Text-Conditioned Sampling Framework for Text-to-Image Generation with Masked Generative Models

Jaewoong Lee*
KAIST
Republic of Korea
hello3196@kaist.ac.kr

Sangwon Jang*
Yonsei University
Republic of Korea
agwmon@gmail.com

Jaehyeong Jo, Jaehong Yoon
KAIST
Republic of Korea
{harryjo97, jaehong.yoon}@kaist.ac.kr

Yunji Kim, Jin-Hwa Kim, Jung-Woo Ha
NAVER AI Lab
Republic of Korea
{yunji.kim, jinhwa.kim, jungwoo.ha}@navercorp.com

Sung Ju Hwang
KAIST
Republic of Korea
sjhwang82@kaist.ac.kr

Abstract

Token-based masked generative models are gaining popularity for their fast inference time with parallel decoding. While recent token-based approaches achieve competitive performance to diffusion-based models, their generation performance is still suboptimal as they sample multiple tokens simultaneously without considering the dependence among them. We empirically investigate this problem and propose a learnable sampling model, Text-Conditioned Token Selection (TCTS), to select optimal tokens via localized supervision with text information. TCTS improves not only the image quality but also the semantic alignment of the generated images with the given texts. To further improve the image quality, we introduce a cohesive sampling strategy, Frequency Adaptive Sampling (FAS), to each group of tokens divided according to the self-attention maps. We validate the efficacy of TCTS combined with FAS with various generative tasks, demonstrating that it significantly outperforms the baselines in image-text alignment and image quality. Our text-conditioned sampling framework further reduces the original inference time by more than 50% without modifying the original generative model.

1. Introduction

In the flood of generative AI systems for vision domains, text-conditional image generation [26, 33, 42] is coming to the fore in recent years. Although many recent works have achieved success in synthesizing high-quality images [19, 34] with plausible class-alignment in class-conditional cases [4, 8], text-to-image (T2I) generation is more challenging since generating visual outputs that are semantically aligned with input texts is a nontrivial problem. We can roughly categorize the works on text-to-image generation into transformer-based *autoregressive* (AR) [12, 30, 33] and *diffusion-based* [18, 29] approaches. Along with the advancement of language models, AR models using transformers have shown impressive performance in text-to-image generation. Despite their success, they suffer from the problem of *unidirectional bias*, which is undesirable for image generation, and crucially, the sampling process requires over 10 times as much time compared to existing models. Another line of work is diffusion-based methods that aim to generate images by iteratively denoising noisy samples. In particular, several continuous diffusion models [10, 32] have achieved outstanding performance and reduced computational cost. Yet, they require excessive sampling steps to obtain high-quality images during the inference.

Recently, a new family of generative models, called *token-based diffusion model* [6, 23, 41], has emerged as an alternative to tackle the problem of text-to-image generation. Token-based diffusion models quantize the latent features into tokens and apply categorical corruption process in discrete state spaces [1], while conventional diffusion models use Gaussian noise in continuous space. Among various discrete diffusion methods, mask-based diffusion, similar to the absorbing state diffusion used in [1], is mostly used. Compared to existing AR models, this token-based approach is advantageous for speeding up the generation process via simultaneously sampling multiple tokens. Despite the limitation on the reconstruction capacity, these token-based diffusion models significantly outperform the competitors in terms of FID scores, even with fewer sampling steps compared to the continuous diffusion models [5, 37].

However, sampling multiple tokens at once often leads

*Equal contribution.

Figure 1: **Generated samples on MS-COCO dataset and evaluation graph of various sampling methods showing their trade-off.** Uniform sampling is a fixed strategy with notably poor text alignment compared to other methods (FID-40K: 15.61, MID-L: 21.23). Random revoke sampling is a revocable strategy with improved text alignment (FID-40K: 16.81, MID-L: 26.98). Ours is TCTS combined with FAS, where both the image quality and the text alignment are significantly better compared to those of baselines (FID-40K: 13.6, MID-L: 29.5). Metrics are measured on all 40K images with their corresponding single caption. The classifier-free guidance scale was fixed at 5 for all sampling methods.

to inconsistency throughout a generated image [37]. For each location, the generator outputs a probability distribution that are coherent with each other. However, there is no guarantee that every single token sampled from the distributions will perfectly align with one another. In other words, it is still possible for the generator to sample incompatible tokens regardless of the generator's capability, leading to potentially nonsensical outputs [24, 37]. This results in a trade-off between the sampling steps and generation quality. Reducing the sampling steps leads to faster generation but results in performance degradation due to a large number of simultaneously sampled tokens. Especially, this problem further stands out in the text-to-image generation tasks since the distribution of text-aligned images is more restricted than the distribution of unconditioned or class-conditioned images.

To address these limitations, we propose a novel sampling approach that refines the images during the diffusion process based on the text condition, which we refer to as the *Text Conditioned Token Selection* (TCTS). TCTS is a sampling strategy that can mask out and resample previously sampled tokens, which we refer to as *revocable*. To find the tokens that do not align with the given text condition, we propose to train a model that selects tokens to be masked out and re-generated, to be well-aligned with the given text. Combining this approach with the revocable sampling scheme, TCTS can generate high-quality images with improved text-alignment in even fewer sampling steps compared to the naive generative model, as shown in Figure 1. We further introduce *Frequency Adaptive Sampling* (FAS) to solve the over-simplification problem that occurs when applying revocable methods for relatively longer steps. FAS leverages the self-attention map of the images and applies a mixed sampling method, which we call *persistent sampling* to prevent the issue. We summarize our contributions as follows:

- We experimentally show that the revocable sampling strategies are crucial for the trade-off between the text-alignment and the image quality and provide in-depth analysis compared to previous fixed sampling methods.
- We propose a novel revocable sampling method based on a learnable token selection model that significantly improves the text alignment as well as the image quality even with fewer sampling steps, without the need of retraining the generator model.
- Moreover, we propose a novel sampling method based on the self-attention map that can be combined with TCTS to solve the over-simplification problem.

2. Related work

Various works have tackled text-to-image generation tasks, and the majority of them are based on Generative Adversarial Networks (GANs) [38, 40, 44, 45, 46, 47, 48]. However, computing directly from the pixel space of an image is challenging, and generating high-resolution images requires substantial computation due to the large number of pixels involved. Therefore, a two-stage approach is often used: models first tokenize the images into a sequence of codes, and then predict the tokenized codes.

Autoregressive models Following the success of Transformer [39] and GPT [28], Autoregressive (AR) models [9, 30, 31, 42] proposed to tokenize images and pose the text-to-image generation problem as a problem of translating textual descriptions into image tokens, much like machine translation of one language into a different language. With the advancement of image tokenization and language models, this approach has achieved very high performance. However, even with the help of this tokenizing stage , there remains

a large number of tokens to predict, resulting in a significant delay in the generation time due to their autoregressive nature. Additionally, errors are accumulated through the irrevocable process, with undesirable unidirectional bias.

Latent diffusion models Recently, diffusion models [18, 36] introduced a new paradigm for text-to-image generation [33]. They generate high-quality images by progressively denoising noise, but require iterative steps and a large amount of computation. Two-stage approaches [10, 32] propose to perform diffusion in latent space successfully reducing computation while maintaining performance. However, most of them still require about 50 inference steps to show plausible performance, and since denoising is applied globally to the entire image during the inference process, locally correcting the image requires complicated processes [32].

Token-based diffusion models Token-based diffusion models [5, 11, 13, 23, 43] have impressive performance in generating complex scenes from the text. Unlike conventional diffusion models, where the Gaussian noise is applied in a continuous space, token-based diffusion models quantize latent features into tokens and apply a categorical corruption process in a discrete state space [1]. Among various discrete diffusion methods, mask-based diffusion is most commonly used due to its attractive properties; they have the advantage of sampling multiple tokens simultaneously, resulting in fewer sampling steps and faster generation time than the latent diffusion model. However, selecting multiple tokens at once leads to inconsistency throughout the image which is often called a joint distribution issue. Moreover, tokens are heavily influenced by the spatially adjacent tokens that have already been modified so that a misaligned token could disperse the errors throughout the image. To tackle this issue, some works [6, 23] modified the sampling strategy while other works [24] introduced new models to find the misaligned tokens.

In particular, Lezama *et al*. [24] proposed a learnable model that removes the misaligned tokens at each step. While most existing works randomly select the tokens to discard, this work examines all sampled tokens. It outputs the score of each location where the score indicates whether the corresponding token is under the real distribution. After the predicting and sampling process of the generator, the model selects the locations according to the scores of the corresponding tokens.

3. Text-conditioned token selection

The diffusion process of the masked image generative models starts with a completely masked image, and samples the tokens of x_t for the whole image to predict $\hat{x}_0$ while re-masking a portion of them to obtain the denoised sample x_{t-1}. Existing masked image generative models use uniform or confidence-based sampling strategy to determine

Figure 2: **Reconstructed images using $\hat{\mathbf{x}}_0^{(\mathbf{t})}$ each step during diffusion using the fixed and revocable method.** Only the revocable method is able to edit the tokens to generate a plane according to the text "*A view of the end of an airplane in the sky over mountains.*".

the locations to re-mask: *Uniform sampling* [13, 23] randomly selects the locations of the tokens to keep among the predicted tokens at all locations, while confidence-based strategy, such as *purity sampling* [37], selects the locations based on the confidence score which is defined as follows:

$$\texttt{CS}(i,t) = \max_{j=1,\cdots,K} p(x_0^i = j | x_t^i), \tag{1}$$

where t is the step number, K is the size of the codebook, and i is the location of the token. Intuitively, a high confidence score means the generator is relatively certain that a token should be sampled at the corresponding location. By keeping these confident tokens while discarding the ambiguous ones, purity sampling shows good performance in practice [5].

However, these strategies are non-revocable, *i.e*., *fixed*, which means tokens that are once sampled cannot be revised afterward, even if they do not match the given text or do not align with other tokens. Once the misaligned tokens are fixed during the process, errors accumulate and spread throughout the image. We observe that fixed strategies negatively affect the text alignment as visualized in Figure 1.

3.1. Random revoke sampling strategy

In order to address this issue, we first introduce a new sampling strategy called *random revoke sampling* and use it as a baseline for our final method. It is similar to uniform sampling in a way that they both determine the locations by sampling from a uniform distribution. However, while uniform sampling is a fixed strategy, random revoke sampling selects the locations from the whole image space regardless of the previously fixed locations, making the process revocable. It randomly selects a certain number of tokens to preserve for the next step and revokes the previously fixed ones. The revocable feature can better align the images to the given text information, especially in complex scenes.

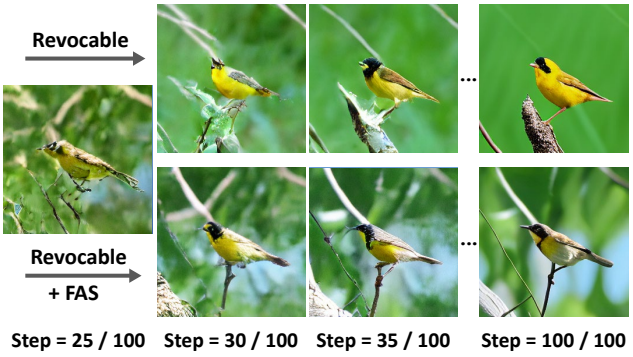

Figure 3: **Reconstructed images using $\hat{\mathbf{x}}_0^{(\mathbf{t})}$ each step during diffusion using the revocable method with and without FAS.** The backgrounds of the images in the top row are over-simplified while our proposed FAS prevents this, as shown in the bottom row. The text is "*This small bird is greyish in color with flecks of yellow on the back and breast, and a bit of white on the belly.*".

To be specific, as shown in the top row of Figure 2, fixed strategies have limitations in recovering from errors or generating diverse images due to their inability to regenerate incorrect tokens or modify missing essential objects or attributes. In contrast, revocable strategies enhance the text alignment by providing an opportunity to correct invalidly generated tokens, as shown in the bottom row of Figure 2. Since revocable strategies have no constraints on which tokens to remove or fix, they have the advantage of mitigating error accumulation.

However, random revoke sampling suffers from two main issues that must be addressed. The first issue is the instability resulting from randomly discarding previously sampled tokens which worked as a given condition of the step. It can lower the quality of the outputs as in Figure 1.

The second issue, namely the *over-simplification*, arises when it is used in the generation process involving longer steps. The low-frequency parts of the output images, such as backgrounds, tend to be over-simplified compared to fixed methods, which generate more realistic backgrounds. The over-simplified samples can be seen in Figure 3, and it has a significant impact on the FID metric as in Table 1 and 2. We observe that this is due to excessive repetition of resampling over a large number of steps, while a fixed schedule does not resample any tokens. More opportunities for resampling make the tokens converge towards safer and simpler patterns. As shown in the top row of Figure 3, the high-frequency areas with rich visual details do not become over-simplified even after multiple resampling processes, but the low-frequency areas become less diverse and more simplified in patterns quickly due to excessive resampling.

To confirm this trade-off between text alignment and image quality, we define a *persistent sampling* as an interpolation between uniform sampling and random revoke sampling. In details, a persistent weight boosts the probability of keeping the tokens that were sampled in the previous step. While random revoke sampling assigns the same probability to all the locations, persistent sampling multiplies the persistent weight to the probability of previously kept locations. Further details are shown in Algorithm 1. If the persistent weight w is 1, all the tokens would have the same probability to be fixed, such as random revoke sampling. If w is sufficiently large, the previously sampled tokens would be kept until the end, such as uniform sampling. In the graph in Figure 1, a trade-off can be seen that as the sampling strategy gets closer to the random revoke sampling, text alignment gets better, while the image quality gets worse. We further analyze the effect of these sampling methods in Section 4.5.

Algorithm 1 Persistent Sampling

Input: k_t: number of tokens sampled at step t, w: persistent weight, N: number of all tokens

$x_T \leftarrow [\text{[MASK]}]_N$
for $t = T, T-1, ..., 1$ **do**
  $\hat{x}_0 = G_\theta(x_t, c)$
  $A_t = \{i | x_t^i = \text{[MASK]}\}$
  $m = k_T + k_{T-1} + ... + k_t$
  $\mathcal{U}_t(I = i | i \in A_t) : \mathcal{U}_t(I = i | i \in A_t^C) = 1 : w$
  $\rightarrow$ uniform distribution + persistent weight
  $i_1, i_2, ..., i_m \sim \mathcal{U}(I) \leftarrow$ sample without replacement
  $x_t \leftarrow [\text{[MASK]}]_N$
  for $i = i_1, i_2, ..., i_m$ **do**
    $x_t^i \leftarrow \hat{x}_0^i$
  end for
end for
Return: Generated image x_0

3.2. Text-conditioned token selection

We propose a learnable model called *Text-Conditioned Token Selection* (TCTS). TCTS is trained to output a score map to detect tokens that are not under the real distribution for the given text condition, and selects the tokens that are not aligned with the text or other tokens in the image. By masking the selected tokens at each step, the diffusion network receives high-quality text-aligned tokens as the input to sample the next set of tokens which alleviates the error accumulation and joint distribution issue of the previous sampling strategies. Moreover, since TCTS is trained to discriminate the well-aligned tokens, it detours the instability problem as mentioned earlier, unlike other revocable methods. From the experimental results in Section 4, we observe that TCTS enhances the text alignment without compromising the image quality.

Frequency adaptive sampling (FAS) The simplest way to solve the over-simplification problem in longer steps is to use the persistent sampling to alleviate the excessive resampling of tokens. However, as previously mentioned, this method

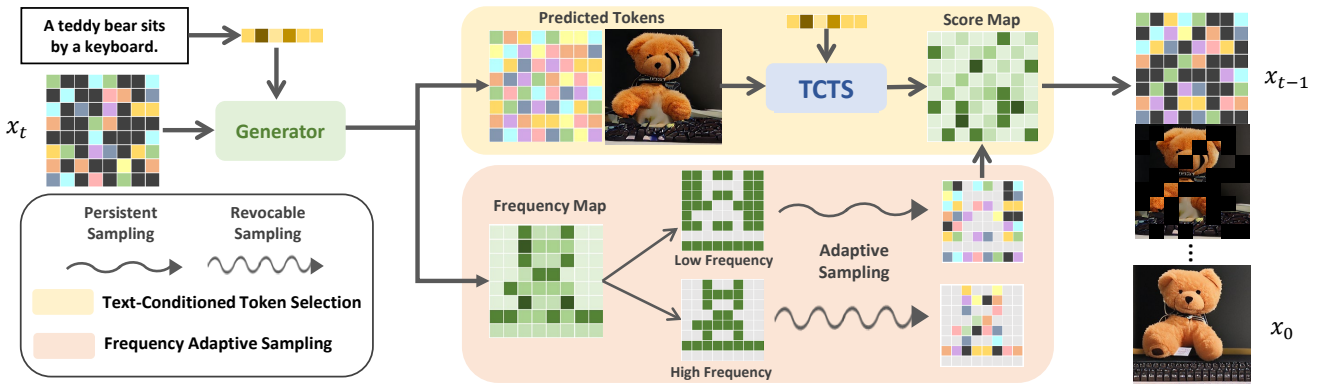

Figure 4: **Overall generation framework of proposed TCTS and FAS.** After the generator predicts the tokens, TCTS exploits the text condition to detect misaligned tokens and outputs the score map. Meanwhile, FAS splits the tokens according to the frequency using the self-attention map from the generator, performing revocable sampling to high-frequency split and persistent sampling to low-frequency split. The adaptive sampling predicts $\hat{x}_0$ and decide a few of the locations to mask according to x_t. The token maps produced by FAS is combined with the score map to predict x_{t-1}. After iterative process, our model removes all the masks and generates x_0.

Step	Method	MID-L ↑	SOA-I ↑	CLIP-S ↑	FID-30K ↓
16	Purity	11.02	72.38	0.2474	19.20
	Uniform	17.94	74.80	0.2500	16.17
	RR	23.60	78.79	0.2526	17.10
	TCTS + FAS	**26.72**	**79.52**	**0.2559**	**14.45**
25	Purity	16.84	75.21	0.2487	18.39
	Uniform	22.27	77.08	0.2524	15.91
	RR	26.77	**81.10**	0.2543	18.43
	TCTS + FAS	**27.79**	80.87	**0.2563**	**15.39**

Table 1: **Quantitative evaluation of sampling methods on MS-COCO dataset.** The ground truth MID on MS-COCO image-caption pairs is 54.73. The classifier-free guidance scale was fixed at 5 for all sampling methods.

Step	Method	MID-L ↑	CLIP-S ↑	FID ↓
16	Purity	-24.21	0.2410	15.21
	Uniform	-25.60	0.2404	16.57
	RR	-25.03	0.2371	17.38
	TCTS + FAS	**-19.88**	**0.246**	**12.35**
25	Purity	-21.26	0.2384	**12.60**
	Uniform	-23.04	0.2396	13.02
	RR	-23.29	0.2364	14.53
	TCTS + FAS	**-18.31**	**0.2409**	13.67

Table 2: **Quantitative evaluation of sampling methods on CUB dataset.** The ground truth MID on CUB image-caption pairs is 15.85. We omit SOA in this table as CUB dataset consists of only one bird per photo, diminishing the metric's significance.

may result in a slight compromise in text alignment and limit the opportunity to correct the image through resampling. To further improve this trade-off, we propose a new method called *Frequency Adaptive Sampling (FAS)*, which can be applied to TCTS.

FAS is a method that utilizes the generator's self-attention map to limit resampling only in the low-frequency areas of the image. In [21], it is known that the generator's self-attention layer contains frequency information of the image, which is also observed in token-based diffusion models. We utilized this to distinguish the frequency areas of the image without additional operations, and applied persistent sampling only to the low-frequency areas using persistent weight. As shown in the bottom row of Figure 3, revocable sampling along with FAS method allows repetitive resampling in areas that require rich visual details, while limiting the number of resampling in relatively simple areas preventing over-simplification. Detailed algorithm is in the supplementary A. The overall framework of our model, including both TCTS and FAS, is illustrated in Figure 4.

4. Experiments

To validate the efficacy of our token selection framework, we modify the transformer from [37] to contain 310M parameters for MS-COCO dataset and 83M parameters for CUB dataset. We extract text features from CLIP ViT-B/32 [27] for a text-conditioned generation. In the training process of TCTS, we freeze the generator's parameters and independently train our model, making it applicable to other types of token-based diffusion models. We use the binary cross-entropy loss for the objective and validate our method and baselines on MS-COCO [25] and CUB [14] datasets by training them for 200 epochs. We leverage classifier-free guidance [20] by stochastic sampling of the guidance strength to improve the generation quality. We further provide the details of the architecture, hyperparameters and additional analysis on the use of classifier-free guidance in the supplementary A.

Model	MID-L ↑	MID-B ↑	SOA-I ↑	FID-30K ↓
GLIDE [26]	1.03	1.00	-	32.08
AttnGAN [40]	-65.20	-8.90	39.01	29.15
DM-GAN [23]	-44.66	3.51	48.03	22.90
DF-GAN [38]	-58.75	-15.21	-	31.75
VQ-Diffusion [41]	-19.63	5.77	-	13.13
LAFITE [47]	6.26	35.17	74.78	**8.03**
TCTS + FAS	**26.02**	**38.98**	**79.52**	9.75

Table 3: **Comparison of the proposed method with recent text-to-image generative models on MS-COCO dataset.** MID-L and MID-B are calculated with ViT-L/14 and ViT-B/32 each. The ground truth MID-L and MID-B on MS-COCO image-caption pairs is 54.73, 41.57 each. Ours are evaluated with 16-step setting and we adopt classifier-free guidance from 3 to 5.

Metrics Since FID [16] metric is known to be problematic for its inability to exactly represent the image quality [5] and unable to consider the text alignment, we additionally use Mutual Information Divergence (MID) [22] to evaluate the text-alignment of generated images which enables sample-wise evaluation and outperforms previous metrics in human Likert-scale judgment correlations. In particular, since MID responds more sensitively to images generated with foiled captions [35], it is more appropriate for analyzing text alignment with complex and challenging captions. Moreover, we use the CLIP score [15] and SOA-I [17], which are widely used in text-to-image synthesis. Note that the CLIP score is calculated with ViT-L/14 throughout the experiments.

4.1. Text-to-image synthesis

Figure 1 Left and Middle visualizes the generated images using improved VQ diffusion [37] with varying sampling methods. Generated examples from uniform token sampling are poorly aligned with the given texts, and often include erratic partitions of the objects since the predicted tokens selected at the same step cannot appropriately reflect the change of others (*e.g.*, *smashed donuts, and imperfect chair and seagull*). Random revoke (RR) sampling, which can re-update previously sampled tokens, seems to improve alignment with text by editing out nonsensical regions and iteratively considering the entire scene at each generation step. Yet, RR selects tokens in a completely random manner without any constraints, which results in suboptimal generation quality and alignment with the texts (*e.g.*, *disappeared man and train*). On the other hand, ours mitigates the problem from RR by selecting text-conditioned tokens at the sampling phase and successfully generates high-quality images which contain a clear semantic connection to the given text captions.

This phenomenon can be quantitatively seen in Table 1. 16-step generation of random revoke sampling shows better text alignment than even 25-step generation of other fixed sampling methods. We can also observe the trade-off between CLIP score and FID by comparing each method to their 25-step versions. Our model outperforms other baselines in most of the metrics especially when comparing MID-L to fixed sampling methods such as Purity and Uniform. Since our model can make corrections during the generation process, it requires fewer steps to match the performance of our baseline sampling methods, see Section 4.5 for further analysis. We also present the results of our experiment on the CUB dataset with fewer parameters in Table 2, which demonstrates satisfactory performance. However, the fact that the CUB dataset contains only single object per image resulted in a tendency that is slightly different from what we mentioned earlier. In Table 3, we compared the performance of our method with other text-to-image generative models. Our model highly exceeds other models in text alignment metrics, especially in MID-L and closely approximate the MID-B value of the ground truth images. More samples generated by our model is in the supplementary B.

Figure 5: **Results of image refinement using TCTS. Top**: Original samples, **Bottom**: Refined images for 8-steps with TCTS.

Method	MID-L ↑	FID ↓
Original	12.10	**16.44**
RR	13.64	17.18
TCTS	**15.64**	16.92
Original	4.54	19.13
Random masked	6.06	19.12
TCTS masked	**6.47**	**18.64**

Table 4: **Qualitative evaluation of the refined images. Top**: Refinement with additional revision steps. **Bottom**: Refinement with masking lowest-scoring tokens.

4.2. Image refinement

The advantage of the masked image generative model is that it enables fast local refinement. We conduct image refinement in two separate methods, inspired by [23] and [24]. First, we use TCTS to apply additional revision steps to images generated with uniform sampling. The overall image

quality is improved by adding additional refinement steps as shown in Figure 5. In order to demonstrate the image refinement performance of our TCTS, we also measure the performance of the image refinement with random revoke sampling. While RR improves does improve sample quality, additional refinement steps increase the FID score, caused by a similar effect to increasing the generation step.

We additionally perform the experiment by masking 60% lowest-scoring tokens with TCTS, and generate the images with uniform sampling. We observe that all metrics of the refined images outperform those of the original images. In Table 4, we compare them with samples refined after randomly masking the tokens without TCTS. Due to the page limit, we describe further details of the two refinement methods and visualize more samples in the supplementary B.

4.3. Mask-free object editing

Since the masked image generative model is capable of local refinement, image editing without manual masking is possible simply by randomly masking a part of the image and resampling it with the new text condition. However, this method requires a low masking ratio and many resampling steps to maintain the overall structure of the image. This can result in significant changes to unnecessary parts or, as mentioned earlier, over-simplification issues. Additionally, even with many steps, editing large objects with small masks can be more challenging due to the significant influence of surrounding tokens rather than the new text condition.

Motivated by the operation of self-attention maps in frequency adaptive sampling, we leverage a cross-attention map corresponding to the word of the object that is to be changed, giving weights to resample tokens so that the corresponding locations can be resampled. Then, it enables efficient image editing with fewer steps and makes it easier to edit larger objects. Although it is similar to DiffEdit [7], ours operates in a more straightforward manner without additional computations, thanks to the capability of local refinement of the masked generative model. In Figure 6, we edit the object by adding 25% noise in only 10-steps. With our method, we can edit more quickly with fewer editing steps, which further does not suffer from the over-simplification problem. More samples are in the supplementary B.

4.4. High-resolution image synthesis

Continuous diffusion models that use encoders and decoders, such as masked image generative models, can generate larger images that are not in the training set. Similar to Bond-Taylor *et al*. [3], we divide the tokens into subsets according to the model input size, individually pass them through the model, and then spatially aggregate tokens to synthesize a high-resolution image. In this way, we are able to generate more realistic and high-resolution samples with the same TCTS model without additional training. Additionally, we propose a new method to generate high-resolution images only with low-resolution TCTS. To be specific, we first generate a small-size image with TCTS and upsample the token map in bicubic mode to the desired size. Then we divide the high-resolution token map into overlapping small-size sections and refine all the sections several times. Then, we can generate high-resolution images with the low-resolution TCTS where we visualize the generated samples in Figure 7.

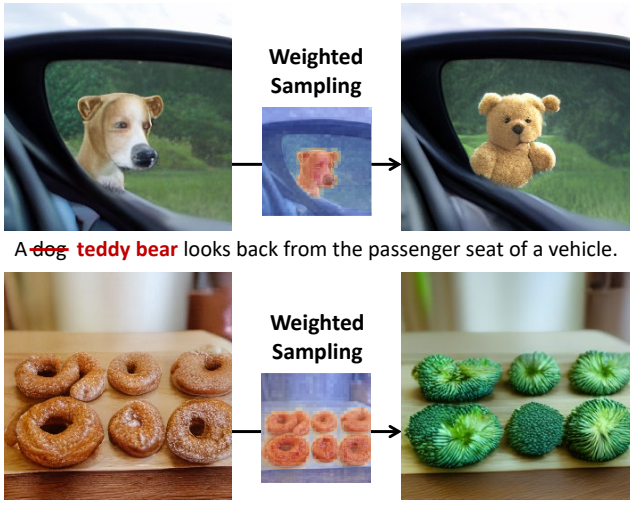

Figure 6: **Examples of mask-free editing samples with cross-attention map.** The cross-attention map is multiplied to the score map of TCTS to perform weighted sampling.

Figure 7: **Examples of high resolution (512^2) samples. Left**: Image generated with all mask tokens, **Right**: Image generated with TCTS as a super-resolution unit.

4.5. Further analysis

Importance of early-stage sampling strategy As mentioned in Section 3.1, by controlling the persistent weight, we can interpolate the two baseline sampling methods: uniform sampling and random revoke sampling. We designed various experiments to find out the impact of the sampling

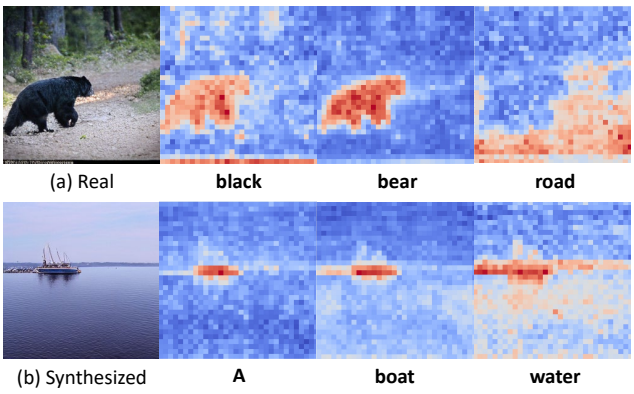

Figure 8: **Visualization of the averaged cross-attention maps for each word from TCTS. Top**: Image from COCO dataset: "*A large black bear walking down a dirty road.*". **Bottom**: Synthesized image: "*A small boat in a narrow body of water.*".

method with respect to the time steps. We switched the sampling methods during the time steps, from uniform sampling to random revoke sampling (U2R) and vice versa (R2U). In the rightmost plot in Figure 1, we can observe that U2R shows better text-alignment performance than R2U.

These results suggest that the early stage sampling, in which the masked ratio is high, substantially influences the text alignment of the final image, which is similarly observed in continuous diffusion models [2]. Generating the whole image and then processing it through a few revision steps like [23] might help the image quality but can not conspicuously enhance the essential text alignment. Therefore, in order to obtain an image that aligns well with the text, sampling must be done carefully in the early stage when not many tokens are generated. This analysis is also shown in the result that the revocable strategy exhibits better performance by aggressively resampling tokens several times. In particular, since weak generators with poor generation performance suffer more from the joint distribution issues, giving a sufficient number of recovery chances to discard tokens and draw new ones at the early stage is the key to the desired text alignment.

Text-conditioned sampling The experiments show that TCTS plays a crucial role in enhancing text alignment in image generation. Although we did not utilize any text-specific loss during training, the model still effectively processes the text condition and achieves optimal performance. To verify the accuracy of TCTS's attention towards text-related parts of the image, we visualized the average cross-attention layer for each word in Figure 8. The results demonstrate TCTS's ability to attend to each word in both real and synthesized images. By helping the selection of relevant tokens during the generating stage, TCTS can boost model performance without the requirement of additional complex text-related loss.

Method	MID-L ↑	SOA-I ↑	CLIP-S ↑	FID-30K ↓
RR	**26.77**	**81.10**	**0.2543**	18.43
RR + FAS	26.32	80.78	0.2539	16.01
RR + pw	24.23	79.14	0.2534	**15.96**
TCTS	**27.82**	**80.93**	**0.2565**	16.69
TCTS + FAS	27.79	80.87	0.2563	15.39
TCTS + pw	25.12	79.51	0.2556	**15.12**

Table 5: **Ablation study of using FAS and persistent weight in revocable methods on MS-COCO dataset.** All models were evaluated on 25-step setting. The ground truth MID-L on MS-COCO image-caption pairs is 54.73.

Step	Model	MID-L ↑	FID-30K ↓	Time
50	Uniform	24.04	16.84	×1
	RR	26.55	21.12	
16	TCTS + FAS	**26.72**	**14.45**	×**0.45**

Table 6: **Inference time relative to MID/FID of our baselines and models on MS-COCO dataset.** We expressed the inference time of our models as a multiple of other baselines. Our model outspeeds other baselines while performing better in both MID-L and FID-30K. Further analysis is in the supplementary C.

Ablation study of FAS As mentioned in Section 3.2, FAS decides whether to apply persistent weight or not to each location. If the whole image is considered low-frequency, FAS would apply the weight to every location, which is the same as the persistent sampling. On the other hand, if the whole image is considered high-frequency, combining FAS would not apply the weight anywhere, which does not change the sampling strategy. To further analyze the effect of FAS, we evaluated it on two revocable methods: random revoke sampling and TCTS. Attaching FAS to a base sampling method can be regarded as a mixture of two methods: the base revocable sampling method and persistent sampling method. Table 5 clearly shows this relation of FAS and persistent weight. While FAS-attached methods do not show the best performance, the performance of them is always in close proximity to the better one of the two comparing methods of each. FAS-attached methods balances well between text-alignment and image quality without any additional training.

Inference time Our TCTS and FAS compute additional operations at each generation step, resulting in marginally increased inference time compared to other token selection baselines. However, regarding the text alignment and the image quality, the relative inference time is considerably decreased. In Table 6, our model outperforms 50-step image generation of baseline sampling methods within only 16 steps in MID-L and FID-30K, which means that our model can synthesize better samples with only $\times 0.45$ time.

5. Conclusion

This paper examines which factors in the masked diffusion process impact the output images and cause the trade-off between image quality and text alignment. We empirically find that the revocable sampling significantly improves the text alignment yet degrades the quality of the generated images. To tackle the problem, we propose a simple token sampling strategy, coined text-conditioned token selection, that combines learnable and revocable methods, pushing the boundary of the trade-off between image quality and text alignment without the necessity to update the large-scale pre-trained generator. We find that collaborative sampling in a persistent and revocable manner surprisingly alleviates over-simplification issues in the generated backgrounds. Our proposed method can be utilized as an image refinement tool for various generative tasks, which is remarkably fast to generate high-quality images within much fewer steps.

6. Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-00075, Artificial Intelligence Graduate School Program (KAIST)) and (No. RS-2022-00187238, Development of Large Korean Language Model Technology for Efficient Pre-training). This work was supported by KAIST-NAVER Hypercreative AI Center.

References

[1] Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems*, pages 17981–17993. Curran Associates, Inc., 2021. 1, 3

[2] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 8

[3] Sam Bond-Taylor, Peter Hessey, Hiroshi Sasaki, Toby P Breckon, and Chris G Willcocks. Unleashing transformers: parallel token prediction with discrete absorbing diffusion for fast high-resolution image generation from vector-quantized codes. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIII*, pages 170–188. Springer, 2022. 7

[4] Arantxa Casanova, Marlene Careil, Jakob Verbeek, Michal Drozdzal, and Adriana Romero Soriano. Instance-conditioned gan. *Advances in Neural Information Processing Systems*, 34:27517–27529, 2021. 1

[5] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 1, 3, 6

[6] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11315–11325, 2022. 1, 3

[7] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance, 2022. 7

[8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 1

[9] 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. 2

[10] Wan-Cyuan Fan, Yen-Chun Chen, DongDong Chen, Yu Cheng, Lu Yuan, and Yu-Chiang Frank Wang. Frido: Feature pyramid diffusion for complex scene image synthesis. *arXiv preprint arXiv:2208.13753*, 2022. 1, 3

[11] Tsu-Jui Fu, Licheng Yu, Ning Zhang, Cheng-Yang Fu, Jong-Chyi Su, William Yang Wang, and Sean Bell. Tell me what happened: Unifying text-guided video completion via multimodal masked video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10681–10692, 2023. 3

[12] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. *arXiv preprint arXiv:2203.13131*, 2022. 1

[13] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10696–10706, 2022. 3

[14] Xiangteng He and Yuxin Peng. Fine-grained visual-textual representation learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(2):520–531, 2019. 5

[15] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528. Association for Computational Linguistics, 2021. 6

[16] 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. 6

[17] Tobias Hinz, Stefan Heinrich, and Stefan Wermter. Semantic object accuracy for generative text-to-image synthesis. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 6

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1, 3

[19] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, 23:47–1, 2022. 1

[20] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 5

[21] Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. Improving sample quality of diffusion models using self-attention guidance, 2022. 5

[22] Jin-Hwa Kim, Yunji Kim, Jiyoung Lee, Kang Min Yoo, and Sang-Woo Lee. Mutual information divergence: A unified metric for multimodal generative models. *arXiv preprint arXiv:2205.13445*, 2022. 6

[23] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Draft-and-revise: Effective image generation with contextual rq-transformer. *arXiv preprint arXiv:2206.04452*, 2022. 1, 3, 6, 8

[24] José Lezama, Huiwen Chang, Lu Jiang, and Irfan Essa. Improved masked image generation with token-critic. In *European Conference on Computer Vision*, pages 70–86. Springer, 2022. 2, 3, 6

[25] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 5

[26] 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. 1, 6

[27] 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. 5

[28] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 2

[29] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1

[30] 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. 1, 2

[31] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019. 2

[32] 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. 1, 3

[33] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 1, 3

[34] Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–10, 2022. 1

[35] Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. Foil it! find one mismatch between image and language caption. *arXiv preprint arXiv:1705.01359*, 2017. 6

[36] 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. 3

[37] Zhicong Tang, Shuyang Gu, Jianmin Bao, Chen Dong, and Fang Wen. Improved vector quantized diffusion models. *arXiv preprint arXiv:2205.16007*, 2022. 1, 2, 3, 5, 6

[38] Ming Tao, Hao Tang, Songsong Wu, Nicu Sebe, Xiao-Yuan Jing, Fei Wu, and Bingkun Bao. Df-gan: Deep fusion generative adversarial networks for text-to-image synthesis. *arXiv preprint arXiv:2008.05865*, 2020. 2, 6

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[40] 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, 6

[41] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021. 1, 6

[42] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. 1, 2

[43] Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, et al. Magvit: Masked generative video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10459–10469, 2023. 3

[44] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017. 2

[45] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stack-

gan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1947–1962, 2018. 2

[46] Yufan Zhou, Chunyuan Li, Changyou Chen, Jianfeng Gao, and Jinhui Xu. Lafite2: Few-shot text-to-image generation. *arXiv preprint arXiv:2210.14124*, 2022. 2

[47] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Towards language-free training for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17907–17917, 2022. 2, 6

[48] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5802–5810, 2019. 2