Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.060252

ARTICLE

# Frequency-Quantized Variational Autoencoder Based on 2D-FFT for Enhanced Image Reconstruction and Generation

**Jianxin Feng**[1,2,*] **and Xiaoyao Liu**[1,2]

[1]School of Information Engineering, Dalian University, Dalian, 116622, China
[2]Key Laboratory of Communication and Networks, Dalian University, Dalian, 116622, China

*Corresponding Author: Jianxin Feng. Email: fengjianxin863@163.com

Received: 28 October 2024; Accepted: 08 February 2025; Published: 16 April 2025

**ABSTRACT:** As a form of discrete representation learning, Vector Quantized Variational Autoencoders (VQ-VAE) have increasingly been applied to generative and multimodal tasks due to their ease of embedding and representative capacity. However, existing VQ-VAEs often perform quantization in the spatial domain, ignoring global structural information and potentially suffering from codebook collapse and information coupling issues. This paper proposes a frequency quantized variational autoencoder (FQ-VAE) to address these issues. The proposed method transforms image features into linear combinations in the frequency domain using a 2D fast Fourier transform (2D-FFT) and performs adaptive quantization on these frequency components to preserve image's global relationships. The codebook is dynamically optimized to avoid collapse and information coupling issue by considering the usage frequency and dependency of code vectors. Furthermore, we introduce a post-processing module based on graph convolutional networks to further improve reconstruction quality. Experimental results on four public datasets demonstrate that the proposed method outperforms state-of-the-art approaches in terms of Structural Similarity Index (SSIM), Learned Perceptual Image Patch Similarity (LPIPS), and Reconstruction Fréchet Inception Distance (rFID). In the experiments on the CIFAR-10 dataset, compared to the baseline method VQ-VAE, the proposed method improves the above metrics by 4.9%, 36.4%, and 52.8%, respectively.

**KEYWORDS:** VAE; 2D-FFT; image reconstruction; image generation

## 1 Introduction

In recent years, the rapid development of deep learning technologies has brought increasing attention to discrete representation learning. Variational Autoencoders (VAE) [1] introduced a probabilistic approach to learning compressed latent representations, laying the foundation for discrete representation learning. The advent of Generative Adversarial Networks (GANs) [2] and Transformers [3] further enriched the downstream tasks of discrete representation learning, achieving stunning results. For instance, the Vector Quantized Generative Adversarial Network (VQGAN) [4] architecture, which combines the strengths of VAEs and GANs, has demonstrated remarkable performance in high-resolution image synthesis. More recently, the popularity of diffusion models has led to significant advancements in generative modeling. By incorporating discrete representation learning, such as in the Vector Quantized Diffusion [5] model, these techniques have not only enhanced the quality of generated samples but also greatly improved model efficiency. The success of these approaches underscores the importance and potential of discrete representation learning in the field of deep learning.

Among the various discrete representation learning methods, Vector Quantized Variational Autoencoders (VQ-VAE) [6] have gained prominence due to their advantages in representative capacity and ease of embedding. VQ-VAE maps continuous feature vectors into discrete spaces through vector quantization and employs a nearest-neighbor algorithm to select the most appropriate discrete vector, simplifying the optimization problem. This approach enables the model to more efficiently capture the fundamental structure and features of the data, leading to its widespread application across multiple domains [7].

Although VQ-VAE has been integrated into various deep learning models as a mature machine learning method, challenges remain regarding the efficiency and accuracy of quantization. In VQ-VAE, the encoder divides the input image into a regular fixed grid, segmented into multiple units, considered "tokens," to represent the image's features. During the quantization process, these tokens interact with a set of vectors known as a codebook, optimized such that the codebook captures as much of the image's representational information as possible, thereby improving the expressive power of the quantized encoding. However, this process introduces a significant issue: disrupting the global relationships within the image.

Unlike the independence of words in natural language processing, image features are typically interrelated holistically. Processing individual image patches not only smooths out high-frequency information, causing blurring during reconstruction but also increases the difficulty of training for downstream tasks such as image generation. To address these issues, Tian et al. [8] designed a multi-scale VQ-VAE model, which quantizes data at different scales into latent vectors and shares a single codebook. On the other hand, Huang et al. [9] proposed a progressive quantization autoencoder from the perspective of maximizing information in limited discrete codes, aligning the codebook dimensions with the latent feature dimensions generated by the encoder. To some extent, these methods have preserved the global correlation of features, but progressive quantization increases the training complexity, undermining the advantages of parallel training in neural networks. Moreover, since code vectors are directly mapped from globally unseparated features, this leads to an information coupling issue—similar code vectors tend to capture redundant information. This reduces the amount of practical information in the codebook and limits the diversity in downstream tasks such as generation.

The efficiency of the codebook is one of the critical factors influencing the quality of representation learning in models. Several studies have explored the codebook collapse issue [10] to improve the reconstruction quality of models. Codebook collapse refers to a scenario where only a small subset of codevectors are co-optimized with latent features, leaving most of the entries in the codebook unused. The literature [10–12] has proposed various approaches to mitigate this problem, including random initialization, random quantization, and latent clustering. While these methods have improved codebook utilization to some extent, they are still based on spatial domain partitioning quantization, which neglects the global correlation of image features, thus limiting the model's expressiveness and increasing the difficulty of training for downstream tasks such as image generation.

This paper proposes a Frequency-Quantized Autoencoder (FQ-VAE) to preserve images' global relationships through frequency domain feature extraction while achieving high-quality image reconstruction. The core of this method lies in utilizing the 2D Fast Fourier Transform (2D-FFT) to encode the original data as a linear combination of different feature components in the frequency domain, with these components subsequently quantized. To further optimize the codebook, this paper dynamically updates it by considering both the usage frequency and dependency of codevectors. Moving codevectors with low usage frequency ensures that all codevectors are updated, effectively preventing codebook collapse. Codevectors with high dependency are significantly updated to avoid information coupling within the codebook. Additionally, this paper introduces graph convolution theory to design and solve correction weights for the quantized features, which helps refine the quantized features and further reduces quantization loss.

By performing quantization and encoding in the frequency domain, the proposed method effectively avoids the loss of global relationships caused by spatial domain partitioning, significantly enhancing the model's expressive power and reconstruction quality while improving its applicability to downstream tasks.

The main contributions of this paper are as follows:

1. Proposed a novel frequency-quantized variational autoencoder (FQ-VAE) method: This method quantizes image features by transforming them from the spatial domain to the frequency domain, effectively preserving the global relationships among image features.
2. Designed an adaptive quantization encoder based on 2D-FFT: The encoder uses 2D-FFT to convert spatial domain feature maps into a combination of essential features in the frequency domain, enabling the adaptive extraction of more effective feature information.
3. Optimized the structure of the codebook: During the dynamic update of the codebook, both the usage frequency and dependency of code vectors were considered, preventing issues such as codebook collapse and information coupling.
4. Introduced graph convolution theory to refine quantized features: Using a Graph Convolutional Model (GCM), the method obtains correction weights for different frequency features, further reducing the information loss caused by quantization.
5. Validated the method's effectiveness: In image reconstruction tasks, the proposed method was compared with state-of-the-art methods on four benchmark datasets, demonstrating superior performance. Additionally, it proved capable of generating high-quality images in generative tasks.

The remainder of this paper is structured as follows: Section 2 provides an overview of related work, focusing on VQ-VAE and its variants, as well as recent advances in 2D-FFT-based neural networks and graph convolutional networks. Section 3 presents the proposed FQ-VAE in detail, including the adaptive quantization encoder based on 2D-FFT, the codebook optimization strategy, and the graph convolution-based feature refinement module. Section 4 reports the experimental results on image reconstruction and generation tasks, along with relevant ablation studies. Finally, Section 5 summarizes the entire paper and discusses the limitations of the method and potential future research directions.

## 2 Related Work

### *2.1 Vector Quantised Variational Autoencoder*

VQ-VAE is a generative model combining Variational Autoencoders principles with vector quantization techniques. By quantizing continuous feature vectors into discrete representations, it learns and generates using a codebook composed of a finite set of codevectors. This approach not only enables the model to learn compact representations of data but is also widely applied across various downstream tasks, such as image generation [13], video generation [14], audio generation [15], communication systems [16], and recognition [17].

The reconstruction quality of VQ-VAE directly reflects the model's ability to learn from the original data and, at the same time, sets an upper bound for performance in downstream tasks. To improve image reconstruction quality, several research works have proposed different enhancements. Razavi et al. [18] improved the VQ-VAE model by designing a multi-scale hierarchical structure, separating global information at the top layers from local information at the bottom layers, which enabled the generation of globally coherent and locally high-resolution images. Additionally, VQGAN and Vision-Transformer-based VQGAN (ViT-VQGAN) [19] introduced Generative Adversarial Networks and Transformer architectures into the training process, further enhancing reconstruction quality. Residual-Quantized VAE (RQ-VAE) and Modulating Quantized Vectors (MoVQ) [20] improved reconstruction by utilizing multi-channel

representations, though at the cost of increasing the dimensionality of latent features, which reduces the model's compression efficiency.

Recently, several methods have been proposed to address the codebook collapse issue. Williams et al. proposed Hierarchical Quantized Autoencoders (HQ-VAE), while Dhariwal et al. [21] introduced Jukebox. Both methods implement a codebook resetting mechanism, randomly reinitializing unused or infrequently used codebook entries. Takida et al. introduced the Stochastically Quantized Variational Autoencoder (SQ-VAE), which incorporates a self-annealing process to learn an effective codebook from the initial stochastic quantization. Vuong et al. [22] proposed the Vector Quantized Wasserstein Autoencoder (VQ-WAE), which replaces the KL divergence with the Wasserstein distance as a regularization term to ensure a uniform distribution of discrete representations. The most relevant work is the Clustering VQ-VAE (CVQ-VAE) proposed by Zheng et al., which takes into account the variability of features in deep networks by employing running mean updates within training batches to capture the dynamic changes throughout the training process. In contrast, CVQ-VAE achieves better reconstruction results while ensuring 100% codebook utilization. Despite the improvements these methods offer in terms of encoding-decoding and codebook utilization, they still suffer from issues affecting images' global coherence. This paper maps latent features into multi-frequency combinations in the frequency domain, ensuring global information preservation in the images, thereby enhancing the codebook's information capacity and the model's representative capabilities.

### *2.2 Network Based on 2D-FFT*

Fourier transform has been widely applied in neural networks to enhance feature extraction capabilities in recent years. For instance, Lee-Thorp et al. [23] proposed the Filter Networks (FNet), which replaces the self-attention sublayer in the Transformer encoder with a standard, unparameterized Fourier transform, achieving competitive performance while accelerating the training process. Building on this, Sevim et al. [24] further improved the approach by introducing Fast-FNet, significantly enhancing model efficiency. These methods have been extensively applied in the field of natural language processing (NLP), and research on Fourier transforms in the vision domain has also been gradually increasing.

Rao et al. [25] proposed Global Filter Networks (GFNet), which utilize the 2D Discrete Fourier Transform (2D-DFT) to transform features into the frequency domain and multiply them with a global filter, achieving a reasonable balance between accuracy and complexity in image classification tasks. Subsequently, Rao et al. [26] applied this method to visual recognition tasks, improving efficiency while maintaining accuracy. However, the global filters in GFNet are shared parameters used by all samples, which somewhat weakens the model's generalization ability. Wang et al. [27] designed an attention network based on multi-scale fast Fourier transform, extracting global information from images using FFT and combining it with spatial local information, further enhancing network performance. Building on this, Tatsunami et al. [28] developed dynamic global filters, narrowing the gap between Fourier transform-based global filters and the multi-head self-attention mechanism. These studies aim to enhance the feature extraction capability by acquiring a global receptive field through Fourier transform, with the final output in the frequency domain being a singular frequency domain feature map.

In contrast, the method proposed in this paper adaptively extracts multi-dimensional frequency domain features based on dynamic filtering and maps image features as linear combinations of these features. This provides new possibilities for feature quantization and better preserve information from the source data, reducing information loss.

### *2.3 Graph Convolution Networks*

Graph convolution has emerged as a powerful tool for processing non-Euclidean data. By leveraging graph structures, Graph Convolutional Networks (GCNs) can effectively capture the complex relationships and dependencies among nodes in a graph, and have been widely applied in the field of computer vision. In recent years, graph convolution techniques have been increasingly employed in various tasks, including image classification [29], deep image clustering [30], feature fusion [31], and more.

In the context of feature optimization, graph convolution has shown great potential in enhancing learned features. Cai et al. [32] generated directional relationships between features by combining the Cross-Attention Mechanism and graph convolution techniques, thereby preserving image feature information to a higher degree. Yang et al. [33] introduced Spatial Graph Attention (SGA) to encode feature correlations in the spatial dimension, enriching feature representations and achieving significant performance improvements in the single image super-resolution (SISR) task. While these studies primarily apply graph convolution techniques in the spatial dimension, some research has also demonstrated promising results when applying them in the channel dimension. Li et al. [34] proposed a dynamic-channel graph convolutional network to map image channels to the topological space and synthesize the features of each channel on the topological map, addressing the limitation of insufficient channel information utilization and further enhancing image edge information. In the latest research, Xiang's Adaptive Graph Channel Attention (AGCA) [35] introduces graph convolution theory into channel attention, treating each channel as a feature vertex and performing non-local operations on the features, significantly improving feature representation capabilities.

The proposed FQ-VAE method in this paper also leverages graph convolution to refine quantized features. By constructing a graph convolutional module, FQ-VAE learns the correction weights of different frequency components, which helps mitigate the information loss caused by quantization. The graph convolution operation enables the model to capture the relationships among frequency components and adaptively adjust their weights, resulting in more accurate and informative quantized features.

## 3 Method

### *3.1 VQ-VAE*

Given an image $x \in \mathbb{R}^{H \times W \times c}$, VQ-VAE uses an autoencoder to encode the input as $z_e(x) \in \mathbb{R}^{h \times w \times D}$, and then quantizes the latent features into the codebook space $z_q(x) \in \mathbb{R}^{h \times w \times D}$, where $h \times w$ represents the dimensions of the basic blocks to register codebook entries, and $D$ is the dimensionality of each codebook entry. The quantized encoding vectors can be used to generate new images.

VQ-VAE is trained by optimizing the following loss:

$$L = ||x - \hat{x}||_2^2 + ||sg[z_e(x)] - z_q(x)||_2^2 + \beta||sg[z_q(x)] - z_e(x)||_2^2 \tag{1}$$

where $sg[.]$ denotes the stop-gradient operation, and $\beta$ is a weighting hyperparameter. The loss function comprises a reconstruction loss, which measures the difference between the observed $x$ and the reconstructed $\hat{x}$; a codebook loss, which forces the codevectors to approach their original features; and a commitment loss, which brings the encoder's output features closer to the chosen codevectors. In this paper, we focus on optimizing the model from the perspective of quantization methods, hence we follow the VQ-VAE loss function, with the detailed methodology discussed in Section 3.2.

### *3.2 FQ-VAE*

#### *3.2.1 Overview*

Our method is based on the VQ-VAE architecture, where the process from the original image input to the reconstructed image output passes through the encoder, quantizer, and decoder. As shown in Fig. 1, the encoder utilizes a 2D-FFT to transform image features into a representation in the frequency domain, forming a linear combination of frequency components. These components are then quantized in the quantizer. To avoid codebook collapse and mitigate potential information redundancy, we dynamically optimize the codebook by considering the usage frequency and dependencies among codevectors. In the decoder, we apply graph convolution theory to refine the quantized features and minimize quantization loss. The specific implementation details are provided in the subsequent sections.

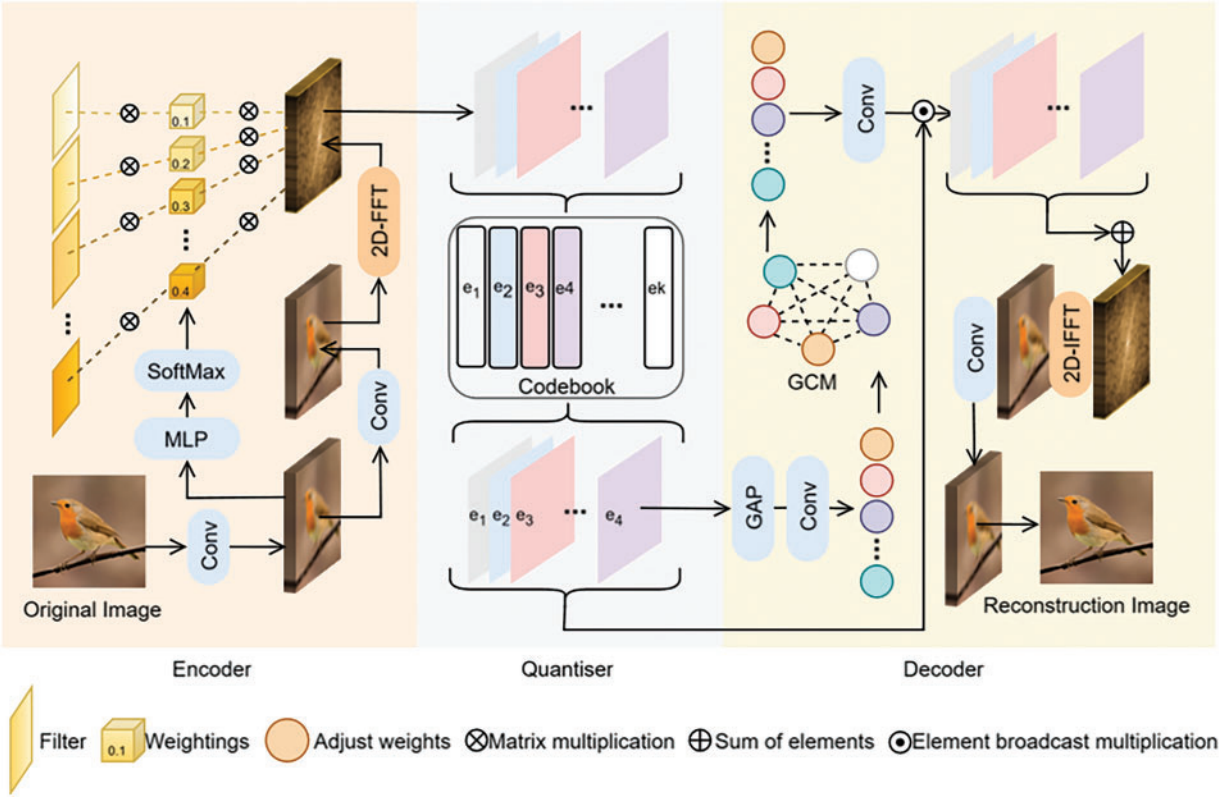


**Figure 1:** Architecture of the frequency quantized variational autoencoder (FQ-VAE)

We utilize 2D-FFT to transfer image features into the frequency domain and extract their components, quantizing these frequency domain features. During decoding, we use 2D-IFFT to return the image features back to the spatial domain, followed by convolution to complete the reconstruction. Additionally, we optimize the codebook and apply a Graph Convolutional Module (GCM) to generate corrective weights, reducing quantization loss. Compared to other VQ-VAE architectures, our method preserves the global relationships within the image, resulting in improved reconstruction performance.

*3.2.2 Encoder*

Compared to traditional methods, which typically map raw data into a set of regular image blocks in the high-dimensional spatial domain and then quantize these blocks, such grid-based partitioning often disrupts the global relationships within the image, leading to information loss. In contrast, the method proposed in this paper preserves the image's global characteristics throughout the encoding process. The encoder encodes the raw data as a linear combination of various frequency components in the frequency domain. Our method is based on the 2D-FFT, but a thorough understanding of the 2D-DFT is essential before implementing the 2D-FFT effectively.

For a given 2D signal $x(h, w)$, the 2D-DFT is defined as:

$$\tilde{x}(h', w') = \frac{1}{\sqrt{HW}} \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} x(h, w) e^{-2\pi j \left( \frac{hh'}{H} + \frac{ww'}{W} \right)} \tag{2}$$

where $H, W \in \mathbb{N}, h, h' \in \{z \in \mathbb{Z} | 0 < z < H\}, w, w' \in \{z \in \mathbb{Z} | 0 < z < W\}$.

The 2D-DFT maps the data from the spatial domain to the frequency domain, where $\tilde{x}(h', w')$ belongs to the frequency domain. The 2D-DFT produces a complete complex matrix; however, due to the conjugate symmetry of real-valued inputs, half of this matrix is redundant, leading to a time complexity of $\mathcal{O}(H^2 W^2)$ for direct implementation. In contrast, the 2D Fast Fourier Transform (2D-FFT) reduces the number of required multiplications and additions by leveraging the symmetry and periodicity inherent in the DFT computation, lowering the time complexity to $\mathcal{O}(HW \log_2(H\mathrm{W}))$.

As illustrated in the encoder module in Fig. 1, our method initially adheres to the conventional process of downsampling through convolution to map raw data into latent space, where the channel dimension is emphasized, resulting in latent features $z(x) \in \mathbb{R}^{D \times h \times w}$, where each channel corresponds to a standard 2D signal. For each channel, distinct frequency transformations are employed to decompose and extract features from the frequency domain, specifically implemented as follows:

$$z_e(x) = \mathcal{K} \odot \mathcal{F}(x) \tag{3}$$

where $\odot$ represents element-wise multiplication. $\mathcal{K} \in \mathbb{C}^{D \times H \times \lceil \frac{W}{2} \rceil}$ is a learnable filter. $\mathcal{F}(x)$ denotes the features after the 2D-FFT, and $z_e(x)$ is the final latent feature of the encoder, where each element serves as the minimum unit for quantization. The purpose of $\mathcal{K}$ is to dynamically generate filters based on data characteristics, capturing important frequency components across different data, thus exhibiting dynamic and adaptive properties.

In this paper, we utilize a multilayer perceptron (MLP) to generate a set of $N$ filter coefficients $m_i$, $(i \in 0, 1, \ldots, N)$, and we define a set of filter bases $K = \{K_1, \ldots, K_N\}$, where $K_1, \ldots, K_N \in \mathbb{C}^{H \times \frac{W}{2}}$. For different frequency-domain feature filters, the following definition applies:

$$\mathcal{K}_{m_i}(x)_{D,:,:} = \left( \frac{e^{m_{i(D-1)N+i}}}{\sum_{n=1}^{N} e^{m_{i(D-1)N+n}}} \right) K_i \tag{4}$$

Finally, the combination of learnable filters is:

$$\mathcal{K} = \sum_{i=1}^{N} \mathcal{K}_{m_i}(x) \tag{5}$$

The encoder extracts frequency-domain features from the channel dimension of the latent feature $z(x)$, and ultimately quantizes the image features with different frequency components from various channels

as the minimum unit. Unlike spatial domain partitioning, this approach preserves the global relationships within the image. Furthermore, the global dynamic filter-based feature extraction scheme, based on the FFT employed in this paper, exhibits lower time complexity compared to other global feature extractors (e.g., Transformer), with a complexity of $\mathcal{O}(HWD\log_2(H\mathrm{W}) + HWD)$.

*3.2.3 Codebook Construction*

To optimize inactive codevectors in the codebook, this paper proposes an improved dynamic codebook updating method. This approach updates the codebook dynamically based on the usage frequency of code vectors and their dependencies. The updated code vector $e_k^{(t)}$ is calculated as follows:

$$e_k^{(t)} = \left(1 - \alpha_k^{(t)}\right) e_k^{(t-1)} + \alpha_k^{(t)} \hat{z}_k^{(t)} \tag{6}$$

where $\hat{z}_k^{(t)}$ is the resampled anchor vector, which provides the correct update direction for the code vector, obtained by identifying the nearest latent vector. $\alpha_k^{(t)}$ is the decay value, calculated using the following formula:

$$\alpha_k^{(t)} = \exp^{-N_k^{(t)}} \delta_k M \frac{10}{1-\gamma} - \epsilon \tag{7}$$

where $M$ is the number of anchor vectors, and $\epsilon$ is a constant (set to 0.001). $N_k^{(t)}$ represents the average usage of the $k$-th entry after step $t$ and is updated according to the following formula:

$$N_k^{(t)} = \gamma N_k^{(t-1)} + (1-\gamma)\frac{n_k^{(t)}}{Bhw} \tag{8}$$

where $n_k^{(t)}$ is the usage count of the $k$-th entry in the mini-batch at step $t$, $N_k^{(t)}$ is the average usage of the $k$-th entry after step $t$, $\gamma$ is the decay hyperparameter (set to 0.99), and $B$ is the batch size.

In addition, to reduce the dependency between codevectors in the codebook and better represent the quantized data with linear combinations, while avoiding excessive information coupling, the approach introduces the Pearson correlation coefficient $\delta$. This term is used to evaluate the linear correlation strength between each code vector and the rest of the vectors in the codebook. For any code vector $e_k$, the Pearson correlation coefficient $\delta_k$ is defined as:

$$\delta_k = \max_{j\in\{0,1,\ldots,M\}, j\neq k} \left(\delta_{k,j}\right) = \max_{j\in\{0,1,\ldots,M\}, j\neq k} \left(\frac{cov\left(e_k, e_j\right)}{\sigma_{e_k}\sigma_{e_j}}\right) \tag{9}$$

where $j$ represents all codevectors in the codebook except $e_k$. $cov(e_k, e_j)$ is the covariance between $e_k$ and $e_j$, and $\sigma_{e_k}\sigma_{e_j}$ are their standard deviations. The covariance $cov(e_k, e_j)$ is expressed as:

$$cov\left(e_k, e_j\right) = \frac{1}{n-1}\sum_{i=1}^{n}\left(e_{k\,i} - \overline{e_k}\right)\left(e_{j_i} - \overline{e_j}\right) \tag{10}$$

where $n$ represents the dimensionality of the codevectors. The magnitude of $\delta_k$ reflects the linear correlation between the code vector and the other vectors in the codebook. The higher the correlation, the larger the update for that codevectors. This dynamic learning process helps the codebook vectors better capture the essential characteristics of the data and enhances their linear independence.

*3.2.4 Decoder*

In traditional quantization methods, information loss is inevitable, primarily due to the distance discrepancy between the codevectors and latent vectors. In our method, the decoder's input is a linear combination of various frequency-domain features. Additionally, through the optimization of correlation strength during codebook training, the codebook entries tend to form a linearly independent set. According to a theorem in linear algebra, if a set of $N$-dimensional vectors is linearly independent and their number exceeds $N$, any arbitrary $N$-dimensional vector can be represented as a linear combination of these vectors. This provides a theoretical foundation for correcting quantized features. By solving for the weight coefficients $w$, we can approximately reconstruct the latent feature $z_e(x)$ without loss. This is expressed as follows:

$$z_e(x) = w_1 e_1 + w_2 e_2 + \ldots + w_n e_n \tag{11}$$

Inspired by recent work of Xiang et al. [35], which models inter-channel relationships using a graph structure, we introduce graph convolution theory to solve for the frequency feature coefficients $w$ in our method.

Specifically, we first apply global average pooling (GAP) to the feature map, reducing its dimension to $U \times 1 \times 1$, where $U = N \times D$ is the product of the number of filter bases and the number of channels, representing the number of frequency features (graph vertices). A subsequent $1 \times 1$ convolution operation is employed to further enhance the feature representation, producing an initial set of feature maps. These feature maps are then processed by a graph convolutional module to obtain the feature weights. The architecture of the graph convolution module is illustrated in Fig. 2. Each frequency feature map is treated as a graph node, and the graph convolution module can be expressed as:

$$\hat{f} = W f(A \times B + C) \tag{12}$$

here, $W$ denotes the weights of different network layers, while $f$ and $\hat{f}$ represent the input and output feature maps, respectively. $A$ is initialized as an $U \times U$ identity matrix, representing the self-relationship of each feature vertex. $B$ is a diagonal $U \times U$ matrix representing the weights of the feature vertices, obtained by applying a one-dimensional convolution followed by a softmax operation. This enables the network to adaptively emphasize or suppress certain features. $C$ is a learnable $U \times U$ adjacency matrix, optimized through backpropagation to represent the relationship between feature vertices.

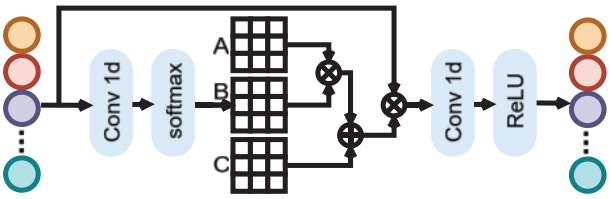


**Figure 2:** Architecture of the Graph Convolutional Module (GCM). After passing the input feature maps through the GCM, the corresponding weights for the quantized feature components are learned, which are used to further minimize the information loss caused by quantization

The feature weights obtained via the graph convolution operation are used to recalibrate the frequency-level response. The weighted frequency features are then transformed back from the frequency domain to the spatial domain using the 2D inverse Fourier transform.

Subsequent operations are similar to those in traditional VQ-VAE. Through convolution operations, the latent features in the spatial domain are restored to their original space, achieving high-quality image reconstruction.

## 4 Experimental Results

In the experimental evaluation, this paper conducts experiments on both image reconstruction and image generation tasks. To ensure fairness, all experiments utilize the same downsampling rates and hyperparameter settings. In the image reconstruction experiments, the superiority of the proposed method is validated by comparing the image reconstruction quality, codebook utilization rate, and codebook information capacity. Ablation studies further confirm the significance and practicality of all components on model performance, while determining the optimal combination of the number of channels and filters. In the image generation task, the proposed method demonstrates advantages in preserving global relationships for downstream tasks. Experimental results show that the proposed method outperforms current state-of-the-art methods in codebook construction, reconstruction performance, and generation quality, achieving significant improvements.

### *4.1 Experimental Setup*

**Datasets and Performance Measures.** The models in this paper are trained on four publicly available datasets: MNIST [36], CIFAR10 [37], CelebA [38], and LSUN Church [39].

The following metrics are utilized as performance evaluation indicators: Structural Similarity Index (SSIM) at the image patch level, Learned Perceptual Image Patch Similarity (LPIPS) [40] at the feature level, and Reconstruction Fréchet Inception Distance (rFID) [41] at the dataset level.

**Implementation Details.** For VQ-VAE, HQ-VAE, SQ-VAE, and CVQ-VAE, this paper utilizes the official code and constructs models for all datasets except LSUN Church. LSUN Church was only applied to the image generation task in the CVQ-VAE paper, however, to demonstrate the model's generalization capability, this paper includes this dataset in the image reconstruction task, using the same configuration parameters as those for CelebA.

The proposed method first maps features through convolutional layers. For the MNIST, CIFAR10, and CelebA datasets, the method modifies only the channel dimensions after resizing the images, without altering their resolution. for the LSUN Church dataset, the $h$ and $w$ are adjusted to $H/2$ and $W/2$, respectively.

Regarding learning rates, a rate of $2 \times 10^{-3}$ is set for MNIST, while a rate of $3 \times 10^{-3}$ is set for the reconstruction tasks on the other datasets. According to the study referenced in [12], the weight hyperparameter $\beta$ in the loss function is set to 0.25, and the decay hyperparameter $\gamma$ is set to 0.99. All experiments are conducted on an NVIDIA A800 PCIe with 80 GB memory.

### *4.2 Image Reconstruction*

This paper conducts image reconstruction tasks on four datasets. For the MNIST and CIFAR10 datasets, the experiments utilize their original resolutions of $28 \times 28$ pixels and $32 \times 32$ pixels, respectively. The CelebA dataset is first center-cropped to $140 \times 140$ pixels and then resized to $64 \times 64$ pixels. The LSUN Church dataset employs the official version and is split into training, validation, and test sets in an 8:1:1 ratio, with the image size adjusted to $128 \times 128$ pixels. The training process continues until the model converges completely. All data in the experiments undergo uniform standardization, with a mean of 0.5 and a variance of 0.5. To ensure fairness, the number of encoded features in the proposed method is consistent with the number of

latent vectors in the compared methods, which is one-quarter of the original resolution, corresponding to a downsampling rate of $f = 4$, applied to all datasets.

**Codebook Utilization.** The codebook utilization rates are calculated and visualized on the MNIST and CIFAR10 validation sets to verify the proposed method's effective improvement of the codebook. Specifically, the total number of times all entries in the codebook are utilized to reconstruct the validation set is counted. As shown in Fig. 3, the orange curve represents the fitted curve of the codebook entries sorted by usage rate, demonstrating that the codebook utilization rate of the proposed method is higher than both the baseline VQ-VAE and the state-of-the-art CVQ-VAE. In the codebook of VQ-VAE, a large portion of the entries remain unused or infrequently used, while a small number of entries are used very frequently, which is a result of the codebook collapse problem. Compared to VQ-VAE, CVQ-VAE addresses this issue by introducing anchor points to update unused or less frequently used codevectors, ensuring that all codevectors are utilized. However, there still exists a subset of codevectors that are used with high frequency. The proposed method demonstrates a more balanced usage of codebook entries compared to CVQ-VAE, indicating a more effective utilization of the codebook.

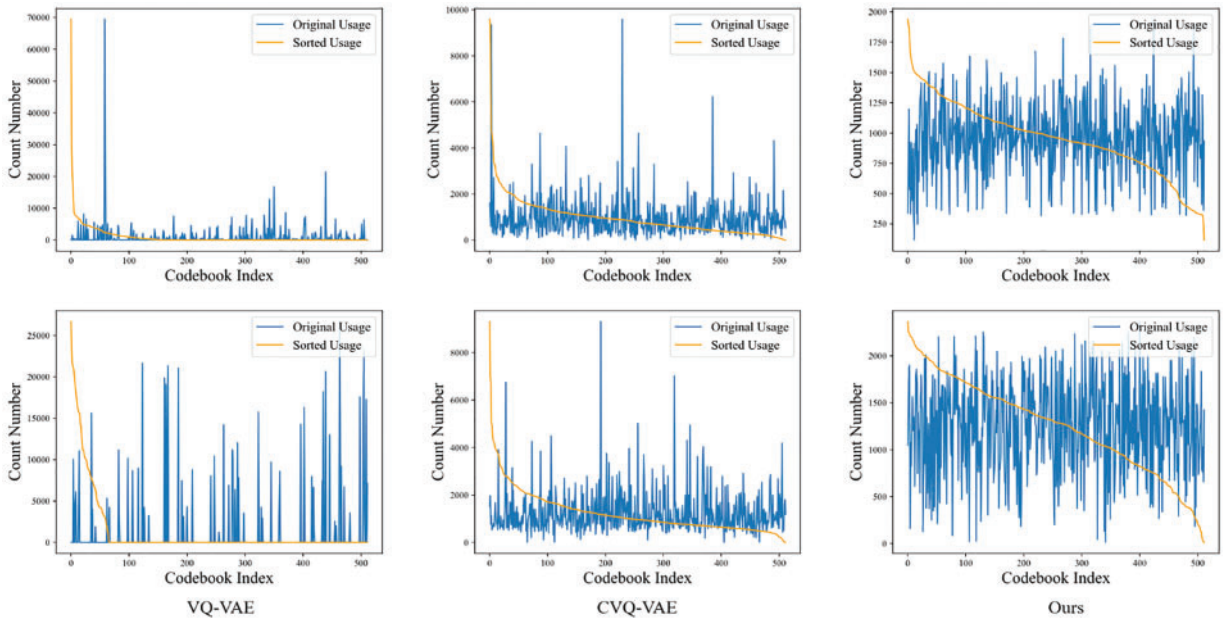


**Figure 3:** Visualization of Codebook Utilization on MNIST (top) and CIFAR-10 (bottom) validation sets. The blue line indicates the usage quantity of different code vectors in the codebook. The orange curve represents the fitted curve of the codebook entries sorted by usage frequency. In VQ-VAE, a large portion of the codebook entries remain unused due to codebook collapse. CVQ-VAE improves this issue by increasing codebook utilization. Our method further ensures a more balanced usage of codebook entries, making the codebook more efficient

The more uniform distribution of codevector usage in the proposed method suggests that it effectively captures a wider range of essential features and structures from the input data. This is achieved through the frequency-domain quantization, which preserves global relationships and extracts more informative features, as well as the dynamic codebook optimization, which prevents the dominance of a small subset of codevectors. As a result, the codebook in the proposed method is more efficient and better equipped to represent the complex characteristics of the data, leading to improved reconstruction and generation performance.

**Unit Information Capacity of Codebook.** Previous studies did not include calculations of the information capacity of individual codebook entries. Referring to the study in [9], this paper calculates the information capacity of codebook vectors by single codevector and measuring the change in mean squared error (MSE) loss between the generated image and the original image before and after the removal. By evaluating the MSE loss difference at the dataset level, the information capacity of individual code vectors is quantified. This metric reflects the effectiveness and compactness of the codebook's information. The measurement results on the CelebA and LSUN Church datasets are presented in Fig. 4, showing that the average information capacity per codebook entry generated by the proposed method is higher than that of prior methods. This demonstrates that the frequency domain features introduced in this paper carry more effective information than traditional image features obtained via spatial domain partitioning, resulting in a more compact and efficient codebook.

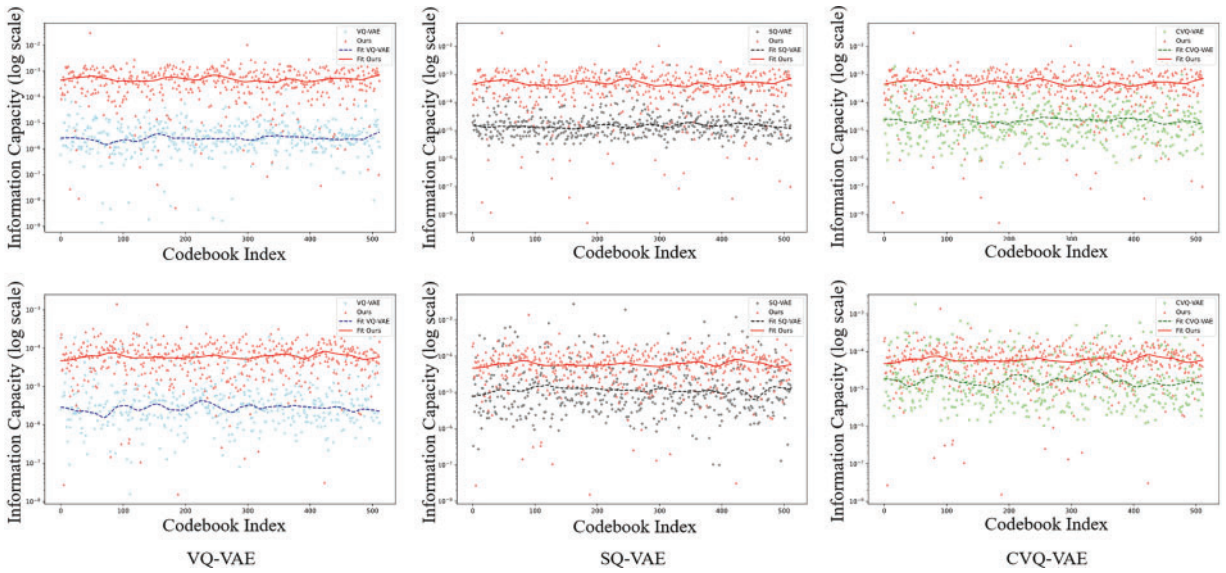


**Figure 4:** Visualization of Codebook Unit Information Comparison on CelebA (top) and LSUN Church (bottom) validation sets. The red dots and curve represent the information content of codebook entries and their fitting curve in the proposed method, respectively. Blue, gray, and green indicate the measurement results of VQ-VAE, SQ-VAE, and CVQ-VAE, respectively. Compared to other methods, the codebook generated by our approach exhibits a higher average information per entry, indicating less information coupling within the codebook. This suggests that our method captures richer and more compact data features in the codebook

**Quantitative Evaluation.** Table 1 presents the quantitative comparison results of image reconstruction across four publicly available datasets: MNIST, CIFAR10, CelebA, and LSUN Church. The results are compared with state-of-the-art quantized autoencoder models, including VQ-VAE, HVQ-VAE, SQ-VAE, and CVQ-VAE. By quantizing frequency domain features, optimizing the codebook structure, and refining the quantization of features, the proposed method significantly enhances image reconstruction quality, outperforms the four baseline models across all performance metrics on all datasets (with the exception of SSIM on MNIST and CelebA). Using the CIFAR-10 dataset as an example, the proposed method surpasses all baselines across all metrics. Its SSIM (0.9012), LPIPS (0.1597), and rFID (18.88) scores are 0.0034, 0.0376, and 6.01 points better than the best-performing baseline, CVQ-VAE, respectively. It is worth noting that in experiments on the CelebA and LSUN Church datasets, the rFID scores of CVQ-VAE increased by 0.43 and 1.00 compared to the baseline VQ-VAE. This can be attributed to the increased information coupling as

the dataset resolution increases, which affects the strategy of updating the codebook solely through anchor points in CVQ-VAE. The proposed method avoids the issue of information coupling through its codebook optimization strategy that reduces dependencies, gaining a significant advantage. It outperforms the second-best method, SQ-VAE, by 15.1% and 18.8% on the rFID metric for the CelebA and LSUN Church datasets, respectively. This demonstrates that the model effectively captures more valuable information from the original images.

**Table 1:** Reconstruction results on CIFAR-10, MNIST, CelebA, and LSUN Church validation sets

| Method | MNIST | | | CIFAR10 | | | CelebA | | | LSUN Church | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | SSIM↑[1] | LPIPS↓[2] | rFID↓ | SSIM↑ | LPIPS↓ | rFID↓ | SSIM↑ | LPIPS↓ | rFID↓ | SSIM↑ | LPIPS↓ | rFID↓ |
| VQ-VAE [6] | 0.9776 | 0.0291 | 3.51 | 0.8595 | 0.2510 | 40.03 | 0.9481 | 0.0965 | 6.26 | 0.9065 | 0.1680 | 8.86 |
| HVQ-VAE [10] | 0.9789 | 0.0268 | 3.15 | 0.8556 | 0.2546 | 41.08 | 0.9484 | 0.0963 | 6.25 | 0.9087 | 0.1668 | 7.90 |
| SQ-VAE [11] | 0.9818 | 0.0243 | 3.21 | 0.8767 | 0.2305 | 37.98 | **0.9502** | 0.0882 | 4.43 | 0.9015 | 0.1642 | 6.93 |
| CVQ-VAE [12] | **0.9833** | 0.0215 | 1.87 | 0.8978 | 0.1973 | 24.89 | 0.9447 | 0.1006 | 6.69 | 0.8978 | 0.1789 | 9.86 |
| Ours | 0.9824 | **0.0183**[3] | **1.38** | **0.9012** | **0.1597** | **18.88** | 0.9500 | **0.0821** | **3.76** | **0.9119** | **0.1415** | **5.63** |

[1]Note: ↑ indicates that higher values correspond to better performance. [2]Note: ↓ indicates that lower values correspond to better performance. [3]Note: Best results are **bolded**.

**Qualitative Evaluation.** Fig. 5 presents some example reconstruction results trained on the LSUN Church dataset. We report reconstruction examples for VQ-VAE, HVQ-VAE, SQ-VAE, CVQ-VAE, and the proposed method, FQ-VAE. It can be observed that the proposed method achieves better visual quality and performs well even in challenging detailed areas.

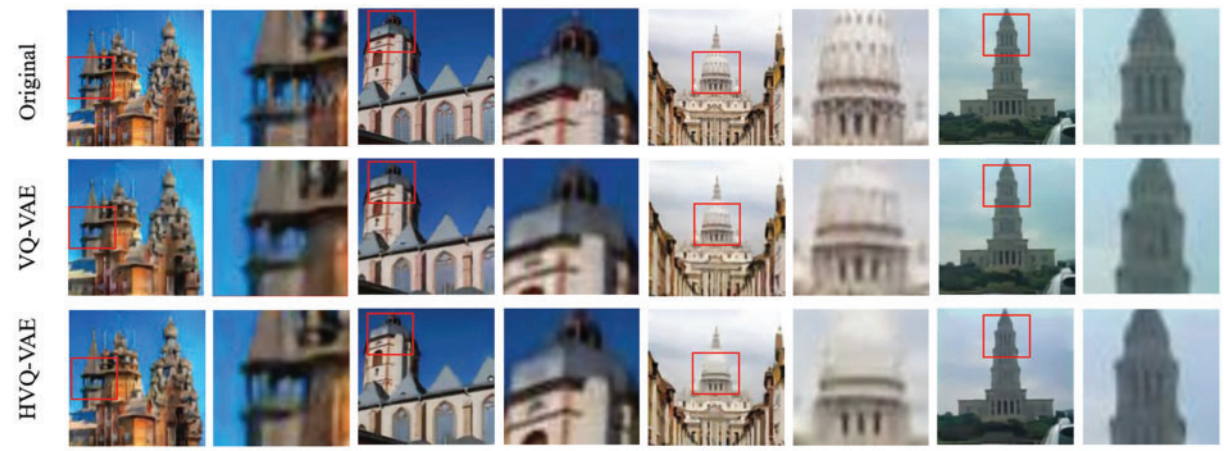


**Figure 5:** (Continued)

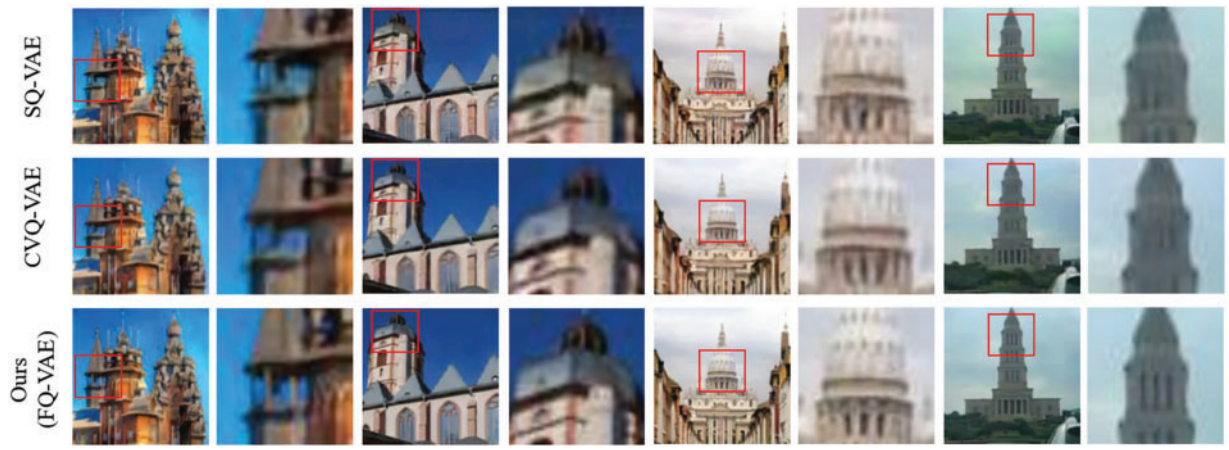


**Figure 5:** Qualitative comparison examples from LSUN Church validation set. The red box highlights areas in the image containing fine details that are difficult to reconstruct. The results demonstrate that our method achieves superior reconstruction performance compared to the baseline model in these regions

### *4.3 Ablation Experiments*

**Core Factors.** The core components of the proposed method are evaluated in Table 2, demonstrating that each designed components contributes significantly to the improvement of reconstruction quality. We begin with the baseline configuration (A) implementing VQ-VAE. In configuration (B), spatial vector quantization is replaced by frequency domain feature quantization, resulting in a notable performance boost, indicating that frequency domain quantization is the key driver behind the method's performance enhancement. Configurations (C) and (D) evaluate the codebook optimization methods. Configuration (C) reinitializes rarely used codevectors by selecting anchor points, yielding significant gains on small datasets such as MNIST and CIFAR-10, with rFID reductions of 29.21% and 33.50%, respectively. However, on larger datasets like CelebA and LSUN Church, the performance across all metrics slightly declines due to more severe information coupling. Configuration (D) considers codebook dependencies on top of (C), further reducing codebook information coupling and outperforming configuration (C) on various datasets. For example, on LSUN Church, configuration (D) improves SSIM, LPIPS, and rFID by 1.01%, 9.67%, and 25.05%, respectively, compared to configuration (C), highlighting the importance of addressing the information coupling issue in the codebook. Configuration (E) employs a Graph Convolutional Module to refine the quantized features and reduce information loss. Compared to the baseline configuration (A), on the MNIST dataset, SSIM improves by 0.17%, while LPIPS and rFID decrease by 9.97% and 12.82%, respectively, demonstrating the effectiveness of the Graph Convolutional Module. Similar improvements are observed on other datasets.

**Table 2:** Evaluation of core component contributions on CIFAR-10, MNIST, CelebA, and LSUN Church validation sets

| Method | MNIST | | | CIFAR10 | | | CelebA | | | LSUN Church | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | SSIM↑ | LPIPS↓ | rFID↓ | SSIM↑ | LPIPS↓ | rFID↓ | SSIM↑ | LPIPS↓ | rFID↓ | SSIM↑ | LPIPS↓ | rFID↓ |
| (A) | 0.9776 | 0.0291 | 3.51 | 0.8595 | 0.2510 | 40.03 | 0.9481 | 0.0965 | 6.26 | 0.9065 | 0.1680 | 8.86 |
| (B) | 0.9791 | 0.0225 | 2.08 | 0.8856 | 0.2017 | 28.10 | 0.9467 | 0.0902 | 5.31 | 0.9096 | 0.1587 | 6.88 |
| (C) | 0.9817 | 0.0236 | 2.23 | 0.8991 | 0.1897 | 26.62 | 0.9445 | 0.1006 | 6.69 | 0.8978 | 0.1789 | 9.86 |
| (D) | 0.9818 | 0.0229 | 2.28 | 0.8995 | 0.1890 | 24.93 | 0.9487 | 0.0913 | 5.44 | 0.9069 | 0.1616 | 7.39 |

(Continued)

**Table 2 (continued)**

| Method | MNIST | | | CIFAR10 | | | CelebA | | | LSUN Church | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | SSIM↑ | LPIPS↓ | rFID↓ | SSIM↑ | LPIPS↓ | rFID↓ | SSIM↑ | LPIPS↓ | rFID↓ | SSIM↑ | LPIPS↓ | rFID↓ |
| (E) | 0.9793 | 0.0262 | 3.06 | 0.8659 | 0.2335 | 34.14 | 0.9483 | 0.0921 | 5.50 | 0.9074 | 0.1628 | 7.88 |
| (F) | 0.9816 | 0.0199 | 1.62 | 0.8963 | 0.1760 | 21.77 | 0.9492 | 0.0854 | 4.62 | 0.9110 | 0.1495 | 6.01 |
| (G) | 0.9811 | 0.0207 | 1.95 | 0.8893 | 0.1916 | 25.41 | 0.9488 | 0.0874 | 4.89 | 0.9101 | 0.1516 | 6.23 |
| (H) | 0.9819 | 0.0213 | 1.99 | 0.9002 | 0.1776 | 22.52 | 0.9491 | 0.0858 | 4.77 | 0.9106 | 0.1500 | 6.12 |
| (I) | **0.9824** | **0.0183** | **1.38** | **0.9012** | **0.1597** | **18.88** | **0.9500** | **0.0821** | **3.76** | **0.9119** | **0.1415** | **5.63** |

Configuration (F) combines the frequency-domain quantization method with the codebook optimization method, achieving better results than individual components. Taking the CelebA dataset as an example, configuration (F) achieves an SSIM of 0.9492, LPIPS of 0.0854, and rFID of 4.62, improving by 0.25%, 5.32%, and 13.00% compared to configuration (B), and by 0.05%, 6.46%, and 15.07% compared to configuration (D), respectively. This indicates that codebook collapse and information coupling problems remain major influencing factors under frequency-domain quantization. Moreover, configuration (G) combines the codebook optimization method with the Graph Convolutional Module, while configuration (H) combines the frequency-domain quantization method with the Graph Convolutional Module. Both configurations outperform their respective individual components, suggesting that the interaction among these components can generate positive gains.

Configuration (I) integrates all the proposed components and, as expected, achieves the best performance across all metrics on all datasets. For instance, on the CIFAR-10 dataset, compared to the second-best configuration (F), configuration (I) improves SSIM, LPIPS, and rFID by 0.55%, 9.26%, and 13.28%, respectively. This fully demonstrates the effectiveness of the proposed components and the advantages of their synergistic collaboration.

**Balance between the Number of Channels and Filters.** The proposed method quantizes raw image data into a linear combination of frequency features across different channels, where the number of latent features $U$ is the product of the number of channels $D$ and the number of global filter bases $N$. To ensure experimental fairness, the number of quantized features is kept consistent with other baseline methods. Results from various combination experiments on MNIST and CIFAR10 are shown in Table 3. For the MNIST dataset, the configuration with D = 1 achieves the highest SSIM score of 0.9875. On the other hand, for the more complex CIFAR-10 dataset, the configuration with D = 2 yields the best performance, with an SSIM score of 0.9012 and an rFID score of 18.88. This indicates that for datasets with higher complexity and diversity, a moderate number of channels combined with an adequate number of filter bases is necessary to effectively represent the data.

**Table 3:** Evaluation of channel and filter count configurations on MNIST and CIFAR-10 validation sets

| Configuration ($D \times N$) | MNIST | | CIFAR10 | |
|---|---|---|---|---|
| | SSIM↑ | rFID↓ | SSIM↑ | rFID↓ |
| $1 \times U$ | **0.9875**[1] | 1.41 | 0.8987 | 20.51 |
| $2 \times U/2$ | 0.9824 | **1.38** | **0.9012** | **18.88** |
| $4 \times U/4$ | 0.9801 | 1.79 | 0.8990 | 19.01 |
| $8 \times U/8$ | 0.9816 | 2.35 | 0.8979 | 22.70 |
| $16 \times U/16$ | 0.9824 | 3.38 | 0.8869 | 30.29 |

[1]Note: Best results are bolded.

It is worth noting that increasing the number of channels beyond a certain point does not always lead to performance improvements. For example, in the CIFAR-10 experiments, the configurations with D = 4 and D = 8 result in lower SSIM scores and higher rFID scores compared to the optimal configuration with D = 2. This suggests that an excessive number of channels may introduce redundancy and hinder the model's ability to learn compact and informative representations. The proposed method achieves its best overall performance when D = 2, demonstrating its effectiveness in learning expressive and efficient discrete representations.

**Computational Efficiency Analysis.** Theoretically, for a feature map of size H × W with D channels, the time complexity of a 2D-FFT-based feature extractor is $\mathcal{O}(HWD\log_2(HW) + HWD)$, which is significantly lower than the $\mathcal{O}((HWD)^2)$ of a transformer. To verify this advantage, we use three different feature extractors in the proposed FQ-VAE architecture: (1) a non-global, convolution-based feature extractor, with its configuration derived from configuration (H) in Table 2, replacing the frequency-domain quantization operation with the convolutional quantization operation from VQ-VAE; (2) a global, transformer-based feature extractor, with parameter configurations mainly referencing ViT-VQGAN; (3) a global, 2D-FFT-based feature extractor, which is the final version proposed in this paper. We evaluate the impact of these three configurations on reconstruction quality (rFID), training time, throughput, and peak memory usage on four datasets: MNIST, CIFAR10, CelebA, and LSUN Church. Training time refers to the time required for the model to complete one training epoch, throughput refers to the number of images the model can process in one second, and peak memory usage refers to the maximum amount of memory consumed by the model during the testing process.

Table 4 shows the quantitative results of different feature extractors on the four datasets. From the perspective of reconstruction quality, the 2D-FFT-based feature extractor achieves the best rFID scores on all datasets, followed by the transformer, with the convolutional operation performing the worst. This indicates that by performing global modeling in the frequency domain, FQ-VAE can better capture the global structure and details of images, thus generating higher-quality reconstructed images. Taking the CIFAR-10 dataset as an example, the training time of 2D-FFT is 3.81 s per epoch, which is only 30.8% of the transformer's, while the throughput reaches 369.48 images per second, which is 1.81 times that of the transformer. This advantage is even more pronounced on the higher-resolution CelebA and LSUN Church datasets. It is worth noting that 2D-FFT does not introduce additional memory overhead while significantly reducing training time. Compared to the transformer, 2D-FFT has lower peak memory usage on all datasets. For example, on LSUN Church, the peak memory usage of 2D-FFT is only 456.94 MB, while that of the transformer is as high as 10,087.05 MB, with the former being only 4.5% of the latter. Overall, the 2D-FFT-based frequency-domain quantization method significantly improves reconstruction quality while avoiding the complexity

issues introduced by global feature extractors such as transformers, exhibiting significant computational efficiency advantages.

**Table 4:** Quantitative analysis results of computational efficiency

| Type | Dataset | rFID | TrainingTime (s/epoch) | Throughput (img/s) | Peak memory (MB) |
|---|---|---|---|---|---|
| + Convolution | **MNIST** | 1.99 | **2.13** | **374.90** | **150.93** |
| + Transformer | (28 × 28) | <u>1.79</u>[1] | 11.61 | 233.32 | 309.39 |
| + 2D-FFT | | **1.38**[2] | <u>3.36</u> | <u>371.77</u> | <u>161.69</u> |
| + Convolution | **CIFAR10** | 22.52 | **2.43** | **373.13** | **158.76** |
| + Transformer | (32 × 32) | <u>20.24</u> | 12.38 | 204.23 | 360.22 |
| + 2D-FFT | | **18.88** | <u>3.81</u> | <u>369.48</u> | <u>170.90</u> |
| + Convolution | **CelebA** | 4.77 | **25.99** | **51.62** | **210.72** |
| + Transformer | (64 × 64) | <u>4.01</u> | 194.50 | 35.87 | 978.84 |
| + 2D-FFT | | **3.76** | <u>39.51</u> | <u>48.81</u> | <u>238.01</u> |
| + Convolution | **LSUN Church** | 6.12 | **43.25** | **44.78** | **407.32** |
| + Transformer | (128 × 128) | <u>5.73</u> | 768.40 | 24.72 | 10,087.05 |
| + 2D-FFT | | **5.63** | <u>73.56</u> | <u>40.65</u> | <u>456.94</u> |

[1]Note: The second best results are <u>underlined.</u>
[2]Note: Best results are bolded.

### *4.4 Image Generation*

The model is trained on the CelebA dataset and applied to downstream generation tasks using the latent diffusion model (LDM) [42]. Since the proposed method is largely consistent with the traditional VQ-VAE architecture, it can replace the VQ-VAE component in LDM directly. Specifically, following the settings in LDM, a cosine learning rate scheduler is used in the generation tasks, with a warm-up period set to 500 steps, and the weight decay for the Adam optimizer is set to $1 \times 10^{-6}$. Additionally, the number of sampling steps is set to 1000, and mixed-precision training with fp16 is employed to enhance computational efficiency. The batch size is set to 64, and the model is trained for 100 epochs in total.

Fig. 6 presents several examples generated by the baseline CVQ-VAE and the proposed method. As expected, the images generated by the proposed method are sharper and more realistic compared to those generated through spatial vector quantization. Benefiting from frequency feature quantization, which preserves the global relationships within the image, the proposed method produces images with stronger overall coherence and avoids the distortions commonly seen in other methods, such as misaligned facial features, local ghosting, and facial warping.


**Figure 6:** Qualitative comparison examples of generation experiments between CVQ-VAE and the proposed method. The results indicate that the images generated by our method are more likely to preserve global relationships within the image, resulting in a more realistic overall appearance

## 5 Conclusion

This paper introduces a frequency-quantized variational autoencoder, which utilizes 2D-FFT to transform image features into the frequency domain for quantization, effectively preserving global image relationships. The method also dynamically optimizes the codebook by considering the frequency and dependency of code vector usage, preventing codebook collapse and reducing information coupling. Moreover, the quantized features are refined using graph convolutional theory to minimize quantization loss. Ablation experiments validate the effectiveness of each component, further confirming the key roles of frequency feature quantization, codebook optimization strategies, and corrected quantized features in enhancing model performance. Reconstruction and generation experiments show that the proposed method excels in image reconstruction and generation tasks, surpassing existing state-of-the-art approaches.

However, since each codebook entry in the proposed method maintains global information, the dimensionality of the codebook vectors increases dramatically as the image resolution rises. Unfortunately, downstream tasks often prefer lower-dimensional codevectors. This may hinder the scalability of the method on high-resolution datasets. Future work will explore more effective codebook compression techniques to reduce the codebook dimensionality while maintaining reconstruction quality, adapting to larger-scale data requirements and more downstream tasks.

Furthermore, maintaining global relationships within data has been a long-standing challenge in video and audio generation tasks. By addressing the above issues and adapting the method to capture temporal information for video data or 1D signals for audio data, the FQ-VAE could be extended to these modalities. Successful application of the proposed method to video and audio generation could lead to more coherent and globally consistent outputs, advancing the state-of-the-art in these domains.

**Acknowledgement:** The authors would like to express their gratitude for the valuable feedback and suggestions provided by all the anonymous reviewers and the editorial team.

**Funding Statement:** This work was supported by the Interdisciplinary project of Dalian University DLUXK-2023-ZD-001.

**Author Contributions:** The authors confirm contribution to the paper as follows: study conception and design: Jianxin Feng, Xiaoyao Liu; data collection: Jianxin Feng, Xiaoyao Liu; analysis and interpretation of results: Jianxin Feng, Xiaoyao Liu; draft manuscript preparation: Xiaoyao Liu. All authors reviewed the results and approved the final version of the manuscript.

**Availability of Data and Materials:** The data used in the study are all available on the OpenDataLab official website (https://opendatalab.com/) (accessed on 07 February 2025).

**Ethics Approval:** Not applicable.

**Conflicts of Interest:** The authors declare no conflicts of interest to report regarding the present study.

## References

1. Kingma DP, Welling M. Auto-encoding variational bayes. arXiv:1312.6114. 2013.
2. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. Commun ACM. 2020;63(11):139–44. doi:10.1145/3422622.
3. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017 Dec 4–9; Long Beach, CA, USA; 2017. Vol. 30, p. 3–19.
4. Esser P, Rombach R, Ommer B. Taming transformers for high-resolution image synthesis. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA. p. 12868–78. doi:10.1109/cvpr46437.2021.01268.
5. Gu S, Chen D, Bao J, Wen F, Zhan B, Chen D, et al. Vector quantized diffusion model for text-to-image synthesis. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 10696–706.
6. Oord A, Vinyals O, Kavukcuoglu K. Neural discrete representation learning. In: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017 Dec 4–9; Long Beach, CA, USA. p. 6309–18.
7. Chen S, Guo W. Auto-encoders in deep learning—a review with new perspectives. Mathematics. 2023;11(8):1777. doi:10.3390/math11081777.
8. Tian K, Jiang Y, Yuan Z, Peng B, Wang L. Visual autoregressive modeling: scalable image generation via next-scale prediction. arXiv:2404.02905. 2024.
9. Huang L, Qiu Q, Sapiro G. PQ-VAE: learning hierarchical discrete representations with progressive quantization. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2024 Jun 17–18; Seattle, WA, USA. p. 7550–8. doi:10.1109/CVPRW63382.2024.00750.
10. Takida Y, Shibuya T, Liao W, Lai CH, Ohmura J, Uesaka T, et al. SQ-VAE: variational bayes on discrete representation with self-annealed stochastic quantization. arXiv:2205.07547. 2022.

11. Williams W, Ringer S, Ash T, MacLeod D, Dougherty J, Hughes J. Hierarchical quantized autoencoders. In: NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems; 2020 Dec 6–12; Vancouver, BC, Canada. p. 4524–35.
12. Zheng C, Vedaldi A. Online clustered codebook. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV); 2023 Oct 1–6; Paris, France. p. 22741–50. doi:10.1109/ICCV51070.2023.02084.
13. Lee D, Kim C, Kim S, Cho M, Han WS. Autoregressive image generation using residual quantization. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 11513–22. doi:10.1109/CVPR52688.2022.01123.
14. Yan W, Zhang Y, Abbeel P, Srinivas A. VideoGPT: video generation using VQ-VAE and transformers. arXiv:2104.10157. 2021.
15. Xie Z, Li B, Xu X, Wu M, Yu K. Enhancing audio generation diversity with visual information. In: ICASSP 2024—2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2024 Apr 14–19; Seoul, Republic of Korea. p. 866–70. doi:10.1109/ICASSP48485.2024.10447384.
16. Grassucci E, Mitsufuji Y, Zhang P, Comminiello D. Enhancing semantic communication with deep generative models: an overview. In: ICASSP 2024—2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2024 Apr 14–19; Seoul, Republic of Korea. p. 13021–5. doi:10.1109/ICASSP48485.2024.10448235.
17. Mao C, Jiang L, Dehghani M, Vondrick C, Sukthankar R, Essa I. Discrete representations strengthen vision transformer robustness. arXiv:2111.10493. 2021.
18. Razavi A, Oord AVD, Vinyals O. Generating diverse high-fidelity images with VQ-VAE-2. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems; 2019 Dec 8–14; Vancouver, BC, Canada. p. 14866–76.
19. Yu J, Li X, Koh JY, Zhang H, Pang R, Qin J, et al. Vector-quantized image modeling with improved VQGAN. arXiv:2110.04627. 2021.
20. Zheng C, Vuong TL, Cai J, Phung D. Movq: modulating quantized vectors for high-fidelity image generation. In: NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems; 2022 Nov 28–Dec 9; New Orleans, LA, USA. p. 23412–225.
21. Dhariwal P, Jun H, Payne C, Kim JW, Radford A, Sutskever I. Jukebox: a generative model for music. arXiv:2005.00341. 2020.
22. Vuong TL, Le T, Zhao H, Zheng C, Harandi M, Cai J, et al. Vector quantized Wasserstein auto-encoder. arXiv:2302.05917. 2023.
23. Lee-Thorp J, Ainslie J, Eckstein I, Ontanon S. FNet: mixing tokens with Fourier transforms. arXiv:2105.03824. 2021.
24. Sevim N, Özyedek EO, Şahinuç F, Koç A. Fast-FNet: accelerating transformer encoder models via efficient Fourier layers. arXiv:2209.12816. 2022.
25. Rao Y, Zhao W, Zhu Z, Lu J, Zhou J. Global filter networks for image classification. In: NIPS'21: Proceedings of the 35th International Conference on Neural Information Processing Systems; 2021 Dec 6–14; Online. p. 980–93.
26. Rao Y, Zhao W, Zhu Z, Zhou J, Lu J. GFNet: global filter networks for visual recognition. IEEE Trans Pattern Anal Mach Intell. 2023;45(9):10960–73. doi:10.1109/TPAMI.2023.3263824.
27. Wang Z, Zhao Y, Chen J. Multi-scale fast Fourier transform based attention network for remote-sensing image super-resolution. IEEE J Sel Top Appl Earth Obs Remote Sens. 2023;16:2728–40. doi:10.1109/JSTARS.2023.3246564.
28. Tatsunami Y, Taki M. FFT-based dynamic token mixer for vision. Proc AAAI Conf Artif Intell. 2024;38(14):15328–36. doi:10.1609/aaai.v38i14.29457.
29. Hong D, Gao L, Yao J, Zhang B, Plaza A, Chanussot J. Graph convolutional networks for hyperspectral image classification. IEEE Trans Geosci Remote Sens. 2020;59(7):5966–78. doi:10.1109/TGRS.2020.3015157.
30. Xu Y, Huang D, Wang CD, Lai JH. Deep image clustering with contrastive learning and multi-scale graph convolutional networks. Pattern Recognit. 2024;146(2):110065. doi:10.1016/j.patcog.2023.110065.
31. Chen Z, Fu L, Yao J, Guo W, Plant C, Wang S. Learnable graph convolutional network and feature fusion for multi-view learning. Inf Fusio. 2023;95(7):109–19. doi:10.1016/j.inffus.2023.02.013.

32. Cai W, Wei Z. Remote sensing image classification based on a cross-attention mechanism and graph convolution. IEEE Geosci Remote Sens Lett. 2020;19:1–5.
33. Yang Y, Qi Y, Qi S. Relation-consistency graph convolutional network for image super-resolution. Vis Comput. 2024;40(2):619–35. doi:10.1007/s00371-023-02805-1.
34. Li Y, Zhang Y, Cui W, Lei B, Kuang X, Zhang T. Dual encoder-based dynamic-channel graph convolutional network with edge enhancement for retinal vessel segmentation. IEEE Trans Med Imaging. 2022;41(8):1975–89. doi:10.1109/TMI.2022.3151666.
35. Xiang X, Wang Z, Zhang J, Xia Y, Chen P, Wang B. AGCA: an adaptive graph channel attention module for steel surface defect detection. IEEE Trans Instrum Meas. 2023;72:5008812. doi:10.1109/TIM.2023.3248111.
36. LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE. 2002;86(11):2278–324. doi:10.1109/5.726791.
37. Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images. In: Technical report. Toronto, ON, Canada: University of Toronto; 2009.
38. Liu Z, Luo P, Wang X, Tang X. Deep learning face attributes in the wild. In: 2015 IEEE International Conference on Computer Vision (ICCV); 2015 Dec 7–13; Santiago, Chile. p. 3730–8. doi:10.1109/ICCV.2015.425.
39. Yu F, Seff A, Zhang Y, Song S, Funkhouser T, Xiao J. LSUN: construction of a large-scale image dataset using deep learning with humans in the loop. arXiv:1506.03365. 2015.
40. Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The unreasonable effectiveness of deep features as a perceptual metric. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. p. 586–95. doi:10.1109/CVPR.2018.00068.
41. Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017 Dec 4–9; Long Beach, CA, USA. p. 6629–40.
42. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-resolution image synthesis with latent diffusion models. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 10674–85. doi:10.1109/CVPR52688.2022.01042.