Improved Masked Image Generation with Token-Critic Supplementary Material

1 Comparison to related work on ImageNet 512x512

1.1 Base models

In Figures 1, 2, 3 and 4 we compare the result of sampling from Token-Critic with one competing GAN, BigGAN [1], and one diffusion model, ADM with classifier guidance (ADM+G) [2]. We compare on ImageNet 512x512 as this is the more challenging case.

Our goal here is to directly compare the performance of the original models in capturing the class-conditional distributions of 512x512 real images. Thus, we do not include classifier rejection for Token-Critic or upsampling for ADM, as the resulting samples would depend on a separate process.

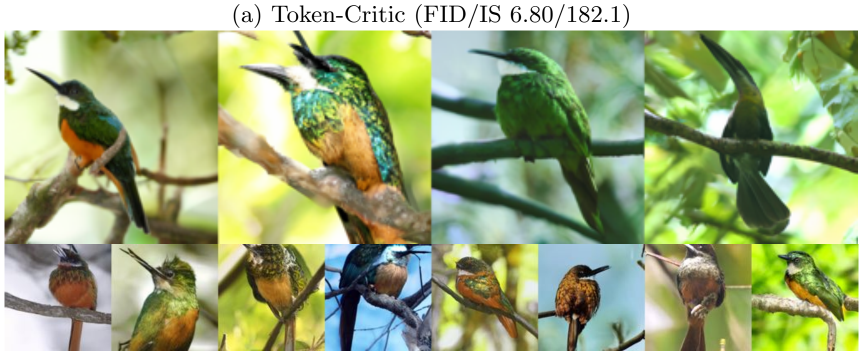
Results for ADM+G [2] were obtained using the authors' publicly available source code[1]. Results for BigGAN [1] were obtained using the authors' implementation. Note that BigGAN uses one step, ADM+G 1000 steps, and Token-Critic 18 forward steps and 18 critic steps.

1.2 Combined models

In Figure 5 we compare the models that obtain better FID and Inception scores in Table 2 by leveraging an external process. For Token-Critic, the external process is classifier-based rejection sampling using a ResNet50 classifier. For ADM with guidance and upsampling (ADM+G+U) [2], the external process consists in using an upsampling diffusion model to rescale samples from 128x128 to 512x512. Results for [2] were obtained using the authors' publicly available source code. Note that ADM+G+U uses 250 steps for 128x128 generation and 250 steps for upsampling. Token-Critic with rejection sampling with 20% acceptance rate uses five times 18 forward steps and 18 critic steps.

[1] https://github.com/openai/guided-diffusion

(a) Token-Critic (FID/IS 6.80/182.1)



(b) ADM+G (FID/IS 7.72/172.7)

(c) BigGAN (FID/IS 8.43/177.9)

Fig. 1. Comparison on 512x512 class-conditional image generation on ImageNet class "jacamar" (95).

(a) Token-Critic (FID/IS 6.80/182.1)

(b) ADM+G (FID/IS 7.72/172.7)

(c) BigGAN (FID/IS 8.43/177.9)

Fig. 2. Comparison on 512x512 class-conditional image generation on ImageNet class "white wolf" (270).

(a) Token-Critic (FID/IS 6.80/182.1)

(b) ADM+G (FID/IS 7.72/172.7)

(c) BigGAN (FID/IS 8.43/177.9)

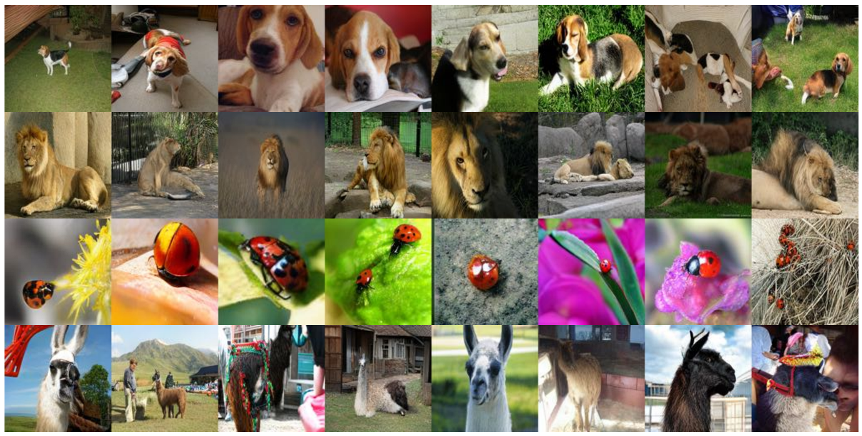
Fig. 3. Comparison on 512x512 class-conditional image generation on ImageNet class "llama" (355).

(a) Token-Critic (FID/IS 6.80/182.1)

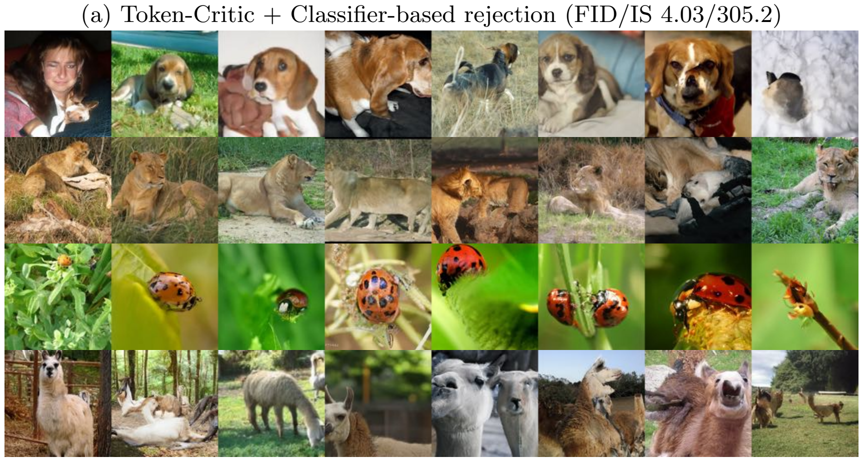
(b) ADM+G (FID/IS 7.72/172.7)

(c) BigGAN (FID/IS 8.43/177.9)

Fig. 4. Comparison on 512x512 class-conditional image generation on ImageNet class "schooner" (780).



(a) Token-Critic + Classifier-based rejection (FID/IS 4.03/305.2)



(b) ADM + Guidance + Upsampling (FID/IS 3.85/221.7)

Fig. 5. Comparison on 512x512 class-conditional image generation with ADM+G+U [2], for ImageNet classes “beagle” (162), “lion” (291), “ladybug” (301) and “llama” (355).

2 On Token-Critic training objective.

As motivated in the main manuscript, we seek to match the distributions of 1) real masked images and 2) masked images obtained by the method, after estimating $\mathbf{x}_0$ with the generator G_θ and selecting the mask with Token-Critic. The masking rate is indicated by t. Next we show that the Token-Critic training objective approximates optimizing the KL divergence between these two distributions.

$$
\begin{aligned}
KL\left(q(\mathbf{x}_t) \| p_{\theta,\phi}(\mathbf{x}_t)\right) &= -\mathbb{E}_{q(\mathbf{x}_t)} \log \frac{p_{\theta,\phi}(\mathbf{x}_t)}{q(\mathbf{x}_t)} && (1)\\
&= -\mathbb{E}_{q(\mathbf{x}_t)} \log \sum_{\mathbf{x}'_t} \sum_{\hat{\mathbf{x}}_0} \frac{p_{\theta,\phi}(\mathbf{x}_t, \hat{\mathbf{x}}_0, \mathbf{x}'_t)}{q(\mathbf{x}_t)} d\hat{\mathbf{x}}_0 d\mathbf{x}'_t && (2)\\
&= -\mathbb{E}_{q(\mathbf{x}_t)} \log \sum_{\mathbf{x}'_t} \sum_{\hat{\mathbf{x}}_0} \frac{p_\phi(\mathbf{x}_t|\hat{\mathbf{x}}_0) p_\theta(\hat{\mathbf{x}}_0|\mathbf{x}'_t) q(\mathbf{x}'_t)}{q(\mathbf{x}_t)} d\hat{\mathbf{x}}_0 d\mathbf{x}'_t && (3)\\
&\leq -\mathbb{E}_{q(\mathbf{x}_t)} \mathbb{E}_{q(\mathbf{x}'_t)} \mathbb{E}_{p_\theta(\hat{\mathbf{x}}_0|\mathbf{x}'_t)} \log \frac{p_\phi(\mathbf{x}_t|\hat{\mathbf{x}}_0)}{q(\mathbf{x}_t)} && (4)\\
&\approx -\mathbb{E}_{q(\mathbf{x}_t)} \mathbb{E}_{p_\theta(\hat{\mathbf{x}}_0|\mathbf{x}_t)} \log p_\phi(\mathbf{x}_t|\hat{\mathbf{x}}_0) + C, && (5)\\
&= -\mathbb{E}_{q(\mathbf{x}_t)} \mathbb{E}_{p_\theta(\hat{\mathbf{x}}_0|\mathbf{x}_t)} \log p_\phi(\mathbf{m}_t|\hat{\mathbf{x}}_0) + C, && (6)
\end{aligned}
$$

where C is constant with respect to Token-Critic parameters ϕ. In (5) we used Jensen's inequality and in (6) we approximate the expectation by choosing $\mathbf{x}'_t = \mathbf{x}_t$, noting that for most random pairs of $\mathbf{x}$ and $\mathbf{x}'_t$ in the dataset this quantity will be very small. Finally, the last step results from $\mathbf{x}_t$ being completely determined by $\hat{\mathbf{x}}_0$ and $\mathbf{m}_t$.

References

1. Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096 (2018)
2. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in Neural Information Processing Systems **34** (2021)