

# Supplementary Materials

Paper Submission #16091

## 1 Implementation Details

**Sythetic Dataset** Our synthetic dataset includes 10 clusters, each with 1,000 data points sampled from a Gaussian distribution and standardized using Standard Scaler. Our VQ-VAE[5] comprises an MLP-based encoder/decoder with three linear layers and uses the ReLU activation function. Training is facilitated by the AdamW optimizer, with a learning rate of 0.001. Additional specifications include a codebook size of 128, a hidden dimension of 32, a batch size of 256, a beta of 0.25, and a decay rate ( $\gamma$  for EMA) of 0.9. In experiments, the autoencoder is trained for 100 epochs. The fine-tuned VQ-VAE and the original VQ-VAE are trained for 100 and 200 epochs respectively.

**CIFAR-10** For CIFAR-10, our VQ-VAE adopts downsampling using a CNN with a downsample channel of 128, and the model includes two residual blocks with a hidden channel size of 64. The codebook size is set at 512 with a token dimension of 64. The learning rate is  $3e-4$ , using the Adam optimizer with amsgrad set to true. The beta is 0.25 and the decay rate is 0.99. The codebook size in experiments is varied from 16 to 65,536, with embedding and token adopting the size of 32. And we pretrain AE for 150 epochs and fine-tune the VQ-VAE for 150 epochs.

**ImageNet-100 & ODIR** As described in the main text, we modified the tokenizer by removing one downsampling layer along with its corresponding upsampling layer and reducing the backbone's channel size to 64. We employed both ReduceLROnPlateau and Cosine annealing schedulers to train the tokenizer. Detailed configurations can be found in the provided codebase (*config.yaml*). The initial learning rate was set to  $1 \times 10^{-4}$ , with a minimum of  $1 \times 10^{-6}$ , and early stopping was applied. For ImageNet-100, we trained selected the checkpoint with the best FID on the validation set for downstream tasks on ImageNet-100, while the final checkpoint was used for the ODIR[3] dataset. VAR[4] and MaskGIT[1] were trained with ReduceLROnPlateau scheduling and early stopping.

It is important to note that when the codebook size is large, it is infeasible to initialize it using embeddings from a single batch. Therefore, on CIFAR-10, ImageNet-100, and ODIR, we initialize the codebook using embeddings collected from multiple batches. Specifically, we maintain an embedding-to-token ratio of 10:1 on CIFAR-10, and 2:1 on ImageNet-100 and ODIR.

## 2 Visualization of Embedding on Synthetic dataset

In our synthetic experiments, we find that the outputs of an untrained encoder are predominantly concentrated between 0.1 and 0.4 (Fig.1 a) and exhibit only 6 distinct peaks, despite the input dataset containing 10 modes.

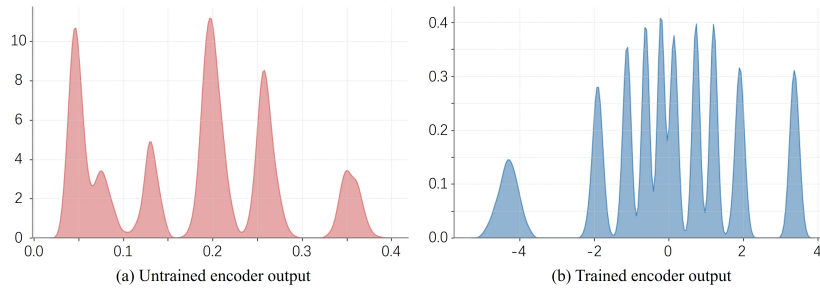


Figure 1: **Distribution of untrained and trained encoder's output.** (a) Untrained encoder's output has fewer peaks than 10 peaks of input and shrinks into a relatively small range. (b) Trained encoder's output displays 10 peaks, which is the same as the input.

### 3 VQGAN based Generation Results

Table 1: **Evaluation of tokenizers trained with GAN loss on the ImageNet-100 dataset.** "Shrink" indicates whether token representation shrinkage is present (✓) or mitigated using our proposed method (✗). MSE and LPIPS values in the table are scaled. To recover the actual values, multiply them by  $10^{-4}$  and  $10^{-3}$ , respectively.

Tokenizer	Shrink	r-FID ↓	MSE ↓	LPIPS ↓	Cosine. ↑	Perp. ↑
MaskGIT	✗	<b>5.28</b>	<b>3.96</b>	<b>2.40</b>	<b>0.96</b>	<b>5575.92</b>
	✓	6.40	4.58	2.70	0.67	920.95
VAR	✗	2.13	<b>2.59</b>	<b>1.73</b>	<b>0.97</b>	<b>7143.41</b>
	✓	<b>2.09</b>	3.00	1.87	0.66	2814.91

Table 2: **Evaluation of tokenizers trained with GAN loss on the ODIR dataset.** To recover the actual MSE and LPIPS values, multiply them by  $10^{-5}$  and  $10^{-4}$ , respectively.

Tokenizer	Shrink	r-FID ↓	MSE ↓	LPIPS ↓	Cosine. ↑	Perp. ↑
VAR	✗	<b>6.90</b>	<b>2.19</b>	<b>7.15</b>	<b>0.90</b>	<b>5451.76</b>
	✓	8.12	2.66	9.39	0.62	801.07

Table 3: **ImageNet-100 generation**

Model	Shrink	g-FID ↓	Pixel Dist. ↑
MaskGIT	✗	<b>10.85</b>	<b>79.17</b>
	✓	12.25	78.83
VAR	✗	8.30	<b>77.10</b>
	✓	<b>7.83</b>	73.37

Table 4: **ODIR generation**

Model	Shrink	g-FID ↓	Pixel Dist. ↑
VAR	✗	29.65	<b>50.56</b>
	✓	<b>27.89</b>	45.14

As shown in Tab.1 and Tab.2, high cosine similarity and low perplexity indicate the presence of token representation shrinkage in VQGAN[2]. And token representation shrinkage still degrades reconstruction quality. For generative models, Tab.3 and Tab.4 further demonstrate that shrinkage impairs generative creativity. In particular, the average of pixel distance is consistently lower under shrinkage, reflecting reduced diversity in generated samples. However, on the ImageNet-100 dataset, we occasionally found better r-FID and g-FID scores in the presence of shrinkage. We hypothesize that this is due to GAN loss might enhance the decoder's expressive capacity, which enables it to compensate for the effects of token representation shrinkage.

### 4 Results for MaskGIT on Medical dataset

Table 5: **Tokenizer and MaskGIT performance on medical dataset.**  $\mathcal{L}_{GAN}$  denotes whether GAN loss was used during training. "Shrink" indicates whether token representation shrinkage is present (✓) or mitigated (✗). To recover the actual MSE and LPIPS values, multiply them by  $10^{-5}$  and  $10^{-4}$ .

$\mathcal{L}_{GAN}$	Shrink	r-FID ↓	MSE ↓	LPIPS ↓	Cosine. ↑	Perp. ↑	g-FID ↓	Pixel Dist. ↑
No	✗	<b>10.33</b>	<b>3.16</b>	<b>0.90</b>	<b>0.95</b>	<b>3346.04</b>	36.98	47.54
	✓	11.96	4.19	1.16	0.68	438.36	<b>31.57</b>	<b>49.96</b>
Yes	✗	<b>10.47</b>	<b>3.67</b>	<b>1.02</b>	<b>0.95</b>	<b>3211.72</b>	38.41	47.79
	✓	12.72	4.64	1.30	0.68	432.07	<b>31.87</b>	<b>49.85</b>

As shown in Tab 5, when token representation shrinkage occurs, the MaskGIT's tokenizer also exhibits a decline in reconstruction quality on the ODIR dataset. However, the generative model unexpectedly demonstrates better creativity under shrinkage conditions. We hypothesize that this counterintuitive result may be attributed to inadequate medical data to train and evaluate the model.

## 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*, pages 11315–11325, 2022.
- [2] 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*, 2021.
- [3] A. Maranhão. Ocular disease intelligent recognition (odir). <https://www.kaggle.com/datasets/andrewmvd/ocular-disease-recognition-odir5k>, 2020.
- [4] K. Tian, Y. Jiang, Z. Yuan, B. Peng, and L. Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024.
- [5] A. Van Den Oord, O. Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 2017.