**MaskGIT: Masked Generative Image Transformer**  
**SUPPLEMENTARY**

**Contents**

A. MaskGIT’s Performance on Image Reconstruction  
B. Additional Class-conditional Image Generation Results  
C. Class-conditional Image Editing  
D. Image Outpainting For Panorama Synthesis  
E. Image Outpainting Comparisons with SOTA Transformer-based Approaches  
F. Image Inpainting and Outpainting Comparisons with SOTA GAN-based Methods  
G. Limitations and Failure Cases

**A. MaskGIT’s Performance on Image Reconstruction**

In Sec 4 of the main paper, we primarily evaluate the performance of MaskGIT on class-conditional image generation. Here, we study its performance on image reconstruction. We first randomly sample input mask $M$ with a mask ratio $r$ of the visual tokens masked out, and then run MaskGIT’s iterative decoding algorithm to reconstruct images. Figure 1 shows the PSNR and LPIPS [15] of the reconstructed samples as functions of $r$, whereas Figure 2 visualizes this process with $r$ ranging from 95% to 75%.

![Figure 1](image1.png)

Figure 1. Reconstruction quality and diversity measured by PSNR and LPIPS [15].

We observe that MaskGIT can reconstruct holistic information (e.g. pose and shape of the foreground objects) even with a very high percentage (e.g. 95%) of tokens masked out. More importantly, we find that there exists an inflection point around 90%: as shown in Figure 1, both reconstruction quality and consistency improve drastically as the mask ratio decreases until it hits 90%, beyond which further improvements are slowed down. This observation is corroborated by the large jump in the visual similarity between reconstruction samples and the original image from 95% to 90% in Figure 2 (e.g. the fence in front of the tiger and the car’s color are consistently captured once the mask ratio is below 90%).
Figure 2. **Examples of MaskGIT on Image Reconstruction.** MaskGIT takes in masked tokens extracted from original images (row one) using random input masks (row two, with unknown tokens marked in light gray), and outputs reconstructed images (row three). We then randomly sample 100 masks with the same mask ratio, and illustrate the median of the 100 reconstructed samples in row four.

In other words, visual tokens are highly redundant. Only a very small portion (e.g. 10%) is essential for a holistic reconstruction, while the remaining ones only benefit the recovery of the appearance or finer details. This echoes the intuition behind our masking design laid out in Sec 3.3 that the prediction of the first few tokens is key to image generation. Similar observations on the spatial redundancy of images are discussed in a concurrent paper MAE [6]. In their work, they find that masking a high proportion of the input image yields a nontrivial and meaningful self-supervised task for image representation learning.

### B. Additional Class-conditional Image Generation Results

In this section, we report additional results on class-conditional image generation.

In Table 1, we report Precision and Recall scores calculated using Inception features [11]. In contrast to the VGG [10] feature-based scores, which we report in the main paper for a more direct comparison with prior work [4, 7], we find that the Inception feature-based scores are more consistent with our qualitative observations that VQGAN’s samples are more diverse than BigGAN’s. Under both measures, MaskGIT’s recall scores outperform those of BigGAN and VQGAN. We also report CAS evaluated on classifiers trained without augmentation from RandAugment [3]. Consistent with our main results, MaskGIT outperforms BigGAN and our baseline VQGAN by a large margin.

Finally, we show a few comparisons of the class-conditional samples generated by MaskGIT with the samples generated by BigGAN-deep and VQVAE-2 in Figure 3, 4, and 5.

<table>
<thead>
<tr>
<th>Model</th>
<th>Inception-Prec ↑</th>
<th>Inception-Rec ↑</th>
<th>CAS ×100 ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Top-1 (73.1)</td>
</tr>
<tr>
<td>BigGAN-deep [1]</td>
<td><strong>0.82</strong></td>
<td>0.27</td>
<td>42.65</td>
</tr>
<tr>
<td>VQ-GAN*</td>
<td>0.61</td>
<td>0.47</td>
<td>47.50</td>
</tr>
<tr>
<td>MaskGIT (Ours)</td>
<td>0.78</td>
<td><strong>0.50</strong></td>
<td><strong>58.20</strong></td>
</tr>
</tbody>
</table>

Table 1. More quantitative comparison with BigGAN-deep and our baseline VQGAN on ImageNet 256×256. * denotes the model we train with the same architecture and setup with ours.
Figure 3. More diversity comparisons between BigGAN-deep with truncation 1.0, VQVAE-2 [9], and our proposed method MaskGIT on ImageNet. † represents extracted samples from the paper.
Figure 4. More diversity comparisons between BigGAN-deep with truncation 1.0, VQVAE-2 [9], and our proposed method MaskGIT on ImageNet. † represents extracted samples from the paper.
Figure 5. More Diversity Comparisons among BigGAN-deep with truncation 1.0, VQVAE-2 [9], and our proposed method MaskGIT on ImageNet. † represents extracted samples from the paper.
C. Class-conditional Image Editing

Figure 6. More Examples of Class-conditional Image Editing. In each column, the bottom images are synthesized using the image on the top, ImageNet class labels on the left, and a bounding box of the main object downsampled into latent space (as shown in the second row).
D. Image Outpainting For Panorama Synthesis

Figure 7. More Samples of Horizontal Image Extrapolation (from $512 \times 256$ to $512 \times 2304$). The synthesized “panoramas” are created by repeatedly applying MaskGIT’s outpainting abilities horizontally in both directions.
E. Image Outpainting Comparisons with SOTA Transformer-based Approaches

In Figure 8 and 9, we show a few outpainting comparisons to ImageGPT [2] and VQGAN [5]. In each set of images, we show the groundtruth (left), extrapolated samples using only the top half of the groundtruth (middle), and extrapolated samples using only the bottom half of the groundtruth (right).

While ImageGPT can only run on a maximum resolution of $192 \times 192$, VQGAN and MaskGIT can perform on higher resolutions by taking advantage of tokenization and thus achieve higher fidelity than ImageGPT’s. Since ImageGPT and VQGAN are both autoregressive approaches, their models can only handle outpainting in one direction. In comparison, MaskGIT is more flexible and can outpaint in arbitrary directions.

<table>
<thead>
<tr>
<th>Groundtruth</th>
<th>—— Outpaint bottom 50% ——</th>
<th>—— Outpaint top 50% ——</th>
</tr>
</thead>
<tbody>
<tr>
<td>ImageGPT</td>
<td><img src="image1" alt="ImageGPT" /></td>
<td><img src="image2" alt="ImageGPT" /></td>
</tr>
<tr>
<td>VQGAN</td>
<td><img src="image3" alt="VQGAN" /></td>
<td><img src="image4" alt="VQGAN" /></td>
</tr>
<tr>
<td>MaskGIT</td>
<td><img src="image5" alt="MaskGIT" /></td>
<td><img src="image6" alt="MaskGIT" /></td>
</tr>
</tbody>
</table>

Figure 8. Outpainting comparisons with the pixel-based approach ImageGPT [2] and the transformer-based approach VQGAN [5].
Figure 9. Outpainting comparisons with the pixel-based approach ImageGPT [2] and the transformer-based approach VQGAN [5].
F. Image Inpainting and Outpainting Comparisons with SOTA GAN-based Methods

In Sec 4.3 of the main paper, we show quantitative comparison of MaskGIT’s performance on inpainting and outpainting tasks with several GAN-based methods dedicated to these tasks. In this section, we show more qualitative comparisons with state-of-the-art image completion methods in Figure 10 and Figure 11.

Compared to prior GAN-based methods, we find that MaskGIT demonstrates a stronger capability of completing structures coherently, and that MaskGIT’s samples contain fewer artifacts. In Figure 11, MaskGIT completes the bridge in row two and the building in the second to last row, which all GAN methods struggle to do in comparison.

Figure 10. More visual comparisons on image inpainting on Places2 [17] with state-of-the-art GAN methods.
In addition, we compare with the state-of-the-art image completion method CoModGAN. The quantitative scores including FID and IS of both methods are close: MaskGIT achieves better FID of 6.78 and IS of 11.69 on outpainting 50% to the right (vs CoModGAN’s FID=7.67, IS=9.09), and slightly worse FID on inpainting with a 50% × 50% mask (FID=7.92, IS=22.95 vs FID=7.13, IS=21.82).

In addition, we compare with CoModGAN on image completion tasks with large masking ratios, i.e. conditioning on the center 50% × 50% and the center 31.25% × 31.25% respectively, which are challenging cases for traditional GANs. Examples are shown in Fig 12.

Figure 11. More visual comparisons on image outpainting, with state-of-the-art GAN methods. * samples are graciously provided by the authors.

Figure 12. Visual comparisons of outpainting with CoModGAN [16] on large outpainting mask.
G. Limitations and Failure Cases

In Figure 13, we show several limitations and failure cases of our approach. (A) and (B) are examples of semantic and color shifts in MaskGIT’s outpainting results. Due to its limited attention size, MaskGIT may “forget” the synthesized semantics or color from one end when it’s outpainting the other end. (C) and (D) show cases where our approach may sometimes ignore or modify objects on the boundary when applied to outpainting and inpainting. (E) showcases MaskGIT’s failure mode in which it causes oversmoothing or creates undesired artifacts on complex structures such as human faces, text and symmetric objects. The improvement for these circumstances remains future work.

Figure 13. Limitations and Failure Cases.
References


[6] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021.


