXiv:2302.10893*, 2023. 1

[14] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2, 3, 4

[15] Inhwa Han, Serin Yang, Taesung Kwon, and Jong Chul Ye. Highly personalized text embedding for image manipulation by stable diffusion. *arXiv preprint arXiv:2303.08767*, 2023. 3, 4

[16] Jiyeon Han, Hwanil Choi, Yunjey Choi, Junho Kim, Jung-Woo Ha, and Jaesik Choi. Rarity score: A new metric to evaluate the uncommonness of synthesized images. *arXiv preprint arXiv:2206.08549*, 2022. 2, 3

[17] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 7, 3

[18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 7

[19] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2, 4

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 3, 1

[21] Chunsan Hong, ByungHee Cha, and Tae-Hyun Oh. Cas: A probability-based approach for universal condition alignment score. In *International Conference on Learning Representations*, 2024. 3

[22] Gaofeng Huang and Amir Hossein Jafari. Enhanced balancing gan: Minority-class image generation. *Neural computing and applications*, 35(7):5145–5154, 2023. 2

[23] Jeongsol Kim, Geon Yeong Park, Hyungjin Chung, and Jong Chul Ye. Regularization by texts for latent diffusion inverse solvers. *arXiv preprint arXiv:2311.15658*, 2023. 5

[24] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2

[25] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. 2023. 7

[26] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in Neural Information Processing Systems*, 32, 2019. 7

[27] Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models, 2024. 2

[28] Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. *arXiv preprint arXiv:2303.16203*, 2023. 1

[29] Shanchuan Lin, Anran Wang, and Xiao Yang. Sdxl-lightning: Progressive adversarial diffusion distillation. *arXiv preprint arXiv:2402.13929*, 2024. 2, 6, 8

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 6

[31] Zinan Lin, Hao Liang, Giulia Fanti, Vyas Sekar, Rahul Anand Sharma, Elahe Soltanaghaei, Anthony Rowe, Hun Namkung, Zaoxing Liu, Daehyeok Kim, et al. Raregan: Generating samples for rare classes. *arXiv preprint arXiv:2203.10674*, 2022. 2

[32] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022. 4, 5

[33] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023. 3, 4

[34] Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning*, pages 7176–7185. PMLR, 2020. 7

[35] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[36] Geon Yeong Park, Jeongsol Kim, Beomsu Kim, Sang Wan Lee, and Jong Chul Ye. Energy-based cross attention for bayesian context update in text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3, 4

[37] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 3

[38] Yiming Qin, Huangjie Zheng, Jiangchao Yao, Mingyuan Zhou, and Ya Zhang. Class-balancing diffusion models. *arXiv preprint arXiv:2305.00562*, 2023. 2

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4

[40] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. 2

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3, 6

[42] Seyedmorteza Sadat, Jakob Buhmann, Derek Bradely, Otmar Hilliges, and Romann M Weber. Cads: Unleashing the diversity of diffusion models through condition-annealed sampling. *arXiv preprint arXiv:2310.17347*, 2023. 2, 4, 6, 8, 5, 10

[43] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 2, 6

[44] Vikash Sehwag, Caner Hazirbas, Albert Gordo, Firat Ozgenel, and Cristian Canton. Generating high fidelity data from low-density regions using diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11492–11501, 2022. 2, 6

[45] Joan Serrà, David Álvarez, Vicenç Gómez, Olga Slizovskaia, José F Núñez, and Jordi Luque. Input complexity and out-of-distribution detection with likelihood-based generative models. *arXiv preprint arXiv:1909.11480*, 2019. 7

[46] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1, 4, 6, 7, 5, 8, 9, 10

[47] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019. 2, 3

[48] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 7, 5

[49] Yonglong Tian, Lijie Fan, Kaifeng Chen, Dina Katabi, Dilip Krishnan, and Phillip Isola. Learning vision from models rivals learning vision from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15887–15898, 2024. 2

[50] Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[51] Soobin Um and Jong Chul Ye. Don't play favorites: Minority guidance for diffusion models. *arXiv preprint arXiv:2301.12334*, 2023. 2, 6

[52] Soobin Um and Jong Chul Ye. Self-guided generation of minority samples using diffusion models. *arXiv preprint arXiv:2407.11555*, 2024. 2, 5, 6, 7, 8, 4, 9

[53] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation, 2023. 7, 3

[54] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative

adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018. 2

[55] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 2

[56] Ning Yu, Ke Li, Peng Zhou, Jitendra Malik, Larry Davis, and Mario Fritz. Inclusive gan: Improving data and minority coverage in generative models. In *European Conference on Computer Vision*, pages 377–393. Springer, 2020. 2

[57] Oz Zafar, Lior Wolf, and Idan Schwartz. Iterative object count optimization for text-to-image diffusion models. *arXiv preprint arXiv:2408.11721*, 2024. 3, 4

[58] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5, 2