

Decoupling High and Low Frequencies for Faithful Image Generation with Fine Details

Tejaswini Medi¹, Hsien-Yi Wang¹, Arianna Rampini², and Margret Keuper^{1,3}

¹ University of Mannheim, Germany

² Autodesk AI Lab

³ MPI for Informatics, Saarland Informatics Campus, Saarbrücken, Germany
`{tejaswini.medi,keuper}@uni-mannheim.de`

Abstract. Latent generative models compress images into learned embeddings prior to synthesis, and the generation quality critically depends on how faithfully these embeddings preserve visual detail. We observe that while such embeddings are effective at reconstructing low frequency structure, they struggle to recover sharp high frequency details that are essential for perceptual realism. Conventional reconstruction objectives implicitly prioritize coarse structural information over high frequency content, which can lead to overly smoothed outputs and degraded visual quality in textured regions. Motivated by this observation, we propose DeBaT, a Decoupled frequency Band Tokenizer that explicitly separates the learning of low and high frequency band embeddings. This decoupling enables accurate reconstruction of fine details while preserving global coherence. Integrated into a latent diffusion based generative model, DeBaT allows for sharper and more realistic samples than previous latent tokenizers, confirming that the explicit decoupling of high and low frequency bands eases the preservation of visual details in learned embedding spaces.

Keywords: Latent Tokenizer · High-Frequency Details

1 Introduction

Latent generative modeling [4, 9, 27, 61] has become a central paradigm in modern content generation, enabling the synthesis of high fidelity visual data. These models typically operate in compressed latent spaces learned through variational autoencoders such as VAEs [22, 25], where generation quality depends critically on the expressiveness and reconstruction fidelity of the latent representations. Increasing latent dimensionality can enhance representational capacity, but often yields diminishing visual generation improvements together with significantly higher computational cost. VAVAE [51] highlights this as an optimization dilemma inherent in latent modeling: pixel level reconstruction objectives favor faithful reconstruction as capacity increases, whereas generative objectives require semantically structured latent spaces. By aligning latents with pretrained



Fig. 1: Reconstruction quality for two examples. We compare the reconstruction of VA-VAE to our approach DeBaT. Highlighted regions emphasize textures and edges. DeBaT better preserves high-frequency details and sharp structures, producing reconstructions visually closer to the input. Arrows indicate reconstruction errors.

vision foundation models, VAVAE improves convergence and semantic realism. However, despite these advances, VAVAE often lacks to faithfully reconstruct sharp textures and fine details, particularly in regions dominated by high frequency content such as text, as shown in Figure 1. In this work, we propose a simple yet effective approach to learn detail preserving latent embeddings, Decoupled frequency Band Tokenizer (DeBaT).

Our proposed approach is motivated by a comprehensive frequency based analysis of the reconstruction behavior of state-of-the-art latent tokenizers, which serve as the foundation of modern latent generative pipelines [25, 43, 46, 47, 51]. For this analysis, we focus on latent diffusion and autoregressive models, which represent the current frontier of high quality visual generation. These methods typically follow a two stage design: a latent tokenizer, often implemented as a VAE variant, learns quantized or continuous embeddings, and a probabilistic generative model is subsequently trained within this latent space.

Our spectral analysis uncovers a consistent limitation. While different tokenizers vary in their ability to reconstruct low frequency structure, their high frequency reconstruction error remains strikingly similar across architectures, *i.e.* they become ever better at encoding semantic properties and global coherence. At the same time, high frequency components, which encode perceptually critical information such as textures and fine structural details and are therefore equally important for perceptual realism, are systematically under-reconstructed. This reveals a fundamental imbalance in current latent tokenizers: they are optimized for structural faithfulness but not for fine detail preservation.

To faithfully encode fine details, *i.e.* high-frequency information, our proposed DeBaT (Decoupled frequency Band Tokenizer) explicitly separates and independently optimizes low and high frequency components via wavelet decomposition. Unlike conventional tokenizers that operate purely in the spatial domain, DeBaT performs an explicit frequency band decomposition and learns distinct latent embeddings for low and high frequency components. This enables faithful reconstruction of fine textures while preserving global structural coherence. Figure 1 illustrates a visual comparison with VA-VAE, demonstrating improved preservation of sharp structures and high frequency details. By integrating DeBaT into a lightweight latent diffusion model, LightningDiT, we

observe that improved spectral reconstruction directly translates into high quality realistic generative samples.

Our main contributions are summarized as follows:

- ✓ We present a comprehensive frequency-based reconstruction evaluation of latent tokenizers used in modern generative pipelines. Our analysis reveals that improvements in overall reconstruction performance are primarily driven by enhanced low-frequency reconstruction, while high-frequency details remain under-preserved across both spatial and spectral latent tokenizers.
- ✓ We propose DeBaT, a Decoupled frequency Band Tokenizer that explicitly separates and independently optimizes low- and high-frequency components via wavelet decomposition. By mitigating spectral optimization coupling, DeBaT learns frequency-aware and more expressive latent representations that preserve both global structure and fine details.
- ✓ We integrate DeBaT into LightningDiT and demonstrate improved generative quality on ImageNet. Our results highlight the importance of frequency-aware latent modeling for achieving sharper and more perceptually realistic image reconstruction and synthesis.

2 Related Work

2.1 Latent Embeddings for Image Generation

Latent representation learning models, or visual tokenizers, are widely used to learn compressed latent representations that capture high-dimensional visual data and enable efficient generation. Encoder-Decoder frameworks such as VAEs [22] and their variants form the foundation for learning these latent representations, which are vital for modeling latent generative models. Notably, Latent Diffusion Models (LDMs) [37] operate over learned latent embeddings to facilitate high-quality synthesis, highlighting the importance of effective latent space modeling. Discrete tokenization techniques, such as VQ-VAE [30, 45], convert images into latent code sequences that can be effectively modeled by transformers. VQGAN [1] further improves this framework by incorporating adversarial losses to learn perceptually meaningful tokens, while RQ-VAE [24] enhances the expressiveness of embeddings through stacked residual quantizers. Recent methods such as TiTok [53], FlexTok [2], and OneDPiece [31] explore compact and adaptive tokenization strategies, aiming for better compression and controllable generation quality.

To enhance reconstruction quality of visual tokenizers, hierarchical and regularized latent representation learning models have been proposed. NVAE [44] introduces spectral regularization within residual hierarchies for stable training, while HQ-VAE [42] and VAR [43] use multi-scale vector quantization to better model local and global structure in latent spaces. EQ-VAE [23] imposes equivariance constraints for robustness to geometric changes. Additionally, frequency-aware losses such as Fourier-based regularization [3] and scale-equivariant decod-

ing [39] have been used to control the spectral properties of the latent reconstructions. A parallel direction focuses on aligning latent spaces with external vision encoders. For example, REPA [54] introduces denoising-time regularization based on foundation model features to accelerate convergence. VA-VAE [51] builds on this by jointly aligning tokenizer latents with pretrained features and optimizing for downstream generation quality. ReaLS [48] introduces semantic priors directly into the latent space to improve representation capacity. Our work builds on VA-VAE but identifies that such alignment alone does not sufficiently preserve high frequency fine details. We propose an explicit decoupling and optimization of low- and high-frequencies during tokenization, leading to improved fidelity in both latent space representation, and by consequence in reconstruction and generation quality.

2.2 High-Frequency Content in Image Reconstruction & Synthesis

High-frequency content is essential for perceptual realism but challenging to reconstruct, particularly for diffusion-based models. Prior analyses show that frequency components are not reconstructed uniformly across the denoising trajectory [49], resulting in a loss of textures and sharp structures. To address this, various frequency-aware approaches have been developed across generative paradigms. In GANs, SWAGAN [10] and WaveFill [55] integrate wavelet transforms into their architectures to improve generation fidelity. Focal frequency loss [19] and WINE [20] address frequency imbalance by reweighting reconstruction errors and correcting low-frequency bias in inversion, respectively. Diffusion models have incorporated frequency priors at both training and inference time. WaveDiff [35], Spatial-Frequency UNet [56], and spectral diffusion approaches [13, 49] integrate wavelet-domain frequency components or gating mechanisms into the network to better optimize the spectral content. Training-free techniques like DiffuseHigh [21] and frequency aggregation [36] inject wavelet-based frequency priors during inference for improved detail preservation. Additional strategies include replacing convolutions with filtered variants in FouriScale [15], and autoregressive frequency modeling in NFIG [16]. Adaptations also exist for domain-specific tasks such as high-resolution [57], underwater [59], or low-light image generation [18]. Beyond images, wavelet representations have further proven useful in 3D generation with diffusion and autoregressive models [17, 29], suggesting the utility of wavelet decompositions as a general tool for capturing fine-scale structure across visual modalities. While effective, these approaches typically introduce additional architectural components or post-processing steps to compensate for frequency degradation.

Prior wavelet based decomposition approaches such as WF-VAE [26] (for video generation) and LiteVAE [38] incorporate wavelet transforms into the encoding and decoding process to enhance multi scale feature extraction and improve reconstruction efficiency. Leveraging the compression properties of wavelets, these methods optimize frequency components jointly within a shared latent space and under a unified objective. As a result, they inherit the same frequency optimization coupling observed in spatial domain tokenizers, where low

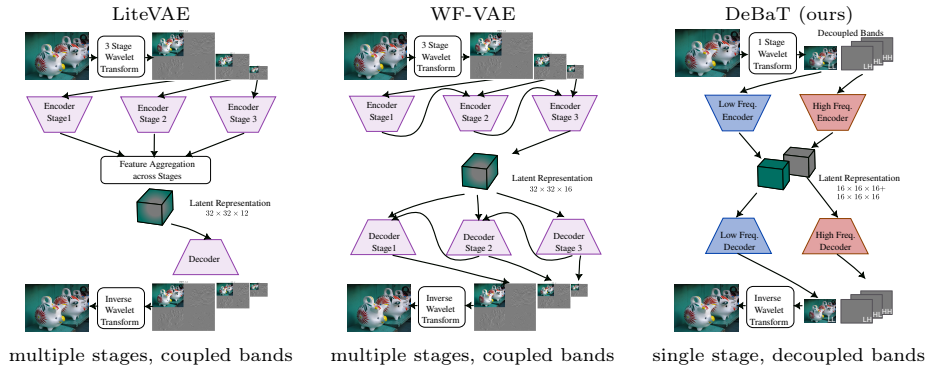


Fig. 2: Comparison of wavelet based latent VAE tokenizer designs. Previous works for image (VA-VAE) and video (WF-VAE) generation leverage spatial compression properties of multiple wavelet stages, while low and high frequency bands are coupled in a joint latent code. We propose DeBaT, which preserves fine details by learning explicit latent codes for low and for high frequency bands through decoupled VAEs.

frequency components dominate over the learning of fine detail. In contrast, our approach DeBaT explicitly decouples the learning dynamics of low and high frequency bands by allocating separate latent embeddings and optimizing them independently prior to fusion. Figure 2 compares DeBaT to different frequency based latent tokenizer models that use wavelet decomposition. By modeling and optimizing frequency subbands (low and high) separately during tokenization, we improve the frequency aware expressivity in the latent embeddings. These latent embeddings improve both reconstruction fidelity and downstream synthesis quality when integrated with the latent generative pipeline without requiring any architectural changes at generation time.

3 Method

In this section, we present the key components of our approach. First, we conduct a frequency-based reconstruction evaluation of latent embeddings used in modern latent image generative pipelines. Our analysis shows that improvements in reconstruction performance primarily arise from better modeling of low-frequency structures, while high-frequency components remain comparatively underrepresented. This imbalance indicates that latent tokenizers jointly optimized over both low- and high-frequency information tend to bias the optimization toward low-frequency content, often resulting in the loss of fine-grained image details. Second, we introduce DeBaT, a novel framework that leverages wavelet decomposition to decouple latent representations into low- and high-frequency components, enabling them to be optimized independently. Third, we propose a simple yet effective fusion strategy that combines these frequency-aware representations into a unified latent embedding, which is subsequently

integrated into the lightweight LightningDiT pipeline [51]. This design enables improved image generation by explicitly modeling frequencies in the latent space. An overview of the proposed DeBaT framework is shown in Figure 3.

3.1 Frequency Reconstruction Evaluation of Latent Tokenizers

In most latent generative models, the first stage involves learning compact latent representations using visual tokenizers such as VAEs or VQ-VAEs and their architectural variants. These tokenizers learn latent embeddings $\mathbf{z} \in \mathcal{Z}$ that encode the underlying structure of the input data $\mathbf{x} \in \mathcal{X}$. A common measure of the representational quality of these latent embeddings is their reconstruction performance, typically evaluated using mean squared error (MSE) or the ℓ_2 -reconstruction loss. The reconstruction objective is defined as:

$$\mathcal{L}_{\text{rec}} = \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2,$$

where $\hat{\mathbf{x}} = D(E(\mathbf{x}))$ denotes the reconstructed image obtained by encoding the input $\mathbf{x}$ into a latent embedding $\mathbf{z} = E(\mathbf{x})$ with encoder E and subsequently decoding it using the decoder D . While this objective provides a measure of overall reconstruction quality, it does not explicitly evaluate how well different frequency components of the image are reconstructed. In particular, the reconstruction error does not distinguish between low-frequency structural information and high-frequency details such as edges and textures. As a result, improvements in reconstruction performance may primarily reflect better recovery of dominant low-frequency content, while the fidelity of high-frequency components remains underexplored. To assess the fidelity of reconstructions across frequency bands, we perform a frequency-aware reconstruction analysis using the discrete wavelet transform (DWT) with a Haar wavelet filter. Specifically, we decompose both the input image and its reconstruction into low- and high-frequency representations. Let $\mathcal{W}(\cdot)$ denote the wavelet decomposition operator:

$$\mathcal{W}(\mathbf{x}) = (\mathbf{x}_L, \mathbf{x}_H), \quad \mathcal{W}(\hat{\mathbf{x}}) = (\hat{\mathbf{x}}_L, \hat{\mathbf{x}}_H),$$

where $\mathbf{x}_L$ and $\hat{\mathbf{x}}_L$ correspond to the low-frequency (approximation) components, while $\mathbf{x}_H$ and $\hat{\mathbf{x}}_H$ correspond to the high-frequency (detail) components of the input and reconstructed images, respectively. We then compute frequency-specific reconstruction losses as follows:

$$\mathcal{L}_L = \|\mathbf{x}_L - \hat{\mathbf{x}}_L\|_2^2, \quad \mathcal{L}_H = \|\mathbf{x}_H - \hat{\mathbf{x}}_H\|_2^2.$$

These metrics provide a more fine-grained evaluation of latent representations by measuring their ability to reconstruct structural information and fine details independently. Unlike conventional reconstruction objectives that aggregate errors across all spatial frequencies into a single pixel-space metric, our analysis explicitly separates the contribution of low- and high-frequency components. This allows us to reveal frequency-specific reconstruction in latent tokenizers that remain hidden under standard pixel based reconstruction losses. In particular, it enables a more principled understanding of whether improvements in

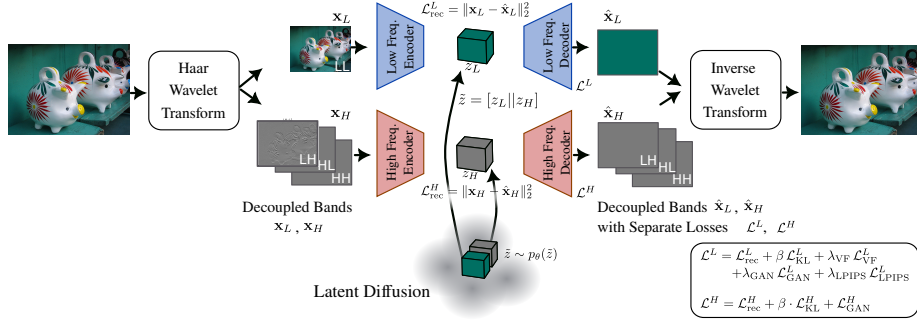


Fig. 3: Overview of the proposed **DeBaT** framework. An input image x is first decomposed using a discrete wavelet transform into low- and high-frequency components. Each frequency band is processed independently using dedicated encoder-decoder pairs (E_L, D_L) and (E_H, D_H) to learn specialized latent embeddings that capture global structure and fine details, respectively. The reconstructed frequency components are then fused through an inverse wavelet transform to obtain the final image $\hat{x}$.

reconstruction quality arise from better modeling of global structures or from enhanced preservation of high-frequency details.

Beyond pixel-wise reconstruction loss, we also evaluate perceptual quality using the Learned Perceptual Image Patch Similarity (LPIPS) metric [58]. LPIPS measures perceptual similarity between input and reconstructed images using deep feature representations and better reflects human visual perception than pixel-space metrics. To further quantify the overall quality of latent reconstructions, we compute the reconstruction Fréchet Inception Distance (rFID) [14]. This metric compares the distributions of original and reconstructed images in the feature space of a pretrained Inception network [41].

$$\text{rFID} = \|\mu_{\mathbf{x}} - \mu_{\hat{\mathbf{x}}}\|_2^2 + \text{Tr}\left(\Sigma_{\mathbf{x}} + \Sigma_{\hat{\mathbf{x}}} - 2(\Sigma_{\mathbf{x}}\Sigma_{\hat{\mathbf{x}}})^{1/2}\right),$$

where $(\mu_{\mathbf{x}}, \Sigma_{\mathbf{x}})$ and $(\mu_{\hat{\mathbf{x}}}, \Sigma_{\hat{\mathbf{x}}})$ denote the mean and covariance of Inception features extracted from the original and reconstructed image distributions, respectively. Together, the frequency-aware reconstruction losses $(\mathcal{L}_L, \mathcal{L}_H)$, perceptual similarity $\mathcal{L}_{\text{LPIPS}}$, and distributional alignment measured by rFID provide a comprehensive evaluation of reconstruction quality. In particular, this combination enables us to analyze both the perceptual fidelity and the frequency-specific reconstruction behavior of latent embeddings.

3.2 Decoupled Band Tokenizer (DeBaT)

To enable frequency-aware optimization in latent tokenizers, we extend the standard VAE-based tokenizer by operating directly on frequency-decomposed representations instead of pixel-space inputs. Specifically, we introduce **DeBaT**, a Decoupled Band Tokenizer that explicitly separates low- and high-frequency components of the input image and learns dedicated latent embeddings for each

band. Given an input image $\mathbf{x} \in \mathcal{X}$, we apply a discrete wavelet transform $\mathcal{W}(\cdot)$ using the Haar filter to obtain its low and high-frequency representations. $\mathcal{W}(\mathbf{x}) = (\mathbf{x}_L, \mathbf{x}_H)$, where $\mathbf{x}_L$ and $\mathbf{x}_H$ denote the low- and high-frequency components, respectively. We adopt the normalization strategy proposed in [32] to normalize both components prior to encoding.

To model low and high frequency specific information, we employ two independent encoder-decoder pairs, (E_L, D_L) and (E_H, D_H) , each responsible for learning latent embeddings $(\mathbf{z}_L, \mathbf{z}_H)$ corresponding to the low- and high-frequency inputs. This allows to capture the full frequency spectrum more effectively than standard VAEs, where low and high spatial frequencies are optimized jointly. To ensure a fair comparison with baseline VAEs, we allocate the latent dimensions of the low and high frequencies such that the concatenated latent vector has the same dimensionality as that of the standard baseline VAE, i.e. $16 \times 16 \times 32$ (see Figure 2),

$$\mathbf{z}_L = E_L(\mathbf{x}_L), \quad \mathbf{z}_H = E_H(\mathbf{x}_H),$$

where $\mathbf{z}_L$ and $\mathbf{z}_H$ represents the low and high frequency latent embeddings.

Low Frequency Objective. To learn latent embeddings that capture the global structural information of images, we optimize the low-frequency component using the VAVAE training objective [51]. VAVAE has demonstrated strong performance among latent tokenizers; it improves reconstruction fidelity through alignment with vision foundation models and enhanced perceptual supervision.

In our framework, the low-frequency representation $\mathbf{x}_L$ obtained from the wavelet decomposition is encoded into a latent embedding $\mathbf{z}_L$, which is reconstructed through the decoder while being optimized under a combination of complementary objectives. Specifically, the training objective consists of the standard VAE reconstruction loss $\mathcal{L}_{\text{rec}}^L$ and KL regularization $\mathcal{L}_{\text{KL}}^L$, together with additional supervision terms introduced in VAVAE [51], including a vision foundation alignment loss $\mathcal{L}_{\text{VF}}^L$, an adversarial regularization objective $\mathcal{L}_{\text{GAN}}^L$ inspired by VQGAN [1], and a perceptual similarity loss $\mathcal{L}_{\text{LPIPS}}^L$ [58]. The complete low-frequency objective is given by $\mathcal{L}^L$.

$$\mathcal{L}^L = \mathcal{L}_{\text{rec}}^L + \beta \mathcal{L}_{\text{KL}}^L + \lambda_{\text{VF}} \mathcal{L}_{\text{VF}}^L + \lambda_{\text{GAN}} \mathcal{L}_{\text{GAN}}^L + \lambda_{\text{LPIPS}} \mathcal{L}_{\text{LPIPS}}^L.$$

where,

$$\begin{aligned} \mathcal{L}_{\text{rec}}^L &= \mathbb{E}_{q(\mathbf{z}_L | \mathbf{x}_L)} \left[\|\mathbf{x}_L - D_L(\mathbf{z}_L)\|_2^2 \right], \\ \mathcal{L}_{\text{KL}}^L &= D_{\text{KL}}(q(\mathbf{z}_L | \mathbf{x}_L) \| p(\mathbf{z}_L)), \end{aligned}$$

where $\mathcal{L}_{\text{rec}}^L$ is the low-frequency reconstruction loss, $\mathcal{L}_{\text{KL}}^L$ is the KL divergence between the encoder’s approximate posterior $q(\mathbf{z}_L | \mathbf{x}_L)$ and the Gaussian prior $p(\mathbf{z}_L) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. The vision foundation loss $\mathcal{L}_{\text{VF}}^L$ aligns the latent codes with the feature space of a pretrained foundation model (DINOv2 [33]). The adversarial loss $\mathcal{L}_{\text{GAN}}^L$ introduces a discriminator to distinguish real from reconstructed low-frequency inputs. Finally, the perceptual loss $\mathcal{L}_{\text{LPIPS}}^L$ encourages perceptual

similarity between input and reconstruction using features from a pretrained Inception network [41]. All these loss components are commonly used in modern VAE frameworks. This overall objective encourages the latent representation to preserve global structure through reconstruction and KL regularization, while perceptual and adversarial supervision with vision foundation alignment loss improve visual fidelity and semantic alignment of the reconstructed images.

High-Frequency Objective. In contrast to low frequency optimization, high frequency representations are optimized without any supervision from pretrained models, as these pretrained models tend to be trained jointly with low-and high spatial frequencies. To mitigate any external biases, we simplify our training objective focusing on reconstruction objective along with KL and adversarial regularization. The overall high-frequency objective is given by $\mathcal{L}^H$.

$$\mathcal{L}^H = \mathcal{L}_{\text{rec}}^H + \beta \cdot \mathcal{L}_{\text{KL}}^H + \mathcal{L}_{\text{GAN}}^H,$$

where,

$$\mathcal{L}_{\text{rec}}^H = \mathbb{E}_{q(\mathbf{z}_H | \mathbf{x}_H)} [\|\mathbf{x}_H - D_H(\mathbf{z}_H)\|_2],$$

$$\mathcal{L}_{\text{KL}}^H = D_{\text{KL}}(q(\mathbf{z}_H | \mathbf{x}_H) \| p(\mathbf{z}_H))$$

$\mathcal{L}_{\text{rec}}^H$ is the high-frequency reconstruction loss, $\mathcal{L}_{\text{KL}}^H$ is the KL divergence between the encoder’s approximate posterior $q(\mathbf{z}_H | \mathbf{x}_H)$ and the Gaussian prior $p(\mathbf{z}_H) = \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Inference. At inference time, both decoders (D_L, D_H) are used to reconstruct the respective frequency bands, and the final image is synthesized via inverse wavelet transform:

$$\hat{\mathbf{x}}_L = D_L(\mathbf{z}_L), \quad \hat{\mathbf{x}}_H = D_H(\mathbf{z}_H), \quad \hat{\mathbf{x}} = \mathcal{W}^{-1}(\hat{\mathbf{x}}_L, \hat{\mathbf{x}}_H).$$

This frequency-aware latent representation learning via DeBaT yields expressive latent embeddings by preserving both coarse structure and fine details across the full spectrum of spatial frequencies.

3.3 Latent Fusion for Generative Modeling

To enable frequency-aware generation, we integrate the proposed DeBaT tokenizer into the latent diffusion framework LightningDiT [51]. Given an input image $\mathbf{x}$, wavelet decomposition produces low- and high-frequency components ($\mathbf{x}_L, \mathbf{x}_H$), which are encoded into frequency-specific latent embeddings.

$$\mathbf{z}_L = E_L(\mathbf{x}_L), \quad \mathbf{z}_H = E_H(\mathbf{x}_H),$$

where $\mathbf{z}_L \in \mathbb{R}^{H \times W \times C_L}$ captures global structure and $\mathbf{z}_H \in \mathbb{R}^{H \times W \times C_H}$ encodes fine details. The two embeddings are fused through channel-wise concatenation yielding a unified latent representation $\tilde{\mathbf{z}} \in \mathbb{R}^{H \times W \times (C_L + C_H)}$. The diffusion model then learns the latent distribution $p_\theta(\tilde{\mathbf{z}})$. During generation, the sampled latent $\tilde{\mathbf{z}}$ is separated into respective low and high frequency embeddings.

$$\hat{\mathbf{z}}_L = \tilde{\mathbf{z}}_{1:C_L}, \quad \hat{\mathbf{z}}_H = \tilde{\mathbf{z}}_{C_L+1:C_L+C_H},$$

which are decoded independently $\hat{\mathbf{x}}_L = D_L(\hat{\mathbf{z}}_L)$, and $\hat{\mathbf{x}}_H = D_H(\hat{\mathbf{z}}_H)$, and the image is reconstructed via inverse wavelet transform $\hat{\mathbf{x}} = \mathcal{W}^{-1}(\hat{\mathbf{x}}_L, \hat{\mathbf{x}}_H)$. This enables the diffusion model to operate on a frequency-aware latent space while preserving modularity with existing latent generative pipelines.

4 Experiments

To evaluate DeBaT, we first assess the reconstruction performance of visual tokenizers in both spatial and spectral domains. Then, we integrate the latent embeddings into standard image synthesis pipeline, comparing DeBaT combined with LightningDiT [51] to competing models. All experiments are conducted on ImageNet-1K [8] and use 50k ImageNet validation samples for evaluation. Further architectural and implementation details of DeBaT are provided in Appendix section A & section B.

4.1 Performance Evaluation of Latent Tokenizers

To evaluate the reconstruction performance of latent embeddings in both the spatial and spectral (frequency) domains, we consider visual tokenizers commonly used in modern latent generative models. These tokenizers are different variants of variational autoencoding architectures, each aiming to learn compact and informative latent embedding representations, which depict the underlying input data distribution. For comprehensive analysis, we include recent discrete [1, 43, 53] and continuous latent tokenizers [7, 24, 25, 37, 51, 53]. We compare these latent tokenizers with DeBaT. This enables comparison across various latent tokenizers with varying latent dimensionalities and feature resolutions as shown in Table 1. We use the metrics discussed in subsection 3.1 for evaluation.

Our analysis in Table 1 reveals that KL-regularized VAE from [25] with right parameterizations consistently outperform both traditional VAE and vector quantization (VQ-VAE) baselines across all reconstruction metrics. Also KL-regularized VAE show efficient reconstruction performance at much better compression rate with lower dimensionality. This improvement stems from the absence of quantization artifacts, which often degrade reconstruction quality in discrete latent spaces. Additionally, the effectiveness of KL-VAE is closely tied to the quality of their latent parameterization schemes. Notably, KL-VAE employed in recent autoregressive models [6]. VA-VAE [51] further aligns its latent space using KL-VAE parameterization with foundation models like DI-NOv2 [33] achieving strong performance across low and high-frequency reconstruction, along with improved perceptual quality. However, these models still rely on a joint optimization of spatial frequencies, which can lead to trade-offs and suboptimal detail preservation, especially in the high-frequency spectrum. In contrast, our proposed approach DeBaT builds on the VA-VAE framework by explicitly decoupling and modeling low- and high-frequency components in learning of the compact and frequency preserving expressive latent embeddings. This separation enables focused representation learning for each frequency band,

Visual Tokenizer	Tokenizer Latent	Rec.	LF	HF	LPIPS	rFID
DC-AE [7]	8x8x32	0.0194	0.0484	0.0097	0.1580	0.7450
DC-AE [7]	4x4x128	0.0207	0.0527	0.0100	0.1667	0.7623
DC-AE [7]	4x4x512	0.0225	0.0581	0.0106	0.1805	0.7912
SD-VAE [37]	16x16x16	0.0180	0.0422	0.0100	0.1743	0.6213
KL-VAE [25]	16x16x16	0.0148	0.0326	0.0089	0.1355	0.5318
MS-VQ-VAE [43]	16x16x32 (4096)	0.0195	0.0549	0.0076	0.1890	0.6981
RQ-VAE [24]	8x8x16 (16384)	0.0416	0.1219	0.0149	0.2712	0.9095
TiTok-VQ-VAE [53]	128x64 (8192)	0.0226	0.0561	0.0114	0.2082	0.7550
TiTok-VQ-VAE [53]	64x64 (8192)	0.0339	0.0947	0.0137	0.2657	0.8823
TiTok-VQ-VAE [53]	32x64 (8192)	0.0450	0.1344	0.0152	0.2949	0.9416
TiTok-VAE [53]	256x16	0.0332	0.0923	0.0134	0.2232	0.8640
TiTok-VAE [53]	128x16	0.0461	0.1380	0.0155	0.2682	0.9187
TiTok-VAE [53]	64x16	0.0617	0.1961	0.0170	0.3182	0.9761
VQ-VAE [1]	16x16x256 (1024)	0.0492	0.1478	0.0163	0.3064	0.9102
VQ-VAE [1]	16x16x256 (16384)	0.0438	0.1262	0.0164	0.2784	0.8826
LiteVAE [38]	32x32x12	0.0086	0.0179	0.0050	0.0970	0.2250
VA-VAE [51]	16x16x32	0.0105	0.0200	0.0074	0.0975	0.2777
DeBaT (Ours)	16x16x32	0.0044	0.0114	0.0020	0.0940	0.2156

Table 1: Quantitative comparison of various latent tokenizers based on their reconstruction quality across multiple metrics. Tokenizer latent represents latent dimensionality (height $\times$ width $\times$ channel), in case of quantized latents latent dimensionality is represented by (height $\times$ width $\times$ channel (codebook size)). For TikTok versions as the latent space is 1-D, latent dimension represents (number of tokens $\times$ channels). Metrics: **Rec.** (reconstruction loss), **LF** (low-frequency loss), **HF** (high-frequency loss), **LPIPS** and **rFID**. Lower is better. DeBaT achieves a substantial improvement over previous methods in terms of reconstruction quality.

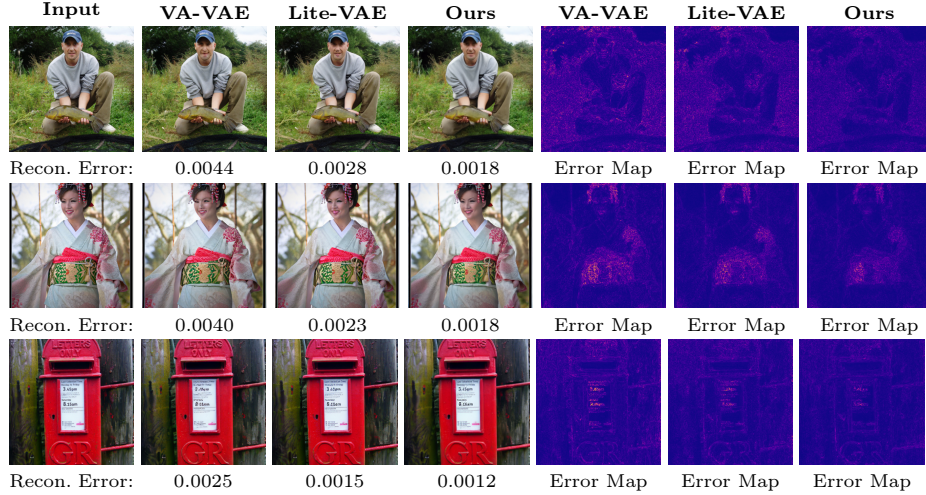


Fig. 4: Comparison with respect to reconstructions and error maps with best baseline VA-VAE and the previous wavelet base method LiteVAE. DeBaT Error Maps have consistently lower magnitudes than VA-VAE and LiteVAE Error Maps, visible as fine details that are correctly reconstructed in particular in structured regions like person faces, the texture of the dress or the printed text. LiteVAE performs better than VA-VAE but makes more errors in fine details than DeBaT.

Method	Tokenizer	Epochs	#Params	Generation w/o CFG					Generation w/ CFG				
				gFID	sFID	IS	Pre.	Rec.	gFID	sFID	IS	Pre.	Rec.
Autoregressive (AR)													
MaskGIT [5]	VQ-VAE	555	227M	6.18	-	182.1	0.80	0.51	-	-	-	-	-
LlamaGen [40]	VQ-VAE	300	3.1B	9.38	8.24	112.9	0.69	0.67	2.18	5.97	263.3	0.81	0.58
VAR [43]	MS-VQ-VAE	350	2.0B	-	-	-	-	-	1.80	-	365.4	0.83	0.57
MagViT-v2 [52]	VQ-VAE	1080	307M	3.65	-	200.5	-	-	1.78	-	319.4	-	-
MAR [25]	KL-VAE	800	945M	2.35	-	227.8	0.79	0.62	1.55	-	303.7	0.81	0.62
Latent Diffusion Models													
MaskDiT [60]	SD-VAE	1600	675M	5.69	10.34	177.9	0.74	0.60	2.28	5.67	276.6	0.80	0.61
DiT [34]	SD-VAE	1400	675M	9.62	6.85	121.5	0.67	0.67	2.27	4.60	278.2	0.83	0.57
SiT [28]	SD-VAE	1400	675M	8.61	6.32	131.7	0.68	0.67	2.06	4.50	270.3	0.82	0.59
FasterDiT [50]	SD-VAE	400	675M	7.91	5.45	131.3	0.67	0.69	2.03	4.63	264.0	0.81	0.60
MDT [11]	SD-VAE	1300	675M	6.23	5.23	143.0	0.71	0.65	1.79	4.57	283.0	0.81	0.61
MDTV2 [12]	SD-VAE	1080	675M	-	-	-	-	-	1.58	4.52	314.7	0.79	0.65
REPA [54]	SD-VAE	800	675M	5.90	-	-	-	-	1.42	4.70	305.7	0.80	0.65
LightningDiT [51]	Lite-VAE	64	675M	7.97	5.13	128.1	0.70	0.64	2.27	4.50	260.0	0.80	0.57
LightningDiT [51]	VA-VAE	64	675M	5.14	4.22	130.2	0.76	0.62	2.11	4.16	252.3	0.81	0.58
LightningDiT [51]	DeBaT	64	675M	3.24	4.09	193.7	0.83	0.69	1.52	4.07	317.4	0.83	0.65

Table 2: Comparison of generation performance of autoregressive and latent diffusion models with and without classifier-free guidance (CFG) vs. LightningDiT with our DeBaT tokenizer. Missing values are indicated by -. We report metrics available in the respective works. DeBaT provides consistent improvement over previous works, in particular over VA-VAE where the direct comparison in terms of generation method (LightningDiT), epochs and number of parameters is straight-forward.

leading to more precise reconstructions of both global structures and fine details. For additional quantitative evaluation results, please refer to Table 1 and Table 2 in the Appendix section C. Figure 4 shows visual comparison with best baselines VAAE, LiteVAE and our approach with reconstruction errors and error maps, where our reconstructions are much detailed with lower reconstruction error and lower magnitude in error maps overall, showcasing preservation of high frequency details. In comparison to previous work leveraging wavelet transforms [38], we observe that DeBaT achieves a substantially lower reconstruction loss, particularly prominent in the high frequency loss. Notably, LiteVAE employs wavelets to facilitate highly compressed latent codes, not fully leveraging their ability to decouple frequency bands as proposed in DeBaT.

4.2 Latent Generation Results

We evaluate the generative quality of our learned latent embeddings by training a diffusion-based generative model on top of the DeBaT latent representations. Specifically, we use LightningDiT [51], a fast-converging latent diffusion backbone to model the generation process, and compare against a wide range of state-of-the-art autoregressive and latent diffusion models, both with and without classifier-free guidance (CFG). Table 2 reports results across standard metrics: generation FID score (gFID), spatial FID (sFID), Inception Score (IS), Precision, and Recall. Lower gFID and sFID values indicate better image fidelity and diversity, while higher IS, precision, and recall reflect better sample quality. sFID is equivalent to FID but uses intermediate spatial features in the inception network rather than the spatially-pooled features in standard FID.



Fig. 5: Class-conditional generation results of classes : Dog (class id : 207), mushroom (class id : 947), Daisy (class id : 985), duck(class id : 97) of LightningDiT with DeBaT.

Table 2 shows comparison of LightningDiT with DeBaT with other standard latent generation baselines. Our approach, LightningDiT with DeBaT, achieves the best gFID among all diffusion based generation methods, along with highly competitive sFID and strong scores across IS, precision, and recall with and without the CFG setting. Compared to auto-regressive models like MaskGIT [5], MAR [25], VAR [43], MagViT-v2 [52] and LlamaGen [40], as well as latent diffusion models such as MaskDiT [60], DiT [34], SiT [28], FasterDiT [50], MDF [11] and REPA [54], our method consistently delivers a better fidelity and diversity. Notably, it outperforms the LightningDiT baseline trained on VA-VAE embeddings across all metrics, demonstrating the advantage of our frequency-aware latent space in enhancing generation quality. These results validate the effectiveness of our DeBaT tokenizer for learning high-quality, detail-preserving representations that generalize well in generative pipelines. The qualitative results of our generated visualizations are shown in Figure 5. Figure shows the conditional generation samples of our approach when integrated with LightningDiT, showcasing diversity and quality in generated samples. Further quantitative evaluations with longer generative training are provided in Appendix subsection C.3.

5 Ablation Study

Ablation on Frequency Decoupling. To further investigate the contribution of decoupling low- and high-frequency components in our frequency-aware

Tokenizer	Rec.↓	LF↓	HF↓	PSNR↑	SSIM↑	LPIPS↓	rFID↓
KL-VAE	0.1146	0.0645	0.1313	25.07	0.74	0.2151	0.7820
MS-VQ-VAE	0.0238	0.0705	0.0082	21.70	0.65	0.2331	0.8427
VA-VAE	0.0125	0.0263	0.0098	26.01	0.78	0.1071	0.6509
DeBaT	0.0044	0.0114	0.0020	30.03	0.90	0.0940	0.4156

Table 3: Reconstruction comparison of tokenizers using wavelet representations instead of pixel inputs. Metrics include reconstruction loss (Rec.), low-frequency loss (LF), high-frequency loss (HF), PSNR, SSIM, LPIPS, and rFID. Lower is better for Rec., LF, HF, LPIPS, and rFID, while higher is better for PSNR and SSIM.

training strategy, we conduct an ablation study utilizing standard KL-VAE [25], multi-scale VQ-VAE [43] and VAAE [51] with coupled low and high frequency representation as input after wavelet transformation.

Specifically, we retrain these models by coupling the low- and high-frequency components in a shared frequency space, using the same wavelet-based losses employed in our DeBaT framework. In this setup, each tokenizer is optimized jointly on both low- and high-frequency representations obtained via the Discrete Wavelet Transform (DWT) using Haar wavelet filter.

The results, presented in Table 3, demonstrate that our frequency-aware DeBaT with decoupling frequency representations as input consistently outperforms the coupled-frequency variants across all metrics, including overall reconstruction error (Rec.), low-frequency (LF) and high-frequency (HF) fidelity error, perceptual similarity loss (LPIPS), and reconstruction FID (rFID) with higher PSNR and SSIM. This indicates that explicitly decoupling frequencies during training leads to more accurate and perceptually faithful reconstructions. For a fair comparison, all models are trained with similar combined latent dimensionality as used in the DeBaT configuration. Additional ablations with varying low and high frequency channels are provided in Appendix subsection C.4.

High-resolution evaluation.

We further evaluate DeBaT on FFHQ at 1024^2 resolution using LightningDiT. As shown in Table 4, DeBaT consistently improves both reconstruction quality (PSNR) and generation quality (gFID and precision) over VA-VAE while maintaining the same latent representation ($32 \times 32 \times 64$), demonstrating the scalability of the proposed tokenizer to high-resolution generation.

Tokenizer	PSNR↑	gFID↓	Pre.↑	Rec.↑
VA-VAE	33.84	6.92	0.69	0.31
DeBaT	34.27	6.86	0.71	0.31

Table 4: High-resolution evaluation on FFHQ (1024^2) using LightningDiT. DeBaT achieves better reconstruction and generation quality than VA-VAE demonstrating the scalability to high-resolution generation.

Training complexity and efficiency. Table 5 compares DeBaT with representative tokenizers in terms of reconstruction quality, architectural complexity, and runtime efficiency. Under the same latent size ($16 \times 16 \times 32$) as VA-VAE, DeBaT achieves substantially better reconstruction quality (30.03 dB PSNR vs. 26.26 dB) while maintaining comparable GPU memory consumption (5892 MB vs. 5694 MB) and practical inference speed (50.08 img/s, 19.97 ms/img). Although DeBaT employs more parameters due to its dual encoder-decoder architecture, it remains more efficient than the considerably larger WF-VAE, achieving both

Tokenizer	PSNR $\uparrow$	# Params $\downarrow$ (M)	Memory $\downarrow$ (MB)	Throughput $\uparrow$ (img/s)	Latency $\downarrow$ (ms/img)
SD-VAE	24.91	83.65	8040	31.78	31.47
VA-VAE	26.26	72.60	5694	57.83	17.29
LiteVAE	28.59	59.96	6922	45.96	21.76
WF-VAE	29.00	317.08	9571	27.48	36.39
DeBaT	30.03	137.91	5892	50.08	19.97

Table 5: Training complexity and efficiency comparison. Throughput, latency, and GPU memory are measured at batch size 32 and 256×256 resolution. Higher PSNR/throughput and lower parameters, memory, and latency are preferred.

higher PSNR and faster inference. These results demonstrate that the proposed frequency-decoupled design provides a favorable trade-off between reconstruction quality and computational efficiency.

6 Conclusion

In this work, we investigate frequency awareness in the latent embeddings of tokenizer models used within latent generative pipelines. We find that jointly optimizing low- and high-frequency components leads to a frequency bias favoring low-frequencies thereby degrading the reconstruction of fine-grained, high-frequency details and overall perceptual quality. To address this limitation, we introduce DeBaT, a simple yet effective VAE framework that explicitly decouples low and high-frequency bands via wavelet decomposition, processes them independently, and fuses them into a unified latent representation. DeBaT achieves state-of-the-art performance across frequency-aware reconstruction metrics, demonstrating improved fidelity and perceptual quality. When used within a diffusion-based generation framework, our approach preserves the spectral richness of the data, enabling more accurate, high fidelity generation of images with coarse structures as well as fine details.

Discussion & Future Direction. After training, DeBaT’s latent code is composed of separate codes for high and low frequencies. A potential future direction is to explore the combination of the learned latent codes with other, complementary latent encoders, where we would hope to improve fine details in generated data without full training. Furthermore, it would be interesting to explore DeBaT’s use in conjunction with different tokenizer architectures. In this work, we focused on VAAE as a baseline model because it recently provided state-of-the-art performance while being computationally efficient. Last, it would be interesting to extend DeBaT to more wavelet decomposition stages. Our current model only leverages a single wavelet decomposition stage which provides only limited efficiency gains. Multi-Stage decompositions similar to the setting recently proposed in WF-VAE [26] could be combined with DeBaT, to facilitate fully decoupled multi-stage encoding, potentially with further benefits in efficiency as well as in detail preservation.

Acknowledgments. The authors acknowledge support by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant INST 35/1597-1 FUGG.

References

1. Esser et al., P.e.a.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12873–12883 (June 2021)
2. Bachmann, R., Allardice, J., Mizrahi, D., Fini, E., Kar, O.F., Amirloo, E., El-Nouby, A., Zamir, A., Dehghan, A.: Flextok: Resampling images into 1d token sequences of flexible length. In: Forty-second International Conference on Machine Learning (2025)
3. Björk, S., Myhre, J.N., Johansen, T.H.: Simpler is better: Spectral regularization and up-sampling techniques for variational autoencoders. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3778–3782. IEEE (2022)
4. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: Proc. CVPR 2023 (2023)
5. Chang, H., Zhang, H., Jiang, L., Liu, C., Freeman, W.T.: Maskgit: Masked generative image transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11315–11325 (2022)
6. Chen, H., Han, Y., Chen, F., Li, X., Wang, Y., Wang, J., Wang, Z., Liu, Z., Zou, D., Raj, B.: Masked autoencoders are effective tokenizers for diffusion models. In: Forty-second International Conference on Machine Learning (2025)
7. Chen, J., Cai, H., Chen, J., Xie, E., Yang, S., Tang, H., Li, M., Han, S.: Deep compression autoencoder for efficient high-resolution diffusion models. In: International Conference on Learning Representations. vol. 2025, pp. 96539–96560 (2025)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
9. Dong, Y., Zuo, Q., Gu, X., Yuan, W., Zhao, Z., Dong, Z., Bo, L., Huang, Q.: Gpld3d: Latent diffusion of 3d shape generative models by enforcing geometric and physical priors. In: Proc. CVPR 2024. p. 56–66 (2024)
10. Gal, R., Hochberg, D.C., Bermano, A., Cohen-Or, D.: Swagan: A style-based wavelet-driven generative model. ACM Transactions on Graphics (TOG) **40**(4), 1–11 (2021)
11. Gao, S., Zhou, P., Cheng, M.M., Yan, S.: Masked diffusion transformer is a strong image synthesizer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23164–23173 (2023)
12. Gao, S., Zhou, P., Cheng, M.M., Yan, S.: Mdtv2: Masked diffusion transformer is a strong image synthesizer. arXiv preprint arXiv:2303.14389 (2023)
13. Guth, F., Coste, S., De Bortoli, V., Mallat, S.: Wavelet score-based generative modeling. In: NeurIPS (2022)
14. 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: Advances in Neural Information Processing Systems (NeurIPS). pp. 6626–6637 (2017)
15. Huang, L., Fang, R., Zhang, A., Song, G., Liu, S., Liu, Y., Li, H.: Fouriscale: A frequency perspective on training-free high-resolution image synthesis. In: European conference on computer vision. pp. 196–212. Springer (2024)
16. Huang, Z., Qiu, X., Ma, Y., Zhou, Y., Chen, J., Zhang, H., Zhang, C., Li, X.: NFIG: Multi-scale autoregressive image generation via frequency ordering. In: NeurIPS (2025)

17. Hui, K.H., Sanghi, A., Rampini, A., Malekshan, K.R., Liu, Z., Shayani, H., Fu, C.W.: Make-a-shape: a ten-million-scale 3d shape model. In: Forty-first International Conference on Machine Learning (2024)
18. Jiang, H., Luo, A., Fan, H., Han, S., Liu, S.: Low-light image enhancement with wavelet-based diffusion models. *ACM Transactions on Graphics (TOG)* **42**(6), 1–14 (2023)
19. Jiang, L., Dai, B., Wu, W., Loy, C.C.: Focal frequency loss for image reconstruction and synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 13919–13929 (October 2021)
20. Kim, C., Moon, S.J., Park, G.M.: Wine: Wavelet-guided gan inversion and editing for high-fidelity refinement. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 4523–4532 (2025)
21. Kim, Y., Hwang, G., Zhang, J., Park, E.: Diffusehigh: Training-free progressive high-resolution image synthesis through structure guidance. In: Proceedings of the AAAI conference on artificial intelligence. pp. 4338–4346 (2025)
22. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013)
23. Kouzelis, T., Kakogeorgiou, I., Gidaris, S., Komodakis, N.: Eq-vae: Equivariance regularized latent space for improved generative image modeling. *arXiv preprint arXiv:2502.09509* (2025)
24. Lee, D., Kim, C., Kim, S., Cho, M., Han, W.S.: Autoregressive image generation using residual quantization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11523–11532 (June 2022)
25. Li, T., Tian, Y., Li, H., Deng, M., He, K.: Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems* **37**, 56424–56445 (2024)
26. Li, Z., Lin, B., Ye, Y., Chen, L., Cheng, X., Yuan, S., Yuan, L.: Wf-vae: Enhancing video vae by wavelet-driven energy flow for latent video diffusion model. In: Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR) (2025)
27. Luo, S., Tan, Y., Huang, L., Li, J., Zhao, H.: Latent consistency models: Synthesizing high-resolution images with few-step inference. In: Proc. ICLR 2024 (2024)
28. Ma, N., Goldstein, M., Albergo, M.S., Boffi, N.M., Vanden-Eijnden, E., Xie, S.: Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In: European Conference on Computer Vision. pp. 23–40. Springer (2024)
29. Medi, T., Rampini, A., Reddy, P., Jayaraman, P.K., Keuper, M.: 3d-wag: Hierarchical wavelet-guided autoregressive generation for high-fidelity 3d shapes. *arXiv preprint arXiv:2411.19037* (2024)
30. Medi, T., Tayyub, J., Sarmad, M., Lindseth, F., Keuper, M.: Fullformer: Generating shapes inside shapes. In: Köthe, U., Rother, C. (eds.) *Pattern Recognition*. pp. 147–162. Springer Nature Switzerland, Cham (2024)
31. Miwa, K., Sasaki, K., Arai, H., Takahashi, T., Yamaguchi, Y.: One-d-piece: Image tokenizer meets quality-controllable compression. *arXiv preprint arXiv:2501.10064* (2025)
32. Mulcahy, C.: Image compression using the haar wavelet transform. *Spelman Science and Mathematics Journal* **1**(1), 22–31 (1997)
33. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023)

34. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4195–4205 (2023)
35. Phung, H., Dao, Q., Tran, A.: Wavelet diffusion models are fast and scalable image generators. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10199–10208 (June 2023)
36. Qian, Y., Cai, Q., Pan, Y., Li, Y., Yao, T., Sun, Q., Mei, T.: Boosting diffusion models with moving average sampling in frequency domain. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8911–8920 (2024)
37. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (June 2022)
38. Sadat, S., Buhmann, J., Bradley, D., Hilliges, O., Weber, R.M.: Litevae: Lightweight and efficient variational autoencoders for latent diffusion models. *Advances in Neural Information Processing Systems* **37**, 3907–3936 (2024)
39. Skorokhodov, I., Girish, S., Hu, B., Menapace, W., Li, Y., Abdal, R., Tulyakov, S., Siarohin, A.: Improving the diffusability of autoencoders. *arXiv preprint arXiv:2502.14831* (2025)
40. Sun, P., Jiang, Y., Chen, S., Zhang, S., Peng, B., Luo, P., Yuan, Z.: Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525* (2024)
41. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2818–2826 (2016)
42. Takida, Y., Ikemiya, Y., Shibuya, T., Shimada, K., Choi, W., Lai, C.H., Murata, N., Uesaka, T., Uchida, K., Liao, W.H., et al.: Hq-vae: Hierarchical discrete representation learning with variational bayes. *arXiv preprint arXiv:2401.00365* (2023)
43. Tian, K., Jiang, Y., Yuan, Z., Peng, B., Wang, L.: Visual autoregressive modeling: Scalable image generation via next-scale prediction. In: *Advances in Neural Information Processing Systems*. vol. 37, pp. 84839–84865 (2024)
44. Vahdat, A., Kautz, J.: NVAE: A deep hierarchical variational autoencoder. In: *Neural Information Processing Systems (NeurIPS)* (2020)
45. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. *Advances in neural information processing systems* **30** (2017)
46. Xie, E., Chen, J., Chen, J., Cai, H., Tang, H., Lin, Y., Zhang, Z., Li, M., Zhu, L., Lu, Y., Han, S.: Sana: Efficient high-resolution image synthesis with linear diffusion transformer (2024), <https://arxiv.org/abs/2410.10629>
47. Xie, E., Chen, J., Zhao, Y., Yu, J., Zhu, L., Lin, Y., Zhang, Z., Li, M., Chen, J., Cai, H., et al.: Sana 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer (2025), <https://arxiv.org/abs/2501.18427>
48. Xu, W., Yue, X., Wang, Z., Teng, Y., Zhang, W., Liu, X., Zhou, L., Ouyang, W., Bai, L.: Exploring representation-aligned latent space for better generation. *arXiv preprint arXiv:2502.00359* (2025)
49. Yang, X., Zhou, D., Feng, J., Wang, X.: Diffusion probabilistic model made slim. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 22552–22562 (2023)
50. Yao, J., Wang, C., Liu, W., Wang, X.: Fasterdit: Towards faster diffusion transformers training without architecture modification. *Advances in neural information processing systems* **37**, 56166–56189 (2024)

51. Yao, J., Yang, B., Wang, X.: Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2025)
52. Yu, L., Lezama, J., Gundavarapu, N.B., Versari, L., Sohn, K., Minnen, D., Cheng, Y., Gupta, A., Gu, X., Hauptmann, A.G., et al.: Language model beats diffusion-tokenizer is key to visual generation. In: *International Conference on Learning Representations*. vol. 2024, pp. 765–783 (2024)
53. Yu, Q., Weber, M., Deng, X., Shen, X., Cremers, D., Chen, L.C.: An image is worth 32 tokens for reconstruction and generation. In: *Advances in Neural Information Processing Systems*. vol. 37, pp. 128940–128966 (2024)
54. Yu, S., Kwak, S., Jang, H., Jeong, J., Huang, J., Shin, J., Xie, S.: Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940* (2024)
55. Yu, Y., Zhan, F., Lu, S., Pan, J., Ma, F., Xie, X., Miao, C.: Wavefill: A wavelet-based generation network for image inpainting. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 14114–14123 (October 2021)
56. Yuan, X., Li, L., Wang, J., Yang, Z., Lin, K., Liu, Z., Wang, L.: Spatial-frequency u-net for denoising diffusion probabilistic models. *arXiv preprint arXiv:2307.14648* (2023)
57. Zhang, J., Huang, Q., Liu, J., Guo, X., Huang, D.: Diffusion-4k: Ultra-high-resolution image synthesis with latent diffusion models. In: *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*. pp. 23464–23473 (June 2025)
58. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 586–595 (2018)
59. Zhao, C., Cai, W., Dong, C., Hu, C.: Wavelet-based fourier information interaction with frequency diffusion adjustment for underwater image restoration. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 8281–8291 (2024)
60. Zheng, H., Nie, W., Vahdat, A., Anandkumar, A.: Fast training of diffusion models with masked transformers. *arXiv preprint arXiv:2306.09305* (2023)
61. Zhou, C., Wang, X., Zhang, M.: Unifying generation and prediction on graphs with latent graph diffusion. In: *Proc. NeurIPS 2024* (2024)