Improved Vector Quantized Diffusion Models

Zhicong Tang[1] **Shuyang Gu**[2] **Jianmin Bao**[3]
Dong Chen[3] **Fang Wen**[3]

[1]Tsinghua University
[2]University of Science and Technology of China
[3]Microsoft Research

tzc21@mails.tsinghua.edu.cn
gsy777@mail.ustc.edu.cn
{jianbao,doch,fangwen}@microsoft.com

Abstract

Vector quantized diffusion (VQ-Diffusion) is a powerful generative model for text-to-image synthesis, but sometimes can still generate low-quality samples or weakly correlated images with text input. We find these issues are mainly due to the flawed sampling strategy. In this paper, we propose two important techniques to further improve the sample quality of VQ-Diffusion. 1) We explore classifier-free guidance sampling for discrete denoising diffusion model and propose a more general and effective implementation of classifier-free guidance. 2) We present a high-quality inference strategy to alleviate the joint distribution issue in VQ-Diffusion. Finally, we conduct experiments on various datasets to validate their effectiveness and show that the improved VQ-Diffusion suppresses the vanilla version by large margins. We achieve an **8.44** FID score on MSCOCO, surpassing VQ-Diffusion by **5.42** FID score. When trained on ImageNet, we dramatically improve the FID score from **11.89** to **4.83**, demonstrating the superiority of our proposed techniques. The code is released at https://github.com/microsoft/VQ-Diffusion

1 Introduction

Denoising Diffusion Probability Models(DDPM) [17] have shown remarkable success in various fields, such as, image generation [16, 28, 33], video generation [20], audio generation [23] and so on. Depending on the signal type, DDPM could be roughly divided into continuous diffusion models [17, 27] and discrete diffusion models [2, 16]. In the forward step, the former adds Gaussian noise on continuous signals, while the latter uses a Markov Transition matrix to obfuscate discrete input tokens.

Extensive prior studies have focused on improving the DDPM from various aspects, including better network architecture [28], hierarchical structure design [33], alternative loss function [21], fast sampling strategy [22], *et al*. The current state-of-the-art models are already capable of producing photo-realistic images, like GLIDE [28] and DALL-E 2 [33] have presented impressive results in text-to-image synthesis. However, most previous improvements are based on continuous diffusion models and very few attempts are on discrete discrete diffusion models.

In this paper, we aim to improve the sample quality of discrete diffusion models, more specifically VQ-Diffusion [16], which leverages the VQVAE [30] to encode images to discrete tokens and then perform diffusion process in discrete space. One of the major strengths of VQ-Diffusion is that we can estimate the probability for each discrete token, thus it achieves high-quality images with

Preprint. Under review.

relatively fewer inference steps. Based on this, we introduce several techniques intended to improve VQ-Diffusion.

Discrete classifier-free guidance. For conditional image generation, suppose the condition information is y, and the generated image is x. The diffusion generative models try to maximize prior probability $p(x|y)$, and assume the generated images x will satisfy the constraints of posterior probability $p(y|x)$. However, we found this assumption may fail and ignore the posterior probability in most cases. We name this as the *posterior issue*. To address this issue, we propose to take both the prior and posterior into consideration simultaneously. Powered by posterior constraint, the generated images are significantly improved in terms of quality and consistency with input conditions. This approach shares the spirit of the previous classifier-free technique [19]. However, our methods are formulated more precisely since our model estimates probability instead of noise. Besides, instead of setting input conditions to zero, we introduce a more general and effective implementation of classifier-free guidance by using a learnable parameter as a condition to approximate $p(x)$. We found it could further improve the performance.

High-quality inference strategy. In each denoising step, we usually sample multiple tokens simultaneously and each token is sampled with estimated probability independently. However, different locations are often associated, thus sampling independently may ignore the dependencies. Assuming a simple dataset with only two samples: AA and BB. Each sample has 50% chances to appear. However, if we sample independently based on the estimated probability of each location, incorrect outputs (AB and BA) will appear, even if these samples never appear during training. We call this the *joint distribution issue*. To alleviate this issue, we introduce a high-quality inference strategy. It is based on two core designs. First, we reduce the number of sampled tokens at each step since more sampled tokens would suffer from the joint distribution issue more heavily. Second, we find that tokens with high confidence tend to be more accurate, thus we introduce purity prior to sample tokens with high confidence.

Powered by these techniques, we improve the sampling quality of VQ-Diffusion by large margins. We conduct experiments on CUB-200, MSCOCO, Conceptual Captions, and an even larger Internet dataset, and find that by fixing the posterior issue and the joint distribution issue, VQ-Diffusion could notably improve its performance. Concretely, we achieve an **8.44** FID score on MSCOCO, surpassing VQ-Diffusion by **5.42** FID score. When trained on class conditional ImageNet dataset, we dramatically improve the FID score from **11.89** to **4.83**. Above all, our key contribution contains three parts:

1. We find adding the posterior constraint will improve the quality of generated images significantly, and introduce a more general and effective implementation of classifier-free guidance.

2. We point out the joint distribution issue in VQ-Diffusion, and propose the High-quality sampling strategy to alleviate it.

3. We validate our approaches on various datasets, demonstrate these techniques improve VQ-Diffusion to achieve the state-of-the-art performance on various tasks.

2 Background: VQ-Diffusion

We first briefly review vector quantized diffusion (VQ-Diffusion) models and analyze the reason for the existence of joint distribution issue. VQ-Diffusion starts with a VQVAE that converts images x to discrete tokens $x_0 \in \{1, 2, ..., K, K+1\}$, K is the size of codebook, and $K+1$ denotes the $[\texttt{MASK}]$ token. Then the forward process of a diffusion model $q(x_t|x_{t-1})$ is a Markov chain that adds noise at each step. The reverse denoising process recovers the sample from a noise state. Specifically, the forward process is given by:

$$q(x_t|x_{t-1}) = \boldsymbol{v}^\top(x_t)\boldsymbol{Q}_t\boldsymbol{v}(x_{t-1}) \tag{1}$$

where $\boldsymbol{v}(x)$ is a one-hot column vector with entry 1 at index x. And $\boldsymbol{Q}_t$ is the probability transition matrix from x_{t-1} to x_t. Specifically, for the mask-and-replace VQ-diffusion strategy,

$$\boldsymbol{Q}_t = \begin{bmatrix} \alpha_t + \beta_t & \beta_t & \beta_t & \cdots & 0 \\ \beta_t & \alpha_t + \beta_t & \beta_t & \cdots & 0 \\ \beta_t & \beta_t & \alpha_t + \beta_t & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma_t & \gamma_t & \gamma_t & \cdots & 1 \end{bmatrix}. \tag{2}$$

given $\alpha_t \in [0, 1]$, $\beta_t = (1 - \alpha_t - \gamma_t)/K$ and γ_t the probability of a token to be replaced with a [MASK] token.

The reverse process is given by the posterior distribution:

$$q(x_{t-1}|x_t, x_0) = \frac{(v^T(x_t)\boldsymbol{Q}_t v(x_{t-1}))(v^T(x_{t-1})\bar{\boldsymbol{Q}}_{t-1} v(x_0))}{v^T(x_t)\bar{\boldsymbol{Q}}_t v(x_0)} \tag{3}$$

where $\bar{\boldsymbol{Q}}_t = \boldsymbol{Q}_t \cdots \boldsymbol{Q}_1$. The cumulative transition matrix $\bar{\boldsymbol{Q}}_t$ and the probability $q(x_t|x_0)$ can be computed in closed form with:

$$\bar{\boldsymbol{Q}}_t \boldsymbol{v}(x_0) = \bar{\alpha}_t \boldsymbol{v}(x_0) + (\bar{\gamma}_t - \bar{\beta}_t)\boldsymbol{v}(K + 1) + \bar{\beta}_t \tag{4}$$

Where $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, $\bar{\gamma}_t = 1 - \prod_{i=1}^{t}(1 - \gamma_i)$, and $\bar{\beta}_t = (1 - \bar{\alpha}_t - \bar{\gamma}_t)/(K + 1)$ can be calculated and stored in advance.

In the inference time, the denoising network p_θ gradually recovers the corrupted input via a fixed Markov chain. Besides, to stable the training and enable a fast inference strategy, VQ-Diffusion proposes a reparameterization trick that the denoising network predicts the denoised token distribution $p_\theta(\tilde{\boldsymbol{x}}_0|\boldsymbol{x}_t)$ at each step. Thus we can compute the reverse transition distribution according to:

$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \sum_{\tilde{\boldsymbol{x}}_0=1}^{K} q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \tilde{\boldsymbol{x}}_0) p_\theta(\tilde{\boldsymbol{x}}_0|\boldsymbol{x}_t). \tag{5}$$

We observed that VQ-Diffusion may suffer from the following two issues: 1) For conditional image generation, *e.g.*, text-to-image generation, the condition information y directly injected into the denoising network $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, y)$, and then the network is hoped to use both $\boldsymbol{x}_t$ and y to recover x_{x-1}. However, the network may ignore y, since x_t already contains sufficient information. Thus the generated image may not correlate to the input y well, causing the posterior issue. 2) For t-th timestep, each location of $\boldsymbol{x_{t-1}}$ is sampled from $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ independently. Thus, it could not model the correspondence among different locations. So sampling from this distribution parallel may be unreasonable, causing the joint distribution issue.

In the next section, we discuss these issues and propose a general classifier-free sampling strategy in the training procedure to solve the posterior constraint issue. Besides, we propose the high-quality inference strategy to alleviating the joint distribution issue in the inference stage.

3 Method

For the mismatch issue between generated image and input text, we propose the discrete classifier-free guidance as a posterior constraint to solve it. We also find a predefined prior on discrete space can help the sampling process and alleviate the joint distribution issue. Furthermore, we introduce the high-quality inference strategy to improve the sampling quality. We describe these techniques in the following section.

3.1 Discrete Classifier-free Guidance

For conditional image generation tasks like text-to-image synthesis, a mandatory requirement is that the generated images should match the condition input. VQ-Diffusion simply injects the condition information into the denoising network and suppose that the network will use both corrupted input and text to recover the original image. However, since the corrupted input usually contains much

more information than the text, the network may ignore the text in the training phase. Thus, we find the VQ-Diffusion may easily generate images with poor correlation with input text, which is often calculated with CLIP score [32].

From the perspective of the optimization target, the diffusion model aims to find x to maximize $p(x|y)$. However, a higher CLIP score also needs $p(y|x)$ as large as possible. Thus, a straight forward solution is to optimize $\log p(x|y) + s \log p(y|x)$, where s is a hyper-parameter to control the degree of posterior constraint. Using Bayes' theorem, we can derive this optimization target as follows:

$$\begin{aligned}
&\operatorname*{argmax}_x[\log p(x|y) + s \log p(y|x)] \\
=&\operatorname*{argmax}_x[(s+1)\log p(x|y) - s \log \frac{p(x|y)}{p(y|x)}] \\
=&\operatorname*{argmax}_x[(s+1)\log p(x|y) - s \log \frac{p(x|y)p(y)}{p(y|x)}] \\
=&\operatorname*{argmax}_x[(s+1)\log p(x|y) - s \log p(x)] \\
=&\operatorname*{argmax}_x[\log p(x) + (s+1)(\log p(x|y) - \log p(x))]
\end{aligned} \tag{6}$$

To predict the unconditional image logits $p(x)$, a direct way is to fine-tune the model with a certain percentage of empty condition inputs, like GLIDE [28] which set the input condition to a "null" text to fine-tune the model. However, we find using a learnable vector instead of text embedding of "null" can better fit the logits of $p(x)$. In the inference stage, we first generate the conditional image logits $p_\theta(x_{t-1}|x_t, y)$, then predict the unconditional image logits $p_\theta(x_{t-1}|x_t)$ by setting conditional input to the learnable vector. The next denoising step samples from:

$$\log p_\theta(x_{t-1}|x_t, y) = \log p_\theta(x_{t-1}|x_t) + (s+1)(\log p_\theta(x_{t-1}|x_t, y) - \log p_\theta(x_{t-1}|x_t)), \tag{7}$$

Compared with previous classifier-free sampling in continuous domain, our posterior constraint at discrete domain has three main difference: (1) First, since VQ-Diffusion leverage the reparameterization trick to predict $p(x|y)$ at unoised state, we may also apply Equation 6 on unnoised state, so it's compatible with other techniques like fast inference strategy [16] or high-quality inference strategy (Sec. 3.2). (2) Second, diffusion models at continuous setting do not predict the probability $p(x|y)$ directly; they use gradient to approximate it. However, discrete diffusion models estimate the probability distribution directly. (3) Third, continuous models set condition to null vector to predict $p(x)$. We find using a learnable vector instead of null vector could further improve the performance.

3.2 High-quality Inference Strategy

Another important issue in VQ-Diffusion is the joint distribution issue that caused by sampling token of different locations independently. This may ignore the correlation between different positions. To alleviate this issue, we propose a high-quality inference strategy, which includes two key techniques.

Fewer tokens sampling. First, we propose to sample fewer tokens at each step. In this way, we model the correlation between different positions by the iterative denoising process rather than ignoring it when sampling multiple tokens independently. Concretely, the number of changed tokens in each step of VQ-Diffusion is uncertain. For simplicity, we set the changed tokens to a certain number of tokens in each step.

For the state of each step, we can count its number of mask and choose the proper timestep as the timestep embedding. Suppose the input is x_t, we have two sets: $A_t := \{i|x_t^i = \texttt{[MASK]}\}$, and $B_t := \{i|x_t^i \neq \texttt{[MASK]}\}$. We aim to recover Δ_z [MASK] tokens from A_t in each step. Thus, the total inference steps are $T' = (H \times W)/\Delta_z$, H and W indicate the spatial resolution of tokens. Current timestep t could be calculated with $\operatorname{argmin}_t ||\frac{|A_t|}{H \times W} - \bar{\gamma}_t||_2$, where $|A|$ denotes number of elements in set A. When $\Delta_z = 1$, it has the same inference speed as autoregressive models. Our fewer tokens sampling is the opposite of previous broadly studied fast sampling strategy [16]. We seek to achieve high sampling quality by sacrificing inference time.

Purity prior sampling. From Equation 3, we can derive the following lemma:

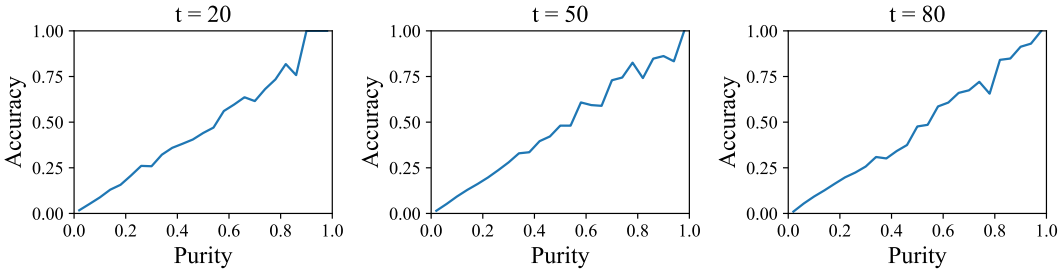

Figure 1: Illustration of the correlation between purity and accuracy of tokens at different timesteps(t=20, 50, and 80). We find high purity usually yields high accuracy.

Lemma 1. *For any position i which satisfied $x_t^i = [\texttt{MASK}]$, then $q(x_{t-1}^i = [\texttt{MASK}]|x_t^i = [\texttt{MASK}], x_0^i) = \bar{\gamma}_{t-1}/\bar{\gamma}_t$ is a constant.*

We leave the proof in the supplementary material. This lemma demonstrates that each position has the same probability to leave the [MASK] state. In other words, the transformation from [MASK] state to non-[MASK] state is position independent. However, we find different positions may have different confidence to leave [MASK] state. Specifically, positions with higher purity usually have higher confidence. We present the correlation between purity and accuracy in Figure 1. We find that a higher purity score usually indicates a more accurate token. Therefore, our key idea is conducting importance sampling rely on the purity score instead of random sampling. By leveraging this purity prior, each step could sample tokens from a more confident region, thus improving the sampling quality. The definition of purity at location i and timestep t is:

$$purity(i,t) = \max_{j=1\ldots K} p(x_0^i = j|x_t^i) \tag{8}$$

Based on these two techniques, we enable a high-quality inference strategy.

4 Experiments

4.1 Implementation details

Datasets To demonstrate the capability of our proposed techniques, we conduct experiments on three commonly used text-to-image synthesis datasets: CUB-200 [41], MSCOCO [26], and Conceptual Captions (CC) [38, 4]. Instead of using all the data from Conceptual Captions dataset, we follow the setting in [16], using a more balanced subset that contains 7M text-image pairs. Besides, to further demonstrate the scalability of our method, we collect 200 million high-quality text-images pairs from the internet. We named it ITHQ-200M dataset.

Backbone For a fair comparison with the original VQ-Diffusion and other previous text-to-image methods under the similar number of parameters, we build two different backbone settings: 1) Improved VQ-Diffusion-B (base), which consists of 370M parameters. We follow the network structure of VQ-Diffusion and directly use its released model as the pretrained model and fine-tune on each database. 2) Improved VQ-Diffusion-L (large), which consists of 1.27B parameters. The image decoder contains 36 transformer blocks with dimensions of 1408. To achieve a more general text-to-image generation model, we train this large model on the ITHQ-200M dataset. We adopt the base size model on other datasets for most experiments.

Evaluation metrics We use four metrics to evaluate the generated images. 1) FID score, which evaluates both quality and diversity of generated images. 2) Clip score [32], which measures the similarity between the generated image and text. 3) Quality Score(QS) [13], which only measures the image quality. A higher QS score denotes higher image quality. 4) Diversity Score(DS), defined as $1 -$ DDS where DDS denotes Diversity Difference Score proposed in [14]. It measures the diversity of generated images and a higher DS denotes more a diverse generated distribution.

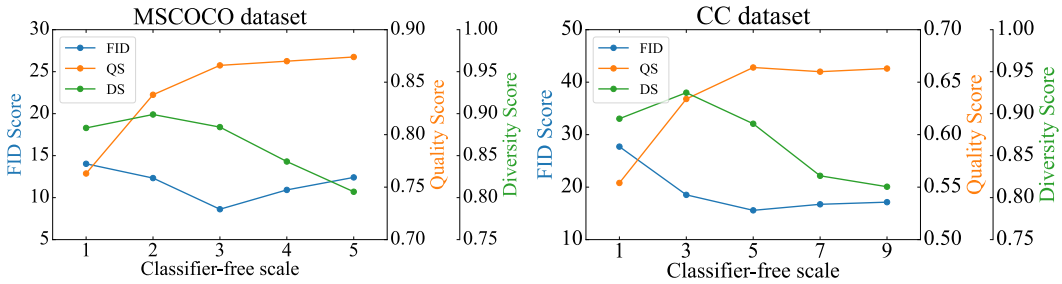

Figure 2: Ablation study on classifier-free scale. Image quality rises as guidance scale increases, while a large scale may cause the loss of image diversity.

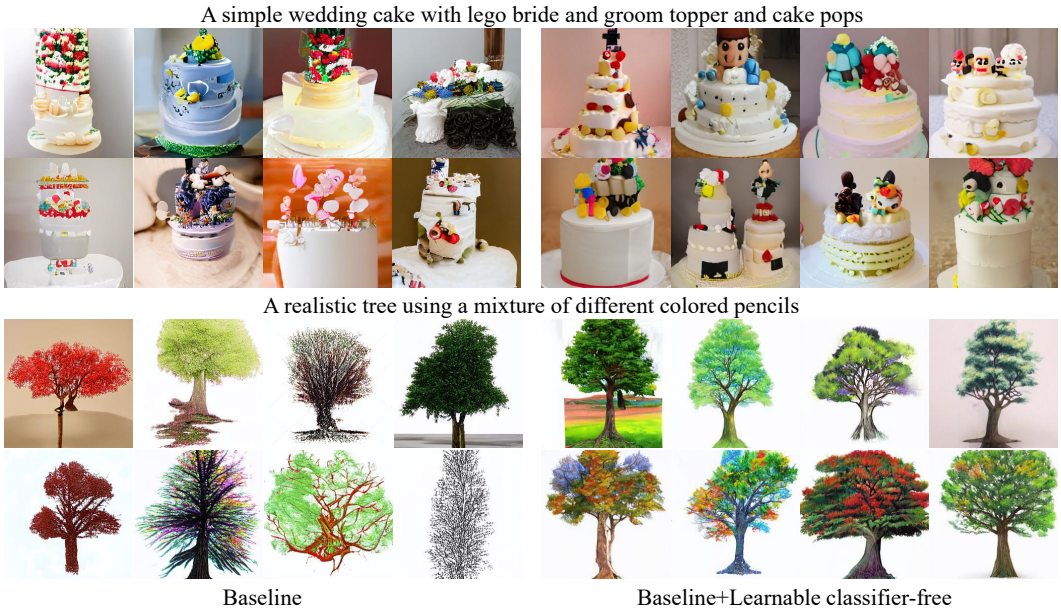

Figure 3: Results of learnable classifier-free sampling on CC dataset.

4.2 Ablation Studies

Discrete classifier-free guidance We investigate how the discrete classifier-free guidance could improve performance. We conduct the experiments on MSCOCO and CC datasets. Specifically, we compare four different settings: 1) The original VQ-Diffusion model. 2) Without fine-tuning the models, we directly set the conditional input to null vector in the inference stage and apply the classifier-free sampling strategy. 3) We set 10% of conditional input to null vector in the fine-tuning stage, and use the classifier-free guidance (Equation 7) to generate images. 4) Instead of setting conditional input to null vector, we set 10% of them to a learnable vector to fit the unconditional image.

We show the results in Table 1. We can find that classifier-free sampling could improve both the FID score and the Quality Score. Besides, the zero-shot classifier-free sampling strategy could improve performance without any training. By fine-tuning this model, the performance could be further improved. And using a learnable vector could achieve better performance than null vectors.

We also investigate how the guidance scale s affects the results. As shown in Fig 2, we conduct experiments on MSCOCO and CC datasets. We find that when s increases, the Quality Score becomes better and the Diversity Score decreases. It demonstrates that the classifier-free sampling is a balance between quality and diversity. Besides, the FID score achieves the best performance when s equals to 3 on MSCOCO, and 5 on CC dataset. Fig 3 provides a qualitative comparison of this posterior constraint. We can find by using the learnable classifier-free sampling strategy, the

quality of generated images is improved significantly, and they are more correlated to the input text. Meanwhile, the diversity decreases.

Table 1: Comparison of different classifier-free methods. Classifier-free outperforms baseline by a large margin and the learnable method further pushes performance.

	MSCOCO			CC		
	FID↓	QS↑	CLIP↑	FID↓	QS↑	CLIP↑
VQ-Diffusion	13.86	0.841	0.267	33.65	0.586	0.257
VQ-Diffusion + zero-shot CF	12.12	0.845	0.284	25.51	0.661	0.292
VQ-Diffusion + CF	8.85	0.864	0.302	16.44	0.647	0.298
VQ-Diffusion + learnable CF	**8.62**	**0.866**	**0.304**	**15.58**	**0.665**	**0.304**

High-quality inference strategy Previous work [16] proposed the fast sampling strategy which adopt inference steps fewer than training. We investigate the high-quality inference strategy in Table 2 which use more inference steps than training to achieve better performance. We perform the experiment on the CUB-200 dataset and evaluate the generated images of 25,50,100,200 inference steps on five models with different training steps. We find by increasing the inference steps, the diffusion model could generate images with a better FID score. And the performance continues to get better as the inference steps increase.

Table 2: FID score of high-quality inference strategy on CUB-200. The shaded part denotes fast inference strategy from [16].

	Training steps					
Inference steps		10	25	50	100	200
	10	32.35	27.62	23.47	19.84	20.96
	25	26.89	18.53	15.25	14.03	16.13
	50	22.70	16.93	13.82	12.45	13.67
	100	20.85	15.76	12.34	11.94	12.27
	200	19.99	15.55	12.20	11.87	11.80

Purity prior sampling We investigate the improvement of purity prior sampling. A token with higher purity demonstrates it has higher confidence to leave the [MASK] state. We conduct experiments on MSCOCO, CUB-200, CC and ITHQ-200M datasets. As shown in Table 3, we find that by adding the prior, all of these results are improved. Especially on larger datasets(CC and ITHQ-200M), the improvement is more significant. Meanwhile, this strategy requires neither training nor additional inference time. So it is an effective sampling strategy to improve the performance.

Table 3: FID score of purity prior sampling strategy.

	MSCOCO	CUB-200	CC	ITHQ-200M
VQ-Diffusion	13.86	10.32	33.65	25.87
VQ-Diffusion + prior	**13.79**	**10.21**	**33.09**	**25.15**

4.3 Compare with state-of-the-art methods

We compare the proposed method with several state-of-the-art text-to-image methods, including DF-GAN [40], XMC-GAN [42], DALL-E [34], GLIDE [28], and VQ-Diffusion [16], on MSCOCO and ITHQ-200M dataset. We evaluate FID score and show the results in Table 4 (a) and (c) respectively. Without any fine-tuning, we may leverage the zero-shot classifier-free sampling strategy and high-quality inference strategy to improve the performance of a well-trained VQ-Diffusion model, which is denoted as "Improved VQ-Diffusion*" in the table. Besides, by fine-tuning this model with the learnable classifier-free strategy, the performance is further improved. We provide the visualized

Figure 4: Qualitative comparison with previous works on MSCOCO dataset.

comparison with previous works on MSCOCO datasets in Fig 4, where our method could generate significantly better results. Besides, we provide the visualization results of the in-the-wild text-to-image synthesis results on VQ-Diffusion-L model in Fig 5. Our method could generate very impressive results.

The proposed improved VQ-Diffusion is general, as it can also be applied to class conditional ImageNet generation tasks. We compare with BigGAN [3], VQGAN [11], ImageBART [10], and VQ-Diffusion [16]. The results are shown in Table 4 (b). Our Improved VQ-Diffusion achieves the best result among all compared methods.

Table 4: Comparison with previous methods on various datasets in terms of FID score. * indicates that we use only zero-shot classifier-free sampling and purity prior sampling.

Methods	FID↓
DFGAN [40]	21.42
GLIDE [28]	12.24
XMC-GAN [42]	9.33
VQ-Diffusion [16]	13.86
Improved VQ-Diffusion*	11.89
Improved VQ-Diffusion	**8.44**

(a) MSCOCO

Methods	FID↓
ImageBART [10]	21.19
VQGAN [11]	15.78
BigGAN [3]	7.53
VQ-Diffusion [16]	11.89
Improved VQ-Diffusion*	7.65
Improved VQ-Diffusion	**4.83**

(b) ImageNet

Methods	FID↓
VQ-Diffusion [16]	25.87
Improved VQ-Diffusion*	22.16
Improved VQ-Diffusion	**19.06**

(c) ITHQ-200M

5 Related Works

Text-to-Image Synthesis In the past few years, Generative Adversarial Networks (GANs) have inspired many advances in image synthesis [35, 43, 6, 15, 44, 12, 24, 42, 36, 9, 5, 39]. For text-to-image generation, most works could generate high fidelity images on single domain datasets, e.g., birds [41] and flowers [29]. For more complex scenes with multiple objects, such as MSCOCO dataset [26], most GAN-based methods struggled to model these complicated distributions.

Other approaches are the Autoregressive Models (ARs). These models convert the complex distributions into multiple conditional distributions relying on previous outputs, thus have a stronger capability to model complex distributions. In recent works, DALL-E [34] adopts a VQVAE to embed images to discrete tokens, and uses an AR model to fit the token distribution. It achieves impressive results on text-to-image generations. Other works like Cogview [8], and M6 [25] have also achieved very promising results based on autoregressive models.

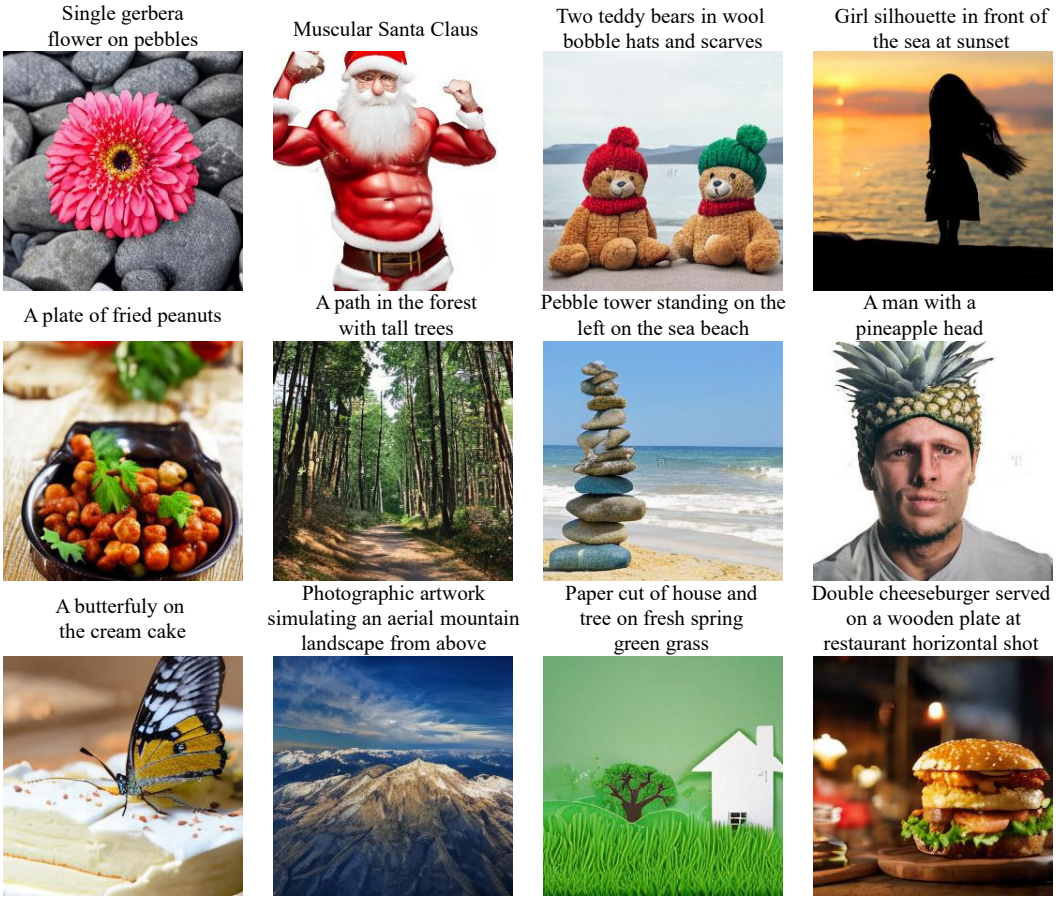

Figure 5: In-the-wild text-to-image synthesis results of Improved VQ-Diffusion.

Denoising Diffusion Probabilistic Model Recently, an emerging generative model, the denoising diffusion probabilistic model (DDPM), has attracted great attention in the community. It has achieved strong results on image generation [17, 27, 18, 7, 1, 37], video generation [20], and audio generation [23, 31]. It was first proposed in [17], and the subsequent work [27] proposes a reparameterization to stabilize the training. [28] proposed to use a classifier to guide the sampling process, which significantly improved the image quality. [19] further discarded the classifier by introducing an additional fine-tune procedure. DALL-E 2 [33] applied DDPM to massive data and achieved very impressive results on text-to-image synthesis tasks.

Meanwhile, other researchers investigate the discrete diffusion models. D3PMs [2] propose to use a transition matrix instead of Gaussian to add noise on discrete distributions. VQ-Diffusion [16] uses a VQVAE to encode images into discrete tokens and leverages the discrete DDPM to model the discrete distribution. It achieved very strong performance on many image synthesis tasks.

6 Conclusion

In this paper, we carefully studied VQ-Diffusion and identify it suffers from two main issues: the posterior issue and the joint distribution issue. To address these issues, we propose two techniques and improve the quality of the generated samples and their consistency with the input text by a large margin. Moreover, our strategies can even benefit VQ-Diffusion without fine-tuning the model. We demonstrate the superiority of our methods on various datasets. We hope our work opens the path for exploring VQ-Diffusion and facilitating future research.

References

[1] Oron Ashual, Shelly Sheynin, Adam Polyak, Uriel Singer, Oran Gafni, Eliya Nachmani, and Yaniv Taigman. Knn-diffusion: Image generation via large-scale retrieval. *arXiv preprint arXiv:2204.02849*, 2022.

[2] Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. *arXiv preprint arXiv:2107.03006*, 2021.

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[4] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021.

[5] Jun Cheng, Fuxiang Wu, Yanling Tian, Lei Wang, and Dapeng Tao. Rifegan: Rich feature generation for text-to-image synthesis from prior knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10911–10920, 2020.

[6] Ayushman Dash, John Cristian Borges Gamboa, Sheraz Ahmed, Marcus Liwicki, and Muhammad Zeshan Afzal. Tac-gan-text conditioned auxiliary classifier generative adversarial network. *arXiv preprint arXiv:1703.06412*, 2017.

[7] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *arXiv preprint arXiv:2105.05233*, 2021.

[8] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *arXiv preprint arXiv:2105.13290*, 2021.

[9] Alaaeldin El-Nouby, Shikhar Sharma, Hannes Schulz, Devon Hjelm, Layla El Asri, Samira Ebrahimi Kahou, Yoshua Bengio, and Graham W Taylor. Tell, draw, and repeat: Generating and modifying images based on continual linguistic instruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10304–10312, 2019.

[10] Patrick Esser, Robin Rombach, Andreas Blattmann, and Björn Ommer. Imagebart: Bidirectional context with multinomial diffusion for autoregressive image synthesis. *arXiv preprint arXiv:2108.08827*, 2021.

[11] 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.

[12] Lianli Gao, Daiyuan Chen, Jingkuan Song, Xing Xu, Dongxiang Zhang, and Heng Tao Shen. Perceptual pyramid adversarial networks for text-to-image synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8312–8319, 2019.

[13] Shuyang Gu, Jianmin Bao, Dong Chen, and Fang Wen. Giqa: Generated image quality assessment. In *European Conference on Computer Vision*, pages 369–385. Springer, 2020.

[14] Shuyang Gu, Jianmin Bao, Dong Chen, and Fang Wen. Priorgan: Real data prior for generative adversarial nets. *arXiv preprint arXiv:2006.16990*, 2020.

[15] Shuyang Gu, Jianmin Bao, Hao Yang, Dong Chen, Fang Wen, and Lu Yuan. Mask-guided portrait editing with conditional gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3436–3445, 2019.

[16] 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. *arXiv preprint arXiv:2111.14822*, 2021.

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arXiv:2006.11239*, 2020.

[18] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *arXiv preprint arXiv:2106.15282*, 2021.

[19] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

[20] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022.

[21] Chin-Wei Huang, Jae Hyun Lim, and Aaron C Courville. A variational perspective on diffusion-based generative models and score matching. *Advances in Neural Information Processing Systems*, 34, 2021.

[22] Zhifeng Kong and Wei Ping. On fast sampling of diffusion probabilistic models. *arXiv preprint arXiv:2106.00132*, 2021.

[23] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020.

[24] Qicheng Lao, Mohammad Havaei, Ahmad Pesaranghader, Francis Dutil, Lisa Di Jorio, and Thomas Fevens. Dual adversarial inference for text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7567–7576, 2019.

[25] Junyang Lin, Rui Men, An Yang, Chang Zhou, Ming Ding, Yichang Zhang, Peng Wang, Ang Wang, Le Jiang, Xianyan Jia, et al. M6: A chinese multimodal pretrainer. *arXiv preprint arXiv:2103.00823*, 2021.

[26] 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.

[27] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. *arXiv preprint arXiv:2102.09672*, 2021.

[28] 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.

[29] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008.

[30] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *arXiv preprint arXiv:1711.00937*, 2017.

[31] Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*, pages 8599–8608. PMLR, 2021.

[32] 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. *arXiv preprint arXiv:2103.00020*, 2021.

[33] 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.

[34] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021.

[35] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International Conference on Machine Learning*, pages 1060–1069. PMLR, 2016.

[36] Shulan Ruan, Yong Zhang, Kun Zhang, Yanbo Fan, Fan Tang, Qi Liu, and Enhong Chen. Dae-gan: Dynamic aspect-aware gan for text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13960–13969, 2021.

[37] Chitwan Saharia, William Chan, Huiwen Chang, Chris A Lee, Jonathan Ho, Tim Salimans, David J Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. *arXiv preprint arXiv:2111.05826*, 2021.

[38] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.

[39] Douglas M Souza, Jônatas Wehrmann, and Duncan D Ruiz. Efficient neural architecture for text-to-image synthesis. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

[40] 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.

[41] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[42] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 833–842, 2021.

[43] 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.

[44] Zizhao Zhang, Yuanpu Xie, and Lin Yang. Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6199–6208, 2018.

A Proof of Lemma 1

Lemma 1. *For any position i which satisfied $x_t^i = [\texttt{MASK}]$, then $q(x_{t-1}^i = [\texttt{MASK}]|x_t^i = [\texttt{MASK}], x_0^i) = \bar{\gamma}_{t-1}/\bar{\gamma}_t$ is a constant.*

Proof. For any position i that satisfied $x_t^i = [\texttt{MASK}]$,

$$\begin{aligned}
&q(x_{t-1}^i = [\texttt{MASK}]|x_t^i = [\texttt{MASK}], x_0^i) \\
&= q(x_{t-1}^i = K+1|x_t^i = K+1, x_0^i) \\
&= \frac{(v^T(x_t^i)\boldsymbol{Q}_t v(x_{t-1}^i))(v^T(x_{t-1}^i)\bar{\boldsymbol{Q}}_{t-1} v(x_0^i))}{v^T(x_t^i)\bar{\boldsymbol{Q}}_t v(x_0^i)} \\
&= \frac{(v^T(K+1)\boldsymbol{Q}_t v(K+1))(v^T(K+1)\bar{\boldsymbol{Q}}_{t-1} v(x_0^i))}{v^T(K+1)\bar{\boldsymbol{Q}}_t v(x_0^i)} \\
&= \frac{v^T(K+1)\bar{\boldsymbol{Q}}_{t-1} v(x_0^i)}{v^T(K+1)\bar{\boldsymbol{Q}}_t v(x_0^i)}.
\end{aligned} \tag{9}$$

Since x_0 is the unnoised state, we know that $x_0^i \neq [\texttt{MASK}]$. So we have

$$\begin{aligned}
&q(x_{t-1}^i = [\texttt{MASK}]|x_t^i = [\texttt{MASK}], x_0^i) \\
&= \frac{v^T(K+1)\bar{\boldsymbol{Q}}_{t-1} v(x_0^i)}{v^T(K+1)\bar{\boldsymbol{Q}}_t v(x_0^i)} \\
&= \frac{\bar{\gamma}_{t-1}}{\bar{\gamma}_t}.
\end{aligned} \tag{10}$$

B High-quality inference strategy Details

B.1 Fewer tokens sampling strategy

Considering β_t in the probability transition matrix $\boldsymbol{Q}_t$ is set to a tiny scale in implementations, we can ignore the replace situation. The detailed fewer tokens sampling strategy is shown in Algorithm 1. For each timestep t, we first denote all the locations with $[\texttt{MASK}]$ as set $A_t = \{i|x_t^i = [\texttt{MASK}]\}$. Then, we choose Δ_z items from A_t by random sampling, we denote this set as C_t. Finally, we sample $\boldsymbol{x}_{0,t}$ from $p_\theta(\tilde{\boldsymbol{x}}_0|\boldsymbol{x}_t, \boldsymbol{y})$ and replace the token $\boldsymbol{x}_t^i$ with $\boldsymbol{x}_{0,t}^i$ for all locations $i \in C_t$. Sampling $\boldsymbol{x}_{0,t}$ from $p_\theta(\tilde{\boldsymbol{x}}_0|\boldsymbol{x}_t, \boldsymbol{y})$ has a similar effect as sampling from $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{y})$, but it could ensure to leave the $[\texttt{MASK}]$ state.

Algorithm 1 Fewer tokens sampling strategy. Assume sampling Δ_z tokens each time, input text s.

$t \leftarrow T$, $\boldsymbol{y} \leftarrow \text{BPE}(s)$
$\boldsymbol{x}_t \leftarrow$ sample from $p(\boldsymbol{x}_T)$
while $t > 0$ **do**
    $A_t = \{i|x_t^i = [\texttt{MASK}]\}$
    $C_t = \{c_1, c_2, \ldots, c_{\Delta_z}\} \leftarrow$ random sample from A_t
    $\boldsymbol{x}_{0,t} \leftarrow$ sample from $p_\theta(\tilde{\boldsymbol{x}}_0|\boldsymbol{x}_t, \boldsymbol{y})$
    for $i = c_1, c_2, \ldots, c_{\Delta_z}$ **do**
        $x_t^i \leftarrow x_{0,t}^i$
    end for
    $t \leftarrow \text{argmin}_t \, ||\frac{|A|-\Delta_z}{H \times W} - \bar{\gamma}_t||_2$
end while
return VQVAE-Decoder($\boldsymbol{x}_t$)

B.2 Purity prior sampling strategy

The detailed purity prior sampling strategy is shown in Algorithm 2. It has similar implementation to the fewer tokens sampling but does not need to sacrifice the inference time, as we can set Δ_z according to normal inference timesteps.

It has two differences with Algorithm 1. 1) It calculates the purity for each location, and chooses locations with importance sampling rather than random sampling. 2) It adjusts the probability $\tilde{\boldsymbol{x}}_0$ with Equation 11 since we find that larger kurtosis for locations with high purity helps improving quality. r is a hyper-parameter and normally the range is $(0.5, 2)$ in our experiments.

$$\tilde{\boldsymbol{x}}_0 = \text{softmax}\left((1 + purity_t \cdot r) \log \tilde{\boldsymbol{x}}_0\right) \tag{11}$$

Algorithm 2 Purity prior sampling strategy. Assume sampling Δ_z tokens each time, input text s, prior scale r.

$t \leftarrow T$, $\boldsymbol{y} \leftarrow \text{BPE}(s)$
$\boldsymbol{x}_t \leftarrow$ sample from $p(\boldsymbol{x}_T)$
while $t > 0$ **do**
  $A_t = \{i | x_t^i = [\texttt{MASK}]\}$
  $purity_t^i \leftarrow \max_{j=1\ldots K} p(\tilde{x}_0^i = j | x_t^i)$ ▷ Eqn. 8
  $C_t = \{c_1, c_2, \ldots, c_{\Delta_z}\} \leftarrow$ importance sample from A_t with $purity_t$
  $\tilde{\boldsymbol{x}}_0 \leftarrow \text{softmax}\left((1 + purity_t \cdot r) \log \tilde{\boldsymbol{x}}_0\right)$ ▷ Eqn. 11
  $\boldsymbol{x}_{0,t} \leftarrow$ sample from $p_\theta(\tilde{\boldsymbol{x}}_0 | \boldsymbol{x}_t, \boldsymbol{y})$
  for $i = c_1, c_2, \ldots, c_{\Delta_z}$ **do**
    $x_t^i \leftarrow x_{0,t}^i$
  end for
  $t \leftarrow \text{argmin}_t \, || \frac{|A| - \Delta_z}{H \times W} - \bar{\gamma}_t ||_2$
end while
return VQVAE-Decoder($\boldsymbol{x}_t$)

C Results

In Figure 6, Figure 7 and Figure 8, we provide more visualization results of in-the-wild text-to-image synthesis on Improved VQ-Diffusion-L model.

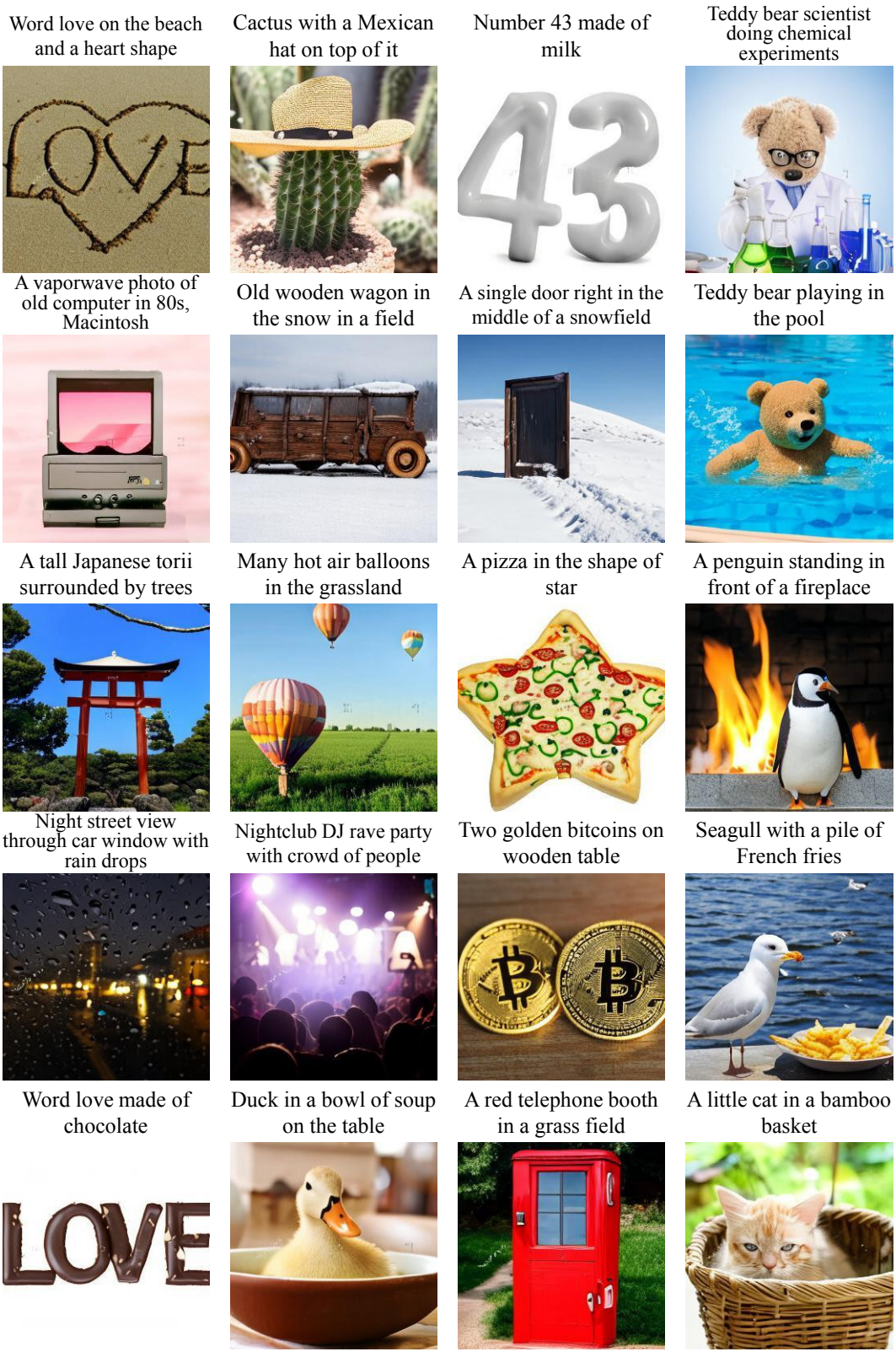

Figure 6: In-the-wild text-to-image synthesis with Improved VQ-Diffusion-L.

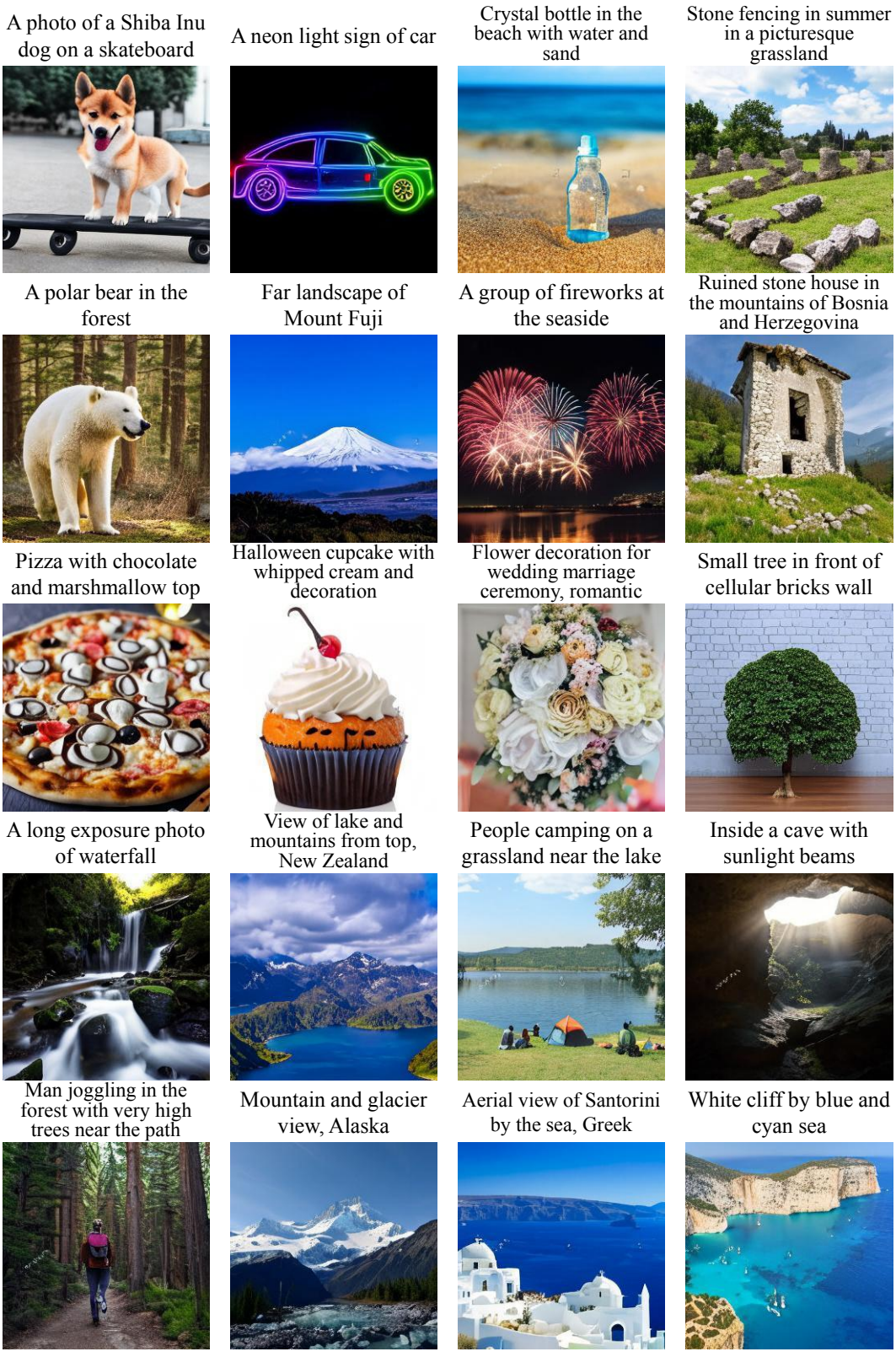

Figure 7: In-the-wild text-to-image synthesis with Improved VQ-Diffusion-L.

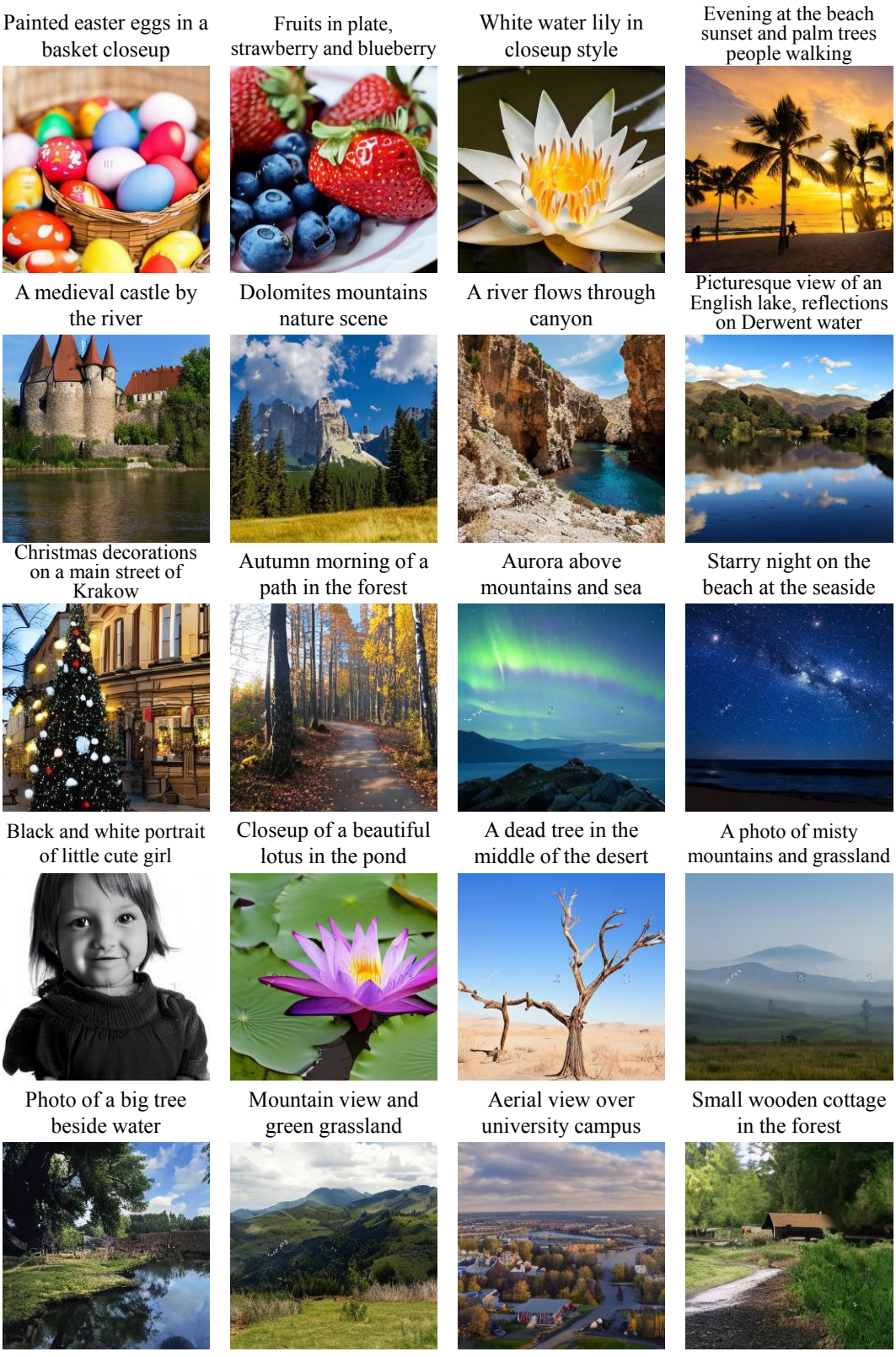

Figure 8: In-the-wild text-to-image synthesis with Improved VQ-Diffusion-L.