A. Implementation Details

A.1. Guidance sampling training

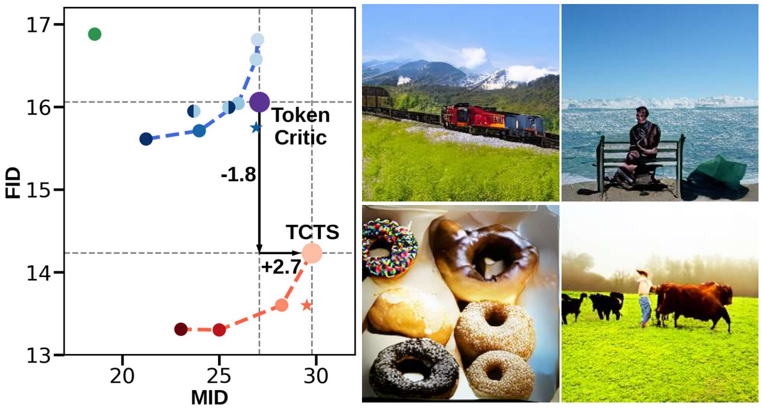


Figure 9: **Left**: Comparison with Token-Critic w/ fixed CF-guidance scale $s = 5$, **Right**: Generated samples with the same text condition in Figure 1, which show issues with the overall quality.

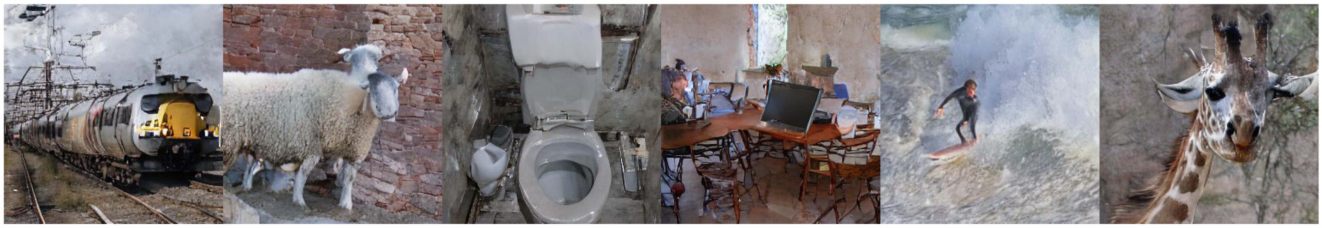


Figure 10: **Generated samples with TCTS trained with no classifier-free guidance (FID-30K: 27.13, MID-L: 12.34).**

Classifier-free guidance [20] is a key factor for the image quality of text-to-image diffusion models [26, 33], and also has been successfully applied to the token-based models [5, 37]. Since classifier-free guidance is used at inference time, it must also be used during the training process. However, determining the guidance scale during training is a difficult problem. We found out that either overly high or low guidance scale can deteriorate the training process of TCTS. The training procedure of TCTS starts with masking random tokens of a real image. Then, a fixed generator reconstructs the image and TCTS is trained to find the originally masked locations. However, since high guidance scale boosts the reconstruction capacity, TCTS suffers from finding the masked locations and tends to output a smooth distribution. In Figure 9, the vanilla *Token Critic* using fixed higher CF guidance suffers from performance degradation. On the other hand, a lowercase with poor reconstruction performance provides the model with diverse and easy samples, making the learning process faster and more stable. However, since a high guidance scale is used during actual inference, the model exhibits very low performance. The samples are in Figure 10. Therefore, we stochastically sample the guidance scale in the training procedure as a regularization of the difficulty of the task. This guidance scale controlling method stabilizes training, improves performance, and enables various guidance scale settings at inference time.

A.2. Frequency Adaptive Sampling

The area where highly detailed information is obtained in Figure 11 has large values in the high-frequency range. It was observed that as the generating process progressed, the values gradually decreased in all areas, especially in the simplified areas. Changes were first observed in the areas that were simplified as a priority. In our FAS method, we divided this area using a threshold, and the visualization of this is shown in Figure 12. This is similar to the object mask or frequency mask that can be found in [21]. The detailed algorithm is in Algorithm 2.

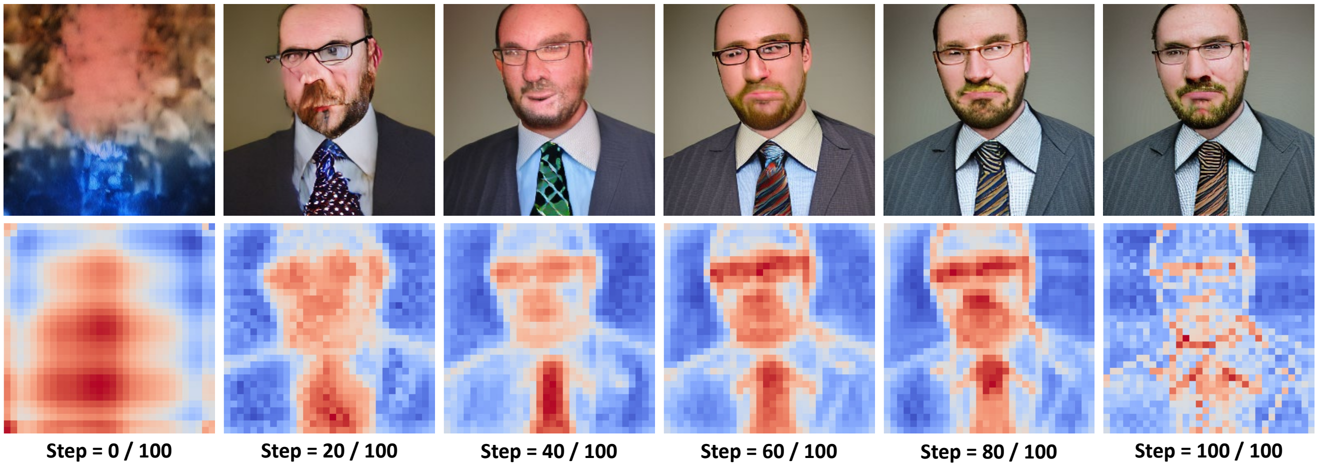


Figure 11: **Visualization of self-attention map**. **Top**: Reconstructed images in each step. **Bottom**: Visualization of self-attention maps for each step.

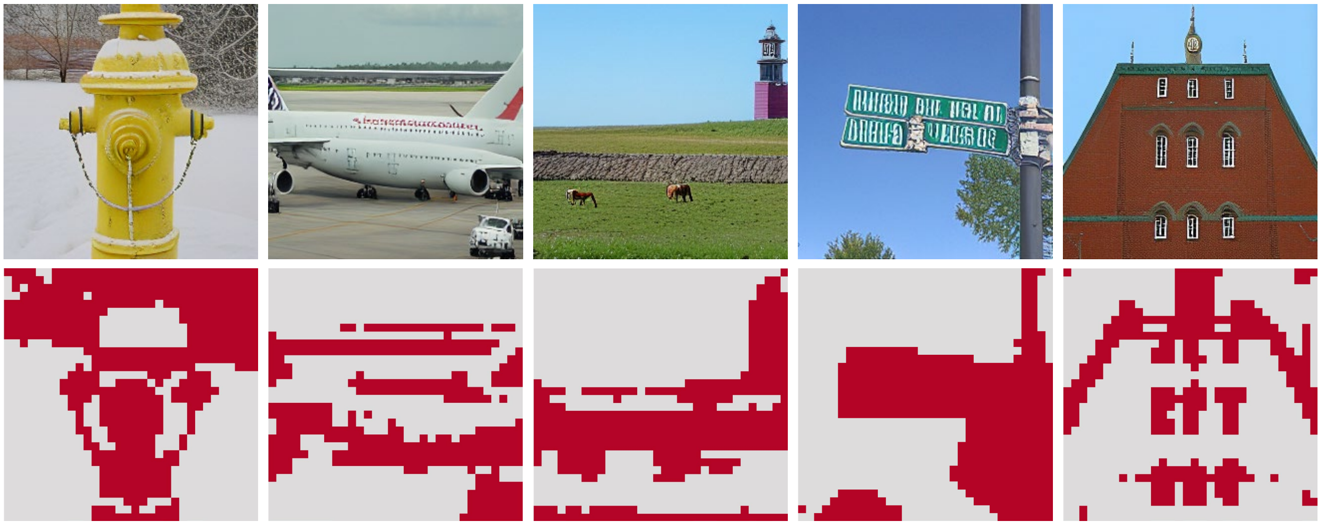


Figure 12: **Visualization of self-attention map with threshold Top**: Synthesized images. **Bottom**: Visualization our self-attention masks. ($\phi = 0.5$)

In addition, unlike [21], the self-attention map here uses sigmoid instead of softmax on the values before the softmax calculation in the original transformer.

$$map_{sa}^{(h)} = sigmoid(Q_t^{(h)}(K_t^{(h)})^T/\sqrt{d}) \tag{2}$$

$$map_{sa} = GAP(map_{sa}^{(h)}) \tag{3}$$

This is because the VQ diffusion model [13] uses a large embedding dimension of 1024, causing the sum of softmax values to decrease too much depending on the location. If we only multiply the persistent weight by the low-frequency location without the process shown in lines 5 and 10 above, the tokens corresponding to the low-frequency location will be more likely to remain probabilistically during the sampling process, which can unintentionally hinder the generation of the object. Therefore,

Algorithm 2 Frequency Adaptive Sampling

```
Input: $w$: persistent weight, $map_{sa}$: self-attention map, $\phi$: self-attention threshold
$G_\theta$: Generator, $D_\gamma$: TCTS model, $c$: text condition embedding
$\hat{x}_0, map_{sa} = G_\theta(x_t, c)$
$map_{tc} = D_\gamma(\hat{x}_0, c) \leftarrow$ TCTS probability map
$A_t = \{i | x_t^i =$[MASK]$\}$
$L_t = \{i | map_{sa}^i < \phi\} \leftarrow$ Low frequency location
$a = 1 + (w - 1) \times (n(A^C) \div N)$
for $i \in A_t^C \cap L_t$ do
    $map_{tc}^i = map_{tc}^i \times w$
end for
for $i \in A_t$ do
    $map_{tc}^i = map_{tc}^i \times a$
end for
Return: $map_{tc}$
```

multiplying weight "a" in high-frequency locations helps maintain the ratio of low and high-frequency tokens while giving the effect of setting the persistent weight to 1.

A.3. Hyper-parameter setting

In our experiment, we set the self-attention threshold (ϕ) to 0.45 and the persistent weight to 15 for hyper-parameter setting. For our learnable TCTS model, we used same architecture in VQ diffusion [13], but we reduced the number of layers from 19 to 16 when using the COCO dataset, and we reduced the hidden embedding size from 1024 to 512 when using the CUB dataset.

B. Additional Samples

B.1. Over-simplification samples

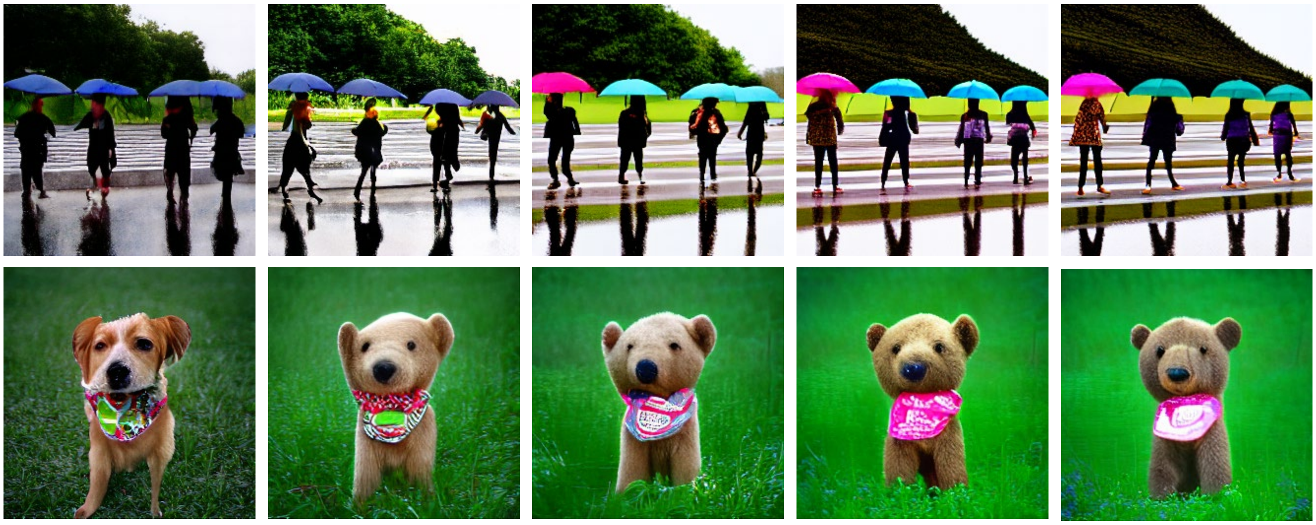


Figure 13: **Over-simplification samples by revocable schedule with long inference steps. Top**: Image generation process with a long inference steps (100 steps). **Bottom**: Mask-free object editing with long steps, "A **dog** with his tongue hanging out in a field" to "A **bear** with his tongue hanging out in a field"

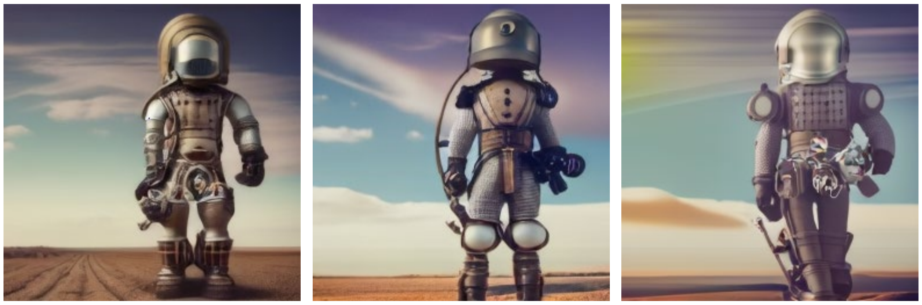


Figure 14: **Oversimplification with long steps in *Paella*.** Recently proposed method *Paella* (Rampas et al. 2023) uses token replacement instead of a mask-based approach and could be considered a revocable method. All samples are generated in 100 steps. The numbers denote the relative number of resampling. **Left**: Custom fixed-like method (×1), **Middle**: Renoising until 50 steps (×6), **Right**: Renoising in all steps (×12). Resampling more tokens make the overall pattern of the background image to be oversimplified. The excessive number of resampling in longer steps results in oversimplification, and these observations agree with our analysis.

B.2. Mask-free object editing

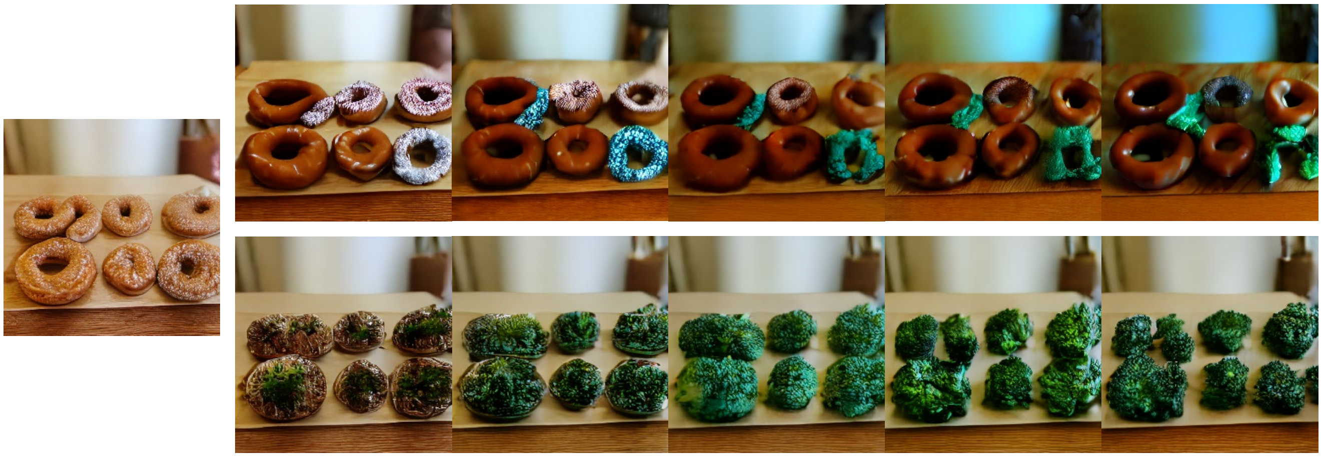


Figure 15: **Mask-free object editing samples with and without cross-attention map weighted sampling**. Starting from the image on the left, the result images every 20 steps of editing with 30% masking ratio. **Top**: Failure case without weighted sampling, **Bottom**: Results with weighted sampling.

In mask-free object editing, it is challenging to change large objects, converting donuts into broccolis for example, with a low masking ratio. See Figure 15. This is because the distributions of token for each objects are entirely different, and even if some parts are masked, the surrounding tokens of the original object can still influence the outcome.

Additionally, in the generating process, our model cannot directly find text-misaligned tokens. Even if the text does not match the current tokens, the TCTS score map is not centered on the misaligned object. Since TCTS does not play a role as a text-misaligned token detector, resampling of the whole image can lead to significant changes in unnecessary parts, such as the background. To address this issue, we propose using a cross-attention map to give more weight to sampling around the object of interest, minimizing unnecessary resampling of backgrounds, which leads to easy editing of the object.

B.3. Image Refinement

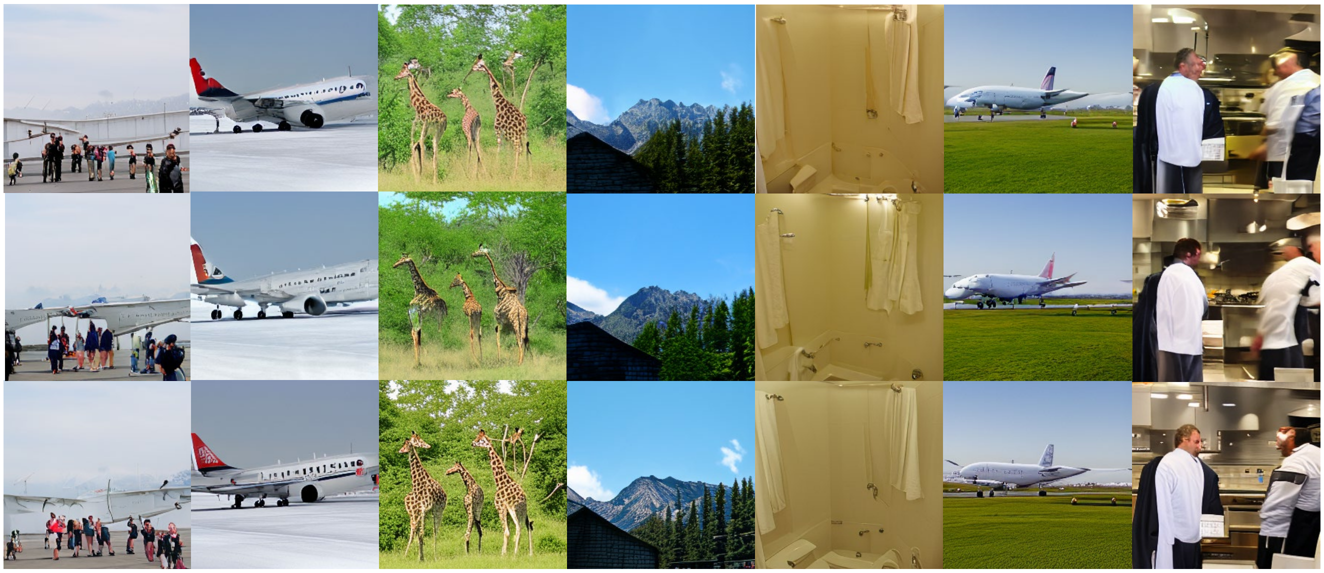


Figure 16: **Random image samples with additional refinement steps. Top**: Original images with *uniform sampling* in 16 steps, **Middle**: Refined images with random noise, **Bottom**: Refined images with TCTS.

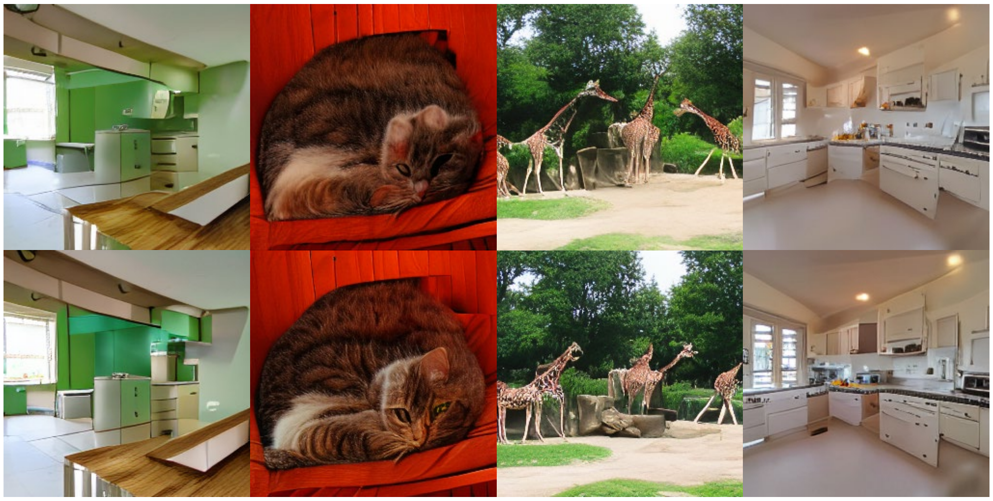
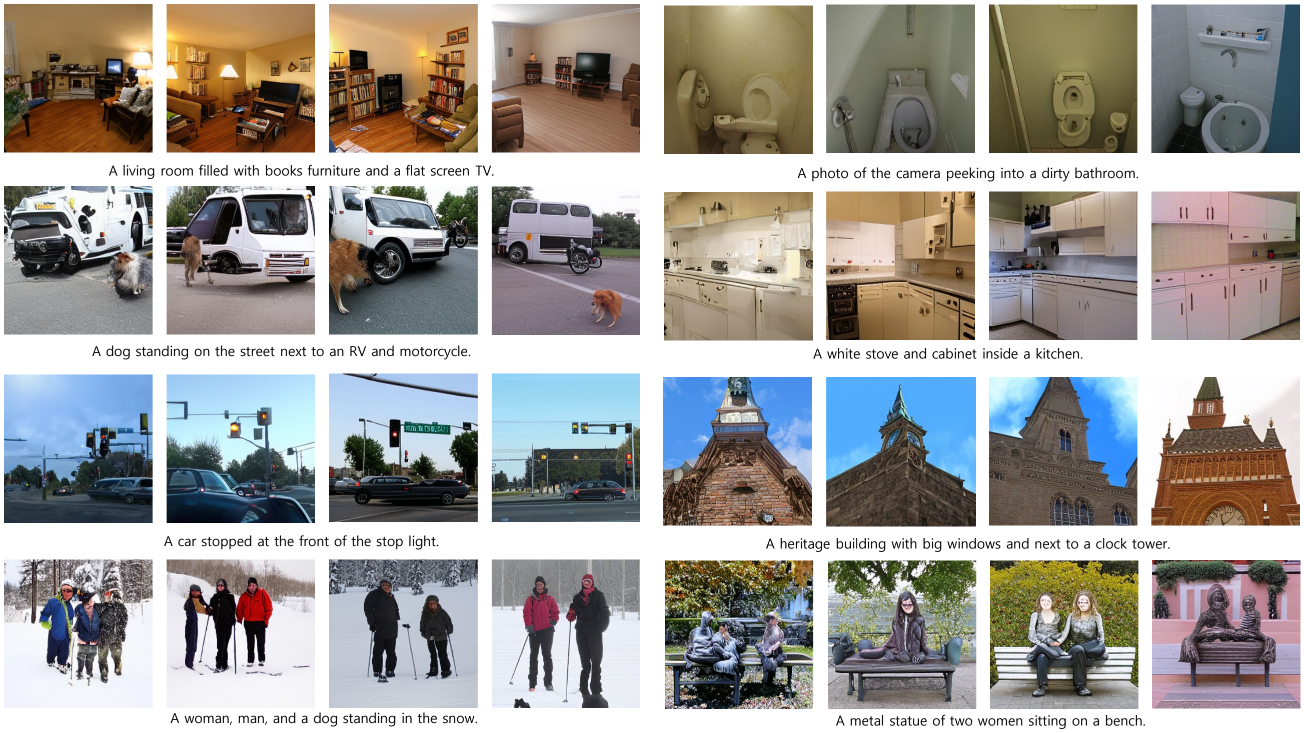


Figure 17: **Random image samples by refinement with masking lowest-scoring tokens. Top**: Original images with *uniform sampling* in 16 steps, **Bottom**: Refined images with masking TCTS lowest-scoring tokens. As mentioned earlier, since it is regenerated using the same uniform sampling method, it is difficult to confirm a noticeable improvement in image quality. However, there was an improvement in performance in terms of FID and MID.



A living room filled with books furniture and a flat screen TV.

A photo of the camera peeking into a dirty bathroom.

A dog standing on the street next to an RV and motorcycle.

A white stove and cabinet inside a kitchen.

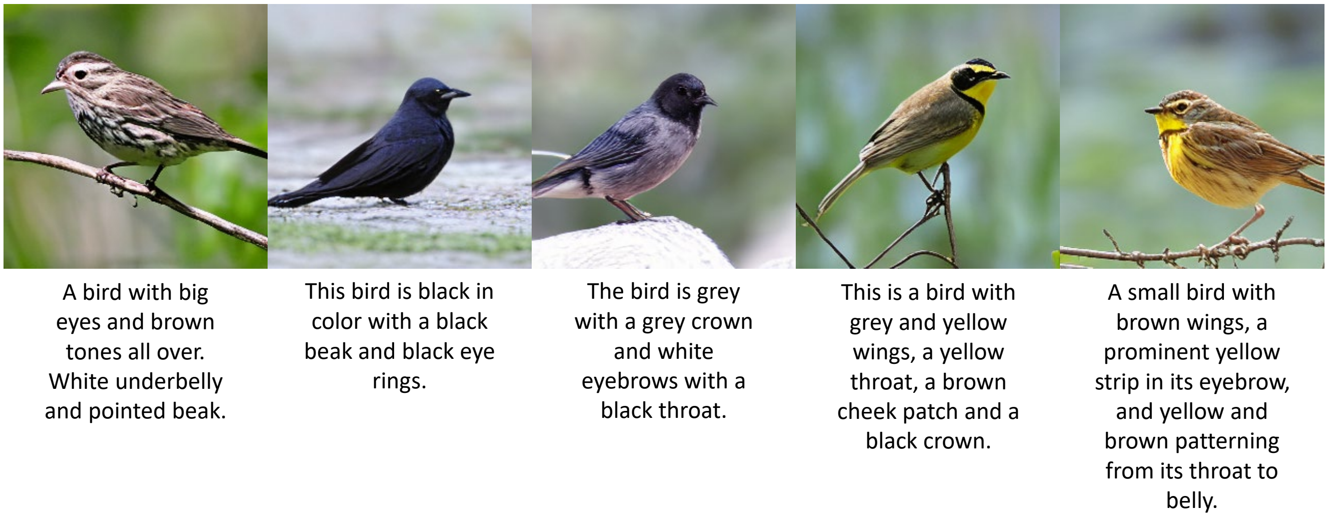
A car stopped at the front of the stop light.

A heritage building with big windows and next to a clock tower.

A woman, man, and a dog standing in the snow.

A metal statue of two women sitting on a bench.

Figure 18: **Samples generated with TCTS. Four images are generated for each text in 8 steps, 16 steps, 25 steps, 50 steps.**



A bird with big eyes and brown tones all over. White underbelly and pointed beak.

This bird is black in color with a black beak and black eye rings.

The bird is grey with a grey crown and white eyebrows with a black throat.

This is a bird with grey and yellow wings, a yellow throat, a brown cheek patch and a black crown.

A small bird with brown wings, a prominent yellow strip in its eyebrow, and yellow and brown patterning from its throat to belly.

Figure 19: **Samples generated by TCTS with FAS in 16 steps.**

C. Further Analysis on Results

C.1. Performance graph over time

Figure 20 demonstrates that TCTS outperforms other baseline methods in terms of MID and CLIP scores, which are relevant metrics for text. RR, TCTS, and FAS exhibit superior performance in SOA than Uniform, thus providing evidence for our analysis that revocable methods offer more opportunities for recovery, leading to the regeneration of missing objects. Furthermore, the figure illustrates the impact of FAS, which significantly enhances TCTS's FID while preserving the alignment between the image and text.

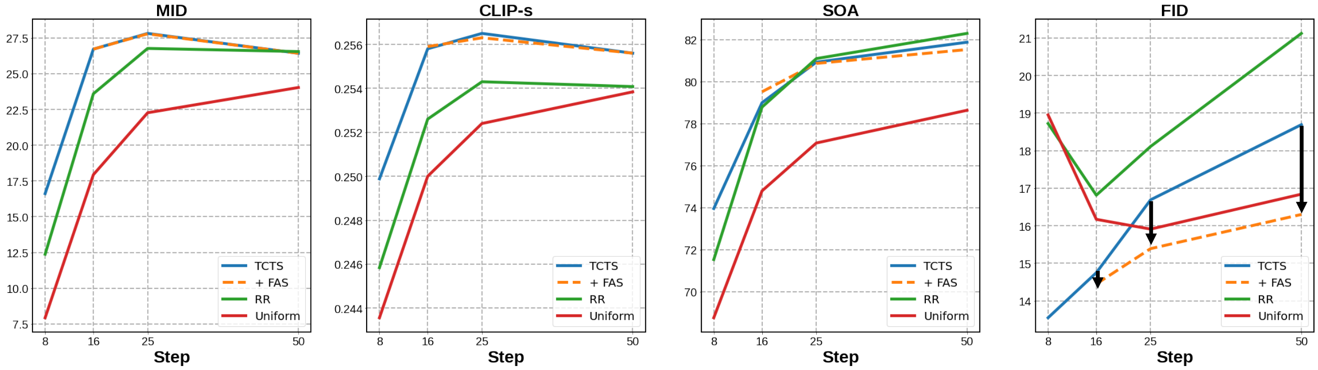


Figure 20: **Performance comparison of each method at different steps.** In our experiments, we fixed classifier-free guidance to 5. When we use FAS method, it was possible to lower the FID score while maintaining text alignment.

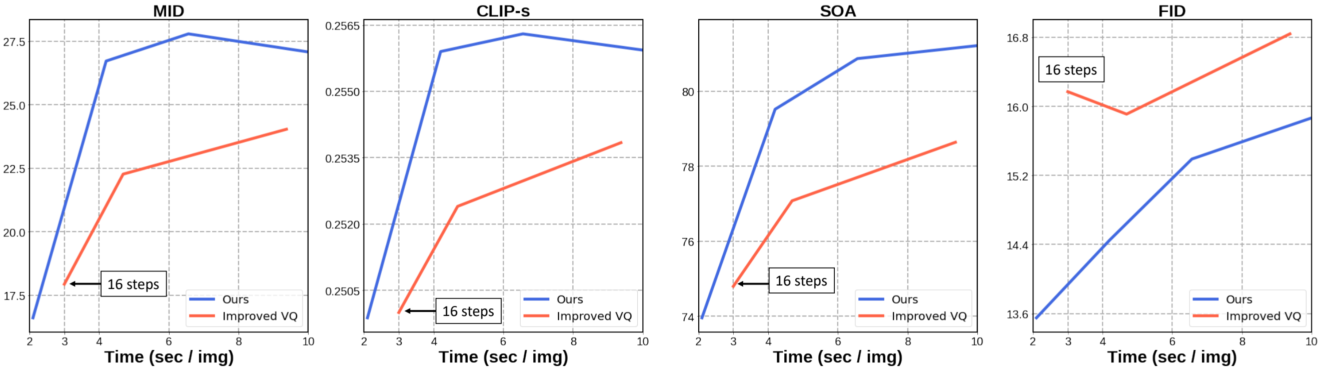


Figure 21: **Comparison of our model and the baseline in performance over generation time.** In our experiments, we fixed classifier-free guidance to 5.

We evaluated the speed and quality of our final model against the baseline, Improved VQ-Diffusion [37]. The baseline method requires three seconds per image to generate an image in 16 steps. As shown in Figure 21, our model surpasses the baseline in all metrics while maintaining the same generation time.

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. ediff-i: 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, 10

[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, 11, 12

[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, 10

[21] Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. Improving sample quality of diffusion models using self-attention guidance, 2022. 5, 10, 11

[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, 10

[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, 10

[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, 10, 16

[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. Stackgan++: 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