# Image Completion using Masked Generative Transformers

**Jaewon Son, Mohammed Azeezulla, Pradyumn Pathak**

June 12, 2025

**Abstract**

We present a token-based image completion framework that integrates a fine-tuned Vector Quantized Generative Adversarial Network (VQGAN) tokenizer with a MaskGIT-style masked generative transformer. The VQGAN is first fine-tuned on the Places365 dataset to produce discrete latent codes that capture meaningful semantic information for natural scene images. A transformer model is then trained to iteratively predict masked tokens in the latent space, enabling efficient and high-quality image reconstruction through parallel decoding. We evaluate our model using both pixel-level (PSNR, SSIM) and perceptual (LPIPS) metrics. While the model successfully restores global structure and semantic coherence, quantitative results reveal opportunities for improvement in fine-grained detail reconstruction. Our results demonstrate the feasibility of token-based masked modeling for semantic image completion and suggest potential directions for future enhancements.

## 1 Introduction

In recent years, remarkable progress has been made in image generation and completion, driven by advances in generative models such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Transformer-based architectures. Despite these developments, many existing models still face limitations, including slow autoregressive decoding, challenges in maintaining global coherence, and dependency on hand-crafted tokenization schemes.

To address these challenges, our project proposes a high-quality image completion pipeline that combines Vector Quantized Generative Adversarial Network (VQGAN) with Masked Generative Image Transformer (MaskGIT). The central idea of our approach is to encode images into discrete latent representations using a VQGAN tokenizer and employ a MaskGIT-style transformer to iteratively predict masked tokens in the latent space. Compared to conventional autoregressive models, MaskGIT enables parallel decoding, allowing for faster and more accurate image reconstruction, particularly well-suited for semantic image inpainting where significant portions of an image may be missing or corrupted.

We utilize the Places365 dataset to benchmark our approach and evaluate reconstruction quality using both pixel-level metrics (PSNR, SSIM) and perceptual metrics (LPIPS). Our pipeline involves fine-tuning the decoder and codebook of the VQGAN on domain-specific data, followed by training a transformer model on masked token sequences to learn effective reconstruction from partial observations.

The key contributions of this work are: (1) integration of MaskGIT with a fine-tuned VQGAN tokenizer for efficient semantic image completion; (2) implementation of iterative masked token decoding from scratch; and (3) comprehensive evaluation of the model's generalization performance across diverse natural scenes.

## 2 Related Works

Early image completion techniques primarily relied on autoregressive models such as PixelCNN, which generate images one pixel at a time by conditioning each pixel on its neighbors. While such methods excel at modeling local dependencies, they suffer from slow inference speed and limited global coherence.

To address these limitations, encoder-decoder architectures such as Context Encoders [3] and DeepFill v2 [2] were introduced. These models leverage contextual and semantic priors to fill missing regions in images. DeepFill further improved performance through gated convolutions, enabling the network to handle irregular masks more effectively. However, since these models operate directly at the pixel level, they often produce blurry or unrealistic results when dealing with large or complex scenes.

More recent approaches have shifted toward discrete visual token representations using vector quantization. VQ-VAE [6] represents a major milestone in this direction by compressing images into discrete latent codes, allowing transformer-based models to capture complex dependencies in the latent space. Building upon this, VQGAN [4] combines vector quantization with adversarial and perceptual losses to generate visually sharp and semantically rich reconstructions. In our work, VQGAN serves as both the tokenizer and decoder.

Parallel to these developments, MaskGIT [1] introduced a masked generative transformer architecture that moves away from traditional autoregressive decoding. By employing a parallel masked token prediction strategy, MaskGIT enables faster generation through iterative refinement of token sequences. While originally applied to object-centric generation tasks, its architecture is well-suited for semantic image inpainting where substantial portions of the image may be missing.

Our work integrates a fine-tuned VQGAN with a custom implementation of the MaskGIT-style transformer to build a token-based image completion pipeline. Unlike prior inpainting approaches, our method predicts missing token indices directly in the discrete latent space and reconstructs complete images using the VQGAN decoder.

Additionally, our design is conceptually inspired by DALL·E [5], which demonstrated the effectiveness of discrete token modeling for text-to-image generation, and Vision Transformers (ViT) [7], which established the capability of pure transformer architectures for visual understanding. Collectively, these prior works motivate our adoption of token-based masked modeling for semantic image completion.

# 3 Preliminary and Background

This section outlines the foundational concepts required to understand our token-based image completion framework. We elaborate on the key components—Vector Quantized Generative Adversarial Networks (VQGAN) and Masked Generative Image Transformers (MaskGIT)—along with their theoretical motivations and architectural details.

## 3.1 Vector Quantized Generative Adversarial Network (VQGAN)

VQGAN extends the Vector Quantized Variational Autoencoder (VQ-VAE) framework by incorporating adversarial training to enhance perceptual quality. The architecture consists of three main components: an encoder, a discrete codebook, and a decoder.

- **Encoder**: Maps the input image $x \in \mathbb{R}^{H \times W \times C}$ into a continuous latent space $z_e(x) \in \mathbb{R}^{h \times w \times d}$, where $h < H$, $w < W$, and $d$ is the latent channel dimension.

- **Vector Quantization**: Each latent vector $z_e(x)_{i,j}$ is replaced with the closest codebook vector from a learned set $\{e_k\}_{k=1}^{K}$, resulting in a discrete token grid $z_q(x)$. This quantization allows image representations to be discretized, which is essential for transformer-based modeling.

- **Decoder**: Reconstructs the image from the quantized latent representation $z_q(x)$, aiming to minimize a perceptual loss (e.g., LPIPS) and adversarial loss provided by a discriminator.

The adversarial component encourages the decoder to generate images that are not only close to the input in pixel space but also perceptually indistinguishable from real images, producing sharper and more realistic reconstructions compared to conventional autoencoder-based models.

## 3.2 Masked Generative Image Transformer (MaskGIT)

MaskGIT is a masked token modeling framework for generative tasks, inspired by BERT-style masked language modeling but adapted to discrete visual tokens. Unlike autoregressive models such as PixelCNN or DALL·E, which decode tokens sequentially, MaskGIT performs parallel decoding and iterative refinement.

- **Token Masking**: A subset of tokens is randomly masked (assigned a special mask token ID), and the model predicts the masked positions based on the visible context.

- **Transformer Decoder**: A multi-layer Transformer processes the full token sequence, incorporating positional embeddings, and outputs logits for all token positions.
- **Iterative Decoding Schedule**: Instead of predicting all masked tokens in a single step, MaskGIT performs multiple decoding iterations. At each iteration, a fraction of tokens with the highest confidence scores are filled, progressively refining the prediction.

This iterative decoding approach enables faster synthesis while preserving high visual quality and global coherence, making it particularly effective for inpainting tasks where substantial regions are missing.

### 3.3 Token-Based Image Completion Pipeline

Our method combines VQGAN and MaskGIT into a unified framework for semantic image inpainting, structured as follows:

- **Tokenization**: Each image is encoded by the VQGAN encoder into a 2D grid of discrete tokens (e.g., $16 \times 16 = 256$ tokens).
- **Masking**: A random subset of tokens (e.g., 15%) is masked to simulate missing regions.
- **Transformer Prediction**: A MaskGIT-style transformer is trained to predict the masked tokens using cross-entropy loss, with ignore indices applied to visible tokens.
- **Iterative Sampling**: During inference, the model iteratively decodes over the masked grid, selecting the most confident predictions at each step.
- **Reconstruction**: The fully completed token grid is passed through the VQGAN decoder to reconstruct the final image.

This design enables the model to generalize to arbitrary inpainting regions while maintaining high visual fidelity and spatial consistency.

## 4 Methodology

In this section, we describe the detailed architecture, training pipeline, and inference procedure of our token-based image completion framework. The entire approach can be divided into two main stages: VQGAN fine-tuning for domain adaptation and MaskGIT-style transformer training for masked token prediction.

First, we fine-tune a pre-trained VQGAN model to adapt both the tokenizer and decoder to our target domain, which is the Places365 dataset. This step ensures that the discrete latent space learned by VQGAN captures meaningful visual patterns relevant to the dataset. The VQGAN encoder maps each input image into a continuous latent representation, which is subsequently quantized into a discrete token grid by mapping each latent vector to its nearest codebook entry. The decoder then reconstructs the image from these quantized tokens by minimizing a combination of perceptual loss and adversarial loss. This fine-tuning step allows the codebook to better reflect the data distribution of our specific domain, improving downstream reconstruction quality.

After the VQGAN fine-tuning is complete, we train a transformer-based model to predict masked tokens in the discrete latent space. During training, each image is first tokenized into discrete codebook indices using the fine-tuned VQGAN encoder and codebook. A random subset of these tokens is masked, typically around 15%, to simulate missing regions within the token grid. The transformer model, which consists of multiple layers of causal attention blocks, processes these partially masked sequences with positional embeddings that preserve the spatial layout of the tokens. The model is trained using a cross-entropy loss function that focuses only on the masked positions, while ignoring the visible tokens.

During inference, the model employs an iterative decoding strategy instead of predicting all masked tokens in a single pass. Initially, a portion of tokens in the input sequence are masked. At each decoding step, the model computes confidence scores for all masked positions based on the softmax output probabilities. A fixed percentage of tokens with the highest confidence scores are then selected and filled with their predicted values. This process is repeated over multiple iterations until all masked positions are filled. Finally, the completed token grid is passed to the fine-tuned VQGAN decoder to reconstruct the full-resolution image.

The entire pipeline consists of several steps: encoding and tokenization using VQGAN, random masking of tokens, transformer-based masked token prediction, iterative decoding with confidence-based filling, and finally image reconstruction using the VQGAN decoder.

To assess the reconstruction quality of our image completion model, we employ both pixel-level and perceptual evaluation metrics. Specifically, we report Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS). PSNR and SSIM measure low-level pixel accuracy, while LPIPS provides a perceptual similarity score that better aligns with human visual perception. These metrics collectively offer a comprehensive assessment of both visual fidelity and semantic consistency.

# 5 Numerical Experiments

In this section, we present the experimental setup, evaluation protocol, and results that demonstrate the effectiveness of our proposed image completion framework.

We conducted all experiments using the Places365 dataset, which contains diverse natural scene images across various categories. All images were resized to 256×256 resolution and normalized to the $[-1, 1]$ range. The dataset was randomly split into training and validation sets. During preprocessing, images were tokenized into 16×16 discrete latent codes using the fine-tuned VQGAN encoder.

The VQGAN model was fine-tuned for 200 epochs using the Adam optimizer with a learning rate of 1e-4. After fine-tuning, the tokenizer and decoder were frozen. The MaskGIT-style transformer was then trained for 100 epochs with a batch size of 32. At each training iteration, 30% of tokens were randomly masked to simulate missing regions. The transformer was trained using cross-entropy loss applied only to masked token positions.

Before training the masked token transformer, we first examined the reconstruction performance of the fine-tuned VQGAN itself. Figure 1 shows a qualitative example comparing the original input image and its VQGAN reconstruction. While the reconstructed image successfully captures the overall structure and semantic information of the scene, some fine details and textures are smoothed or slightly distorted due to quantization bottlenecks inherent to the VQGAN architecture. Nevertheless, the tokenizer and decoder preserve sufficient information for downstream inpainting tasks.

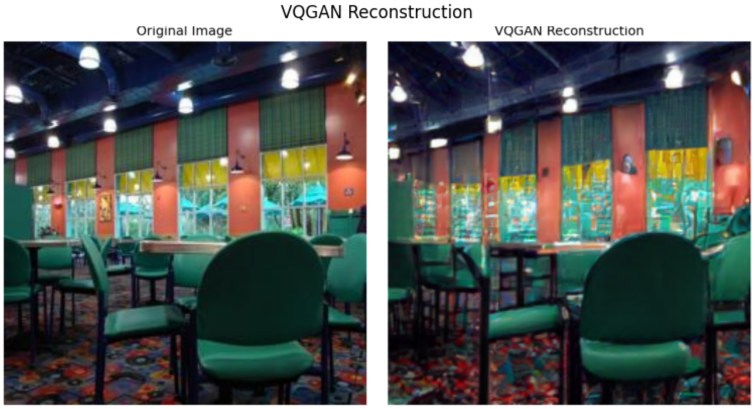

Figure 1: VQGAN reconstruction example. Left: original input image. Right: VQGAN reconstruction.

Throughout training, both training and validation loss steadily decreased, indicating stable convergence. The learning rate followed a cosine annealing schedule, and the gap between training and validation loss remained small, suggesting minimal overfitting. Figure 2 summarizes the overall training dynamics, including loss curves, learning rate schedule, and overfitting monitor.

We evaluated reconstruction quality using PSNR, SSIM, and LPIPS metrics. Table 1 reports the average validation results. The model achieved a mean PSNR of 10.64 dB, SSIM of 0.1795, and LPIPS of 0.7637. Although PSNR and SSIM remain below target benchmarks, LPIPS indicates that the model captures certain perceptual structures, while leaving room for improvement.

The distribution of each metric across the validation set is visualized in Figure 3. The results show sample-wise variability, suggesting that reconstruction performance may depend on image

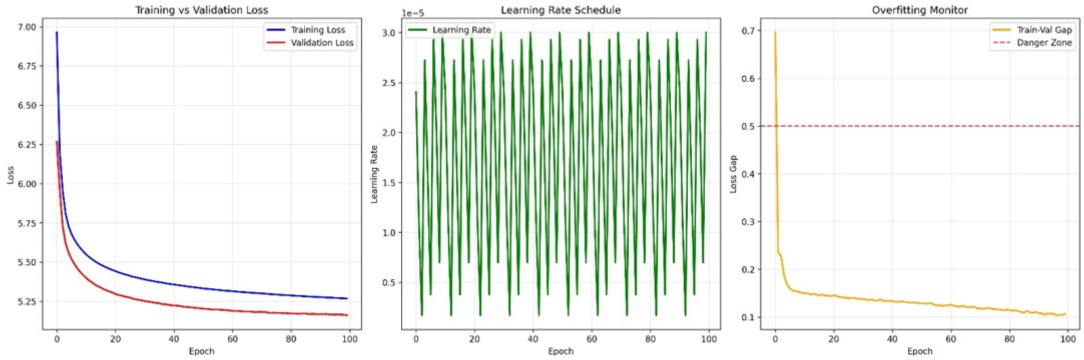

Figure 2: Training curves: loss curves, learning rate schedule, and overfitting monitor.

Table 1: Quantitative evaluation results on the validation set.

| Metric | Mean Value | Target |
|---|---|---|
| PSNR | 10.64 | 15 |
| SSIM | 0.1795 | 0.3 |
| LPIPS | 0.7637 | $< 0.5$ |

complexity and mask patterns.

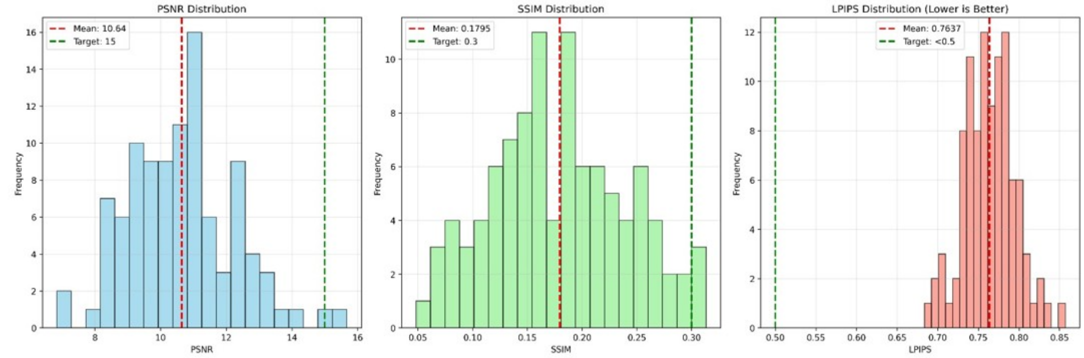

Figure 3: Metric distributions across the validation set: PSNR, SSIM, and LPIPS.

These results demonstrate the feasibility of our token-based image completion framework while identifying directions for future improvements in model architecture and training strategy.

# 6 Conclusion

In this project, we proposed a token-based image completion framework that integrates a fine-tuned VQGAN tokenizer with a MaskGIT-style masked generative transformer. By fine-tuning VQGAN on the Places365 dataset, we ensured that the discrete latent representations effectively captured semantic features of natural scenes. The transformer model was then trained to iteratively predict masked tokens in the latent space, enabling efficient and high-fidelity image reconstruction through parallel decoding.

Our experiments evaluated reconstruction performance using both pixel-level metrics (PSNR, SSIM) and perceptual metrics (LPIPS). While the model successfully recovered semantic structures and maintained global coherence, quantitative results indicated that reconstruction sharpness and fine-grained details can be further improved. Training dynamics also demonstrated stable convergence and minimal overfitting throughout the process.

This study does have several limitations. Due to computational resource constraints, the training was conducted on a limited subset of images, which may have restricted the model's ability to generalize across highly diverse scenes. Additionally, model size and transformer depth were kept relatively small to fit within feasible hardware limitations. These factors likely impacted the model's capacity to capture complex visual structures at higher fidelity.

Overall, our work demonstrates the feasibility of token-based masked modeling for semantic image completion. Future work could explore scaling up the transformer architecture, increasing dataset size and diversity, and leveraging more advanced decoding schedules to further improve reconstruction quality and generalization ability.

# References

[1] H. Chang, H. Zhang, L. Jiang, C. Liu, and W. T. Freeman. MaskGIT: Masked Generative Image Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[2] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. Free-Form Image Inpainting with Gated Convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[3] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context Encoders: Feature Learning by Inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[4] P. Esser, R. Rombach, and B. Ommer. Taming Transformers for High-Resolution Image Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[5] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. Zero-Shot Text-to-Image Generation. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

[6] A. van den Oord, O. Vinyals, and K. Kavukcuoglu. Neural Discrete Representation Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021.