

MGVQ: Could VQ-VAE Beat VAE? A Generalizable Tokenizer with Multi-group Quantization

Mingkai Jia^{1,2} Wei Yin^{2*} Xiaotao Hu^{1,2} Jixin Guo³ Xiaoyang Guo²
 Qian Zhang² Xiao-Xiao Long⁴ Ping Tan¹

¹The Hong Kong University of Science and Technology ²Horizon Robotics
³The Chinese University of Hong Kong ⁴Nanjing University

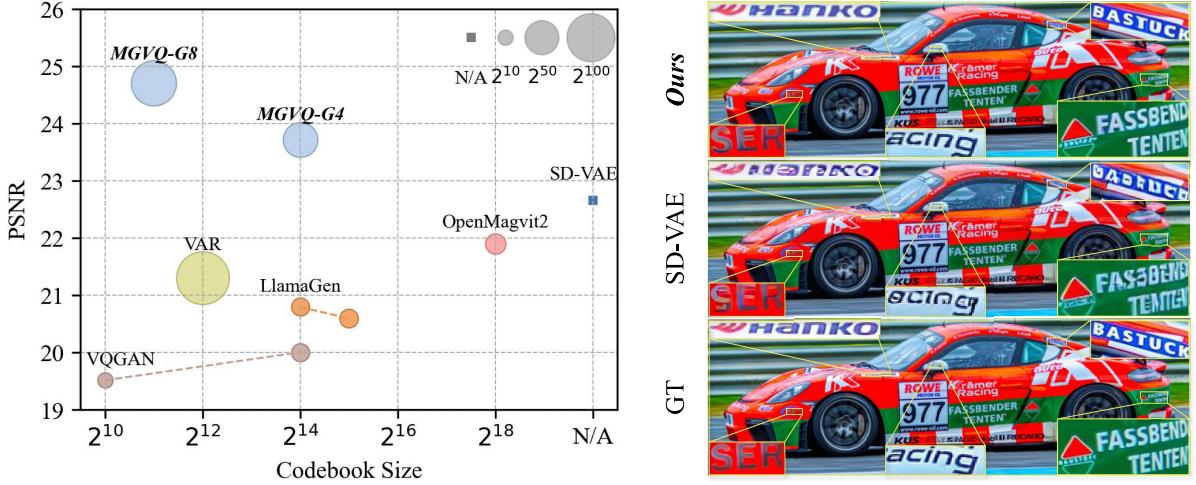


Figure 1. **Reconstruction performance comparison between VQ-VAEs and SD-VAE on ImageNet 256×256 benchmark with 16× downsampling.** The size of bubbles indicates the capacity, that is, the possibility of a token by sampling from the codebook. **MGVQ-G8** with 8 groups achieves a PSNR of 24.70, evidently surpassing all others, with a large capacity of 2^{88} . Qualitative results are illustrated where details are zoomed in for a better view.

Abstract

Vector Quantized Variational Autoencoders (VQ-VAEs) are fundamental models that compress continuous visual data into discrete tokens. Existing methods have tried to improve the quantization strategy for better reconstruction quality, however, there still exists a large gap between VQ-VAEs and VAEs. To narrow this gap, we propose MGVQ, a novel method to augment the representation capability of discrete codebooks, facilitating easier optimization for codebooks and minimizing information loss, thereby enhancing reconstruction quality. Specifically, we propose to retain the latent dimension to preserve encoded features and incorporate a set of sub-codebooks for quantization. Furthermore, we construct comprehensive zero-shot benchmarks featuring resolutions of 512p and 2k to evaluate the reconstruction perfor-

mance of existing methods rigorously. MGVQ achieves the state-of-the-art performance on both ImageNet and 8 zero-shot benchmarks across all VQ-VAEs. Notably, compared with SD-VAE, we outperform them on ImageNet significantly, with **rFID 0.49 v.s. 0.91**, and achieve superior PSNR on all zero-shot benchmarks. These results highlight the superiority of MGVQ in reconstruction and pave the way for preserving fidelity in HD image processing tasks. Code will be publicly available at <https://github.com/MKJia/MGVQ>.

1. Introduction

Vector Quantization (VQ) [13] is a fundamental technique in computer vision that transforms continuous visual data into a set of discrete set of tokens. By mapping high-dimensional data to a finite collection of learned codebook embeddings, VQ creates a compact and efficient latent space. This discrete

*Corresponding author.

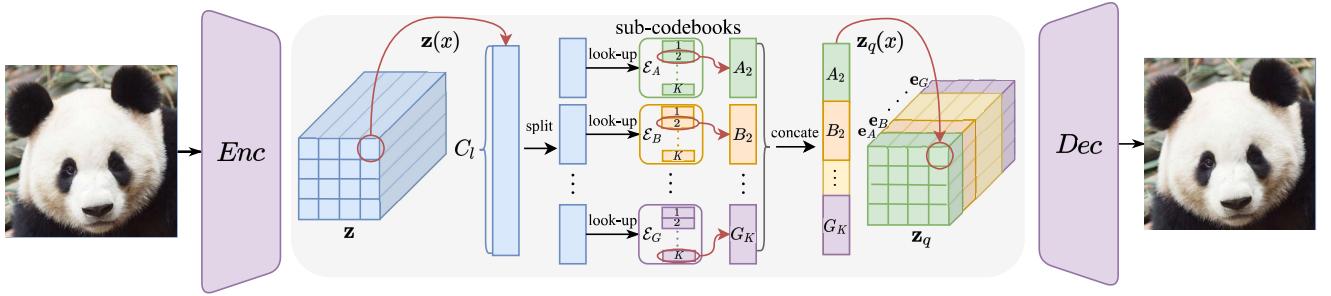


Figure 2. **An overview of *MGVQ* framework.** *MGVQ* keeps a larger dimension C_l of latent \mathbf{z} and split it into G sub-tokens, where each sub-token is quantized individually with sub-codebook \mathcal{E}_i . Sub-tokens are then combined to compose \mathbf{z}_q and for decoding.

representation not only reduces computational complexity but also retains critical semantic information, enabling more efficient data processing and analysis. As a key component in modern vision systems, VQ facilitates tasks such as image compression [1, 2, 41, 52, 55] and feature extraction [26, 28, 35, 57], providing a robust foundation for a wide range of applications [25, 56, 62, 67].

Despite recent advances, current tokenizers still suffer from significant information loss due to the limited representation capacity of codebooks. Consequently, Vector Quantized Variational Autoencoders (VQ-VAEs) [3, 10, 32, 34, 46, 50, 58] tend to exhibit inferior encoding-decoding performance compared to Variational Autoencoders (VAEs) [19], often resulting in decoded images with substantial errors. Simply increasing the number of feature dimensions and enlarging the codebook size typically leads to codebook collapse [47], where many code vectors remain inactive (receiving zero gradients during backpropagation). Thus, merely expanding codebook capacity fails to alleviate reconstruction errors.

Existing methods have explored various strategies to tackle these challenges. For instance, LlamaGen [46] and Emu3 [51] reduce the latent dimension to 8 and 4, respectively, to mitigate codebook collapse [47] while expanding the codebook size to 16384 and 32768. Other approaches, such as MAGVIT-v2 [59], introduce a lookup-free quantization (LFQ) strategy, reducing the code embedding dimension to zero and achieves an extra-large codebook size of 262144. While reducing the latent dimension enhanced the discrete latent space representation capability, it also results in information loss from the reduced dimension, ultimately degrading performance. These trade-offs highlight the need for a more effective approach to expanding representation capacity while preserving information of encoded features and preventing codebook collapse.

To address these issues, we introduce *MGVQ*, a novel method to enhance the representation capability of discrete codebooks. Unlike other methods that aggressively reduce dimensionality [46, 58], we preserve the latent dimension to maintain more expressive features while avoiding informa-

tion loss. Specifically, we split the latent embeddings into sub-tokens, each mapped to an independent sub-codebook, facilitating easier optimization and larger representation capacity. Additionally, we employ a nested masking strategy during training to enforce the encoder to compress the image into the latent in an ordered manner. By exponentially expanding the discrete latent space and reducing information loss, our *MGVQ* achieves higher fidelity and generalization ability across diverse benchmarks. Our experiments demonstrate that *MGVQ* outperforms existing VQ-VAE variants and significantly narrows the gap with VAE-based methods. Notably, *MGVQ* establishes new state-of-the-art results on ImageNet 256p and multiple zero-shot HD datasets, providing a promising direction for advancing the reconstruction performance of VQ-VAE models.

We summarize our contributions as follows.

1. We propose *MGVQ*, a novel method to enlarge the representation capability of discrete codebooks by over a billion-fold, aiming to narrow the reconstruction gap between VAE and VQ-VAE.
2. We preserve the latent dimension and propose to use a group of sub-codebooks with a nested masking training strategy for quantization instead of a single unified one, achieving easier optimization for codebooks, larger representation capacity, and reduced information loss.
3. We construct a zero-shot benchmark of 512p and 2k resolution to comprehensively evaluate the generalization ability for reconstruction. *MGVQ* realizes SOTA performance on both ImageNet 256p benchmark and 7 zero-shot HD datasets. Note that *MGVQ* achieves rFID of 0.49, significantly outperforming SD-VAE [43] of 0.91.

2. Related Works

2.1. Variational Autoencoder

VAEs [19] have been instrumental in enhancing the performance of diffusion models, particularly in the context of image generation tasks. Recent advancements [7, 14, 40, 43, 63, 65, 68] have significantly improved reconstruction quality and efficiency in these models. One of the popular VAEs

is the SD-VAE [43] model, fine-tuning the VAE decoder and trained on a combined dataset of billions of data for notable performance enhancements. Building upon the foundations of SD-VAE [43], SDXL-VAE [40] introduces several architectural and training improvements. By employing larger batch sizes and incorporating exponential moving average weight tracking, it achieves superior performance in various reconstruction metrics compared to its predecessors.

2.2. Vector Quantized Variational Autoencoder

VQ-VAEs extend traditional VAEs by discretizing latent representations into tokens from a learned codebook [50]. Based on VQ-VAE [50], VQGAN [10] further use a discriminator to keep perceptual quality at increased compression rate. Following the framework of VQ-VAE and VQGAN, several methods attempt to enhance the performance in many ways. Some existing methods focus on enhancing the representation capability of discrete codes. ViT-VQGAN [58] discovers that reducing dimension of latent space from 256 to 32 consistently improves reconstruction quality. Then LlamaGen [46] further reduces dimension from 32 to 8, with an enlarged codebook of 16384 that benefits the overall performance. However, they figure out that continuing to increase the codebook size (e.g. 32768) may lead to a degradation of performance and lower usage of discrete codes. This shows the necessity to improve the usage of codes. To deal with this issue, SQ-VAE [47] applies self-annealed stochastic quantization for greater codebook utilization. CVQ-VAE [66] proposes a clustering VQ-VAE that selects encoded features as anchors to update unused codes. SimVQ [69] reparameterizes the codes based on a learnable latent basis to optimize the entire linear space. However, a larger representation capacity remains a pressing need. Recent works turn to different types of quantization to a larger capacity of the discrete latent space. FSQ [37] replace VQ with a simple scheme termed finite scalar quantization, and MAGVIT-v2 [59] proposes a similar lookup-free quantization (LFQ) strategy and achieves a super-large codebook size of 262144 by reducing the dimension of code embedding to zero. Nevertheless, they neglect the information loss with dimension downsampling, which is harmful for retaining the reconstruction quality.

3. Method

3.1. Preliminary

VAE. VAEs [19] are powerful deep generative models widely used to represent complex input data through a smaller latent space learned in an unsupervised manner [12]. A VAE consists of an encoder that parameterizes input data $\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$ to a latent variable $\mathbf{z} \in \mathbb{R}^{\frac{H}{D} \times \frac{W}{D} \times C_l}$, where $H \times W$ are input spatial sizes, D is the downsample scale, and C_l is the latent dimension. A continuous latent \mathbf{z}_s is sampled from the posterior distribution $q(\mathbf{z}|\mathbf{x})$ with a given

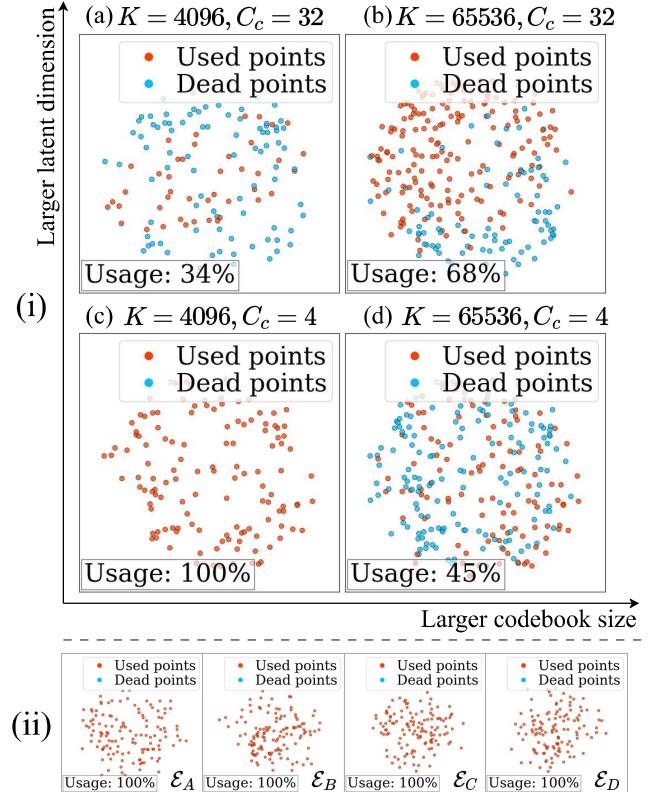


Figure 3. (i) Codebook points of well-trained VQ-VAE models, that have different codebook sizes or latent dimensions. (ii) Sub-codebook points in our proposed MGVQ . The group size is 4. Used points are shown in red, while dead points shown in blue. (i.a) A larger dimension and smaller size may lead to an anisotropic distribution and low usage. (i.b) A larger dimension and larger size show favor of certain directions, resulting in dead points in a specific area (lower right corner). (i.c) A smaller dimension and smaller size could be fully used without dead points. (i.d) A smaller dimension and larger size allow for a uniform spread, but there are more codes than needed, leading to some underutilization.

prior distribution $q(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I})$, and VAE's decoder $p_{\theta}(\mathbf{x}|\mathbf{z})$ with parameters θ reconstructs the data [50]. The posterior distribution of a specific input data \mathbf{x} can be approximated as a Gaussian distribution with approximately diagonal covariance matrix [19]:

$$\log q_{\phi}(\mathbf{z}|\mathbf{x}) = \log \mathcal{N}(\mathbf{z}|\mu, \sigma^2 \mathbf{I}), \quad (1)$$

where ϕ denotes variational parameters within the encoder, and μ, σ are the mean and standard deviation of the approximate posterior. The optimization objective is the evidence lower bound (ELBO), derived from Jensen's inequality [20]:

$$\mathcal{L}_{\theta, \phi}(\mathbf{x}) = \mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x})} [\log p_{\theta}(\mathbf{x}, \mathbf{z}) - \log q_{\phi}(\mathbf{z}|\mathbf{x})]. \quad (2)$$

VQ-VAE. Oord *et al.* [50] reveals that training VAE easily faces posterior collapse and large variance, thus the Vector Quantized Variational Autoencoder (VQ-VAE) is proposed. Instead of sampling from the posterior distribution,

it proposes to encode the inputs into discrete latents and remove the KL term. They set up a learnable codebook $\mathcal{E} = \{\mathbf{e}_i\}_{i=1}^K \subset \mathbb{R}^{K \times C_c}$, where K is the number of learnable vectors, and C_c is the vector's dimension [10]. They employ a channel downsample to C_c first and then use the nearest neighbor lookup method to map the encoder's latent $\mathbf{z} \in \mathbb{R}^{\frac{H}{n} \times \frac{W}{n} \times C_c}$ to the quantized latent $\mathbf{z}_q \in \mathbb{R}^{\frac{H}{n} \times \frac{W}{n} \times C_c}$:

$$\mathbf{z}_q(x) = \mathbf{e}_k, \quad \text{where } k = \arg \min_{\mathbf{e}_i \in \mathcal{E}} \|\mathbf{z}(x) - \mathbf{e}_i\|_2. \quad (3)$$

where x represents a token. The quantized latent \mathbf{z}_q is fed to the decoder for reconstruction. Note that in their implementations [46, 51], they usually employ a small quantization dimension C_c to avoid codebook collapse during training.

3.2. What limits the reconstruction quality of VQ-VAE?

Information loss. Recent research on VAE [6, 11, 15] reveals that a larger latent space dimension (i.e. C_l) significantly boosts reconstruction quality. For example, SD3 [11] increases the latent space channel number from 4 (SD1 [43]) to 16, while DC-AE [6] employs up to 128 channels in their implementations. However, training codebook discretization easily face codebook collapse [47, 66], where codes do not receive gradients, resulting in unoptimized during training and unused in inference, also known as dead points. To avoid this, VQ-VAE employs a much lower latent dimension C_c . However, this may introduces noticeable information loss, leading to inferior reconstruction quality. For example, LlamaGen [46] reduces the latent channel from 256 to 8 for quantization and Emu3 [8] use 4 quantization channels.

Limited latent representation capacity. The representation capacity of a VQ-VAE denotes the sampling possibility from its codebooks. For example, LlamaGen's representation capacity can be denoted as 16384 with a codebook size of 16384. A larger representation capacity means a stronger reconstruction ability. VQ-VAEs utilize a discrete latent space represented by a learned codebook \mathcal{E} with limited vectors, while VAE encodes images into the continuous latent variables \mathbf{z} , resulting in a continuous latent space. Compared with VAE, such a finite codebook lags VQ-VAE's representation capacity, resulting in worse reconstruction results, especially for texture-complex images.

This brings up the question: *Can we simultaneously improve the codebook size K and the quantized latent dimension C_c to boost VQ-VAE's performance?* Unfortunately, we observe that there is an inherent conflict between the latent dimension and enlarging the codebook size during training. Fig. 3 illustrates an experiment of code points. We trained 4 models with different codebook sizes or latent dimensions and showed their codebook points. Compared with the baseline (Fig. 3 (c)), Fig. 3 (a)(b)(d) present that either enlarging the codebook size or increasing latent dimension size will cause significant dead codes, i.e. points not optimized in

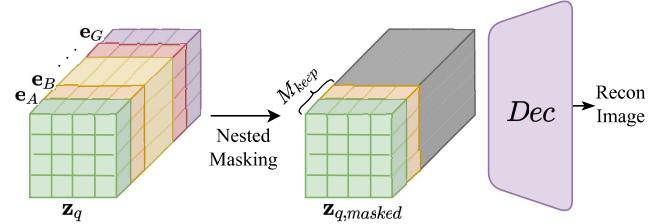


Figure 4. **The process of nested masking.** Gray blocks represents the masked last tokens, and other colors show active sub-groups.

training and unused in inference. In contrast, the codebook usage of the baseline (Fig. 3 (c)) is much higher.

3.3. MGVQ

To tackle the aforementioned conflicts and narrow the reconstruction gap between VAE and VQ-VAE, we propose *MGVQ*, a novel method to enlarge the representation capability of discrete codebooks. To enhance the VQ-VAE performance, our insight is that the key point is increasing the representation capacity instead of the codebook size. Therefore, we attempt to enlarge the representation capacity but maintain a small codebook size for easy optimization. Fig. 2 presents our pipeline. Following VQGAN [10], we employ a convolution-based backbone to extract features $\mathbf{z} \in \mathbb{R}^{\frac{H}{D} \times \frac{W}{D} \times C_l}$. In contrast to existing methods reducing the latent channel to a small one before quantization, we use a larger latent dimension C_l than them to retain as much encoded feature as possible. This can avoid information loss. Instead of using a single unified codebook, we propose to use a group of sub-codebooks for quantization. Specially, for each latent embedding $\mathbf{z}(x) \in \mathbb{R}^{C_l}$, we uniformly split it to G sub-tokens $\{\mathbf{z}(x_i)\}_{i=1}^G$, where each sub-token has the same dimension $\frac{C_l}{G}$. Correspondingly, we set up G non-shared sub-codebooks, each with the same size K . Each sub-token is quantized individually using its associated sub-codebook, $\mathcal{E}_i = \{\mathbf{e}_i^j\}_{j=1}^K \subset \mathbb{R}^{K \times \frac{C_l}{G}}$. Finally, we concatenate all quantized sub-tokens for reconstruction,

$$\mathbf{z}_q(x) = \text{concat}(\mathbf{e}_1^{k_1}, \dots, \mathbf{e}_G^{k_G}), \\ \text{where } k_i = \arg \min_{\mathbf{e}_i^k \in \mathcal{E}_i} \|\mathbf{z}(x_i) - \mathbf{e}_i^k\|_2, \quad (4)$$

where i is the index of sub-codebook, k_i is the lookup index from the codebook.

In our final models, for example, the latent dimension C_l is set to 32, the group size G is set to 4, and the total codebook size is set to 32768 at most. Thus, the codebook dimension is very small, i.e. 8, and sub-codebook size (8192) is much smaller than existing methods, which alleviates codebook collapse, and representation capacity of 8192⁴ is much larger. With this divide-and-conquer approach, we can achieve the following abilities:

- Easier optimization for codebooks. With reduced dimension of each sub-codebook, we can avoid the issue of dead

Method	Quantization	Ratio	Dim	Size	Capacity	Codebook Usage	ImageNet		
							rFID↓	PSNR↑	SSIM↑
SD-VAE ^{ukn.} [43]	-	16	16	-	-	-	0.91	22.65	0.697
VQGAN [10]	VQ	16	256	1024	$1024^1 = 2^{10}$	-	8.30	19.51	0.614
VQGAN [10]	VQ	16	256	16384	$16384^1 = 2^{14}$	-	4.99	20.00	0.629
LlamaGen [46]	VQ	16	8	16384	$16384^1 = 2^{14}$	97%	2.19	20.79	0.675
LlamaGen [46]	VQ	16	8	32768	$32768^1 = 2^{15}$	85%	2.26	20.59	0.663
OpenMagvit2 [32]	LFQ	16	18	262144	$262144^1 = 2^{18}$	100%	1.17	21.90	-
VAR [48]	Multi-scale VQ	16	32	4096	$4096^{10} = 2^{120}$	-	-	21.30	0.647
MGVQ-G4	VQ	16	32	8192×4	$8192^4 = 2^{52}$	100%	<u>0.64</u>	<u>23.71</u>	<u>0.755</u>
MGVQ-G8	VQ	16	32	2048×8	$2048^8 = 2^{88}$	100%	0.49	24.70	0.787
SD-VAE ^{ukn.} [43]	-	8	4	-	-	-	0.74	25.68	0.820
SDXL-VAE ^{ukn.} [40]	-	8	4	-	-	-	0.68	26.04	0.834
VQGAN ^{oim.} [10]	VQ	8	4	256	$256^1 = 2^8$	-	1.44	22.63	0.737
VQGAN ^{oim.} [10]	VQ	8	4	16384	$16384^1 = 2^{14}$	-	1.19	23.38	0.762
LlamaGen [46]	VQ	8	8	16384	$16384^1 = 2^{14}$	-	0.59	24.45	0.813
OpenMagvit2 [32]	LFQ	8	18	262144	$262144^1 = 2^{18}$	100%	0.34	27.02	-
MGVQ-G4	VQ	8	32	8192×4	$8192^4 = 2^{52}$	100%	<u>0.31</u>	<u>28.42</u>	<u>0.898</u>
MGVQ-G8	VQ	8	32	2048×8	$2048^8 = 2^{88}$	100%	0.27	29.96	0.918

Table 1. **Reconstruction evaluation on 256×256 ImageNet benchmark.** MGVQ-G4 indicates our method with 4 sub-codebooks, and MGVQ-G8 indicates 8 sub-codebooks. All models are trained on ImageNet except “oim.” is on OpenImage, “ukn.” is unknown training data. Gray denotes continuous tokenizers.

Method	CelebA		DAVIS		TextOCR		VFHQ		Spring		ENeRF		Ranking	
	rFID↓	PSNR↑	rFID↓	PSNR↑	rFID↓	PSNR↑	rFID↓	PSNR↑	rFID↓	PSNR↑	rFID↓	PSNR↑	rFID↓	PSNR↑
SD-VAE [43]	0.73	30.81	9.30	23.37	2.38	26.30	4.71	30.82	19.29	25.26	5.21	25.13	1.5	2.0
VQGAN [10]	7.19	25.43	26.05	20.23	6.49	21.40	18.58	25.97	52.65	22.34	19.26	21.62	7.0	7.0
LlamaGen [46]	2.45	26.09	19.96	21.20	4.94	22.47	12.15	26.46	37.49	23.05	13.35	22.52	5.7	6.0
OpenMagvit2 [32]	1.35	27.41	15.56	21.64	4.19	23.42	7.27	27.85	32.31	23.48	8.91	23.03	3.2	4.0
VAR [48]	1.73	26.50	17.84	21.41	4.49	23.19	8.46	27.22	34.25	23.26	8.89	22.98	4.0	5.0
MGVQ-G4	0.69	31.25	<u>9.52</u>	26.20	2.30	27.95	3.42	31.96	<u>22.30</u>	26.76	<u>5.83</u>	26.98	1.5	1.0

Table 2. **Zero-shot reconstruction evaluation with a downsample ratio of 16 on 512×512 datasets.** Gray denotes continuous tokenizers. MGVQ-G4 indicates our method with 4 sub-codebooks.

- points [66], as shown in Fig. 3.
- Larger representation capacity. Traditional VQ-VAE methods [10, 46, 51] sample a token from the codebook, limiting the representation capacity of the codebook to K . In contrast, our multi-group design can easily boost the capacity to K^G by over a *billion-fold*, experimentally (see Tab. 1).
 - Reduced information loss. With a larger dimension, we avoid information loss and improve reconstruction quality.

Nested Masking. To avoid sub-codebooks converging to the same one, we propose to mask some sub-tokens in a nested manner [5, 22, 42] during training. Specifically, as illustrated in Fig. 4, given a quantized sub-tokens, we ran-

domly sample the number of units to keep $M_{keep} \in 1, \dots, G$ for reconstruction and remove $G - M_{keep}$ last tokens by masking them out. With such a training method, we enforce the encoder to compress the image into the latent in an ordered manner, and the decoder learns to reconstruct the image in a coarse-to-fine manner. Fig. 6 shows the reconstructed image with nested M_{keep} quantized tokens. The image is reconstructed with finer details when incorporating more quantized sub-tokens.

Training loss. Following LlamaGen [46], we utilize a straight-through gradient estimator [4], and employ the commit loss [49], vq-loss [50], perceptual loss from LPIPS [64], and GAN loss [17], and employ the same weights as [46]. Additionally, we employ Charbonnier loss [16, 24] instead

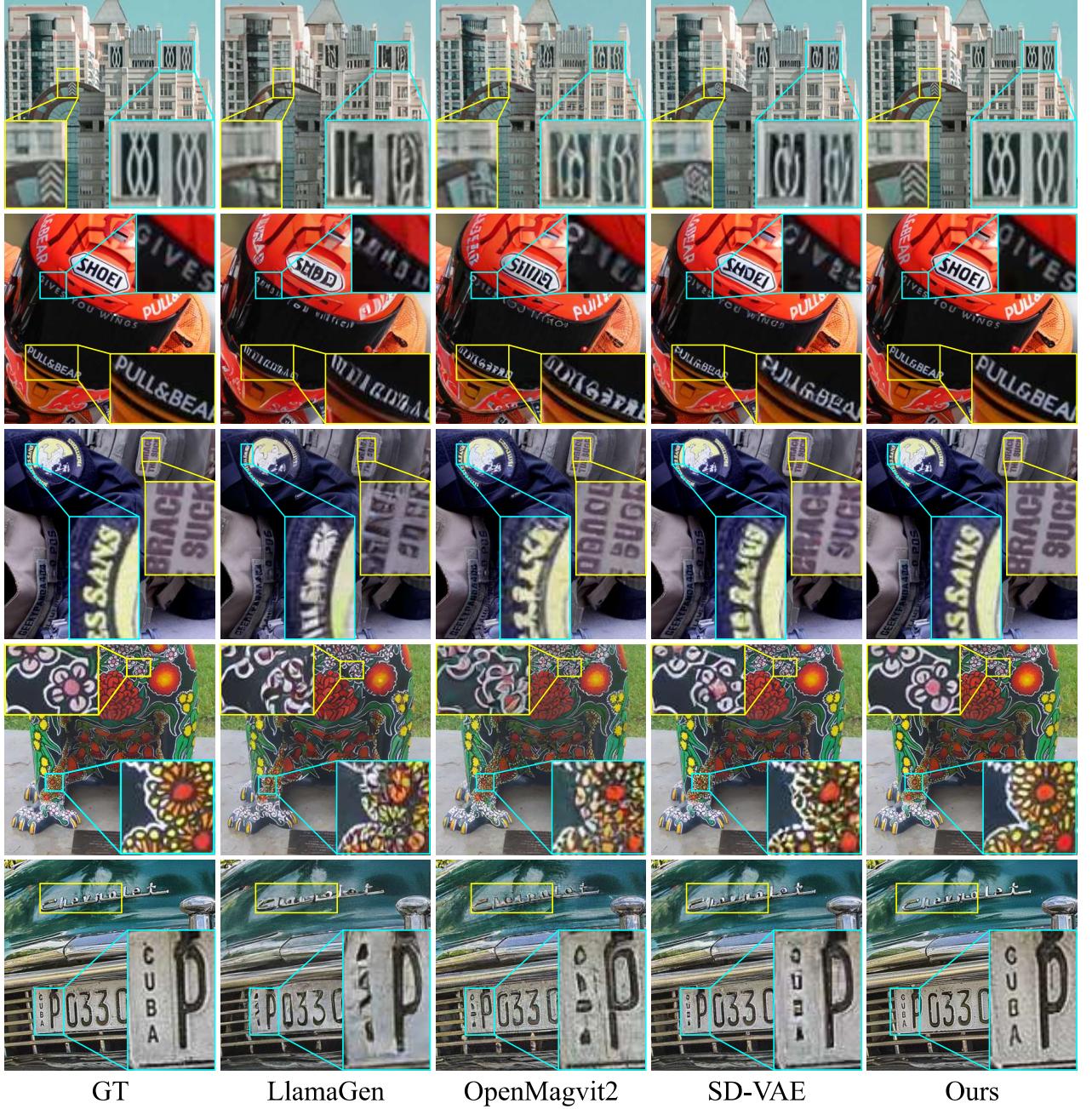


Figure 5. **Qualitative reconstruction images with 16x downsampling on 2560 × 1440 UHDBench dataset.** We crop a 360 × 360 sub-region, and zoom in *detailed textures* using **blue** and **yellow** for better view.

of L1 loss, defined as $L_{Char} = \sqrt{(f(x) - x)^2 + \epsilon^2} \approx |f(x) - x| = L_1$, and L2 loss for reconstruction. The total loss is written as follows:

$$\begin{aligned} \mathcal{L}_{\text{total}} = & \lambda_1 \mathcal{L}_{l2} + \lambda_2 \mathcal{L}_{\text{Char}} + \lambda_3 \mathcal{L}_{\text{Commit}} + \lambda_4 \mathcal{L}_{\text{vq}} \\ & + \lambda_5 \mathcal{L}_{\text{GAN}} + \lambda_6 \mathcal{L}_{\text{Percept}}, \end{aligned} \quad (5)$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5$ denote the weights for different losses. Specifically, we set $\lambda_1 = 2, \lambda_2 = 1, \lambda_3 = 0.25, \lambda_4 = 1, \lambda_5 = 0.5, \lambda_6 = 1$.

4. Implementation Details

Training setup. The sizes of our image tokenizer models are 308M, 210M, and 122M, corresponding to compression ratios of 32, 16, and 8, respectively. All models are trained with a constant learning rate of 10^{-5} , a weight decay of 5×10^{-2} , and the AdamW [31] optimizer ($\beta_1 = 0.9, \beta_2 = 0.95$). When training the model with group 4, the nested masking probability is $\{0.1, 0.1, 0.1, 0.7\}$. On



Figure 6. **Qualitative ablation of our nested masking strategy.** The first figure is the input image. Others show reconstruction results using first M_{keep} sub-groups for $G = 4$ from left to right. The output images gradually gain detailed textures using more sub-groups. The yellow box zooms in *letters* for better view.

the ImageNet benchmark, we strictly follow LlamaGen [46] training setting. On zero-shot benchmarks, we enclose more training data, including FFHQ [18], OpenImages [23], Segment Anything [21], Argoverse2 [53], Mapillary Vistas [38], LAION-HQ [44].

Evaluation datasets and metrics. We evaluate the methods on the ImageNet 50k validation set at a resolution of 256×256 , and on our collected zero-shot benchmark at resolutions of 512×512 and 2560×1440 . Specifically, we construct UHDBench for zero-shot evaluation, which includes 2293 images at 2k resolution sourced from the ground-truth test sets of HRSOD [61], LIU4k [29], UAVid [33], UHDMD [60], and UHRSD [61]. Images smaller than 2560×1440 are excluded. Furthermore, we collect CelebA [30], Davis [39], TextOCR [45], VFHQ [54], Spring [36], ENeRF [27] for more zero-shot evaluation. The image reconstruction quality is measured by PSNR and reconstruction-Fréchet Inception Distance (rFID). We also report SSIM and LPIPS [64] to assess reconstruction quality on the ImageNet benchmark.

5. Experiments

5.1. Main Results

ImageNet Benchmark Reconstruction Comparison. We compare *MGVQ* for two downsample ratios (i.e. 8 and 16), with both discrete tokenizers (VQGAN [10], LlamaGen [46], Open-MAGVIT2 [32], VAR [48]) and continuous tokenizers (SD-VAE [43], SDXL-VAE [40] on ImageNet [9]). In Tab. 1, the representation capacity represents the possibility of a token by sampling from the codebook. Benefiting from our sub-codebooks design, our *MGVQ* could achieve larger representation capacity while eliminating the latent dimension downsampling. We achieve the state-of-the-art performance compared to all existing quantization methods. As reported in Tab. 1, *MGVQ* significantly outperforms the existing discrete tokenizers and even surpasses the continuous tokenizer, SD-VAE [43] and SDXL-VAE, which have been trained on billions of images. Furthermore, we compare 2 different

Method	UHDBench		DAVIS	
	rFID \downarrow	PSNR \uparrow	rFID \downarrow	PSNR \uparrow
SD-VAE [43]	1.07	<u>26.86</u>	1.92	<u>28.02</u>
VQGAN [10]	5.95	22.91	12.74	24.02
LlamaGen [46]	5.59	23.90	11.63	24.90
OpenMagvit2 [32]	4.18	23.91	7.64	25.24
VAR [48]	9.85	21.79	21.85	22.52
<i>MGVQ</i> -G4	<u>1.59</u>	28.27	<u>2.77</u>	29.50

Table 3. **Zero-shot reconstruction evaluation with a downsample ratio of 16 on 2560×1440 datasets.** Gray denotes continuous tokenizers. *MGVQ*-G4 indicates our method with 4 sub-codebooks.

groups, G4 (4 groups) and G8 (8 groups), and their latent dimensions are the same. As G8 has a larger representation capacity, it achieves much better performance.

For VQ-based methods, although LlamaGen [46] increases the codebook size from 16384 to 32768, the performance decreases. We conjunct it with the codebook usage decrease. In contrast, ours has the same codebook size (32768) but a higher representation capacity and 100% usage, resulting in better performance (2.26 v.s. 0.64).

Additionally, Open-MAGVIT2 [59] adopts a different quantization type, i.e. look-up free quantization (LFQ). The codebook size is much higher, but its representation capacity is much lower than ours. We conjunct that this is why our performance is much higher.

Zero-shot Generalization Reconstruction Comparison. To validate the generalization ability, we construct 8 zero-shot benchmarks. We conduct experiments with both 2560×1440 resolution in Tab. 3 on 2 benchmarks and 512×512 resolution in Tab. 2 on 6 benchmarks. Our performance ranks the first on both rFID and PSNR metrics. Note that we even outperform SD-VAE [43] significantly on PSNR. Furthermore, we prepare UHDBench and DAVIS for 2K images reconstruction evaluation. Results are reported in Tab. 3. Our method still achieves the best performance on PSNR in 2K image reconstruction. Thanks to our sub-codebooks design, our *MGVQ* effectively captures the



Figure 7. **Visualization of class-to-image generation.** The classes are dog(Golden Retriever and Husky), cliff, and bald eagle of ImageNet-1k dataset, for each row.

enriched features.

Image Generation Results. We strictly follow LlamaGen’s code [46] and settings in the ImageNet generation experiment, but replace its VQ tokenizer with ours. We apply $cfg = 2.0$ and $topk = 4096$ for each sub-codebook for image generation. As shown in Tab. 4, our generation quality is better than the baseline, especially on IS. Visualization of samples are illustrated in Fig. 7, showing that *MGVQ* is able to generate diverse class-conditional images with high fidelity, reality, and quality. Detailed textures are accurately recovered with ours VQ tokenizer.

Qualitative Comparison of Image Tokenizers. As shown in Fig. 5, we present a qualitative comparison of *MGVQ* with SD-VAE [43], LlamaGen [46], and OpenMagvit2 [32] on our collected UHDBench. Compared to others, *MGVQ* displays a significant improvement in reconstructing high fidelity details, especially for complex textures and tiny texts, surpassing both discrete and continuous tokenizers. In particular, LlamaGen [46] fails to reconstruct fine details. In

Type	Methods	Params	gFID \downarrow	IS \uparrow
VQ	VQGAN [10]	1.4B	15.78	74.3
VQ	RQ-VAE [25]	1.4B	8.71	119.0
VQ	LlamaGen [46]	1.4B	3.09	253.6
VQ	ours	1.4B	3.02	294.1

Table 4. **ImageNet 256p class-to-image generation evaluation.**

row 3, the letters’ color is falsely reconstructed from green to white. We conjecture that this is caused by the limited latent representation capacity. By extending the codebook size, OpenMagvit2 [32] enhances the ability for more realistic colors and details than LlamaGen [46]. However, there remains a failure in accurately reconstructing tiny letters. Note that SD-VAE [43], with a continuous latent space, has the highest ability for reconstruction among other VQ-VAEs. It achieves a relatively good reconstruction of details. However, the edges are blurred and the reconstruction of tiny details is ineffective. In contrast, ours *MGVQ* preserves la-

SC	NM	rFID \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	Usage \uparrow
		2.49	20.49	0.599	0.212	80%
✓		0.75	23.41	0.747	0.119	93%
✓	✓	0.64	23.71	0.755	0.110	100%

Table 5. **Ablation study of proposed strategies.** SC indicates sub-codebooks. NM indicates nested masking.

Method	rFID \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
Ours ($M_{keep} = 1$)	196.43	12.83	0.194	0.602
Ours ($M_{keep} = 2$)	80.00	15.55	0.369	0.438
Ours ($M_{keep} = 3$)	4.56	20.99	0.629	0.202
Ours ($M_{keep} = 4$)	0.64	23.71	0.755	0.110

Table 6. **Ablation study of nested masking.** M_{keep} indicates the number of used groups to keep in the nested masking strategy.

tent dimension to retain more encoded features and learns a coarse-to-fine representation by sub-codebooks, enabling us to reconstruct these tiny textures accurately.

5.2. Ablation Study

5.2.1. Ablation on nested masking and sub-codebooks

This ablation study investigates the contribution of two key factors: the nested masking strategy and the sub-codebook design. A codebook size of 32768 is maintained throughout, and for models using sub-codebooks, the group parameter $G = 4$. We ablation on ImageNet [9].

As shown in Tab. 5, the model without sub-codebooks shows lower PSNR and higher rFID, probably due to its limited representation capacity of only 2^{15} . When the nested masking is applied, a clear improvement in reconstruction fidelity is observed, confirming the strategy’s effectiveness.

Tab. 6 investigates how the nested masking works. Specifically, by analyzing different values of M_{keep} , each sub-codebook’s impact can be isolated. We can conclude that consistent performance gains when additional sub-codebooks are utilized. The nested masking process enables a *coarse-to-fine reconstruction*. As shown in Fig. 6, even with one fewer sub-codebook, coarse textures can still be recovered, while using all sub-codebooks enables finer details. This suggests that the encoder is guided to organize latent representations in a hierarchical manner, and the decoder leverages this hierarchy to progressively refine the output.

Ablation on sub-codebook number, sub-token dimension, and codebook size We ablate on ImageNet. To decouple the encoding, we fix the encoded latent dimension $C_l = 32$. The number of sub-codebooks G varies from 2 to 16, and the sub-token dimension C_l/G is adjusted from 16 to 2. The codebook size varies from 4096 to 262144. Fig. 8 shows the PSNR of different settings. With the same codebook size, the performance is better with increasing sub-codebook groups number and decreasing sub-codebook dimensions. This is because larger sub-codebooks group numbers im-

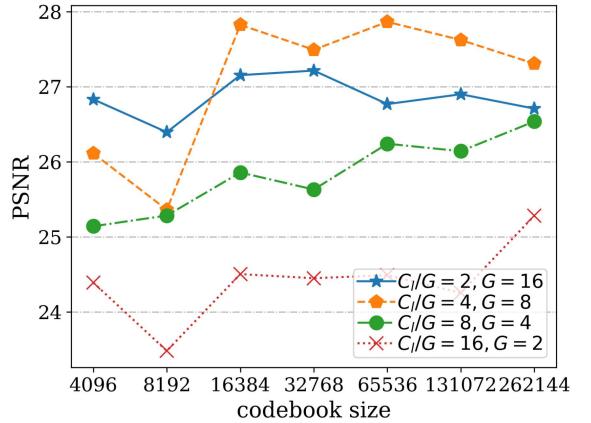


Figure 8. **Reconstruction quality of different latent dimensions and codebook size.** G indicates the number of sub-codebooks. C_l indicates the latent dimension of the encoded features and is fixed at 32. C_l/G indicates the dimension of each sub-token.

prove the representation capacity, while lower sub-codebook dimensions ease the optimization. However, when the sub-codebook groups number increases to 16, the performance is worse than that of 8. We conject that this is because each sub-token dimension shrinks to 2, reducing each code’s representation ability. As shown in Fig. 8, when the codebook size is lower than 16384, decreasing the codebook size will cause inferior performance. However, when the codebook size is larger than 16384, the performance will saturate.

6. Conclusion

In conclusion, *MGVQ* enhances VQ-VAE performance by expanding the representation capacity of discrete codebooks with reduced information loss. By preserving the latent dimension and employing multiple sub-codebooks, *MGVQ* addresses the limitations and narrows the gap between VQ-VAEs and VAEs. Experimental evidence on both ImageNet 256p and zero-shot HD benchmarks confirms its state-of-the-art reconstruction quality, surpassing strong baselines such as SD-VAE in key metrics. These contributions highlight the capability of *MGVQ* in HD image compression and reconstruction and facilitate future research into more efficient discrete latent representations and downstream tasks.

References

- [1] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025. 2
- [2] Eirikur Agustsson, Fabian Mentzer, Michael Tschannen, Lukas Cavigelli, Radu Timofte, Luca Benini, and Luc V Gool. Soft-to-hard vector quantization for end-to-end learning compressible representations. *NeurIPS*, 30, 2017. 2
- [3] Zechen Bai, Jianxiong Gao, Ziteng Gao, Pichao Wang, Zheng

- Zhang, Tong He, and Mike Zheng Shou. Factorized visual tokenization and generation. *arXiv preprint arXiv:2411.16681*, 2024. 2
- [4] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013. 5
- [5] Mu Cai, Jianwei Yang, Jianfeng Gao, and Yong Jae Lee. Matryoshka multimodal models. In *Workshop on Video-Language Models@ NeurIPS 2024*, 2024. 5
- [6] Junyu Chen, Han Cai, Junsong Chen, Enze Xie, Shang Yang, Haotian Tang, Muyang Li, Yao Lu, and Song Han. Deep compression autoencoder for efficient high-resolution diffusion models. *arXiv preprint arXiv:2410.10733*, 2024. 4
- [7] Liuhan Chen, Zongjian Li, Bin Lin, Bin Zhu, Qian Wang, Shenghai Yuan, Xing Zhou, Xinhua Cheng, and Li Yuan. Od-vae: An omni-dimensional video compressor for improving latent video diffusion model. *arXiv preprint arXiv:2409.01199*, 2024. 2
- [8] Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, et al. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807*, 2023. 4
- [9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 7, 9
- [10] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, pages 12873–12883, 2021. 2, 3, 4, 5, 7, 8
- [11] 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 *ICML*, 2024. 4
- [12] Laurent Girin, Simon Leglaive, Xiaoyu Bie, Julien Diard, Thomas Hueber, and Xavier Alameda-Pineda. Dynamical variational autoencoders: A comprehensive review. *arXiv preprint arXiv:2008.12595*, 2020. 3
- [13] Robert Gray. Vector quantization. *IEEE Assp Magazine*, 1(2):4–29, 1984. 1
- [14] Songen Gu, Wei Yin, Bu Jin, Xiaoyang Guo, Junming Wang, Haodong Li, Qian Zhang, and Xiaoxiao Long. Dome: Tampering diffusion model into high-fidelity controllable occupancy world model. *arXiv preprint arXiv:2410.10429*, 2024. 2
- [15] Philippe Hansen-Estruch, David Yan, Ching-Yao Chung, Orr Zohar, Jialiang Wang, Tingbo Hou, Tao Xu, Sriram Vishwanath, Peter Vajda, and Xinlei Chen. Learnings from scaling visual tokenizers for reconstruction and generation. *arXiv preprint arXiv:2501.09755*, 2025. 4
- [16] Xiaotao Hu, Wei Yin, Mingkai Jia, Junyuan Deng, Xiaoyang Guo, Qian Zhang, Xiaoxiao Long, and Ping Tan. Driving-world: Constructingworld model for autonomous driving via video gpt. *arXiv preprint arXiv:2412.19505*, 2024. 5
- [17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 1125–1134, 2017. 5
- [18] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 7
- [19] Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013. 2, 3
- [20] Diederik P Kingma, Max Welling, et al. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019. 3
- [21] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pages 4015–4026, 2023. 7
- [22] Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, et al. Matryoshka representation learning. *NeurIPS*, 35: 30233–30249, 2022. 5
- [23] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020. 7
- [24] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Fast and accurate image super-resolution with deep laplacian pyramid networks. *PAMI*, 41(11):2599–2613, 2018. 5
- [25] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *CVPR*, pages 11523–11532, 2022. 2, 8
- [26] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. Unimo-2: End-to-end unified vision-language grounded learning. *arXiv preprint arXiv:2203.09067*, 2022. 2
- [27] Haotong Lin, Sida Peng, Zhen Xu, Yunzhi Yan, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Efficient neural radiance fields for interactive free-viewpoint video. In *SIGGRAPH Asia Conference Proceedings*, 2022. 7
- [28] Alexander H Liu, SouYoung Jin, Cheng-I Jeff Lai, Andrew Rouditchenko, Aude Oliva, and James Glass. Cross-modal discrete representation learning. *arXiv preprint arXiv:2106.05438*, 2021. 2
- [29] J. Liu, D. Liu, W. Yang, S. Xia, X. Zhang, and Y. Dai. A comprehensive benchmark for single image compression artifact reduction. *IEEE Transactions on Image Processing*, 29: 7845–7860, 2020. 7
- [30] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, pages 3730–3738, 2015. 7
- [31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6
- [32] 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 preprint arXiv:2409.04410*, 2024. 2, 5, 7, 8

- [33] Ye Lyu, George Vosselman, Gui-Song Xia, Alper Yilmaz, and Michael Ying Yang. Uavid: A semantic segmentation dataset for uav imagery. *ISPRS journal of photogrammetry and remote sensing*, 165:108–119, 2020. 7
- [34] 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 preprint arXiv:2502.20321*, 2025. 2
- [35] Chengzhi Mao, Lu Jiang, Mostafa Dehghani, Carl Vondrick, Rahul Sukthankar, and Irfan Essa. Discrete representations strengthen vision transformer robustness. *arXiv preprint arXiv:2111.10493*, 2021. 2
- [36] Lukas Mehl, Jenny Schmalfuss, Azin Jahedi, Yaroslava Nalivayko, and Andrés Bruhn. Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In *CVPR*, pages 4981–4991, 2023. 7
- [37] Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: Vq-vae made simple. *arXiv preprint arXiv:2309.15505*, 2023. 3
- [38] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kortschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *ICCV*, pages 4990–4999, 2017. 7
- [39] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, pages 724–732, 2016. 7
- [40] 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. 2, 3, 5, 7
- [41] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *NeurIPS*, 32, 2019. 2
- [42] Oren Rippel, Michael Gelbart, and Ryan Adams. Learning ordered representations with nested dropout. In *ICML*, pages 1746–1754. PMLR, 2014. 5
- [43] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2, 3, 4, 5, 7, 8
- [44] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 35:25278–25294, 2022. 7
- [45] Amanpreet Singh, Guan Pang, Mandy Toh, Jing Huang, Wojciech Galuba, and Tal Hassner. Textocr: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In *CVPR*, pages 8802–8812, 2021. 7
- [46] 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 preprint arXiv:2406.06525*, 2024. 2, 3, 4, 5, 7, 8
- [47] Yuhta Takida, Takashi Shibuya, WeiHsiang Liao, Chieh-Hsin Lai, Junki Ohmura, Toshimitsu Uesaka, Naoki Murata, Shusuke Takahashi, Toshiyuki Kumakura, and Yuki Mitsu-fuji. Sq-vae: Variational bayes on discrete representation with self-annealed stochastic quantization. *arXiv preprint arXiv:2205.07547*, 2022. 2, 3, 4
- [48] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *NeurIPS*, 37:84839–84865, 2025. 5, 7
- [49] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *NeurIPS*, 30, 2017. 5
- [50] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *NeurIPS*, 30, 2017. 2, 3, 5
- [51] 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*, 2024. 2, 4, 5
- [52] Will Williams, Sam Ringer, Tom Ash, David MacLeod, Jamie Dougherty, and John Hughes. Hierarchical quantized autoencoders. *NeurIPS*, 33:4524–4535, 2020. 2
- [53] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*, 2023. 7
- [54] Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *CVPR*, pages 657–666, 2022. 7
- [55] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 2
- [56] Tackgeun You, Saehoon Kim, Chiheon Kim, Doyup Lee, and Bohyung Han. Locally hierarchical auto-regressive modeling for image generation. *NeurIPS*, 35:16360–16372, 2022. 2
- [57] 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*, 2021. 2
- [58] 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*, 2021. 2, 3
- [59] Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, et al. Language model beats diffusion-tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023. 2, 3, 7
- [60] Xin Yu, Peng Dai, Wenbo Li, Lan Ma, Jiajun Shen, Jia Li, and Xiaojuan Qi. Towards efficient and scale-robust ultra-high-definition image demoiréing. In *ECCV*, pages 646–662. Springer, 2022. 7
- [61] Yi Zeng, Pingping Zhang, Jianming Zhang, Zhe Lin, and Huchuan Lu. Towards high-resolution salient object detection. In *ICCV*, pages 7234–7243, 2019. 7

- [62] Jiahui Zhang, Fangneng Zhan, Christian Theobalt, and Shijian Lu. Regularized vector quantization for tokenized image synthesis. In *CVPR*, pages 18467–18476, 2023. [2](#)
- [63] Kaiwen Zhang, Zhenyu Tang, Xiaotao Hu, Xingang Pan, Xiaoyang Guo, Yuan Liu, Jingwei Huang, Li Yuan, Qian Zhang, Xiao-Xiao Long, et al. Epona: Autoregressive diffusion world model for autonomous driving. *arXiv preprint arXiv:2506.24113*, 2025. [2](#)
- [64] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. [5](#), [7](#)
- [65] Sijie Zhao, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Muyao Niu, Xiaoyu Li, Wenbo Hu, and Ying Shan. Cv-vae: A compatible video vae for latent generative video models. *NeurIPS*, 37:12847–12871, 2025. [2](#)
- [66] Chuanxia Zheng and Andrea Vedaldi. Online clustered codebook. In *ICCV*, pages 22798–22807, 2023. [3](#), [4](#), [5](#)
- [67] Chuanxia Zheng, Tung-Long Vuong, Jianfei Cai, and Dinh Phung. Movq: Modulating quantized vectors for high-fidelity image generation. *NeurIPS*, 35:23412–23425, 2022. [2](#)
- [68] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024. [2](#)
- [69] Yongxin Zhu, Bocheng Li, Yifei Xin, and Linli Xu. Addressing representation collapse in vector quantized models with one linear layer. *arXiv preprint arXiv:2411.02038*, 2024. [3](#)