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
Fig. 1. Comparison on 512x512 class-conditional image generation on ImageNet class “jacamar” (95).
Fig. 2. Comparison on 512x512 class-conditional image generation on ImageNet class “white wolf” (270).
Fig. 3. Comparison on 512x512 class-conditional image generation on ImageNet class “llama” (355).
Fig. 4. Comparison on 512x512 class-conditional image generation on ImageNet class “schooner” (780).
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 $x_0$ with the generator $G_\theta$ 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.

\[
KL(q(x_t) \| p_{\theta,\phi}(x_t)) = -E_{q(x_t)} \log \frac{p_{\theta,\phi}(x_t)}{q(x_t)}
\]

\[
= -E_{q(x_t)} \log \sum_{x_t'} \sum_{\hat{x}_0} \frac{p_{\theta,\phi}(x_t, \hat{x}_0, x_t')}{q(x_t)} d\hat{x}_0 d\hat{x}_t
\]

\[
= -E_{q(x_t)} \log \sum_{x_t'} \sum_{\hat{x}_0} p_{\phi}(x_t | \hat{x}_0) p_{\theta}(\hat{x}_0 | x_t') q(x_t') \frac{q(x_t)}{q(x_t)}
\]

\[
\leq -E_{q(x_t)} E_{q(x_t')} E_{p_{\theta}(\hat{x}_0 | x_t')} \log \frac{p_{\phi}(x_t | \hat{x}_0)}{p_{\phi}(x_t | \hat{x}_0)}
\]

\[
\approx -E_{q(x_t)} E_{p_{\theta}(\hat{x}_0 | x_t)} \log p_{\phi}(x_t | \hat{x}_0) + C,
\]

\[
= -E_{q(x_t)} E_{p_{\theta}(\hat{x}_0 | x_t)} \log p_{\phi}(m_t | \hat{x}_0) + C,
\]

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