

# WE TOK: POWERFUL DISCRETE TOKENIZATION FOR HIGH-FIDELITY VISUAL RECONSTRUCTION

Shaobin Zhuang<sup>1</sup> ♪ Yiwei Guo<sup>3</sup> ♪ Canmiao Fu<sup>2</sup> Zhipeng Huang<sup>2</sup>

Zeyue Tian<sup>4</sup> ♪ Ying Zhang<sup>2</sup> Chen Li<sup>2</sup> Yali Wang<sup>3,5,\*</sup>

<sup>1</sup> Shanghai Jiao Tong University <sup>2</sup> WeChat Vision, Tencent Inc.

<sup>3</sup> Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences

<sup>4</sup> Hong Kong University of Science and Technology

<sup>5</sup> Shanghai AI Laboratory

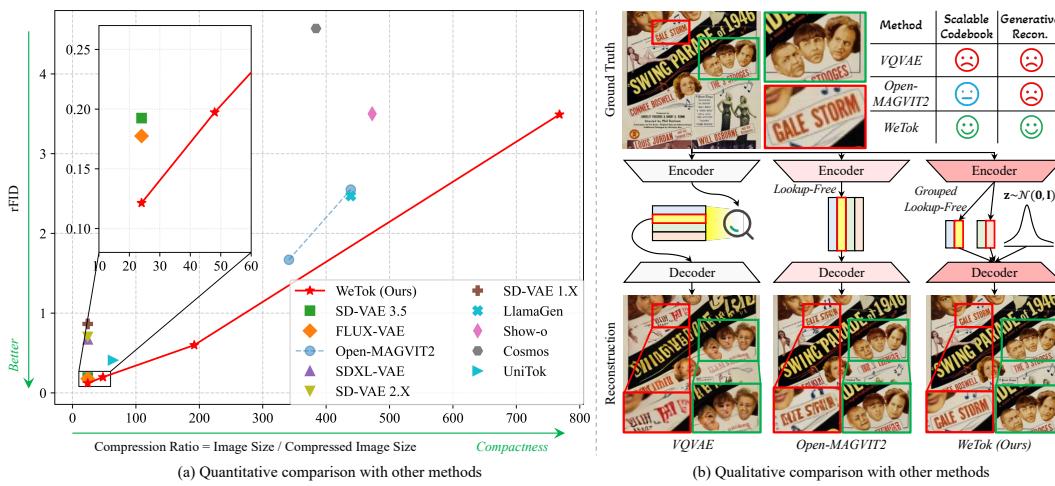


Figure 1: **Zero-shot reconstruction comparison with state-of-the-art tokenizers.** (a) Our WeTok establishes a new state-of-the-art trade-off between compression ratios and reconstruction performance among the compared methods. (b) WeTok achieves a significant improvement in reconstruction quality over previous discrete tokenizers such as VQ-VAE and Open-MAGVIT2.

## ABSTRACT

Visual tokenizer is a critical component for vision generation. However, the existing tokenizers often face unsatisfactory trade-off between compression ratios and reconstruction fidelity. To fill this gap, we introduce a powerful and concise **WeTok** tokenizer, which surpasses the previous leading tokenizers via two core innovations. (1) Group-wise lookup-free Quantization (GQ). We partition the latent features into groups, and perform lookup-free quantization for each group. As a result, GQ can efficiently overcome memory and computation limitations of prior tokenizers, while achieving a reconstruction breakthrough with more scalable codebooks. (2) Generative Decoding (GD). Different from prior tokenizers, we introduce a generative decoder with a prior of extra noise variable. In this case, GD can probabilistically model the distribution of visual data conditioned on discrete tokens, allowing WeTok to reconstruct visual details, especially at high compression ratios. Extensive experiments on mainstream benchmarks show superior performance of our WeTok. On the ImageNet 50k validation set, WeTok achieves a record-low zero-shot rFID (**WeTok: 0.12** vs. **FLUX-VAE: 0.18** vs. **SD-VAE 3.5: 0.19**). Furthermore, our highest compression model achieves a zero-shot rFID of 3.49 with a compression ratio of 768, outperforming Cosmos (384) 4.57 which has only **50% compression rate** of ours. Code and models are available: <https://github.com/zhuangshaobin/WeTok>.

\*Corresponding author. ♪ Work done as interns at WeChat Vision, Tencent Inc.

---

## 1 INTRODUCTION

In visual generation, the high computational cost of pixel-based data is a central challenge (Chen et al., 2020b; Rombach et al., 2022b). Visual tokenizers are a key solution that uses an encoder to compress an image into a compact latent representation and a decoder to reconstruct it (Kingma & Welling, 2013; Rezende et al., 2014), allowing generative models to operate efficiently in this latent space (Rombach et al., 2022a). These tokenizers are broadly divided into two categories: *continuous* (Kingma & Welling, 2013) and *discrete* (Van Den Oord et al., 2017). Continuous tokenizers map images to a continuous latent space, while discrete tokenizers employ a quantizer to produce a finite set of codes. This architectural difference introduces a critical trade-off. Discrete tokenizers can achieve a higher compression ratio, but this efficiency often comes at the cost of lower reconstruction fidelity compared to continuous methods. This leads to a natural question: *Can we build a discrete tokenizer that can maintain high compression as well as achieve high-fidelity reconstruction?*

To achieve this goal, two critical issues must be resolved. (1) **Scalable Codebook.** To minimize quantization error of discrete tokenizers, the existing methods attempt to enlarge the codebook (Yu et al., 2024a; Zhao et al., 2024b; Sun et al., 2024). In particular, the Lookup-Free Quantization (LFQ) (Yu et al., 2024a) quantizes the latent features directly, which largely increases the codebook size for better reconstruction. However, the substantial memory and computational overhead required to manage such a large codebook during training hinders further scalability. (2) **Generative Modeling.** Discrete tokenizers are inherently deterministic. Rather than modeling data distribution of images, decoder is trained to reconstruct the expected value of images (Esser et al., 2020), corresponding to the latent codes from encoder. Such a manner is limited to capture rich diversity and fine details in the original images, leading to unsatisfactory reconstruction, particularly at high compression ratios.

To fill the gap, we introduce a powerful discrete tokenizer, namely **WeTok**, as shown in Fig. 1 (b). Specifically, we introduce two concise designs to solve the issues above. First, we develop a Group-Wise Lookup-Free Quantization (GQ), which partitions the latent feature of each image into groups, and performs lookup-free quantization for each group. As shown in Tab. 1, our GQ addresses the challenge in LFQ where entropy loss (Chang et al., 2022; Jansen et al., 2019) causes memory usage to grow with the codebook size, while yielding superior reconstruction performance. Second, we introduce a Generative Decoder (GD), which mimics the GAN-style generator by adding a prior of extra noise variable. As shown in Fig. 7, GD can effectively model data distribution of input images, allowing to reconstruct visual details at high compression ratios.

Finally, we conduct extensive experiments on mainstream benchmarks, via scaling WeTok across group size, model size, and training data size. Moreover, we pre-train our WeTok on a 400M general-domain dataset across multiple compression ratios. As illustrated in Fig. 1 (a), WeTok consistently outperforms the state-of-the-art continuous and discrete tokenizers, e.g., rFID on ImageNet 50k validation set: **WeTok: 0.12 vs. FLUX-VAE: 0.18** (Batifol et al., 2025) **vs. SD-VAE 3.5: 0.19** (Esser et al., 2024a). Furthermore, our highest compression model also achieves the superior reconstruction performance, e.g., rFID on ImageNet 50k validation set: WeTok: 3.59 **vs. Cosmos: 4.57** (Agarwal et al., 2025), while Cosmos (384) only has **50% compression ratio** of our WeTok (768), showing effectiveness and efficiency of WeTok.

## 2 RELATED WORK

### 2.1 CONTINUOUS TOKENIZER

Generative modeling in the pixel space typically requires extensive compute resources (Chen et al., 2020a; Ho et al., 2020). Subsequent works (Rombach et al., 2022b; Podell et al., 2023; Peebles & Xie, 2023; Esser et al., 2024b; Batifol et al., 2025; Zhuang et al., 2025; 2024; Esser et al., 2024a; Zha et al., 2025; Chen et al., 2025) adopt VAE (Kingma & Welling, 2013), which projects visual content from pixels to latent features, achieving efficient and photo-realistic visual generation at high resolution. FLUX-VAE (Batifol et al., 2025) shows the state-of-the-art performance in both reconstruction quality and generalization ability across all continuous tokenizers. However, continuous tokenizer is criticized for its low compression rate, because latent features are usually stored and calculated in `float 32` or `bfloat16`. Therefore, discrete tokenizers that can store data in `int` or `bool` seem to be more promising in terms of compression capabilities.

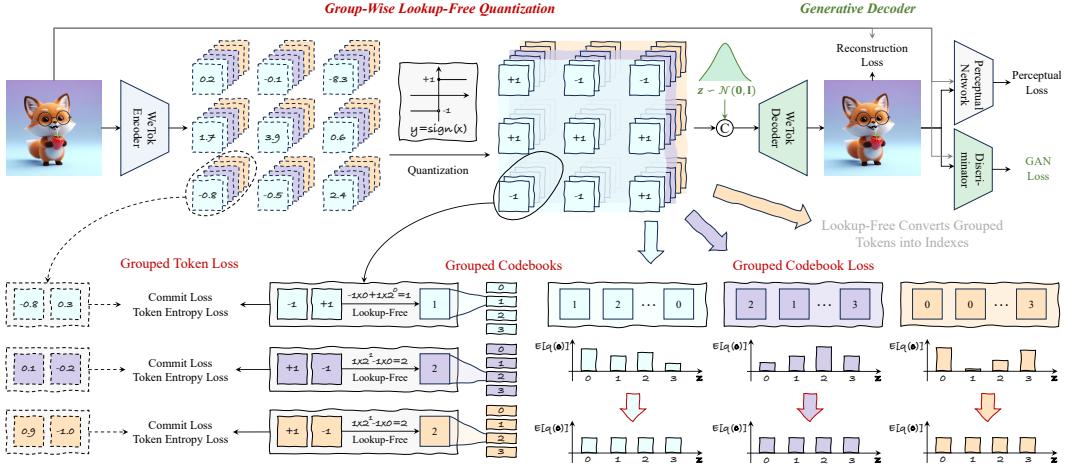


Figure 2: **WeTok with Group-Wise Lookup-Free Quantization and Generative Decoder.**

## 2.2 DISCRETE TOKENIZER

Traditional works like VQVAE (Van Den Oord et al., 2017) and VQGAN (Esser et al., 2021) employ vector-quantization (VQ) to transform visual input into discrete tokens. But they both suffer from low reconstruction quality caused by quantization error and instability of the codebook utilization. To overcome these drawbacks, subsequent works have explored several paths. One line of work introduces specific optimization strategies or modules to improve performance (Lee et al., 2022b; Shi et al., 2024; Zhu et al., 2024; Yu et al., 2024b). Another line of work focuses on mitigating the training instability when scaling up the codebook size by using grouped codebooks (Ma et al., 2025; Jia et al., 2025; Zhang et al., 2025). These methods split the input feature into groups along the channel dimension, where each group is then looked up using a sub-codebook. However, despite these improvements, VQ-based tokenizers still introduce additional inference and training costs due to the lookup operation (Yu et al., 2021b; Lee et al., 2022b; Fang et al., 2025). MAGVIT-v2 (Yu et al., 2024a) introduces Lookup-Free Quantization (LFQ) to address this extra cost and proposes the entropy loss (Chang et al., 2022; Jansen et al., 2019) to ensure the utilization of the codebook. However, the entropy loss causes unaffordable memory cost as it scales linearly with the codebook size, limiting the further expansion of the codebook. Binary Spherical Quantization (BSQ) (Zhao et al., 2024a) is proposed to mitigate this memory issue by assuming independence between the bits of the binary code, while this strong assumption can lead to performance degradation compared to LFQ. Concurrent to the above work, the GQ in our WeTok does not rely on explicit codebooks, and can eliminate the memory usage caused by entropy loss while having better performance than LFQ.

## 2.3 AUTOREGRESSIVE VISUAL GENERATION

The autoregressive (AR) modeling paradigm, which underpins modern Large Language Models (LLMs) (Vaswani et al., 2017), has been successfully adapted for visual generation (Chen et al., 2020a), where models learn to predict sequences of discrete tokens for images (Ramesh et al., 2021b; Ding et al., 2021; Liu et al., 2024) and videos (Hong et al., 2022; Kondratyuk et al., 2023). Recent AR models (Sun et al., 2024; Team, 2024; Wu et al., 2025a; Wang et al., 2024b; Liu et al., 2025) achieve remarkable image quality, highlighting the significant potential of this paradigm. Notably, the success of AR models is critically dependent on the visual tokenizer. Therefore, we adopt our WeTok to the LlamaGen (Sun et al., 2024) framework to enable high-fidelity and efficient autoregressive generation. This shows that our WeTok is not only capable of compression, but its compressed features are also suitable for generative models.

## 3 METHOD

In this section, we first establish the necessary preliminaries for discrete tokenization. We then introduce the Group-Wise Lookup-Free Quantization (GQ) to unify Lookup-Free Quantization (LFQ)

(Yu et al., 2024a) and Binary Spherical Quantization (BSQ) (Zhao et al., 2024b). Finally, we present a Generative Decoder (GD) specifically engineered for high-compression scenarios to reconstruct high-fidelity outputs from the compact representations generated by our GQ.

### 3.1 PRELIMINARIES

**Vector Quantized Variational Autoencoder (Esser et al., 2020).** Given an image  $\mathcal{I} \in \mathbb{R}^{H \times W \times 3}$ , VQVAE first compresses it into latent feature  $\mathcal{U} = \mathcal{E}(\mathcal{I})$ ,  $\mathcal{U} \in \mathbb{R}^{h \times w \times d}$ , through the encoder  $\mathcal{E}$ . Then it is quantized into latent codes  $\mathcal{Q}$  by searching the nearest neighbor in the codebook  $\mathcal{C} \in \mathbb{R}^{K \times d}$ ,

$$\mathcal{Q}[i, j] = \arg \min_{\mathbf{c}_k \in \mathcal{C}} \|\mathcal{U}[i, j] - \mathbf{c}_k\|^2, \quad (1)$$

$$\mathcal{U}_{\mathcal{Q}} = \mathcal{U} + \text{sg}[\mathcal{Q} - \mathcal{U}], \quad (2)$$

where  $\text{sg}[\cdot]$  is stop-gradient operation. Finally, it is reconstructed into image space  $\hat{\mathcal{I}} = \mathcal{G}(\mathcal{U}_{\mathcal{Q}})$  through the decoder  $\mathcal{G}$ . The loss function consists of five parts,

$$\mathcal{L}_{\text{VQVAE}} = \underbrace{\|\mathcal{I} - \hat{\mathcal{I}}\|^2}_{\text{Recon. Loss}} + \underbrace{\|\mathcal{Q} - \text{sg}[\mathcal{U}]\|^2}_{\text{Codebook Loss}} + \alpha \underbrace{\|\mathcal{U} - \text{sg}[\mathcal{Q}]\|^2}_{\text{Commitment Loss}} + \beta \underbrace{\mathcal{L}_{\text{LPIPS}}(\mathcal{I}, \hat{\mathcal{I}})}_{\text{Perceptual Loss}} + \gamma \underbrace{\mathcal{L}_{\text{GAN}}(\mathcal{U}_{\mathcal{Q}})}_{\text{GAN Loss}}, \quad (3)$$

where perceptual loss (Zhang et al., 2018) and GAN loss are introduced for better visual quality.

**Lookup-Free Quantization (Yu et al., 2024a).** LFQ introduces an implicit and non-learnable codebook  $\mathcal{C}_{\text{LFQ}} = \{-1, 1\}^d$  to perform lookup-free quantization on each channel of latent feature,

$$\mathcal{Q}[i, j, k] = \text{sign}(\mathcal{U}[i, j, k]). \quad (4)$$

Since the codebook in LFQ is fixed, there is no need for codebook loss during LFQ training. To address the issue of codebook utilization collapse in VQVAE, LFQ introduces entropy loss,

$$\mathcal{L}_{\text{Entropy}}(\mathcal{U}) = \underbrace{\frac{1}{hw} \sum_{i=1}^h \sum_{j=1}^w H(q(\mathbf{c}|\mathcal{U}[i, j]))}_{\text{Token Entropy Loss}} - \zeta \underbrace{\frac{1}{hw} \sum_{i=1}^h \sum_{j=1}^w q(\mathbf{c}|\mathcal{U}[i, j])}_{\text{Codebook Entropy Loss}}. \quad (5)$$

**Binary Spherical Quantization (Zhao et al., 2024b).** When increasing the codebook size, *i.e.*, increasing  $d$ , the  $H(\cdot)$  calculation in the  $\mathcal{L}_{\text{entropy}}$  of LFQ will lead to significant memory consumption. To alleviate this issue, BSQ first rewrites the token entropy loss as  $\frac{1}{hw} \sum_{i=1}^h \sum_{j=1}^w \sum_{k=1}^d H(q_B(\mathbf{c}[k]|\mathcal{U}[i, j, k]))$ . Next, BSQ rewrites the codebook entropy loss as  $\sum_{k=1}^d H(q_B(\mathbf{c}[k]|\mathcal{U}[i, j, k]))$  by assuming the approximation  $\frac{1}{hw} \sum_{i=1}^h \sum_{j=1}^w q(\mathbf{c}|\mathcal{U}[i, j]) \approx \prod_{k=1}^d \frac{1}{hw} \sum_{i=1}^h \sum_{j=1}^w q(\mathbf{c}[k]|\mathcal{U}[i, j, k])$ . Both operations reduce the variable space of the  $H(\cdot)$  calculation from  $\{-1, 1\}^d$  to the linear combination of  $d \{-1, 1\}$ , significantly decreasing memory consumption. However, due to the approximation proposed by BSQ for the codebook entropy loss, the entropy loss calculation in BSQ introduces errors. As a result, under the same codebook size, the performance of BSQ is inferior to that of LFQ.

### 3.2 GROUP-WISE LOOKUP-FREE QUANTIZATION

As shown in Eq. 5, the computational cost of  $H(\cdot)$  increases linearly with the codebook size. To simultaneously address it and the optimization error issue of BSQ, we propose grouping the latent features first and then performing lookup-free quantization. As shown in Fig. 2, we group the latent features from the channel dimension, reshape  $\mathcal{U}$  into  $\mathcal{U}_G \in \mathbb{R}^{h \times w \times g \times d'}$ , where  $d = gd'$  and  $g$  is a newly added dimension representing the number of groups and  $d'$  is group channel. For  $k$ -th group, there is a non-learnable grouped codebook  $\mathcal{C}_{\text{GQ}, k} = \{-1, 1\}^{d'}$ . The conditional probability after grouping can be formulated as

$$q(\mathbf{c}|\mathcal{U}[i, j]) = \prod_{k=1}^g q_G(\mathbf{c}_k|\mathcal{U}_G[i, j, k]), \quad (6)$$

---

where  $\mathbf{c}_k$  refers to the  $k$ -th latent code after  $\mathbf{c}$  is divided into  $g$  parts, *i.e.*,  $\mathbf{c}_k = \mathbf{c}[(k-1)d'+1 : kd']$ . Considering the additivity of entropy and Eq. 6, we can rewrite the token entropy loss term as

$$\mathcal{L}_{\text{Token Entropy Loss}} = \frac{1}{hw} \sum_{i=1}^h \sum_{j=1}^w H(q(\mathbf{c}|\mathcal{U}[i,j])) = \frac{1}{hw} \sum_{i=1}^h \sum_{j=1}^w \sum_{k=1}^g H(q_G(\mathbf{c}_k|\mathcal{U}_G[i,j,k])). \quad (7)$$

We transform the original token entropy loss calculation from the  $\{-1, 1\}^d$  space into grouped token entropy loss which is a linear combination of entropy from  $g \{-1, 1\}^{d'}$  spaces. This change eliminates the token entropy loss as the memory bottleneck.

However, for the codebook entropy loss, the  $H(\sum \cdot)$  operation prevents us from leveraging the additivity of entropy to decompose the  $\{-1, 1\}^d$  space into a linear combination of multiple subspaces. We propose the assumption that

$$\sum_{i=1}^h \sum_{j=1}^w q(\mathbf{c}|\mathcal{U}[i,j]) \approx \prod_{k=1}^g \sum_{i=1}^h \sum_{j=1}^w q_G(\mathbf{c}_k|\mathcal{U}_G[i,j,k]). \quad (8)$$

Thus, we leverage the additivity of entropy to transform the codebook entropy loss into

$$\mathcal{L}_{\text{Codebook Entropy Loss}} = H\left(\frac{1}{hw} \sum_{i=1}^h \sum_{j=1}^w q(\mathbf{c}|\mathcal{U}[i,j])\right) = \sum_{k=1}^g H\left(\frac{1}{hw} \sum_{i=1}^h \sum_{j=1}^w q_G(\mathbf{c}_k|\mathcal{U}_G[i,j,k])\right). \quad (9)$$

Similar to the token entropy loss, we transform the calculation of the codebook entropy loss from the  $\{-1, 1\}^d$  space into a grouped codebook entropy loss, which becomes a linear combination of entropy from  $g \{-1, 1\}^{d'}$  spaces.

Governed by the group number  $g$ , GQ provides a tunable trade-off between approximation accuracy and memory cost. When  $g = 1$ , GQ degenerates to the LFQ formulation, which uses no approximation and thus achieves high precision at the cost of substantial memory consumption. Conversely, when  $g = d$ , GQ mirrors the BSQ approach, which employs a strong approximation to minimize memory usage but consequently suffers from significant optimization error. Experiments in Sec. 4.2 demonstrate that by selecting an appropriate intermediate value for  $g$ , GQ can significantly reduce memory overhead while introducing only minimal optimization error. This allows GQ to surpass even the standard LFQ baseline. Furthermore, this tunable design provides the flexibility to effectively scale the codebook to a virtually unlimited size.

### 3.3 GENERATIVE DECODER

Unlike the continuous tokenizer, the decoders in previous discrete (Esser et al., 2021; Yu et al., 2024a) tokenizers fit deterministic transformations. Although GAN loss is employed during training as shown in Eq. 3 and its specific form is as follows

$$\mathcal{L}_{\text{GAN}}(\mathcal{U}_Q) = \log(1 - \mathcal{D}(\mathcal{G}(\mathcal{U}_Q))), \quad (10)$$

where  $\mathcal{D}$  is the discriminator. However, the GAN loss only serves to assist in improving the perceptual quality. In scenarios with high compression rates, the correspondence between  $\mathcal{U}_Q$  and the ground truth is likely not unique. In such cases, we need the decoder to become a generative model. As shown in Fig. 2, we randomly sample  $\mathbf{z} \in \mathcal{N}(\mathbf{0}, \mathbf{I})$ , concatenate it with  $\mathcal{U}_Q$  along the channel dimension, and then feed the result into the decoder. In this way, the GAN loss is subsequently reformulated as

$$\mathcal{L}_{\text{GAN}}(\mathcal{U}_Q) = \mathbb{E}_{\mathbf{z} \in \mathcal{N}(\mathbf{0}, \mathbf{I})} [\log(1 - \mathcal{D}(\mathcal{G}(\mathbf{z}, \mathcal{U}_Q)))], \quad (11)$$

where  $\mathcal{U}_Q$  serves as a condition for the generation process. The development from Eq. 10 to Eq. 11 is not merely concatenating  $\mathbf{z}$  to the input of the decoder. More importantly, the learning objective of the decoder shifts to modeling the transformation from Gaussian noise conditioned on  $\mathcal{U}_Q$  to the ground truth distribution, turning it into a GD.

To ensure training stability, we employ a two-stage training process. In the first stage, we train our WeTok with the reconstruction loss, *i.e.*, Eq. 3, 7 and 9. In the second stage, we adapt the model for generative tasks. Specifically, we expand the channel dimension of the `conv_in` layer in decoder to accept  $\mathbf{z}$  as additional input. To preserve the powerful reconstruction capabilities learned in the

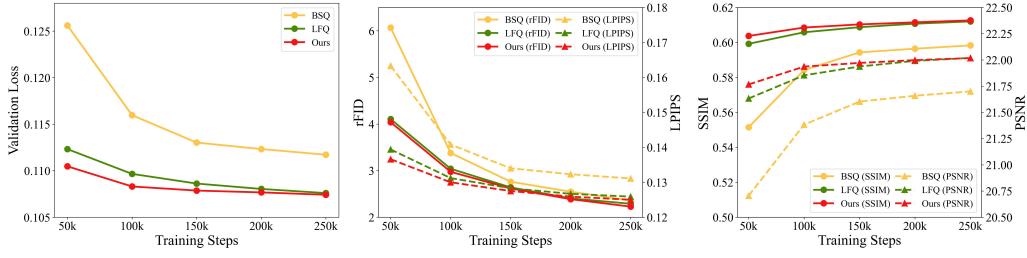


Figure 3: **Quantization method ablation.** GQ and LFQ are significantly better than BSQ.

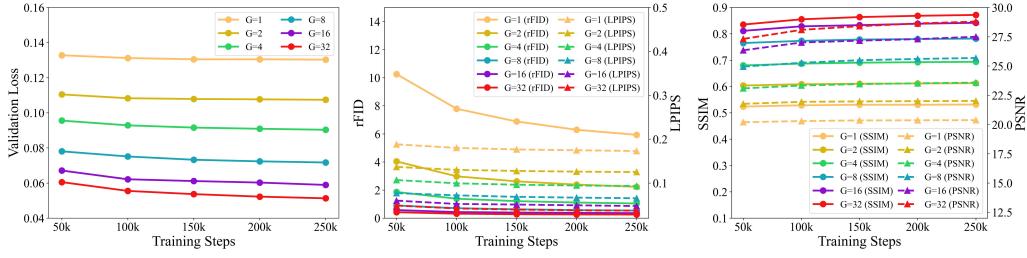


Figure 4: **Number of group ablation.**  $G$  refers to the number of group. The reconstruction performance of the model increases significantly with the increase of  $G$ .

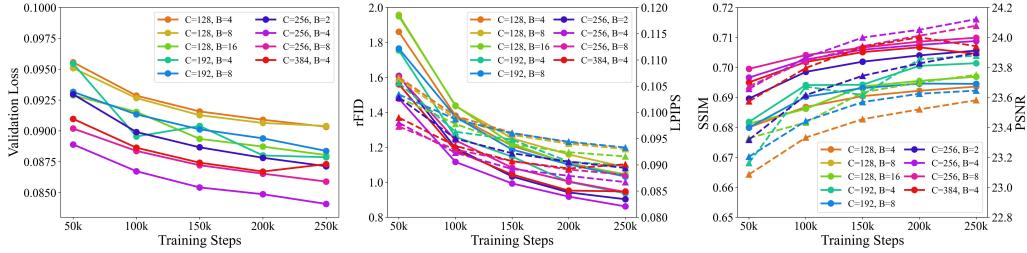


Figure 5: **Model architecture ablation.**  $C$  and  $B$  refer to the number of base channel and residual block respectively.  $C = 256$  and  $B = 4$  achieve the best reconstruction performance.

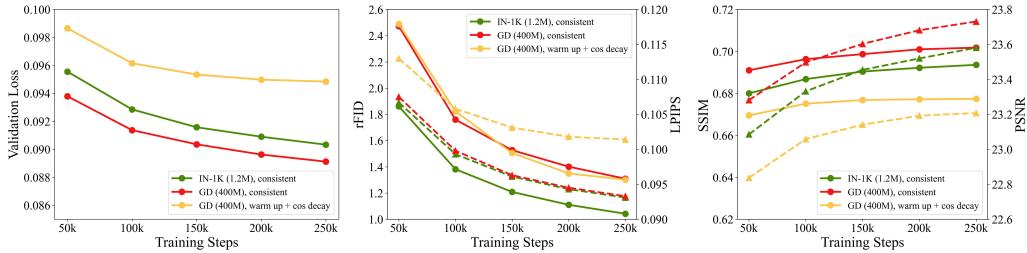


Figure 6: **Training data and learning rate schedule ablation.** GD refers to general-domain data. The model trained on general-domain data is not as good as the model trained on in-distribution data in terms of distribution fitting metrics, but has better generalization on PSNR and SSIM. The effect of consistent learning schedule is significant compared with warm up + cosine decay.

first stage, the newly added weights for these channels are zero-initialized. This strategy ensures that at the beginning of the second stage, the decoder’s behavior is identical to its pre-trained state. Consequently, this versatile approach allows our GD to be effectively integrated with various pre-trained discrete tokenizers that share a similar architectural foundation.

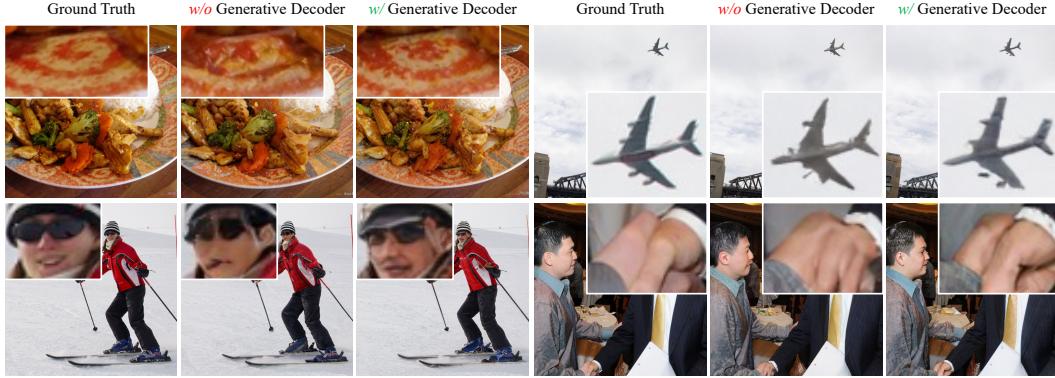


Figure 7: **Qualitative ablation of GD on MS-COCO val2017.** The images reconstructed by the model with GD are obviously more fidelity and natural.

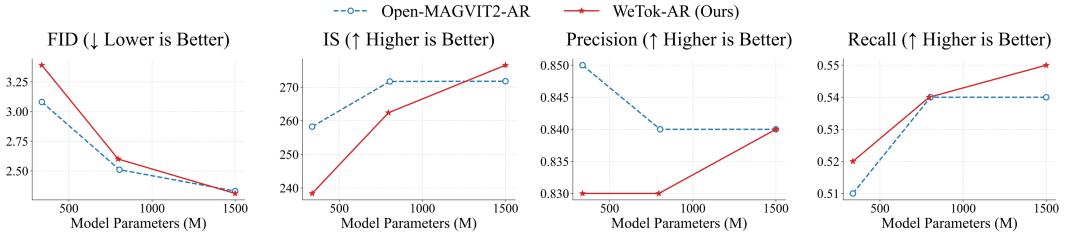


Figure 8: **Parameter ablation experiments of the Autoregressive model based on WeTok.** Compared with Open-MAGVIT2, the performance of the AR model based on WeTok improves more significantly with the increase in the number of parameters.

Table 1: **Memory usage ablation.** We set Table 2: **Generative decoder modeling ablation.**  $d'=8$  in GFQ. **OOM** refers to *out of memory*. **Stage2** refers to generative decoder modeling.

Method	$d = 8$	$d = 16$	$d = 24$	$d = 32$	$d = 40$	<i>Stage1</i>	<i>Stage2</i>	$rFID \downarrow$	LPIPS $\downarrow$	SSIM $\uparrow$	PSNR $\uparrow$
LFQ	10.5 GB	10.6 GB	<b>OOM</b>	<b>OOM</b>	<b>OOM</b>	✓	✗	5.37	0.17	0.54	20.53
BSQ	10.5 GB	10.5 GB	10.6 GB	10.6 GB	10.6 GB	✓	✓	<b>3.90</b>	<b>0.16</b>	<b>0.55</b>	<b>20.72</b>
GFQ	10.5 GB	10.6 GB	10.6 GB	10.6 GB	10.6 GB						

## 4 EXPERIMENTS

### 4.1 IMPLEMENTATION

**Datasets.** To thoroughly evaluate our WeTok, we perform large-scale training on two datasets: **(i) 1.2M ImageNet** (Russakovsky et al., 2014) training set; **(ii) 400M** general-domain dataset. To make state-of-the-art comparison, we then evaluate WeTok performance on the ImageNet 50k validation set and MS-COCO 2017 validation set (Lin et al., 2014b). This multi-benchmark evaluation ensures a robust assessment of our model’s capabilities across different domains and data distributions. Unless otherwise stated, we conduct a series of ablation studies on the ImageNet training set. Besides, we train the class-to-image model on the ImageNet training set and test it on the validation set.

**Settings.** WeTok adopts the architecture proposed in Open-MAGVIT2 (Luo et al., 2024), employing the CNN architecture for encoder, decoder, and discriminator. Images are randomly cropped to  $256 \times 256$  for training. For the ablation study, all models are trained for 250K steps using a consistent set of hyperparameters. Including a fixed learning rate of 1e-4, the most global batch size is 128, and the Adam (Kingma & Ba, 2014) optimizer with  $\beta_1 = 0.5$  and  $\beta_2 = 0.9$ . For large-scale training, we used the same Adam optimizer settings, but increased the most global batch size to 1024. In this phase, other hyperparameters were individually tuned for each model to achieve optimal performance. For class-to-image generation, we adopt the generative transformer architecture in LlamaGen (Sun et al., 2024). All experiments were carried out using the H20 GPU with Pytorch. More details in Sup. A.

Table 3: **Reconstruction evaluation on  $256 \times 256$  ImageNet 50K validation set.** All models are trained on ImageNet. WeTok achieves SOTA results on different downsampling rates.

Method	Token Type	Tokens	Ratio	Train Resolution	Codebook Size	rFID $\downarrow$	PSNR $\uparrow$	Codebook Usage $\uparrow$
VQGAN (Esser et al., 2020)	2D	$16 \times 16$	16	$256 \times 256$	1024	8.30	19.51	—
VQGAN (Esser et al., 2020)	2D	$16 \times 16$	16	$256 \times 256$	16384	4.99	20.00	—
SD-VQGAN (Rombach et al., 2022b)	2D	$16 \times 16$	16	$256 \times 256$	16384	5.15	—	—
MaskGIT (Chang et al., 2022)	2D	$16 \times 16$	16	$256 \times 256$	1024	2.28	—	—
ReVQ (Zhang et al., 2025)	2D	$16 \times 16$	16	$256 \times 256$	65536	2.57	21.69	—
LlamaGen (Sun et al., 2024)	2D	$16 \times 16$	16	$256 \times 256$	32768	2.26	20.59	85%
LlamaGen (Sun et al., 2024)	2D	$16 \times 16$	16	$256 \times 256$	16384	2.19	20.79	97%
ReVQ (Zhang et al., 2025)	2D	$16 \times 16$	16	$256 \times 256$	262144	2.05	21.96	—
VAR (Tian et al., 2024)	2D	$16 \times 16$	16	$256 \times 256$	4096	—	21.30	97%
IBQ (Shi et al., 2025)	2D	$16 \times 16$	16	$256 \times 256$	16384	1.37	22.35	96%
Open-MAGVIT2 (Luo et al., 2024)	2D	$16 \times 16$	16	$256 \times 256$	$2^{18}$	1.17	22.64	100%
IBQ (Shi et al., 2025)	2D	$16 \times 16$	16	$256 \times 256$	262144	1.00	20.30	84%
VFMTok (Zheng et al., 2025)	1D	256	—	$256 \times 256$	16384	0.89	—	100%
GigaTok (Xiong et al., 2025)	1D	256	—	$256 \times 256$	16384	0.79	21.65	—
AliTok (Wu et al., 2025c)	1D	273	—	$256 \times 256$	4096	0.84	—	—
MGVQ (Jia et al., 2025)	2D	$16 \times 16$	16	$256 \times 256$	$2^{52}$	0.64	23.71	100%
<b>WeTok (Ours)</b>	2D	$16 \times 16$	16	$256 \times 256$	$2^{32}$	<b>0.61</b>	<b>24.50</b>	<b>100%</b>
ViT-VQGAN (Yu et al., 2021a)	2D	$32 \times 32$	8	$256 \times 256$	8192	1.28	—	—
OmiTokenizer-VQ (Wang et al., 2024a)	2D	$32 \times 32$	8	$256 \times 256$	8192	1.11	—	—
LlamaGen (Sun et al., 2024)	2D	$32 \times 32$	8	$256 \times 256$	16384	0.59	24.45	—
Open-MAGVIT2 (Luo et al., 2024)	2D	$32 \times 32$	8	$128 \times 128$	$2^{18}$	0.34	27.02	100%
BSQ (Zhao et al., 2024a)	1D	1024	—	$256 \times 256$	$2^{18}$	1.14	25.36	100%
BSQ (Zhao et al., 2024a)	1D	1024	—	$256 \times 256$	$2^{36}$	0.45	28.14	100%
MGVQ (Jia et al., 2025)	2D	$32 \times 32$	8	$256 \times 256$	$2^{52}$	0.31	28.42	100%
<b>WeTok (Ours)</b>	2D	$32 \times 32$	8	$256 \times 256$	$2^{32}$	<b>0.19</b>	<b>29.69</b>	<b>100%</b>

#### 4.2 ABLATION STUDY

We conducted a comprehensive ablation study to validate the key components of WeTok. We first verified the effectiveness of our two important algorithms, GQ and GD. Then we ablate the performance improvements brought to WeTok by various dimensions, including the number of groups  $g$  in GQ, model architecture, training data, and learning rate schedule. In addition, we conduct parameter ablation experiments on the AR model based on WeTok. More implementation details in Sup. A.1. We employ validation loss, rFID (Heusel et al., 2017), LPIPS (Zhang et al., 2018), SSIM (Wang et al., 2004), and PSNR to evaluate the quality of ablation study.

**Quantization method.** We conduct an ablation study of quantization methods under the same compression ratio, *i.e.*,  $d = gd' = 16$ . As shown in Tab. 1, our GQ not only has almost no increase in GPU memory usage due to quantization like BSQ. In Fig. 3, we set the same compression ratio for 3 different quantization methods (LFQ:  $g=1$ ,  $d'=16$ ; BSQ:  $g=16$ ,  $d'=1$ ; GQ:  $g=2$ ,  $d'=8$ ), GQ performs better than LFQ and far exceeds BSQ.

**Generative decoder.** When the performance of the model is saturated after the first stage of reconstruction training, we conduct the second stage of generative training. As shown in Tab. 2, the results show that GD can continue to improve the reconstruction performance of the model, especially in rFID. In addition, we present qualitative ablation results in Fig. 7. After converting the decoder to a generative model, the reconstructed images are more realistic and consistent with visual logic, demonstrating the effectiveness of GD.

**Number of group in GQ.** We increase the number of groups  $g$  by a power of 2. As shown in Fig. 4, the results show that as  $g$  increases, the reconstruction performance of the model continues to increase significantly and does not encounter the memory bottleneck like LFQ.

**Model architecture.** We performed ablation studies to scale up our discrete tokenizer, focusing on the number of base channels and residual blocks in the encoder and decoder. As shown in Fig. 5, across 9 different settings, a configuration with 256 base channels and 4 residual blocks achieves the best reconstruction performance. The encoder and decoder for this optimal architecture contain 198M and 261M parameters, respectively.

**Table 4: Zero-shot reconstruction performance on ImageNet 50k validation set and MS-COCO val2017.** The tokenizers are trained with large scale general-domain datasets. Text in gray signifies the results directly from Cosmos (Agarwal et al., 2025) report. Our WeTok achieves the best performance on both resolution settings.

Method	Tokenizer Type	Training Data	Ratio	Compression Ratio↑	MS-COCO 2017			Imagenet-1k		
					rFID↓	PSNR↑	SSIM↑	rFID↓	PSNR↑	SSIM↑
<i>Resize 256 × 256</i>										
DALL-E dVAE (Ramesh et al., 2021a)	Discrete	103M	18	118	48.60	<b>26.97</b>	0.08	32.63	<b>27.31</b>	<b>0.79</b>
Cosmos (Agarwal et al., 2025)	Discrete	-	16	384	11.97	19.22	0.48	4.57	19.93	0.49
Show-o (Xie et al., 2024)	Discrete	35M	16	473	9.26	20.90	0.59	3.50	21.34	0.59
Open-MAGVIT2-I-PT (Luo et al., 2024)	Discrete	100M	16	439	7.93	22.21	0.62	2.55	22.21	0.62
LlamaGen (Sun et al., 2024)	Discrete	70M	16	439	8.40	20.28	0.55	2.47	20.65	0.54
Open-MAGVIT2-I-PT (Luo et al., 2024)	Discrete	100M	16	341	<b>6.76</b>	22.31	<b>0.65</b>	<b>1.67</b>	22.70	0.64
<b>WeTok (Ours)</b>	Discrete	400M	32	<b>768</b>	8.94	20.31	0.55	3.49	20.77	0.55
BSQ (Zhao et al., 2024b)	Discrete	1B	-	219	-	-	-	3.81	24.12	0.66
QLIP-B (Zhao et al., 2025)	Discrete	1B	-	219	-	-	-	3.21	23.16	0.63
QLIP-L (Zhao et al., 2025)	Discrete	1B	-	168	-	-	-	1.46	<b>25.36</b>	0.69
SD-VAE 1.x (Rombach et al., 2022a)	Continuous	1B	8	24	5.94	23.21	0.69	1.22	23.54	0.68
SD-VAE 1.x (Rombach et al., 2022a)	Discrete	1B	8	110	5.75	24.17	0.70	1.13	24.48	0.69
<b>WeTok (Ours)</b>	Discrete	400M	16	<b>192</b>	<b>4.41</b>	<b>24.44</b>	<b>0.74</b>	<b>0.60</b>	24.77	<b>0.73</b>
QLIP-B (Zhao et al., 2025)	Discrete	1B	-	55	-	-	-	0.70	26.79	0.79
SD-VAE 2.x (Rombach et al., 2022a)	Continuous	6B	8	24	4.26	26.62	0.77	0.70	26.90	0.76
SDXL-VAE (Podell et al., 2023)	Continuous	>6B	8	24	3.93	27.08	0.80	0.67	27.37	0.78
UniTok (Ma et al., 2025)	Discrete	1B	16	64	-	-	-	0.41	-	-
<b>WeTok (Ours)</b>	Discrete	400M	8	<b>48</b>	<b>2.18</b>	<b>29.49</b>	<b>0.89</b>	<b>0.20</b>	<b>29.63</b>	<b>0.88</b>
SD-VAE 3.5 (Esser et al., 2024b)	Continuous	-	8	24	1.66	31.08	0.90	0.19	31.19	0.90
FLUX-VAE (Labs, 2024)	Continuous	-	8	24	<b>1.35</b>	<b>32.32</b>	0.93	0.18	<b>32.74</b>	0.92
<b>WeTok (Ours)</b>	Discrete	400M	8	<b>24</b>	1.43	32.00	<b>0.93</b>	<b>0.12</b>	32.06	<b>0.93</b>
<i>Original Resolution</i>										
DALL-E dVAE (Ramesh et al., 2021a)	Discrete	103M	18	118	55.07	<b>25.15</b>	<b>0.75</b>	36.84	<b>25.46</b>	<b>0.74</b>
Cosmos (Agarwal et al., 2025)	Discrete	-	16	384	7.51	20.45	0.52	1.93	20.56	0.51
Cosmos (Agarwal et al., 2025)	Discrete	-	16	384	7.23	20.45	0.53	2.52	20.49	0.52
Open-MAGVIT2-I-PT (Luo et al., 2024)	Discrete	100M	16	439	6.65	21.61	0.57	1.39	21.74	0.56
Open-MAGVIT2-I-PT (Luo et al., 2024)	Discrete	100M	16	341	<b>5.10</b>	22.18	0.60	<b>0.78</b>	22.24	0.59
<b>WeTok (Ours)</b>	Discrete	400M	32	<b>768</b>	7.83	20.47	0.52	2.03	20.54	0.51
SD-VAE 1.x (Rombach et al., 2022a)	Continuous	1B	8	24	5.94	21.68	0.64	1.35	21.99	0.63
SD-VAE 1.x (Rombach et al., 2022a)	Discrete	1B	8	110	6.07	22.54	0.65	1.23	22.82	0.64
<b>WeTok (Ours)</b>	Discrete	400M	16	<b>192</b>	<b>3.80</b>	<b>23.70</b>	<b>0.67</b>	<b>0.40</b>	<b>23.75</b>	<b>0.67</b>
SD-VAE 2.x (Rombach et al., 2022a)	Continuous	6B	8	24	4.63	24.82	0.72	0.78	25.08	0.71
SDXL-VAE (Podell et al., 2023)	Continuous	>6B	8	24	4.23	25.11	0.74	0.72	25.38	0.73
<b>WeTok (Ours)</b>	Discrete	400M	8	<b>48</b>	<b>2.09</b>	<b>27.50</b>	<b>0.82</b>	<b>0.18</b>	<b>27.54</b>	<b>0.82</b>
SD-VAE 3.5 (Esser et al., 2024b)	Continuous	-	8	24	1.64	28.35	0.86	0.24	28.39	0.86
<b>WeTok (Ours)</b>	Discrete	400M	8	<b>24</b>	<b>1.46</b>	<b>29.47</b>	<b>0.88</b>	<b>0.12</b>	<b>29.51</b>	<b>0.88</b>

**Training data.** We compare the model trained on the 1.2M ImageNet training set and the other trained on a larger 400M general-domain dataset. As shown in Fig. 6, the model trained on the general-domain dataset achieved a better validation loss and SSIM and PSNR scores. However, it yielded worse rFID and LPIPS scores. We attribute it to the data distribution gap. The ImageNet training and validation sets are in-distribution, whereas the general-domain data are significantly more diverse. This reveals the trade-off between generalization and performance on specific in-distribution evaluation metrics.

**Learning rate schedule.** A learning rate schedule composed of a warm-up phase followed by a cosine decay is widely adopted in training models. However, we found that this convention may be suboptimal for the training of discrete tokenizers. As illustrated in Fig. 6, the model trained with a constant learning rate demonstrates markedly better performance. Based on this, we exclusively adopted the constant learning rate schedule for all subsequent large-scale training in WeTok.

**Parameter of autoregressive model.** As shown in Fig. 8, We conduct ablation experiments on the parameter size of the WeTok-based AR model and compared it with an Open-MAGVIT2-based AR model. The experimental results show that the performance of the WeTok-based AR model is slightly inferior to that of the Open-MAGVIT2-based AR model when the parameter size is small. However, as the parameter size increases, the WeTok-based AR model surpasses the Open-MAGVIT2-based AR model in all indicators.

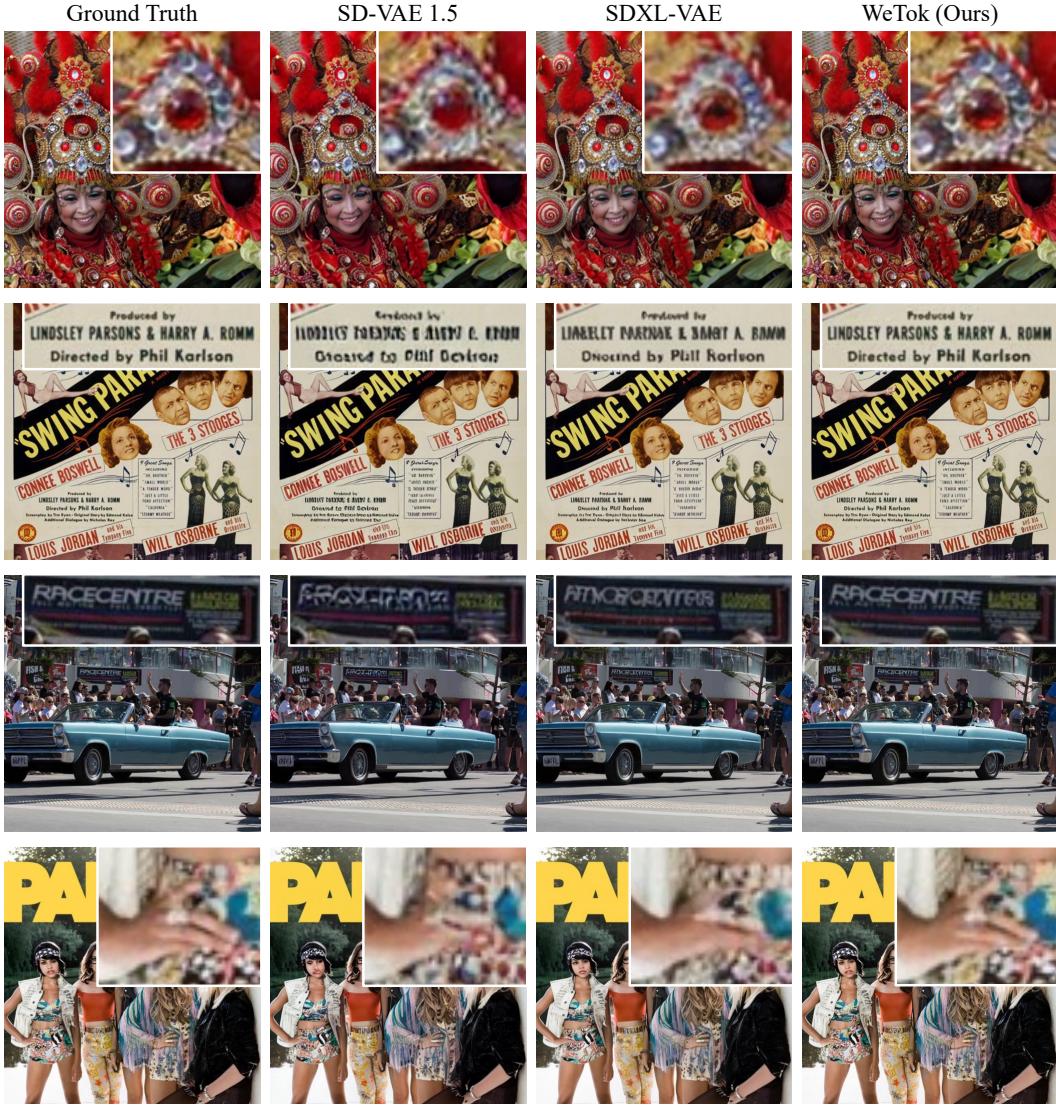


Figure 9: Qualitative comparison of  $512 \times 512$  image reconstruction on TokBench.

#### 4.3 COMPARISON WITH STATE-OF-THE-ART

**Visual Reconstruction.** We first evaluate WeTok’s performance in an in-distribution setting, where both the training and validation data originate from the ImageNet dataset. This establishes our model’s core effectiveness before addressing the more complex challenge of distribution gaps. As shown in Tab. 3, WeTok achieves state-of-the-art reconstruction performance, outperforming existing methods across different downsampling ratios. Subsequently, we compare WeTok against other methods, with all models being trained on a large-scale general-domain dataset. To ensure a fair and unbiased comparison, we have carefully filtered this dataset to exclude any data samples in ImageNet and MS-COCO Lin et al. (2014a). This crucial step prevents potential data leakage and ensures that the evaluation accurately measures the models’ ability to generalize from a broad data distribution to the specific target domain. As shown in Tab. 4, our WeTok shows the state-of-the-art reconstruction performance in a wide range of compression ratio scenarios. It not only shows the strongest performance among discrete tokenizers, but also even surpasses the current strongest continuous tokenizers, FLUX-VAE (Labs, 2024) and SD-VAE 3.5 (Esser et al., 2024a). In addition, we present qualitative comparison results on the TokBench (Wu et al., 2025b) in Fig. 9. Compared to

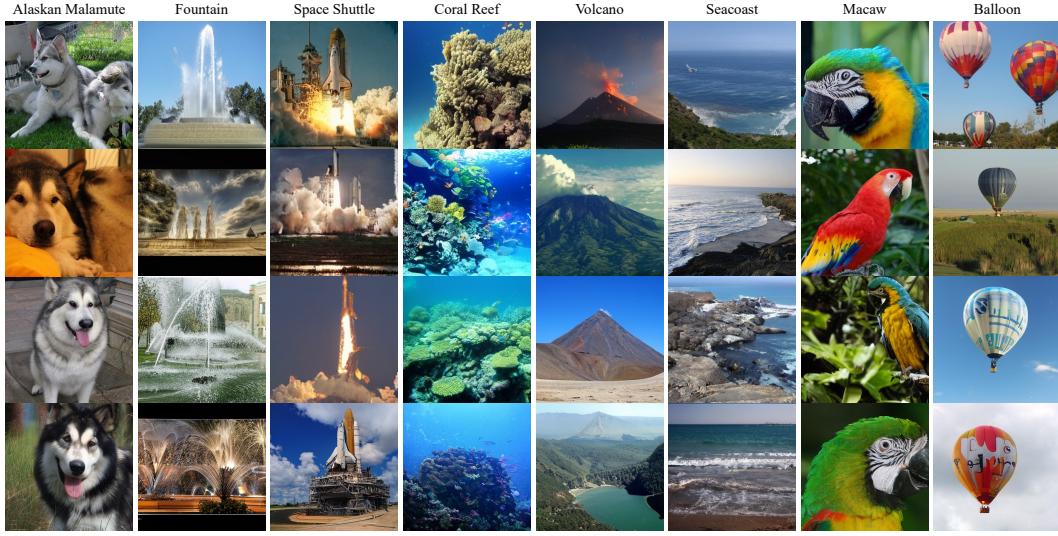


Figure 10: WeTok-AR-XL generated samples at  $256 \times 256$  resolution.

Table 5: **Class-conditional generation on  $256 \times 256$  ImageNet.** \* specifies the generated images are  $384 \times 384$  and are resized to  $256 \times 256$  for evaluation.

Type	Model	#Para.	FID $\downarrow$	IS $\uparrow$	Precision $\uparrow$	Recall $\uparrow$
Diffusion	ADM (Dhariwal & Nichol, 2021)	554M	10.94	101.0	0.69	0.63
	CDM (Ho et al., 2021)	—	4.88	158.7	—	—
	LDM-4 (Rombach et al., 2022b)	400M	3.60	247.7	—	—
	DiT-XL/2 (Peebles & Xie, 2022)	675M	2.27	278.2	0.83	0.57
AR	VQGAN (Esser et al., 2020)	227M	18.65	80.4	0.78	0.26
	VQGAN (Esser et al., 2020)	1.4B	15.78	74.3	—	—
	VQGAN-re (Esser et al., 2020)	1.4B	5.20	280.3	—	—
	ViT-VQGAN (Yu et al., 2021a)	1.7B	4.17	175.1	—	—
	ViT-VQGAN-re (Yu et al., 2021a)	1.7B	3.04	227.4	—	—
	RQTran. (Lee et al., 2022a)	3.8B	7.55	134.0	—	—
	RQTran.-re (Lee et al., 2022a)	3.8B	3.80	323.7	—	—
VAR	VAR-d16 (Tian et al., 2024)	310M	3.30	274.4	0.84	0.51
	VAR-d20 (Tian et al., 2024)	600M	2.57	302.6	0.83	0.56
	VAR-d24 (Tian et al., 2024)	1.0B	2.09	312.9	0.82	0.59
	VAR-d30 (Tian et al., 2024)	2.0B	1.92	323.1	0.82	0.59
AR	LlamaGen-L* (Sun et al., 2024)	343M	3.07	256.06	0.83	0.52
	LlamaGen-XL* (Sun et al., 2024)	775M	2.62	244.08	0.80	0.57
	LlamaGen-XXL* (Sun et al., 2024)	1.4B	2.34	253.90	0.80	0.59
	LlamaGen-L (Sun et al., 2024)	343M	3.80	248.28	0.83	0.51
	LlamaGen-XL (Sun et al., 2024)	775M	3.39	227.08	0.81	0.54
	LlamaGen-XXL (Sun et al., 2024)	1.4B	3.09	253.61	0.83	0.53
	UniTok (Ma et al., 2025)	1.4B	2.51	216.7	0.82	0.57
	Open-MAGVIT2-AR-B (Luo et al., 2024)	343M	3.08	258.26	0.85	0.51
	Open-MAGVIT2-AR-L (Luo et al., 2024)	804M	2.51	271.70	0.84	0.54
WeTok-AR-XL (Ours)	Open-MAGVIT2-AR-XL (Luo et al., 2024)	1.5B	2.33	271.77	0.84	0.54
	WeTok-AR-XL (Ours)	1.5B	2.31	276.55	0.84	0.55

the widely used SDXL-VAE (Podell et al., 2023) and SD-VAE 1.5 (Rombach et al., 2022b), WeTok’s reconstruction performance is significantly superior under the same compression ratio.

**Visual Generation.** To evaluate WeTok’s capabilities beyond image reconstruction, we extend its application to visual generation. We adopt an autoregressive model modified from LlamaGen as Open-MAGVIT2 for this task, primarily due to the well-established scalability of such models. Specifically, we employ the in-distribution WeTok with  $16 \times$  downsampling in Tab. 3 as the image tokenizer. More training details can be found in Sup. A.2. As shown in Tab. 5, our WeTok-based generative model achieves state-of-the-art performance on the ImageNet 50K validation set. This

---

result demonstrates that our WeTok is a highly effective tokenizer not only for image reconstruction but also for high-fidelity visual generation tasks. As shown in Fig. 10, we show the realistic and diverse image generation results of our WeTok-AR-XL.

## 5 CONCLUSION

In this paper, we introduce WeTok, a powerful discrete visual tokenizer designed to resolve the long-standing conflict between compression efficiency and reconstruction fidelity. Our primary contributions are the Group-Wise Lookup-Free Quantization (GQ) method, which provides a scalable and memory-efficient solution for codebooks, and Generative Decoder (GD) that excels at producing high-fidelity images even from highly compressed representations. Through extensive experiments, we demonstrated that WeTok consistently outperforms existing state-of-the-art discrete and continuous tokenizers in both in-distribution and zero-shot reconstruction tasks across a wide range of compression ratios. Furthermore, by integrating WeTok into an autoregressive framework, we achieved state-of-the-art performance in class-conditional image generation, confirming that its learned tokens are highly effective for downstream generative tasks. WeTok establishes a new performance standard, proving that discrete tokenizers can achieve superior reconstruction quality without compromising their inherent advantage in compression.

---

## REFERENCES

- Nvidia Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaoqiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, Yunhao Ge, Jinwei Gu, Siddharth Gururani, Ethan He, Jiahui Huang, Jacob Samuel Huffman, Pooya Jannaty, Jingyi Jin, Seung Wook Kim, Gergely Kl'ar, Grace Lam, Shiyi Lan, Laura Leal-Taixé, Anqi Li, Zhaoshuo Li, Chen-Hsuan Lin, Tsung-Yi Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Arsalan Mousavian, Seungjun Nah, Sriharsha Niverty, David Page, Despoina Paschalidou, Zeeshan Patel, Lindsey Pavao, Morteza Ramezanali, Fitsum A. Reda, Xiao-Shuai Ren, Vasanth Rao Naik Sabavat, Ed Schmerling, Stella Shi, Bartosz Stefaniak, Shitao Tang, Lyne P. Tchapmi, Przemek Tredak, Wei-Cheng Tseng, Jibin Rajan Varghese, Hao Wang, Haoxiang Wang, Hengyi Wang, Tingwei Wang, Fangyin Wei, Xinyue Wei, Jay Zhangjie Wu, Jiashu Xu, Wei Yang, Lin Yen-Chen, Xiaohui Zeng, Yuan Zeng, Jing Zhang, Qinsheng Zhang, Yuxuan Zhang, Qingqing Zhao, and Artur Zolkowski. Cosmos world foundation model platform for physical ai. *ArXiv*, abs/2501.03575, 2025.
- Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv e-prints*, pp. arXiv–2506, 2025.
- Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. Maskgit: Masked generative image transformer. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11305–11315, 2022.
- Hao Chen, Yujin Han, Fangyi Chen, Xiang Li, Yidong Wang, Jindong Wang, Ze Wang, Zicheng Liu, Difan Zou, and Bhiksha Raj. Masked autoencoders are effective tokenizers for diffusion models. In *Forty-second International Conference on Machine Learning*, 2025.
- Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International conference on machine learning*, pp. 1691–1703. PMLR, 2020a.
- Mark Chen, Alec Radford, Jeff Wu, Heewoo Jun, Prafulla Dhariwal, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International Conference on Machine Learning*, 2020b.
- Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *ArXiv*, abs/2105.05233, 2021.
- Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in neural information processing systems*, 34:19822–19835, 2021.
- Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12868–12878, 2020.
- Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.
- Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024a.
- Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning*, 2024b.
- Xianghong Fang, Litao Guo, Hengchao Chen, Yuxuan Zhang, Dingjie Song, Yexin Liu, Hao Wang, Harry Yang, Yuan Yuan, Qiang Sun, et al. Enhancing vector quantization with distributional matching: A theoretical and empirical study. *arXiv preprint arXiv:2506.15078*, 2025.

- 
- Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Neural Information Processing Systems*, 2017.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- Jonathan Ho, Chitwan Saharia, William Chan, David J. Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, 23:47:1–47:33, 2021.
- Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pre-training for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.
- Aren Jansen, Daniel P. W. Ellis, Shawn Hershey, R. Channing Moore, Manoj Plakal, Ashok Popat, and Rif A. Saurous. Coincidence, categorization, and consolidation: Learning to recognize sounds with minimal supervision. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 121–125, 2019.
- Mingkai Jia, Wei Yin, Xiaotao Hu, Jiaxin Guo, Xiaoyang Guo, Qian Zhang, Xiao-Xiao Long, and Ping Tan. Mgvq: Could vq-vae beat vae? a generalizable tokenizer with multi-group quantization. *arXiv preprint arXiv:2507.07997*, 2025.
- Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.
- Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013.
- Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, et al. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*, 2023.
- Black Forest Labs. Flux, 2024. URL <https://github.com/black-forest-labs/flux>.
- Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11513–11522, 2022a.
- Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11523–11532, 2022b.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014a.
- Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014b. URL <https://api.semanticscholar.org/CorpusID:14113767>.
- Dongyang Liu, Shitian Zhao, Le Zhuo, Weifeng Lin, Yi Xin, Xinyue Li, Qi Qin, Yu Qiao, Hongsheng Li, and Peng Gao. Lumina-mgpt: Illuminate flexible photorealistic text-to-image generation with multimodal generative pretraining. *arXiv preprint arXiv:2408.02657*, 2024.
- Dongyang Liu, Shitian Zhao, Le Zhuo, Weifeng Lin, Yi Xin, Xinyue Li, Qi Qin, Yu Qiao, Hongsheng Li, and Peng Gao. Lumina-mgpt: Illuminate flexible photorealistic text-to-image generation with multimodal generative pretraining, 2025.
- Zhuoyan Luo, Fengyuan Shi, Yixiao Ge, Yujiu Yang, Limin Wang, and Ying Shan. Open-magvit2: An open-source project toward democratizing auto-regressive visual generation. *ArXiv*, abs/2409.04410, 2024.

- 
- Chuofan Ma, Yi Jiang, Junfeng Wu, Jihan Yang, Xin Yu, Zehuan Yuan, Bingyue Peng, and Xiaojuan Qi. Unitok: A unified tokenizer for visual generation and understanding. *ArXiv*, abs/2502.20321, 2025.
- William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.
- William S. Peebles and Saining Xie. Scalable diffusion models with transformers. *2023 IEEE/CVF International Conference on Computer Vision*, pp. 4172–4182, 2022. URL <https://api.semanticscholar.org/CorpusID:254854389>.
- Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.
- Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *ArXiv*, abs/2102.12092, 2021a.
- Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pp. 8821–8831. Pmlr, 2021b.
- Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, 2014.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022a.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022b.
- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211 – 252, 2014.
- Fengyuan Shi, Zhuoyan Luo, Yixiao Ge, Yujiu Yang, Ying Shan, and Limin Wang. Scalable image tokenization with index backpropagation quantization. *arXiv preprint arXiv:2412.02692*, 2024.
- Fengyuan Shi, Zhuoyan Luo, Yixiao Ge, Yujiu Yang, Ying Shan, and Limin Wang. Scalable image tokenization with index backpropagation quantization, 2025.
- Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *ArXiv*, abs/2406.06525, 2024.
- Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.
- Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *ArXiv*, abs/2404.02905, 2024.
- Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- Junke Wang, Yi Jiang, Zehuan Yuan, Binyue Peng, Zuxuan Wu, and Yu-Gang Jiang. Omnitokenizer: A joint image-video tokenizer for visual generation. *ArXiv*, abs/2406.09399, 2024a.

---

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024b.

Zhou Wang, Alan Conrad Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13: 600–612, 2004.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12966–12977, 2025a.

Junfeng Wu, Dongliang Luo, Weizhi Zhao, Zhihao Xie, Yuanhao Wang, Junyi Li, Xudong Xie, Yuliang Liu, and Xiang Bai. Tokbench: Evaluating your visual tokenizer before visual generation. *arXiv preprint arXiv:2505.18142*, 2025b.

Pingyu Wu, Kai Zhu, Yu Liu, Longxiang Tang, Jian Yang, Yansong Peng, Wei Zhai, Yang Cao, and Zheng-Jun Zha. Alitok: Towards sequence modeling alignment between tokenizer and autoregressive model. *arXiv preprint arXiv:2506.05289*, 2025c.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *ArXiv*, abs/2408.12528, 2024.

Tianwei Xiong, Jun Hao Liew, Zilong Huang, Jiashi Feng, and Xihui Liu. Gigatok: Scaling visual tokenizers to 3 billion parameters for autoregressive image generation, 2025.

Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. *ArXiv*, abs/2110.04627, 2021a.

Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021b.

Lijun Yu, José Lezama, Nitesh B. Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, Alexander G. Hauptmann, Boqing Gong, Ming-Hsuan Yang, Irfan Essa, David A. Ross, and Lu Jiang. Language model beats diffusion – tokenizer is key to visual generation, 2024a.

Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. An image is worth 32 tokens for reconstruction and generation. *Advances in Neural Information Processing Systems*, 37:128940–128966, 2024b.

Kaiwen Zha, Lijun Yu, Alireza Fathi, David A Ross, Cordelia Schmid, Dina Katabi, and Xiuye Gu. Language-guided image tokenization for generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15713–15722, 2025.

Borui Zhang, Qihang Rao, Wenzhao Zheng, Jie Zhou, and Jiwen Lu. Quantize-then-rectify: Efficient vq-vae training. *arXiv preprint arXiv:2507.10547*, 2025.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Yue Zhao, Yuanjun Xiong, and Philipp Krahenbuhl. Image and video tokenization with binary spherical quantization. *ArXiv*, abs/2406.07548, 2024a.

Yue Zhao, Yuanjun Xiong, and Philipp Krähenbühl. Image and video tokenization with binary spherical quantization. *arXiv preprint arXiv:2406.07548*, 2024b.

---

Yue Zhao, Fuzhao Xue, Scott Reed, Linxi Fan, Yuke Zhu, Jan Kautz, Zhiding Yu, Philipp Krahenbuhl, and De-An Huang. Qlip: Text-aligned visual tokenization unifies auto-regressive multi-modal understanding and generation. *ArXiv*, abs/2502.05178, 2025.

Anlin Zheng, Xin Wen, Xuanyang Zhang, Chuofan Ma, Tiancai Wang, Gang Yu, Xiangyu Zhang, and Xiaojuan Qi. Vision foundation models as effective visual tokenizers for autoregressive image generation. *arXiv preprint arXiv:2507.08441*, 2025.

Lei Zhu, Fangyun Wei, Yanye Lu, and Dong Chen. Scaling the codebook size of vq-gan to 100,000 with a utilization rate of 99%. *Advances in Neural Information Processing Systems*, 37:12612–12635, 2024.

Shaobin Zhuang, Kunchang Li, Xinyuan Chen, Yaohui Wang, Ziwei Liu, Yu Qiao, and Yali Wang. Vlogger: Make your dream a vlog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8806–8817, 2024.

Shaobin Zhuang, Zhipeng Huang, Ying Zhang, Fangyikang Wang, Canmiao Fu, Binxin Yang, Chong Sun, Chen Li, and Yali Wang. Video-gpt via next clip diffusion. *arXiv preprint arXiv:2505.12489*, 2025.

## A MORE IMPLEMENTATION DETAILS

### A.1 ABLATION STUDY

**Quantization method.** As shown in Tab. 6, 7 and 8.

Table 6: **GFQ training setting.** Table 7: **LFQ training setting.** Table 8: **BSQ training setting.**

config	GFQ	config	LFQ	config	BSQ
training data	IN-1K training set	training data	IN-1K training set	training data	IN-1K training set
image size	[256, 256]	image size	[256, 256]	image size	[256, 256]
data augmentation	random crop	data augmentation	random crop	data augmentation	random crop
downsample	16 × 16	downsample	16 × 16	downsample	16 × 16
ema	True	ema	True	ema	True
$g$ (group number)	2	$g$ (group number)	1	$g$ (group number)	16
$d'$ (group channel)	8	$d'$ (group channel)	16	$d'$ (group channel)	1
optimizer	Adam	optimizer	Adam	optimizer	Adam
optimizer momentum	$\beta_1, \beta_2=0.5, 0.9$	optimizer momentum	$\beta_1, \beta_2=0.5, 0.9$	optimizer momentum	$\beta_1, \beta_2=0.5, 0.9$
weight decay	0	weight decay	0	weight decay	0
learning rate schedule	consistent	learning rate schedule	consistent	learning rate schedule	consistent
learning rate	1e-4	learning rate	1e-4	learning rate	1e-4
warmup steps	0	warmup steps	0	warmup steps	0
cos decay end ratio	1	cos decay end ratio	1	cos decay end ratio	1
total steps	250250	total steps	250250	total steps	250250
channel_mult	[1,1,2,2,4]	channel_mult	[1,1,2,2,4]	channel_mult	[1,1,2,2,4]
channel	128	channel	128	channel	128
num_res_blocks	4	num_res_blocks	4	num_res_blocks	4
generative decoder	False	generative decoder	False	generative decoder	False
per GPU batchsize	16	per GPU batchsize	16	per GPU batchsize	16
global batchsize	128	global batchsize	128	global batchsize	128
GPU number	8 H20	GPU number	8 H20	GPU number	8 H20

**Generative decoder.** As shown in Tab. 9 and 10.

Table 9: **Stage-1 training setting.**

config	Stage-1
training data	general domain dataset
image size	[256, 256]
data augmentation	random crop
downsample	32 × 32
ema	True
$g$ (group number)	4
$d'$ (group channel)	8
optimizer	Adam
optimizer momentum	$\beta_1, \beta_2=0.5, 0.9$
weight decay	0
learning rate schedule	consistent
learning rate	1e-4
warmup steps	0
cos decay end ratio	1
total steps	550550
channel_mult	[1,1,2,2,4,8]
channel	256
num_res_blocks	4
generative decoder	False
per GPU batchsize	6
global batchsize	1056
GPU number	176 H20

Table 10: **Stage-2 training setting.**

config	Stage-2
training data	general domain dataset
image size	[256, 256]
data augmentation	random crop
downsample	32 × 32
ema	True
$g$ (group number)	4
$d'$ (group channel)	8
optimizer	Adam
optimizer momentum	$\beta_1, \beta_2=0.5, 0.9$
weight decay	0
learning rate schedule	consistent
learning rate	1e-4
warmup steps	0
cos decay end ratio	1
total steps	550550
channel_mult	[1,1,2,2,4,8]
channel	256
num_res_blocks	4
generative decoder	True
per GPU batchsize	6
global batchsize	1056
GPU number	176 H20

**Number of group in GQ.** As shown in Tab. 11, 12, 13, 14, 15 and 16.

**Model architecture.** As shown in Tab. 17, 18, 19, 20, 21, 22, 23, 24, and 25.

**Training data.** As shown in Tab. 26 and 27.

**Learning rate schedule.** As shown in Tab. 28.

**Parameter of autoregressive model.** As shown in Tab. 29, 30 and 31.

Table 11: **1 group training setting.**

config	1 group
training data	IN-1K training set
image size	[256, 256]
data augmentation	random crop
downsample	$16 \times 16$
ema	True
$g$ (group number)	1
$d'$ (group channel)	8
optimizer	Adam
optimizer momentum	$\beta_1, \beta_2=0.5, 0.9$
weight decay	0
learning rate schedule	consistent
learning rate	$1e-4$
warmup steps	0
cos decay end ratio	1
total steps	250250
channel_mult	[1,1,2,2,4]
channel	128
num_res_blocks	4
generative decoder	False
per GPU batchsize	16
global batchsize	128
GPU number	8 H20

Table 14: **8 group training setting.**

config	8 group
training data	IN-1K training set
image size	[256, 256]
data augmentation	random crop
downsample	$16 \times 16$
ema	True
$g$ (group number)	8
$d'$ (group channel)	8
optimizer	Adam
optimizer momentum	$\beta_1, \beta_2=0.5, 0.9$
weight decay	0
learning rate schedule	consistent
learning rate	$1e-4$
warmup steps	0
cos decay end ratio	1
total steps	250250
channel_mult	[1,1,2,2,4]
channel	128
num_res_blocks	4
generative decoder	False
per GPU batchsize	16
global batchsize	128
GPU number	8 H20

Table 12: **2 group training setting.**

config	2 group
training data	IN-1K training set
image size	[256, 256]
data augmentation	random crop
downsample	$16 \times 16$
ema	True
$g$ (group number)	2
$d'$ (group channel)	8
optimizer	Adam
optimizer momentum	$\beta_1, \beta_2=0.5, 0.9$
weight decay	0
learning rate schedule	consistent
learning rate	$1e-4$
warmup steps	0
cos decay end ratio	1
total steps	250250
channel_mult	[1,1,2,2,4]
channel	128
num_res_blocks	4
generative decoder	False
per GPU batchsize	16
global batchsize	128
GPU number	8 H20

Table 15: **16 group training setting.**

config	16 group
training data	IN-1K training set
image size	[256, 256]
data augmentation	random crop
downsample	$16 \times 16$
ema	True
$g$ (group number)	16
$d'$ (group channel)	8
optimizer	Adam
optimizer momentum	$\beta_1, \beta_2=0.5, 0.9$
weight decay	0
learning rate schedule	consistent
learning rate	$1e-4$
warmup steps	0
cos decay end ratio	1
total steps	250250
channel_mult	[1,1,2,2,4]
channel	128
num_res_blocks	4
generative decoder	False
per GPU batchsize	16
global batchsize	128
GPU number	8 H20

Table 13: **2 group training setting.**

config	4 group
training data	IN-1K training set
image size	[256, 256]
data augmentation	random crop
downsample	$16 \times 16$
ema	True
$g$ (group number)	4
$d'$ (group channel)	8
optimizer	Adam
optimizer momentum	$\beta_1, \beta_2=0.5, 0.9$
weight decay	0
learning rate schedule	consistent
learning rate	$1e-4$
warmup steps	0
cos decay end ratio	1
total steps	250250
channel_mult	[1,1,2,2,4]
channel	128
num_res_blocks	4
generative decoder	False
per GPU batchsize	16
global batchsize	128
GPU number	8 H20

Table 16: **32 group training setting.**

config	32 group
training data	IN-1K training set
image size	[256, 256]
data augmentation	random crop
downsample	$16 \times 16$
ema	True
$g$ (group number)	32
$d'$ (group channel)	8
optimizer	Adam
optimizer momentum	$\beta_1, \beta_2=0.5, 0.9$
weight decay	0
learning rate schedule	consistent
learning rate	$1e-4$
warmup steps	0
cos decay end ratio	1
total steps	250250
channel_mult	[1,1,2,2,4]
channel	128
num_res_blocks	4
generative decoder	False
per GPU batchsize	16
global batchsize	128
GPU number	8 H20

Table 17: **128 channel 4 block training setting.**

config	128 channel 4 block
training data	IN-1K training set
image size	[256, 256]
data augmentation	random crop
downsample	$16 \times 16$
ema	True
$g$ (group number)	4
$d'$ (group channel)	8
optimizer	Adam
optimizer momentum	$\beta_1, \beta_2=0.5, 0.9$
weight decay	0
learning rate schedule	consistent
learning rate	1e-4
warmup steps	0
cos decay end ratio	1
total steps	250250
channel_mult	[1,1,2,2,4]
channel	128
num_res_blocks	4
generative decoder	False
per GPU batchsize	16
global batchsize	128
GPU number	8 H20

Table 20: **128 channel 8 block training setting.**

config	128 channel 8 block
training data	IN-1K training set
image size	[256, 256]
data augmentation	random crop
downsample	$16 \times 16$
ema	True
$g$ (group number)	4
$d'$ (group channel)	8
optimizer	Adam
optimizer momentum	$\beta_1, \beta_2=0.5, 0.9$
weight decay	0
learning rate schedule	consistent
learning rate	1e-4
warmup steps	0
cos decay end ratio	1
total steps	250250
channel_mult	[1,1,2,2,4]
channel	128
num_res_blocks	8
generative decoder	False
per GPU batchsize	8
global batchsize	128
GPU number	16 H20

Table 23: **256 channel 2 block training setting.**

config	256 channel 2 block
training data	IN-1K training set
image size	[256, 256]
data augmentation	random crop
downsample	$16 \times 16$
ema	True
$g$ (group number)	4
$d'$ (group channel)	8
optimizer	Adam
optimizer momentum	$\beta_1, \beta_2=0.5, 0.9$
weight decay	0
learning rate schedule	consistent
learning rate	1e-4
warmup steps	0
cos decay end ratio	1
total steps	250250
channel_mult	[1,1,2,2,4]
channel	256
num_res_blocks	2
generative decoder	False
per GPU batchsize	8
global batchsize	128
GPU number	16 H20

Table 18: **192 channel 4 block training setting.**

config	192 channel 4 block
training data	IN-1K training set
image size	[256, 256]
data augmentation	random crop
downsample	$16 \times 16$
ema	True
$g$ (group number)	4
$d'$ (group channel)	8
optimizer	Adam
optimizer momentum	$\beta_1, \beta_2=0.5, 0.9$
weight decay	0
learning rate schedule	consistent
learning rate	1e-4
warmup steps	0
cos decay end ratio	1
total steps	250250
channel_mult	[1,1,2,2,4]
channel	192
num_res_blocks	4
generative decoder	False
per GPU batchsize	8
global batchsize	128
GPU number	16 H20

Table 21: **128 channel 16 block training setting.**

config	128 channel 16 block
training data	IN-1K training set
image size	[256, 256]
data augmentation	random crop
downsample	$16 \times 16$
ema	True
$g$ (group number)	4
$d'$ (group channel)	8
optimizer	Adam
optimizer momentum	$\beta_1, \beta_2=0.5, 0.9$
weight decay	0
learning rate schedule	consistent
learning rate	1e-4
warmup steps	0
cos decay end ratio	1
total steps	250250
channel_mult	[1,1,2,2,4]
channel	128
num_res_blocks	16
generative decoder	False
per GPU batchsize	4
global batchsize	128
GPU number	32 H20

Table 22: **192 channel 8 block training setting.**

Table 19: **256 channel 4 block training setting.**

config	256 channel 4 block
training data	IN-1K training set
image size	[256, 256]
data augmentation	random crop
downsample	$16 \times 16$
ema	True
$g$ (group number)	4
$d'$ (group channel)	8
optimizer	Adam
optimizer momentum	$\beta_1, \beta_2=0.5, 0.9$
weight decay	0
learning rate schedule	consistent
learning rate	1e-4
warmup steps	0
cos decay end ratio	1
total steps	250250
channel_mult	[1,1,2,2,4]
channel	256
num_res_blocks	4
generative decoder	False
per GPU batchsize	4
global batchsize	128
GPU number	32 H20

Table 23: **192 channel 8 block training setting.**

Table 24: **384 channel 4 block training setting.**

config	256 channel 8 block
training data	IN-1K training set
image size	[256, 256]
data augmentation	random crop
downsample	$16 \times 16$
ema	True
$g$ (group number)	4
$d'$ (group channel)	8
optimizer	Adam
optimizer momentum	$\beta_1, \beta_2=0.5, 0.9$
weight decay	0
learning rate schedule	consistent
learning rate	1e-4
warmup steps	0
cos decay end ratio	1
total steps	250250
channel_mult	[1,1,2,2,4]
channel	256
num_res_blocks	8
generative decoder	False
per GPU batchsize	2
global batchsize	128
GPU number	64 H20

Table 26: **ImageNet training setting.**

config	1.2M
training data	IN-1K training set
image size	[256, 256]
data augmentation	random crop
downsample	$16 \times 16$
ema	True
$g$ (group number)	4
$d'$ (group channel)	8
optimizer	Adam
optimizer momentum	$\beta_1, \beta_2=0.5, 0.9$
weight decay	0
learning rate schedule	consistent
learning rate	1e-4
warmup steps	0
cos decay end ratio	1
total steps	550550
channel_mult	[1,1,2,2,4]
channel	128
num_res_blocks	4
generative decoder	False
per GPU batchsize	16
global batchsize	128
GPU number	8 H20

Table 27: **General-domain training setting.**

config	400M
training data	general domain dataset
image size	[256, 256]
data augmentation	random crop
downsample	$16 \times 16$
ema	True
$g$ (group number)	4
$d'$ (group channel)	8
optimizer	Adam
optimizer momentum	$\beta_1, \beta_2=0.5, 0.9$
weight decay	0
learning rate schedule	consistent
learning rate	1e-4
warmup steps	0
cos decay end ratio	1
total steps	550550
channel_mult	[1,1,2,2,4]
channel	128
num_res_blocks	4
generative decoder	False
per GPU batchsize	16
global batchsize	128
GPU number	8 H20

Table 28: **Warm up + cosine decay learning rate schedule training setting.**

config	warm up + cosine decay
training data	general domain dataset
image size	[256, 256]
data augmentation	random crop
downsample	$16 \times 16$
ema	True
$g$ (group number)	4
$d'$ (group channel)	8
optimizer	Adam
optimizer momentum	$\beta_1, \beta_2=0.5, 0.9$
weight decay	0
learning rate schedule	warm up + cosine decay
learning rate	1e-4
warmup steps	10000
cos decay end ratio	0.01
total steps	550550
channel_mult	[1,1,2,2,4]
channel	128
num_res_blocks	4
generative decoder	False
per GPU batchsize	16
global batchsize	128
GPU number	8 H20

Table 29: **LlamaGen Base** training setting.

config	LlamaGen Base
training data	IN-1K training set
image size	[256, 256]
data augmentation	random crop
downsample	16 × 16
ema	False
optimizer	AdamW
optimizer momentum	$\beta_1, \beta_2=0.9, 0.95$
weight decay	5e-2
learning rate schedule	warm up + linear decay
learning rate	3e-4
warmup epochs	6
linear decay end ratio	0.1
total epochs	1000
dim	1024
num_head	16
trans_layers	24
cond_dim	1024
factorized_layers	2
factorized_k	4
token_drop	0.1
residual_drop	0.1
per GPU batchsize	64
global batchsize	3072
GPU number	48 H20

Table 30: **LlamaGen Large** training setting.

config	LlamaGen Large
training data	IN-1K training set
image size	[256, 256]
data augmentation	random crop
downsample	16 × 16
ema	False
optimizer	AdamW
optimizer momentum	$\beta_1, \beta_2=0.9, 0.95$
weight decay	5e-2
learning rate schedule	warm up + linear decay
learning rate	3e-4
warmup epochs	6
linear decay end ratio	0.1
total epochs	1000
dim	1280
num_head	20
trans_layers	36
cond_dim	1280
factorized_layers	3
factorized_k	4
token_drop	0.1
residual_drop	0.1
per GPU batchsize	32
global batchsize	3072
GPU number	96 H20

Table 31: **LlamaGen X-Large** training setting.

config	LlamaGen X-Large
training data	IN-1K training set
image size	[256, 256]
data augmentation	random crop
downsample	16 × 16
ema	False
optimizer	AdamW
optimizer momentum	$\beta_1, \beta_2=0.9, 0.95$
weight decay	5e-2
learning rate schedule	warm up + linear decay
learning rate	3e-4
warmup epochs	6
linear decay end ratio	0.1
total epochs	1000
dim	1536
num_head	24
trans_layers	48
cond_dim	1536
factorized_layers	4
factorized_k	4
token_drop	0.1
residual_drop	0.1
per GPU batchsize	16
global batchsize	3072
GPU number	192 H20

## A.2 COMPARISON WITH STATE-OF-THE-ART

**Visual Reconstruction.** The settings of in-distribution comparison are shown in Tab. 32 and 33. The settings of general-domain comparison are shown in Tab. 34, 35, 36 and 37, .

**Visual Generation.** As shown in Tab. 31.

Table 32: Large scale training on ImageNet training set at 16 downsample ratio.

config	IN-1K 16× SOTA
training data	IN-1K training set
image size	[256, 256]
data augmentation	random crop
downsample	16 × 16
ema	True
$g$ (group number)	4
$d'$ (group channel)	8
optimizer	Adam
optimizer momentum	$\beta_1, \beta_2=0.5, 0.9$
weight decay	0
learning rate schedule	consistent
learning rate	1e-4
warmup steps	0
cos decay end ratio	1
total steps	400400
channel_mult	[1,1,2,2,4]
channel	256
num_res_blocks	4
generative decoder	False
per GPU batchsize	8
global batchsize	1024
GPU number	128 H20

Table 33: Large scale training on ImageNet training set at 8 downsample ratio.

config	IN-1K 8× SOTA
training data	IN-1K training set
image size	[256, 256]
data augmentation	random crop
downsample	8 × 8
ema	True
$g$ (group number)	4
$d'$ (group channel)	8
optimizer	Adam
optimizer momentum	$\beta_1, \beta_2=0.5, 0.9$
weight decay	0
learning rate schedule	consistent
learning rate	1e-4
warmup steps	0
cos decay end ratio	1
total steps	350350
channel_mult	[1,2,2,4]
channel	256
num_res_blocks	4
generative decoder	False
per GPU batchsize	8
global batchsize	1024
GPU number	128 H20

**Table 34: Large scale training on general-domain dataset at 768 compression ratio.**

config	768 compression ratio SOTA		
training data	general domain dataset		
image size	[256, 256]		
data augmentation	random crop		
downsample	$32 \times 32$		
ema	True		
$g$ (group number)	4		
$d'$ (group channel)	8		
optimizer	Adam		
optimizer momentum	$\beta_1, \beta_2=0.5, 0.9$		
weight decay	0		
learning rate schedule	consistent		
learning rate	1e-4	1e-4	1e-5
warmup steps	0		
cos decay end ratio	1		
total steps	550550	550550	330330
channel_mult	[1,1,2,2,4,8]		
channel	256		
num_res_blocks	4		
generative decoder	False	True	True
per GPU batchsize	6		
global batchsize	1056		
GPU number	176 H20		

**Table 36: Large scale training on general-domain dataset at 48 compression ratio.**

config	48 compression ratio SOTA		
training data	general domain dataset		
image size	[256, 256]		
data augmentation	random crop		
downsample	$8 \times 8$		
ema	True		
$g$ (group number)	4		
$d'$ (group channel)	8		
optimizer	Adam		
optimizer momentum	$\beta_1, \beta_2=0.5, 0.9$		
weight decay	0		
learning rate schedule	consistent		
learning rate	1e-4	1e-5	1e-6
warmup steps	0		
cos decay end ratio	1		
total steps	200200	200200	20020
channel_mult	[1,2,2,4]		
channel	256		
num_res_blocks	4		
generative decoder	False	True	True
per GPU batchsize	8		
global batchsize	1024		
GPU number	128 H20		

**Table 35: Large scale training on general-domain dataset at 192 compression ratio.**

config	192 compression ratio SOTA		
training data	general domain dataset		
image size	[256, 256]		
data augmentation	random crop		
downsample	$16 \times 16$		
ema	True		
$g$ (group number)	4		
$d'$ (group channel)	8		
optimizer	Adam		
optimizer momentum	$\beta_1, \beta_2=0.5, 0.9$		
weight decay	0		
learning rate schedule	consistent		
learning rate	1e-4	1e-5	1e-6
warmup steps	0		
cos decay end ratio	1		
total steps	470470	90090	10010
channel_mult	[1,1,2,2,4]		
channel	256		
num_res_blocks	4		
generative decoder	False	False	False
per GPU batchsize	8		
global batchsize	1024		
GPU number	128 H20		

**Table 37: Large scale training on general-domain dataset at 24 compression ratio.**

config	24 compression ratio SOTA		
training data	general domain dataset		
image size	[256, 256]		
data augmentation	random crop		
downsample	$8 \times 8$		
ema	True		
$g$ (group number)	8		
$d'$ (group channel)	8		
optimizer	Adam		
optimizer momentum	$\beta_1, \beta_2=0.5, 0.9$		
weight decay	0		
learning rate schedule	consistent		
learning rate	1e-4	1e-5	1e-5
warmup steps	0		
cos decay end ratio	1		
total steps	300300	50050	400400
channel_mult	[1,2,2,4]		
channel	256		
num_res_blocks	4		
generative decoder	False	True	True
per GPU batchsize	8		
global batchsize	1024		
GPU number	128 H20	128 H20	512 H20