LiteVAE: Lightweight and Efficient Variational Autoencoders for Latent Diffusion Models

Seyedmorteza Sadat[1], Jakob Buhmann[2], Derek Bradley[2], Otmar Hilliges[1], Romann M. Weber[2]

[1]ETH Zürich, [2]DisneyResearch|Studios

{seyedmorteza.sadat,otmar.hilliges}@inf.ethz.ch
{jakob.buhmann,derek.bradley,romann.weber}@disneyresearch.com

Abstract

Advances in latent diffusion models (LDMs) have revolutionized high-resolution image generation, but the design space of the autoencoder that is central to these systems remains underexplored. In this paper, we introduce LiteVAE, a new autoencoder design for LDMs, which leverages the 2D discrete wavelet transform to enhance scalability and computational efficiency over standard variational autoencoders (VAEs) with no sacrifice in output quality. We investigate the training methodologies and the decoder architecture of LiteVAE and propose several enhancements that improve the training dynamics and reconstruction quality. Our base LiteVAE model matches the quality of the established VAEs in current LDMs with a six-fold reduction in encoder parameters, leading to faster training and lower GPU memory requirements, while our larger model outperforms VAEs of comparable complexity across all evaluated metrics (rFID, LPIPS, PSNR, and SSIM).
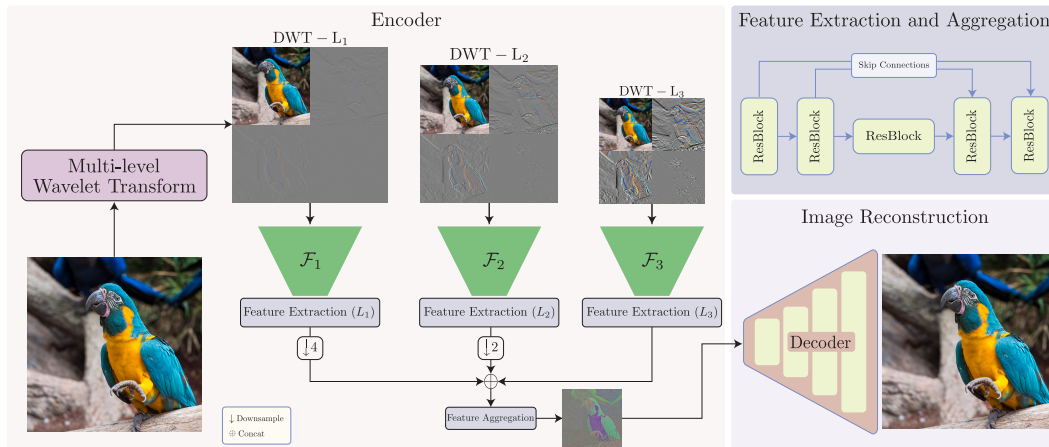
Figure 1: An overview of LiteVAE. The input image is first decomposed into multi-level wavelet coefficients, and each wavelet sub-band is separately processed via a feature-extraction network. The features are then combined via a feature-aggregation module to compute the final latent code, which is then transformed back into the image space by the decoder. We use a lightweight UNet architecture (top right) without spatial down/upsampling for feature extraction and aggregation. The decoder is a fully convolutional network similar to that in the Stable Diffusion VAE [55]. LiteVAE's design allows it to be significantly more efficient than standard VAEs in LDMs while maintaining high reconstruction quality.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

1 Introduction

Latent diffusion models (LDMs) [55] have recently assumed dominance in the field of high-resolution image generation, primarily due to their scalability and training stability over pixel-space diffusion. The training process of LDMs involves two separate stages. In the first, an expressive variational autoencoder (VAE) is trained to transform the raw pixels of an image into a more compact latent representation. In the second, a diffusion model is trained on the latent representations of training images. While numerous studies have investigated the scalability and dynamics of the diffusion component in LDMs [48, 33], the autoencoder element has received far less attention.

The VAE in LDMs is not only computationally demanding to train but also affects the efficiency of the diffusion training phase due to the resource requirements of querying a large encoder network for computing the latent codes. For example, as the autoencoder operates on high-resolution images, the VAE encoder of Stable Diffusion 2.1 uses 135.59 GFLOPS compared with 86.37 for the diffusion UNet.[1] This becomes an even greater concern for video diffusion models, as the encoder then needs to provide the latents for a batch of frames instead of a single image [3].

A common workaround for this resource burden is to precompute and cache the latent codes for the entire dataset to avoid having to use the autoencoder during diffusion training. However, in addition to its initial overhead, this approach eliminates the possibility of using on-the-fly techniques, such as data augmentation, which have been shown to improve the training and performance of diffusion models [32]. Using a large encoder also adds noticeable overhead in applications that are based on pretrained latent diffusion models. For example, when training 3D models through score distillation of 2D LDMs [51], the process necessitates backpropagating gradients through the LDM encoder, which is computationally intensive [38]. Beyond the computational aspects, improving the reconstruction quality of the autoencoder also improves the quality of generated images, as the autoencoder provides an upper bound on the generation quality [50, 13].

With these issues in mind, we investigated improving the efficiency of LDMs through their core VAE component with the goal of preserving overall quality. We show that with the help of the 2D discrete wavelet transform (DWT), we can considerably simplify the encoder network in LDMs. This leads to our proposal of LiteVAE, a new autoencoder design for LDMs, which has superior compute/quality trade-offs compared with standard VAEs.

LiteVAE consists of a lightweight feature-extraction module to compute features from the wavelet coefficients and a feature-aggregation module to combine these multiscale features into a unified latent code. A decoder then converts the latent code back to an image. An overview of the LiteVAE pipeline is shown in Figure 1.

We chose the wavelet transform due to its proven ability to represent rich, compact image features [43], and we argue that the wavelet decomposition simplifies the encoder's task by facilitating the learning of meaningful features. We examine the design space of LiteVAE in depth and propose several variations on the network architecture and training setup that further boost reconstruction quality and training efficiency.

Through extensive experimentation, we show that LiteVAE considerably reduces the computational cost of the standard VAE encoder while maintaining the same level of reconstruction quality. In addition, LiteVAE provides better reconstruction quality when compared with a VAE of comparable complexity. We also perform an analysis on the latent space learned by LiteVAE and show that it is similar to that of a regular VAE.

To summarize, our main contributions in this paper are as follows: (i) We introduce LiteVAE, a more efficient and lightweight VAE for LDMs with similar reconstruction quality. This leads to faster training of the autoencoder and higher throughput when training latent diffusion models. (ii) We explore the design space of LiteVAE and propose variations that further enhance reconstruction quality and improve its training dynamics. (iii) We perform extensive experimental analyses on the design choices and computational efficiency of LiteVAE and empirically verify its superior compute efficiency compared to a regular VAE.

---

[1]Result of processing a single $256 \times 256 \times 3$ image and its corresponding $32 \times 32 \times 4$ latent representation.

2

2 Related work

Diffusion models and LDMs   Score-based diffusion models [64, 65, 24, 66] are a class of generative models that learn the data distribution by reversing a forward destruction process that gradually adds Gaussian noise to the data. These models have recently achieved state-of-the-art generation performance on a number of diverse tasks, including unconditional and conditional image generation [46, 10, 32], text-to-image synthesis [53, 60, 55, 1, 13], video generation [4, 3, 19], image-to-image translation [59, 39], and audio generation [7, 36, 28].

While diffusion models were originally proposed for operating in the ambient image space, Rombach et al. [55] advocated for following the same methodology in the latent space of a frozen, pre-trained VAE. Following this, a number of advancements have been proposed to enhance latent diffusion models, including architecture improvements [48, 16, 33], training setups [21, 32], and sampling techniques [23, 25, 58]. In contrast to these proposed methods, our work focuses on the first stage of LDMs and aims at improving the architecture and efficiency of the VAE component.

Zhu et al. [76] recently proposed an improved decoder for the Stable Diffusion VAE that better preserves the details of conditional inputs for tasks such as in-painting. In contrast, our focus in this paper is mainly on the efficiency and properties of the *entire* VAE in LDMs, and our method is not restricted to conditional scenarios. Dai et al. [9] also introduced FFT features as input to the VAE for better reconstruction quality. However, their work does not address efficiency, and it can be seen as complementary to ours since FFT features can be combined with our DWT approach to further refine the encoder's initial representation.

Wavelet transform   The wavelet transformation [5, 42] is a classic spatial-frequency decomposition of a signal that has gained popularity in numerous computer vision tasks, including denoising [6, 45], image and video compression [62, 67, 54, 41], super-resolution [18, 27], and image restoration [14, 40, 73]. More recently, wavelets have been integrated into generative adversarial networks [15, 71] and pixel-space diffusion models for high-resolution image synthesis [26, 49, 20]. Building on these advancements, we investigate the use of DWT to enhance the efficiency and characteristics of VAEs in LDMs, addressing an underexplored area in the literature.

3 Background

This section includes a brief overview of deep autoencoders and the wavelet transform. A summary of diffusion models is given in Appendix C.

Deep autoencoders   Deep autoencoders consist of an encoder network $\mathcal{E}$ that maps an image to a latent representation and a decoder $\mathcal{D}$ that reconstructs the data from the latent code. More specifically, given an input image $x \in \mathbb{R}^{H \times W \times 3}$, convolutional autoencoders aim to find a latent vector $\mathcal{E}(x) \in \mathbb{R}^{H/f \times W/f \times n_z}$ such that $\mathcal{D}(\mathcal{E}(x)) \approx x$, where f is the spatial downsampling scale and n_z is the number of latent channels.

The training of autoencoders mainly consists of a reconstruction loss $\mathcal{L}_{\text{recon}}(\mathcal{D}(\mathcal{E}(x)), x)$ between the input image and the reconstructed image, and a regularization term $\mathcal{L}_{\text{reg}}(\mathcal{E}(x))$ on the latents. $\mathcal{L}_{\text{recon}}$ is typically a combination of ℓ_1 and perceptual loss [75], and the regularization $\mathcal{L}_{\text{reg}}$ can be enforced via Kullback–Leibler (KL) divergence [35] relative to a reference distribution, typically the standard Gaussian. The regularization term forces the latent space to have a better structure for other applications, such as generative modeling. Following Esser et al. [12], it is also common to train a discriminator D with an adversarial loss $\mathcal{L}_{\text{adv}}$ that differentiates the real images x from the reconstructions $\mathcal{D}(\mathcal{E}(x))$ for more photorealistic outputs. The overall training loss is then equal to

$$\mathcal{L}_{\text{train}} = \mathcal{L}_{\text{recon}} + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}} + \lambda_{\text{adv}} \mathcal{L}_{\text{adv}}, \tag{1}$$

where the λ's are weighting hyperparameters. Esser et al. [12] also proposed an adaptive weighting strategy for λ_{adv} given by

$$\lambda_{\text{adv}} = \frac{1}{2} \left(\frac{\|\nabla \mathcal{L}_{\text{recon}}\|}{\|\nabla \mathcal{L}_{\text{adv}}\| + \delta} \right) \tag{2}$$

for a small $\delta > 0$ to balance the relative gradient norm of the adversarial loss with that of the reconstruction loss.

3

Discrete wavelet transform Wavelet transforms are a signal processing technique for extracting spatial-frequency information from input data. Wavelets are characterized by a low-pass filter L and a high-pass filter H. For 2D signals, four filters are defined via $LL^\top$, $LH^\top$, $HL^\top$, and $HH^\top$. Given an input image x, the 2D wavelet transform decomposes x into a low-frequency sub-band x_L and three high-frequency sub-bands $\{x_H, x_V, x_D\}$ capturing horizontal, vertical, and diagonal details. For an image of size $H \times W$, each wavelet sub-band is of size $H/2 \times W/2$. Multi-resolution analysis is achievable by iteratively applying the wavelet transform to x_L at each level. Wavelet transforms are also invertible, and one can reconstruct the original image x from the sub-bands $\{x_L, x_H, x_V, x_D\}$ using the inverse wavelet transform. Additionally, the Fast Wavelet Transform (FWT) [44] enables the computation of wavelet sub-bands with linear complexity relative to the number of pixels in x. Consistent with the recent literature [15, 49], we use Haar basis as the wavelet filter.

4 Method

In this section, we describe our design of a more efficient VAE for LDMs and discuss our modifications to the network architectures and the training setup that lead to better reconstruction quality and training efficiency. To motivate our approach, Figure 2 shows that when visualizing the latent code learned by the Stable Diffusion VAE (SD-VAE), the code is itself image-like, with a strong similarity to the input. This observation leads us to explore whether the learning of these latent representations can be simplified by applying a fast image-processing function to the input images prior to encoding. We opt for the discrete wavelet

(a) Input image   (b) Latent code

Figure 2: RGB visualization of the first three channels of a SD-VAE latent code.

transform (DWT) as the image-processing function due to its image-like structure, proven effectiveness in extracting rich, compact features from images, and wide applicability in image-processing tasks such as image compression.

4.1 Model design

We now propose LiteVAE, a wavelet-based autoencoder that reaches the reconstruction quality of standard VAEs with much lower complexity. Our method consists of three main components (see also Figure 1):

Wavelet processing: Each image x is first processed via a multi-level DWT to get the corresponding wavelet coefficients $\{x_L^l, x_H^l, x_V^l, x_D^l\}$ at level l. To achieve an $8\times$ downsampling, we use three wavelet levels (i.e., $l \in \{1, 2, 3\}$). These features extract multiscale information from x.

Feature extraction and aggregation: The wavelet coefficients $\{x_L^l, x_H^l, x_V^l, x_D^l\}$ are then separately processed via a feature-extraction module $\mathcal{F}_l$ to compute a multiscale set of feature maps $\mathcal{F}_l(\{x_L^l, x_H^l, x_V^l, x_D^l\})$. The features are then combined via a feature-aggregation module $\mathcal{F}_{\text{agg}}$ that takes in the output of each $\mathcal{F}_l$ and computes the latent z. We use a UNet-based architecture similar to the ADM model [10] without spatial down/upsampling layers for feature extraction and aggregation. (See Appendix B for a discussion of the importance of these learned modules.)

Image reconstruction: Finally, a decoder network $\mathcal{D}$ processes the latent code z and computes the reconstructed image $\hat{x} = \mathcal{D}(z)$. We use the same decoder network as in SD-VAE for $\mathcal{D}$.

The model is then trained end-to-end to learn the parameters of $\{\mathcal{F}_l\}$, $\mathcal{F}_{\text{agg}}$, and $\mathcal{D}$. Because different wavelet levels already contain enough information about the images, we can use lightweight networks for the feature extraction and aggregation steps. Hence, LiteVAE essentially combines the computational benefits of DWT with the expressiveness of a learned encoder. Please refer to Appendices F and G for implementation details.

4.2 Self-modulated convolution

In addition to improving the encoder, we observe that the intermediate feature maps learned by the decoder are relatively imbalanced, with certain areas having significantly stronger magnitudes. An example of this issue is shown in Figure 3. Consistent with Karras et al. [30], we argue that this issue

4

| Group Norm | SMC | Group Norm | SMC |
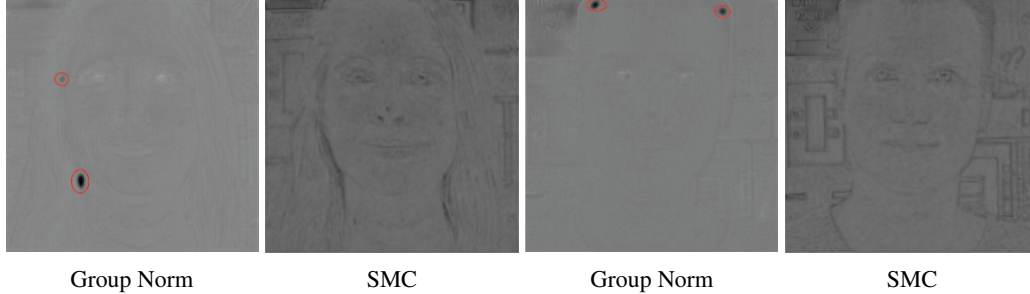
Figure 3: Two examples of the feature maps from the final block of the decoder before and after removing group normalization layers. Using SMC blocks instead of group normalization allows the model to learn more balanced feature maps. The image is best viewed when zoomed in.

is due to excessive group normalization layers [69] in the decoder architectures typically used in autoencoders, since such layers potentially destroy any information found in the magnitudes of the features relative to each other [31].

We propose a modified version of modulated convolution [31] instead of group normalization to avoid imbalances. Instead of modulating the convolution layers via a data-dependent style vector, we allow the convolution layer to learn the corresponding scales for each feature map. We call this operation self-modulated convolution (SMC). SMC modifies the convolution weights w_{ijk} according to

$$w'_{ijk} = \frac{s_i w_{ijk}}{\sqrt{\sum_{i,k}(s_i w_{ijk})^2 + \epsilon}} \tag{3}$$

for $\epsilon > 0$, where s_i is a learnable parameter, and $\{i, j, k\}$ spans the input feature maps, output feature maps, and the spatial dimension of the convolution. Our experiments show that using SMC in the decoder balances the feature maps and also improves the final reconstruction quality due to better training dynamics. Two examples of the decoder feature maps after using SMC are shown in Figure 3.

4.3 Training improvements

Besides the network architecture, we also introduce the following modifications that further enhance the training dynamics and reconstruction quality of LiteVAE. We verify the effect of these modifications in Sections 5 and 6.

Training resolution While the autoencoders in LDMs are typically trained on 256×256 data (similar to SD-VAE), we observe that the bulk of the training of LiteVAE can be effectively conducted at a lower 128×128 resolution. Our experiment suggests that pretraining at this lower resolution followed by a fine-tuning stage at the full resolution achieves similar reconstruction quality while requiring significantly less compute for most of the training. We later show in Appendix D.8 that this improvement is also generally applicable to the standard VAE models.

Improving the adversarial setup We replace the PatchGAN discriminator used in Stable Diffusion with a UNet-based model for pixel-wise discrimination [61]. We also notice that the adaptive weight (Equation (2)) for the adversarial loss update does not introduce any benefit and can be removed for more stable training, especially in mixed-precision setups.

Additional loss functions We also introduce two high-frequency reconstruction loss terms based on the wavelet transform and Gaussian blurring [74]. Let x be the input image and $\hat{x}$ the corresponding reconstruction. For the wavelet term, we compute the Charbonnier loss [2] between the high-frequency DWT sub-bands $\{x_H, x_V, x_D\}$ and $\{\hat{x}_{LH}, \hat{x}_{HL}, \hat{x}_{HH}\}$. For the Gaussian loss, given a Gaussian filter h, we compute the ℓ_1 loss between $x - h(x)$ and $\hat{x} - h(\hat{x})$.

5

Table 1: Comparison between LiteVAE and VAE in terms of reconstruction quality across different datasets and latent dimensions. LiteVAE achieves better or similar reconstruction quality while having considerably fewer parameters in the encoder (34.16M for the VAE and 6.75M for LiteVAE). All models use a downscaling factor of $f = 8$ and are trained from scratch with similar training configs (including the choice of loss functions and discriminator).

Dataset	Latent dim	Model	rFID ↓	LPIPS ↓	PSNR ↑	SSIM ↑
FFHQ 128	16×16×4	VAE	0.88	0.089	28.08	**0.85**
		LiteVAE	**0.74**	**0.085**	**28.36**	**0.85**
FFHQ 256	32×32×4	VAE	0.47	**0.109**	28.16	0.81
		LiteVAE	**0.41**	0.117	**28.33**	**0.82**
ImageNet 128	16×16×4	VAE	4.54	**0.164**	24.25	0.69
		LiteVAE	**4.40**	**0.164**	**24.49**	**0.71**
	16×16×12	VAE	**0.94**	**0.069**	29.25	0.86
		LiteVAE	**0.94**	**0.069**	**29.45**	**0.87**
ImageNet 256	32×32×4	VAE	0.89	0.160	25.83	0.73
		LiteVAE	**0.87**	**0.157**	**26.02**	**0.74**
	32×32×12	VAE	**0.23**	0.073	30.41	0.86
		LiteVAE	**0.23**	**0.072**	**30.91**	**0.88**

(a) Input   (b) Reconstruction   (c) Latent
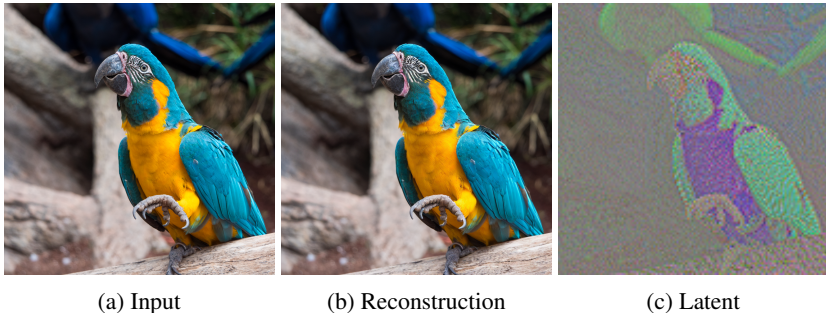
Figure 4: An example of the autoencoder reconstruction alongside the learned latent code by LiteVAE. We observe that LiteVAE maintains the image-like structure of SD-VAE.

5   Experiments

This section presents a comprehensive empirical evaluation of LiteVAE, demonstrating its superior trade-off between computational efficiency and quality relative to standard VAEs. We further explore the properties of LiteVAE along with the changes proposed in Section 4. For each experiment, all models in comparison are trained with the exact same training setup, including the loss functions and the discriminator, to ensure a fair comparison.

Evaluation metrics   We follow the same evaluation pipeline as in Rombach et al. [55] and use reconstruction Fréchet Inception Distance (rFID) [22] as the main metric to measure the quality and realism of autoencoder outputs due to its alignment with human judgment. For completeness, we also report PSNR, SSIM, and LPIPS [75]. As FID is sensitive to small implementation details [47], we recompute the metrics as much as possible based on released checkpoints to have a fair comparison between different models.

Main results   We first demonstrate that LiteVAE matches or exceeds the performance of standard VAEs across various datasets and latent dimensions, as shown in Table 1. Notably, the model employed for this table utilizes approximately one-sixth of the encoder parameters compared to the VAE model (6.75M vs 34.16M) and hence trains faster. Also, one example of the reconstruction quality and the learned latent representation by LiteVAE is given in Figure 4. We notice that LiteVAE maintains the image-like latent codes, similar to the SD-VAE latent in Figure 2.

6

Table 2: Comparison of the scalability of LiteVAE with a standard VAE across different model sizes. (a) LiteVAE matches the performance of the VAE with significantly fewer parameters and outperforms VAEs of similar complexity. (b) A naïve downscaling of the VAE performs worse than LiteVAE. All models use the same decoder. More architecture details are provided in Appendix F.

(a) Scaling LiteVAE ($n_z = 12$ for all models)

Model	Params (M)	rFID ↓	LPIPS ↓	PSNR ↑	SSIM ↑
VAE	34.16	0.95	0.069	29.25	0.86
LiteVAE-S	1.03	1.11	0.075	29.12	0.86
LiteVAE-B	6.75	0.94	0.069	29.55	0.87
LiteVAE-M	32.75	0.79	0.064	29.68	0.87
LiteVAE-L	41.42	**0.74**	**0.062**	**29.94**	**0.88**

(b) Downscaling the VAE ($n_z = 4$ for all models)

Model	Params (M)	rFID ↓	LPIPS ↓	PSNR ↑	SSIM ↑
VAE	34.16	4.54	**0.164**	24.25	0.69
VAE-Small	6.75	5.27	0.175	24.10	0.69
LiteVAE-B	6.75	**4.40**	**0.164**	**24.49**	**0.71**

Table 3: Comparing the complexity of our encoder with the encoder from the Stable Diffusion VAE for a batch size of 32. The values are measured on one Quadro RTX 6000.

Encoder	Params (M)	GPU Memory (MB)	Throughput (img/sec)
VAE	34.16	8860	68
LiteVAE-S	1.03	1324	384
LiteVAE-B	6.75	3155	129
LiteVAE-M	32.75	12130	42.24
LiteVAE-L	41.425	12130	41.6

Increasing model complexity In Table 2 we show the scalability of LiteVAE as we increase the complexity of the feature-extraction and feature-aggregation blocks. We note that the reconstruction performance strictly improves by using more encoder parameters, and our large models outperform a standard VAE of similar complexity across all metrics. Hence, we conclude that LiteVAE offers superior scalability w.r.t. the model size.

Scaling down the encoder in VAEs Table 2b also indicates that the naïve approach of scaling down the encoder in standard VAEs does not perform on par with our method in terms of reconstruction quality. Thus, we conclude that LiteVAE takes better advantage of the encoder parameters than normal VAEs, mainly due to the wavelet processing step that provides the encoder with a rich representation from the beginning.

Computational cost Table 3 presents a comparison of the computational costs between LiteVAE and the Stable Diffusion VAE encoder. LiteVAE-B requires considerably less GPU memory and offers nearly double the throughput. This reduction in computational complexity allows the usage of larger batch sizes when training the autoencoder, as shown to be beneficial by Podell et al. [50], and leads to better hardware utilization for diffusion training in the second stage of LDMs since fewer resources should be devoted to computing the latent input for the diffusion model.

Removing group normalization in the decoder We qualitatively showed in Figure 3 that group normalization in the decoder causes imbalanced feature maps in the network and that SMC can remove such artifacts. Here we also quantitatively show in Table 4 that replacing group normalization with SMC leads to better reconstruction quality. Additionally, we demonstrate in Appendix D.7 that removing the imbalanced feature maps results in less scale dependency in the final model.

Training resolution We next demonstrate the feasibility of pretraining LiteVAE at a lower resolution of 128×128 followed by a fine-tuning step on 256×256 images. To illustrate this, we compare a model trained for 150k steps at full resolution (256-full) with one trained for 100k steps at 128 and an additional 50k steps at 256 (128-tuned). As shown in Table 5, the 128-tuned model even slightly outperforms the model fully trained at the higher resolution. We also note that fine-tuning is essential, as the model trained solely on 128×128 images for 150k steps (128-full) performs worse than the other two. This experiment implies that the model can learn most of the semantics at lower resolutions and recover additional higher-frequency contents in the fine-tuning stage. This pretraining technique reduced the overall wall-clock time of our training runs at 256×256 resolution by more than a factor of two.

Scale dependency Figure 5 demonstrates that compared to the standard VAEs, LiteVAE is less prone to performance degradation when evaluating the model at different resolutions. We hypothesize

7

Table 4: Effect of replacing group normalization with SMC on reconstruction quality based on the ImageNet 128×128 model.

Normalization	rFID ↓	LPIPS ↓	PSNR ↑	SSIM ↑
Group Norm	1.01	**0.07**	29.24	0.86
SMC	**0.97**	**0.07**	**29.32**	**0.87**

Table 5: Effect of pretraining the autoencoder at lower resolutions. We observe that training at 128×128 followed by fine-tuning at 256×256 performs best.

Training Config	rFID ↓	LPIPS ↓	PSNR ↑	SSIM ↑
256-full	0.75	0.153	26.10	0.73
128-full	0.97	0.162	25.90	0.72
128-tuned	**0.73**	**0.147**	**26.22**	**0.74**
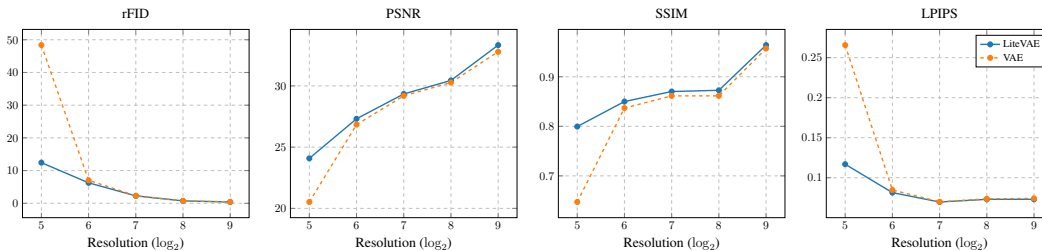
Figure 5: Comparing the performance of LiteVAE with a normal VAE across different resolutions. LiteVAE shows less degradation in all metrics.

Table 6: Comparing MMD between LiteVAE latent space and a standard Gaussian vs SD-VAE latent space for different RBF kernels. LiteVAE is statistically closer to a standard Gaussian.

σ	SD-VAE	LiteVAE
25	8.67±0.10	**1.44±0.28**
50	28.90±0.49	**7.94±0.19**
100	10.77±0.29	**5.14±0.19**
250	1.78±0.06	**1.09±0.04**
500	0.44±0.02	**0.28±0.01**

Table 7: Comparison between diffusion models trained in the latent space of a standard VAE [55] vs the latent space of LiteVAE. We observe that both models perform similarly in terms of generation quality.

Dataset	Encoder	FID ↓
FFHQ [30] (256×256)	LDM	8.11
	LiteVAE	**8.03**
CelebA-HQ [29] (256×256)	LDM	5.92
	LiteVAE	**5.73**

that as our model learns features on top of multi-resolution wavelet coefficients, it is able to learn more scale-independent features compared to a standard encoder and leave the specific details of each scale to the initial wavelet processing step.

Analysis of the LiteVAE latent space We also analyzed the characteristics of the latent space of LiteVAE. Qualitative inspection of Figures 2 and 4, which are representative of the results that hold across our data, show that our latent space and SD-VAE share a similar image-like structure. Separately, we also examined the statistical distance between our model's latent space and pure Gaussian noise. The intuition here is that, since a diffusion model will have to form a path from pure Gaussian noise to our model's latent space, we do not want that path to be longer than the path a diffusion model has to form between Gaussian noise and the Stable Diffusion latent space. To this end, we compute the maximum mean discrepancy (MMD) [17] between latent codes from LiteVAE and samples from a standard Gaussian and compare the result with that observed for the SD-VAE (See Table 6). Here the MMD serves as a proxy measure for the path length between these distributions. In all tested cases, over a variety of RBF kernel bandwidths, our latent space is closer to Gaussian noise than that of SD-VAE.

Lastly, we trained two diffusion models on the FFHQ and CelebA-HQ datasets and compared their performance with standard VAE-based LDMs. The diffusion model architecture used for this experiment is a UNet identical to the original model from Rombach et al. [55]. Table 7 shows that the diffusion models trained in the latent space of LiteVAE perform similarly to (or slightly better than) the standard LDMs. Additionally, Figure 6 includes some generated examples from our FFHQ model. These results suggest that diffusion models are also capable of modeling the latent space of LiteVAE.

8

Figure 6: Generated samples from the FFHQ model.
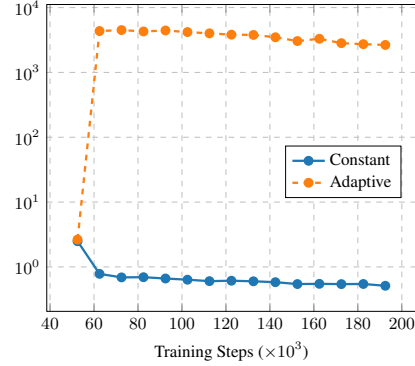
Figure 7: Relative gradient norm of the adversarial and the reconstruction loss.

Table 8: Reconstruction quality after using constant weight for the adversarial loss.

Weight Type	rFID ↓	LPIPS ↓	PSNR ↑	SSIM ↑
Adaptive	**1.01**	**0.07**	29.24	0.86
Constant	**1.01**	**0.07**	**29.33**	**0.87**

Table 9: Effect of using Gaussian and wavelet loss on final reconstruction quality.

Training Config	rFID ↓	LPIPS ↓	PSNR ↑	SSIM ↑
Baseline	0.99	0.070	29.33	0.86
+ Gaussian loss	0.99	0.070	29.64	0.87
+ wavelet loss	**0.96**	**0.069**	**29.73**	**0.88**

Table 10: Reconstruction quality for different discriminators.

Discriminator	rFID ↓	LPIPS ↓	PSNR ↑	SSIM ↑
UNet	**1.01**	0.070	29.24	0.86
StyleGAN	1.38	0.065	29.51	**0.87**
PatchGAN	1.61	**0.063**	**29.61**	**0.87**

6 Ablation studies

We next present our main ablation studies to determine the individual impact of the changes proposed in Section 4. We use the ImageNet 128×128 model with a latent size of $32\times32\times12$ as the baseline for all ablations. Further ablation studies on other design choices in LiteVAE are provided in Appendix D.

Removing adaptive weight for λ_{reg}   Table 8 demonstrates that we can safely remove the adaptive weight for the adversarial loss (Equation (2)) and still slightly improve the metrics. Figure 7 also shows the relative norm of the gradient of the adversarial loss compared to the reconstruction loss for both adaptive and constant λ_{adv}. We observe that using adaptive λ_{adv} leads to more imbalanced gradient ratios, and hence less stable training, especially for mixed-precision scenarios. Accordingly, we exclusively use a constant weight for the adversarial loss in our experiments.

High-frequency loss functions   Table 9 shows the effect of adding high-frequency losses based on Gaussian filtering and the wavelet transform. The addition of these high-frequency loss terms during training consistently improves all reconstruction metrics.

Choice of the discriminator   We finally show that using a UNet-based discriminator [61] outperforms both PatchGAN and StyleGAN discriminators used in previous works [55, 72] in terms of rFID while having comparable performance for other metrics. We also empirically noted that using a UNet discriminator resulted in more stable training across different runs and hyperparameters. The full comparison for this experiment is given in Table 10.

9

7 Conclusion

In this paper, we presented LiteVAE, a new design concept for autoencoders based on the multi-resolution wavelet transform. LiteVAE can match the performance of standard VAEs while requiring significantly less compute. We also analyzed the design space and training of this proposed family of autoencoders and offered several modifications that further improve the final reconstruction quality and training dynamics of the base model. Overall, LiteVAE offers more flexibility in terms of performance/compute trade-off and outperforms the naïve approach of making the VAE encoder smaller. Our current work is focused on improving efficiency in the models responsible for encoding the latent representation of natural images, and whether the efficiency benefits of LiteVAE extend to other domains is a question we leave to follow-up work. Although we introduced LiteVAE in the context of LDMs, we hypothesize that its application is not confined to this scenario. We consider the extension of LiteVAE to other autoencoder-based generative modeling schemes (e.g., tokenization) a promising avenue for further research.

References

[1] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *CoRR*, abs/2211.01324, 2022. doi: 10.48550/arXiv.2211.01324. URL https://doi.org/10.48550/arXiv.2211.01324.

[2] Jonathan T Barron. A general and adaptive robust loss function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4331–4339, 2019.

[3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *CoRR*, abs/2311.15127, 2023. doi: 10.48550/ARXIV.2311.15127. URL https://doi.org/10.48550/arXiv.2311.15127.

[4] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023.

[5] M. E. Brewster. An introduction to wavelets (charles k. chui). *SIAM Rev.*, 35(2):312–313, 1993. doi: 10.1137/1035061. URL https://doi.org/10.1137/1035061.

[6] S Grace Chang, Bin Yu, and Martin Vetterli. Adaptive wavelet thresholding for image denoising and compression. *IEEE transactions on image processing*, 9(9):1532–1546, 2000.

[7] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J. Weiss, Mohammad Norouzi, and William Chan. Wavegrad: Estimating gradients for waveform generation. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=NsMLjcFaO8O.

[8] Xiaojie Chu, Liangyu Chen, and Wenqing Yu. Nafssr: Stereo image super-resolution using nafnet. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1239–1248, June 2022.

[9] Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, et al. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807*, 2023.

[10] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat gans on image synthesis. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 8780–8794, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/49ad23d1ec9fa4bd8d77d02681df5cfa-Abstract.html.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly,

10

Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

[12] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 12873–12883. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.01268. URL https://openaccess.thecvf.com/content/CVPR2021/html/Esser_Taming_Transformers_for_High-Resolution_Image_Synthesis_CVPR_2021_paper.html.

[13] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. *CoRR*, abs/2403.03206, 2024. doi: 10.48550/ARXIV.2403.03206. URL https://doi.org/10.48550/arXiv.2403.03206.

[14] Mário AT Figueiredo and Robert D Nowak. An em algorithm for wavelet-based image restoration. *IEEE Transactions on Image Processing*, 12(8):906–916, 2003.

[15] Rinon Gal, Dana Cohen Hochberg, Amit Bermano, and Daniel Cohen-Or. SWAGAN: a style-based wavelet-driven generative model. *ACM Trans. Graph.*, 40(4):134:1–134:11, 2021. doi: 10.1145/3450626.3459836. URL https://doi.org/10.1145/3450626.3459836.

[16] Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. Masked diffusion transformer is a strong image synthesizer. *CoRR*, abs/2303.14389, 2023. doi: 10.48550/arXiv.2303.14389. URL https://doi.org/10.48550/arXiv.2303.14389.

[17] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[18] Tiantong Guo, Hojjat Seyed Mousavi, Tiep Huu Vu, and Vishal Monga. Deep wavelet prediction for image super-resolution. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1100–1109. IEEE Computer Society, 2017. doi: 10.1109/CVPRW.2017.148. URL https://doi.org/10.1109/CVPRW.2017.148.

[19] Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Li Fei-Fei, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic video generation with diffusion models. *arXiv preprint arXiv:2312.06662*, 2023.

[20] Florentin Guth, Simon Coste, Valentin De Bortoli, and Stéphane Mallat. Wavelet score-based generative modeling. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/03474669b759f6d38cdca6fb4eb905f4-Abstract-Conference.html.

[21] Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Dong Chen, Han Hu, Xin Geng, and Baining Guo. Efficient diffusion training via min-snr weighting strategy. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 7407–7417. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00684. URL https://doi.org/10.1109/ICCV51070.2023.00684.

[22] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6626–6637, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.

[23] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *CoRR*, abs/2207.12598, 2022. doi: 10.48550/arXiv.2207.12598. URL https://doi.org/10.48550/arXiv.2207.12598.

11

[24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

[25] Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. Improving sample quality of diffusion models using self-attention guidance. *CoRR*, abs/2210.00939, 2022. doi: 10.48550/arXiv.2210.00939. URL https://doi.org/10.48550/arXiv.2210.00939.

[26] Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images. *CoRR*, abs/2301.11093, 2023. doi: 10.48550/arXiv.2301.11093. URL https://doi.org/10.48550/arXiv.2301.11093.

[27] Huaibo Huang, Ran He, Zhenan Sun, and Tieniu Tan. Wavelet-srnet: A wavelet-based CNN for multi-scale face super resolution. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 1698–1706. IEEE Computer Society, 2017. doi: 10.1109/ICCV.2017.187. URL https://doi.org/10.1109/ICCV.2017.187.

[28] Qingqing Huang, Daniel S. Park, Tao Wang, Timo I. Denk, Andy Ly, Nanxin Chen, Zhengdong Zhang, Zhishuai Zhang, Jiahui Yu, Christian Havnø Frank, Jesse H. Engel, Quoc V. Le, William Chan, and Wei Han. Noise2music: Text-conditioned music generation with diffusion models. *CoRR*, abs/2302.03917, 2023. doi: 10.48550/arXiv.2302.03917. URL https://doi.org/10.48550/arXiv.2302.03917.

[29] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=Hk99zCeAb.

[30] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4401–4410. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00453. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Karras_A_Style-Based_Generator_Architecture_for_Generative_Adversarial_Networks_CVPR_2019_paper.html.

[31] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8107–8116. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.00813. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Karras_Analyzing_and_Improving_the_Image_Quality_of_StyleGAN_CVPR_2020_paper.html.

[32] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. 2022. URL https://openreview.net/forum?id=k7FuTOWMOc7.

[33] Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models, 2023.

[34] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

[35] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. 2014. URL http://arxiv.org/abs/1312.6114.

[36] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=a-xFK8Ymz5J.

[37] Jie Liang, Hui Zeng, and Lei Zhang. Details or artifacts: A locally discriminative learning approach to realistic image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

12

[38] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[39] Guan-Horng Liu, Arash Vahdat, De-An Huang, Evangelos A. Theodorou, Weili Nie, and Anima Anandkumar. I^{2}sb: Image-to-image schrödinger bridge. *CoRR*, abs/2302.05872, 2023. doi: 10.48550/arXiv.2302.05872. URL https://doi.org/10.48550/arXiv.2302.05872.

[40] Pengju Liu, Hongzhi Zhang, Kai Zhang, Liang Lin, and Wangmeng Zuo. Multi-level wavelet-cnn for image restoration. In *2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 773–782. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPRW.2018.00121. URL http://openaccess.thecvf.com/content_cvpr_2018_workshops/w13/html/Liu_Multi-Level_Wavelet-CNN_for_CVPR_2018_paper.html.

[41] Haichuan Ma, Dong Liu, Ning Yan, Houqiang Li, and Feng Wu. End-to-end optimized versatile image compression with wavelet-like transform. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1247–1263, 2020.

[42] Stéphane Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 11(7):674–693, 1989. doi: 10.1109/34.192463. URL https://doi.org/10.1109/34.192463.

[43] Stéphane Mallat. *A wavelet tour of signal processing*. Elsevier, 1999.

[44] Stephane G Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE transactions on pattern analysis and machine intelligence*, 11(7):674–693, 1989.

[45] S Kother Mohideen, S Arumuga Perumal, and M Mohamed Sathik. Image de-noising using discrete wavelet transform. *International Journal of Computer Science and Network Security*, 8 (1):213–216, 2008.

[46] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8162–8171. PMLR, 2021. URL http://proceedings.mlr.press/v139/nichol21a.html.

[47] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in GAN evaluation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 11400–11410. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01112. URL https://doi.org/10.1109/CVPR52688.2022.01112.

[48] William Peebles and Saining Xie. Scalable diffusion models with transformers. *CoRR*, abs/2212.09748, 2022. doi: 10.48550/arXiv.2212.09748. URL https://doi.org/10.48550/arXiv.2212.09748.

[49] Hao Phung, Quan Dao, and Anh Tran. Wavelet diffusion models are fast and scalable image generators. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 10199–10208. IEEE, 2023. doi: 10.1109/CVPR52729.2023.00983. URL https://doi.org/10.1109/CVPR52729.2023.00983.

[50] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: improving latent diffusion models for high-resolution image synthesis. *CoRR*, abs/2307.01952, 2023. doi: 10.48550/ARXIV.2307.01952. URL https://doi.org/10.48550/arXiv.2307.01952.

[51] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=FjNys5c7VyY.

[52] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR, 2021. URL http://proceedings.mlr.press/v139/ramesh21a.html.

13

[53] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *CoRR*, abs/2204.06125, 2022. doi: 10.48550/arXiv.2204.06125. URL https://doi.org/10.48550/arXiv.2204.06125.

[54] Oren Rippel and Lubomir Bourdev. Real-time adaptive image compression. In *International Conference on Machine Learning*, pages 2922–2930. PMLR, 2017.

[55] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01042. URL https://doi.org/10.1109/CVPR52688.2022.01042.

[56] Negar Rostamzadeh, Emily Denton, and Linda Petrini. Ethics and creativity in computer vision. *CoRR*, abs/2112.03111, 2021. URL https://arxiv.org/abs/2112.03111.

[57] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y. URL https://doi.org/10.1007/s11263-015-0816-y.

[58] Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M. Weber. CADS: Unleashing the diversity of diffusion models through condition-annealed sampling. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=zMoNrajk2X.

[59] Chitwan Saharia, William Chan, Huiwen Chang, Chris A. Lee, Jonathan Ho, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In Munkhtsetseg Nandigjav, Niloy J. Mitra, and Aaron Hertzmann, editors, *SIGGRAPH '22: Special Interest Group on Computer Graphics and Interactive Techniques Conference, Vancouver, BC, Canada, August 7 - 11, 2022*, pages 15:1–15:10. ACM, 2022. doi: 10.1145/3528233.3530757. URL https://doi.org/10.1145/3528233.3530757.

[60] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/ec795aeadae0b7d230fa35cbaf04c041-Abstract-Conference.html.

[61] Edgar Schonfeld, Bernt Schiele, and Anna Khoreva. A u-net based discriminator for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8207–8216, 2020.

[62] Ke Shen and Edward J Delp. Wavelet based rate scalable video compression. *IEEE transactions on circuits and systems for video technology*, 9(1):109–122, 1999.

[63] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016.

[64] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. 37:2256–2265, 2015. URL http://proceedings.mlr.press/v37/sohl-dickstein15.html.

[65] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 11895–11907, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/3001ef257407d5a371a96dcd947c7d93-Abstract.html.

[66] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

14

[67] David S. Taubman and Michael W. Marcellin. *JPEG2000 - image compression fundamentals, standards and practice*, volume 642 of *The Kluwer international series in engineering and computer science*. Kluwer, 2002. ISBN 978-0-7923-7519-7. doi: 10.1007/978-1-4615-0799-4. URL https://doi.org/10.1007/978-1-4615-0799-4.

[68] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *The European Conference on Computer Vision Workshops (ECCVW)*, September 2018.

[69] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[70] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Yingxia Shao, Wentao Zhang, Ming-Hsuan Yang, and Bin Cui. Diffusion models: A comprehensive survey of methods and applications. *CoRR*, abs/2209.00796, 2022. doi: 10.48550/arXiv.2209.00796. URL https://doi.org/10.48550/arXiv.2209.00796.

[71] Mengping Yang, Zhe Wang, Ziqiu Chi, and Yanbing Zhang. Fregan: Exploiting frequency components for training gans under limited data. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/d804cef41362be39d3972c1a71cfc4e9-Abstract-Conference.html.

[72] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved VQGAN. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=pfNyExj7z2.

[73] Yingchen Yu, Fangneng Zhan, Shijian Lu, Jianxiong Pan, Feiying Ma, Xuansong Xie, and Chunyan Miao. Wavefill: A wavelet-based generation network for image inpainting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14114–14123, 2021.

[74] Eduard Zamfir, Marcos V Conde, and Radu Timofte. Towards real-time 4k image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1522–1532, 2023.

[75] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 586–595. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00068. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.html.

[76] Zixin Zhu, Xuelu Feng, Dongdong Chen, Jianmin Bao, Le Wang, Yinpeng Chen, Lu Yuan, and Gang Hua. Designing a better asymmetric VQGAN for stablediffusion. *CoRR*, abs/2306.04632, 2023. doi: 10.48550/ARXIV.2306.04632. URL https://doi.org/10.48550/arXiv.2306.04632.

15

A   Broader impact statement

Our work can significantly reduce the training time and memory requirements of autoencoders in latent diffusion models (LDMs). Given the rising popularity of LDMs, our approach holds promise for positive environmental impacts and significant advancements in generative modeling. It is important to note that while AI-generated content can enhance productivity and creativity, we must remain mindful of the potential risks and ethical concerns involved. For a deeper discussion of ethics and creativity in computer vision, readers are directed to [56].

B   Using a non-learned encoder

This section provides further motivation behind the design of LiteVAE. We investigate using a non-learned (i.e., fixed) encoder in two settings: (1) for simple datasets such as FFHQ [30] and (2) for more diverse datasets such as ImageNet [57]. We use the reconstruction FID (rFID) [22] as our measure of reconstruction quality, aiming to achieve the SD-VAE's downsampling factor of $f = 8$. The analysis leads to two observations. First, the non-learned autoencoder (although efficient) can provide high-quality reconstructions only if we use a larger channel depth for the encoder network compared to SD-VAE. Secondly, the dense latent space learned by the autoencoder provides a better structure for generative modeling. LiteVAE essentially combines the computational benefits of the non-learned encoder with the learned latent space of a regular VAE.

Simple datasets   For simpler datasets like FFHQ, it is possible to completely replace the encoder $\mathcal{E}$ with a predefined function and get similar reconstruction quality. In our case, we used a three-level DWT and only kept the sub-bands of the lowest level. We then trained a decoder to convert the lowest-level sub-bands back to the image. Table 11 shows the results of this approach on two relatively restricted datasets. We observe that this wavelet representation offers a similar reconstruction quality to a learned encoder while reducing the number of encoder parameters from

Table 11: The performance of the DWT-based encoder on simple datasets.

Dataset	Encoder	n_z	rFID ↓
FFHQ	SD-VAE	4	0.85
	DWT	12	0.70
DeepFashion	SD-VAE	4	1.64
	DWT	12	1.71

about 34M to zero. This experiment indicates that with the help of rich image representations from the wavelet transform, we can speed up SD-VAE by reducing the complexity of the encoder.

Complex datasets   The next step is to explore whether this non-learned encoder setup is scalable to more diverse datasets such as ImageNet. Table 12 demonstrates that while the non-learned encoder is effective in simpler scenarios, it falls short of the quality of normal VAEs in more complex settings. This indicates that the information present in the higher frequency sub-bands of the wavelet transform is essential for the decoder to reconstruct more diverse images with higher quality. To

Table 12: The performance of the non-learned encoder on ImageNet.

Dataset	Encoder	n_z	rFID ↓
	SD-VAE	4	0.80
ImageNet	DWT	12	1.65
	DWT-2	48	0.32

validate this hypothesis, we incorporate the information from higher frequency sub-bands via a space-to-depth operation [63] in the encoder and observe that we can recover the high reconstruction quality of the learned encoder (DWT-2 in Table 12). However, this approach is not preferable because the channel dimension is now too high for generative modeling.

Importance of having a learned latent space   Finally, we demonstrate that although it is possible to completely replace the encoder of the VAE with a non-learned wavelet-based latent representation for the FFHQ dataset, the learned latent space in LiteVAE offers a better structure for training diffusion models. Table 13 indicates that training the diffusion model on the learned latent code of Lite-VAE outperforms the non-learned DWT representation. We argue that the sparse nature of wavelets is harmful to

Table 13: Comparison between a non-learned encoder and LiteVAE for training diffusion models.

Dataset	Encoder	FID ↓
FFHQ (256×256)	non-learned	12.51
	LiteVAE	**8.03**

generation quality compared to the dense representation learned by the encoder of LiteVAE.

16

C Summary of diffusion models

Diffusion models learn the data distribution p_{data} by reversing a noising process that gradually converts a data point $\boldsymbol{x}$ into random Gaussian noise. More specifically, diffusion models define a forward process via $\boldsymbol{x}_t = \boldsymbol{x} + \sigma(t)\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I})$. Then, they train a denoiser network $D_{\boldsymbol{\theta}}$ to estimate the clean signal $\boldsymbol{x}$ from the current noisy sample $\boldsymbol{x}_t$. It has been shown that this process corresponds to the following stochastic differential equation (SDE) [66, 32]

$$\mathrm{d}\boldsymbol{x}_t = -\dot{\sigma}(t)\sigma(t)\,\nabla_{\boldsymbol{x}_t}\log p_t(\boldsymbol{x}_t)\,\mathrm{d}t - \beta(t)\sigma(t)^2\,\nabla_{\boldsymbol{x}_t}\log p_t(\boldsymbol{x}_t)\,\mathrm{d}t + \sqrt{2\beta(t)}\sigma(t)\,\mathrm{d}\omega_t, \quad (4)$$

where $\mathrm{d}\omega_t$ is the standard Wiener process, $p_t(\boldsymbol{x}_t)$ is the distribution of noisy samples at time t, and $\beta(t)$ is a term that controls the influence of noise during the sampling process. The denoiser network $D_{\boldsymbol{\theta}}$ effectively approximates the score function $\nabla_{\boldsymbol{x}_t}\log p_t(\boldsymbol{x}_t)$. Given that $p_0 = p_{\text{data}}$ and $p_1 = \mathcal{N}(\boldsymbol{0}, \sigma_{\max}^2\boldsymbol{I})$, sampling new data points is then possible by starting from random Gaussian noise and solving the corresponding SDE reverse in time.

Latent diffusion models [55] follow the same methodology, but instead of performing the forward and backward process in the pixel space, they first convert the data into the latent codes via a pretrained VAE and employ the diffusion process in the latent space. Please refer to Karras et al. [32] and Yang et al. [70] for more details on diffusion models.

D Additional ablation studies

This section contains additional ablation studies on the design space and training dynamic of LiteVAE.

D.1 Training loss functions

We experimented with the following changes to the loss functions during training of the autoencoder to measure whether they lead to any improvement in reconstruction quality.

Changing the VGG loss   Wang et al. [68] proposed a different VGG loss function based on the features *before* the activation layer. We also ablated this choice against the standard LPIPS loss typically used in the VAEs of LDMs. Table 14 indicates that this change has a considerable boost to the reconstruction FID at the cost of lower PSNR. As the perceptual quality is more important for LDMs compared to distortion metrics, we recommend switching to this loss function instead of the LPIPS loss. However, we used the LPIPS loss for the experiments in the main text to have a similar training setup with commonly used VAEs in LDMs.

Including the locally discriminative learning (LDL) loss   Liang et al. [37] introduced the LDL loss function to reduce the artifacts caused by the discriminator in the super-resolution context. We also experimented with this loss term and found that it does not have any noticeable impact on the reconstruction quality of LiteVAE, as shown in Table 15.

Choosing different adversarial loss functions   We also ablated the adversarial loss function for two different setups: a hinge loss, and a non-saturating (logistic) loss. As depicted in Table 16, we observe that the hinge loss generally leads to slightly better rFID while the logistic loss achieves slightly better PSNR. Since the adversarial loss in the autoencoder training is only responsible for increasing the photorealism of the outputs, we conclude that both loss terms work equally well.

D.2 Role of the 1×1 convolution layers

Ramesh et al. [52] showed that using a 1×1 convolution after the output of the encoder and before the input of the decoder improves the approximation accuracy of the evidence lower bound (ELBO) term in the loss function. We ablated this design choice in the context of LiteVAE and found that restricting the receptive field of the latent space with these 1×1 convolution layers might be harmful to the reconstruction quality by enforcing too much KL regularization. Table 17 shows that removing these convolution blocks leads to much better reconstruction quality in our 256×256 model. Accordingly, we suggest removing these 1×1 convolutions from the model (or, equivalently, adjusting the weight of the KL loss) to get better reconstruction.

17

Table 14: Ablation on using different VGG loss functions for the perceptual loss.

Perceptual Loss Type	rFID ↓	LPIPS ↓	PSNR ↑	SSIM ↑
LPIPS [75]	0.99	**0.071**	**29.33**	**0.86**
ESRGAN [68]	**0.78**	**0.071**	28.53	0.84

Table 15: Ablation on the effect of adding the LDL loss [37].

with LDL	rFID ↓	LPIPS ↓	PSNR ↑	SSIM ↑
✗	0.99	**0.071**	**29.33**	0.86
✓	**0.98**	**0.071**	**29.33**	**0.87**

Table 16: Ablation on using different adversarial loss functions.

Adversarial Loss	rFID ↓	LPIPS ↓	PSNR ↑	SSIM ↑
Hinge	**0.99**	0.071	29.33	0.86
Logistic	1.00	**0.068**	**29.67**	**0.88**

Table 17: Ablation on the effect of 1×1 convolution layers.

1×1 Conv	rFID ↓	LPIPS ↓	PSNR ↑	SSIM ↑
✓	2.04	0.190	25.61	0.72
✗	**0.87**	**0.157**	**26.02**	**0.74**

Table 18: Ablation on using NAFNet [8] for feature extraction.

$\mathcal{F}_l$	rFID ↓	LPIPS ↓	PSNR ↑	SSIM ↑
UNet	0.94	**0.069**	**29.55**	**0.87**
NAFNet	**0.93**	**0.069**	29.36	**0.87**

D.3 Different networks for feature extraction

We also experimented with NAFNet [8] instead of the UNet for extracting features from wavelet sub-bands and observed that it performs similarly to the UNet architecture mentioned in the main text. The results are given in Table 18. This experiment indicates that other network choices for the feature-extraction module are indeed possible, and LiteVAE is flexible w.r.t. this design choice. We chose the UNet to keep the setup as close as possible to the standard VAE design in LDMs.

D.4 Sharing the weights of the feature-extraction UNet

We next investigated whether a single UNet could be shared across different wavelet sub-bands to further reduce the encoder's trainable parameters. Table 19 demonstrates that it is indeed possible to share $\mathcal{F}_l$ between different sub-bands. A shared UNet might lead to the post hoc usage of the encoder across different wavelet levels and resolutions at inference. However, as the computational cost (in terms of GFLOPS) does not change with parameter sharing, we did not use this technique for the main experiments.

D.5 Using ViT for feature aggregation

We also explore the use of non-convolutional vision transformer (ViT) blocks [11] for feature aggregation $\mathcal{F}_{\text{agg}}$. As indicated in Table 20, employing ViT achieves comparable reconstruction quality to that of a fully-convolutional encoder, but with fewer parameters. However, it is important to note that incorporating ViT makes the model resolution-dependent. This is a drawback, as the VAE in LDMs is usually required to operate on data with varying resolutions. Hence, we side with the UNet models to make the encoder resolution-independent.

18

Table 19: Ablation on parameter sharing for the feature-extraction module.

Shared UNet	rFID ↓	LPIPS ↓	PSNR ↑	SSIM ↑
✗	**0.75**	**0.153**	**26.10**	**0.73**
✓	0.78	0.154	26.05	**0.73**

Table 20: Ablation on using ViT for feature aggregation.

$\mathcal{F}_{agg}$	Params (M)	rFID ↓	LPIPS ↓	PSNR ↑	SSIM ↑
UNet	1.69	0.94	**0.069**	29.25	0.86
ViT	0.84	**0.92**	0.070	**29.44**	**0.87**

Table 21: Ablation on removing the highest resolution wavelets from feature extraction.

Config	rFID ↓	LPIPS ↓	PSNR ↑	SSIM ↑
All sub-bands	**0.87**	**0.157**	**26.02**	**0.74**
Last two sub-bands	1.20	0.17	26.04	0.74
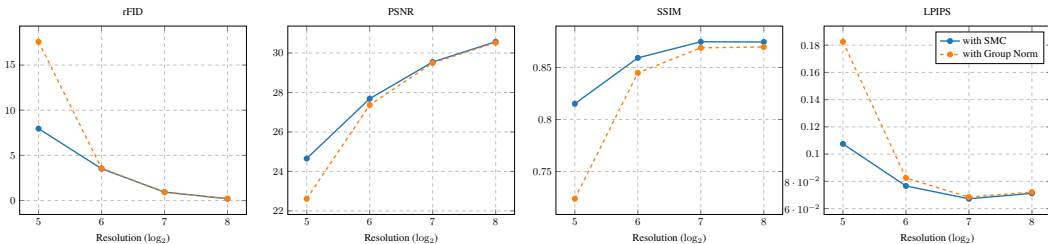
Figure 8: Comparing the performace of LiteVAE with and without Group Normalization. Using SMC instead of Group Norm makes the autoencoder less scale-dependent.

D.6 Importance of using all wavelet levels

We also explored the possibility of performing feature extraction on only a subset of wavelet coefficients rather than across all wavelet levels. As shown in Table 21, this approach negatively impacts reconstruction performance on ImageNet, indicating that incorporating information from all wavelet levels is essential for high-quality reconstruction, particularly with complex datasets.

D.7 Scale dependency of SMC

This section shows that using SMC improves the scale dependency of LiteVAE. The results in Figure 8 indicate that using SMC instead of group normalization leads to less degradation in performance as we change the resolution of the evaluation dataset. We argue that removing the imbalanced feature maps aids the network in learning features that are less scale-dependent.

D.8 Training resolution for the standard VAEs

This experiment validates that the idea of pretraining the autoencoder at 128×128 followed by fine-tuning at 256×256 also works for the standard VAEs. The results of this experiment are given in Table 22. Similar to LiteVAE, the 128-tuned model matches the performance of the 256-full model while requiring considerably less training compute.

E   Additional generated samples

Figure 9 provides additional generated samples from our latent diffusion model trained on FFHQ.

19

Table 22: Effect of pretraining the autoencoder at lower resolutions for a standard VAE model. The 128-tuned model performs similarly to the model trained solely on 256×256 data.

Training Config	rFID ↓	LPIPS ↓	PSNR ↑	SSIM ↑
256-full	**0.67**	**0.150**	**26.01**	**0.74**
128-full	0.89	0.161	25.83	0.73
128-tuned	0.69	0.151	25.97	0.73

Figure 9: Additional uncurated generations from the FFHQ diffusion model

Table 23: Details of the feature-extraction module for different LiteVAE models.

Model	Feature extraction			
	Input dim	Output dim	Channels	Channels multiple
LiteVAE-S	12	12	16	(1, 2, 2)
LiteVAE-B	12	12	32	(1, 2, 3)
LiteVAE-M	12	12	64	(1, 2, 4)
LiteVAE-L	12	12	64	(1, 2, 4)

Table 24: Details of the feature-aggregation module for different LiteVAE models.

Model	Feature aggregation			
	Input dim	Output dim	Channels	Channels multiple
LiteVAE-S	36	Latent dim (n_z)	16	(1, 2, 2)
LiteVAE-B	36	Latent dim (n_z)	32	(1, 2, 3)
LiteVAE-M	36	Latent dim (n_z)	32	(1, 2, 3)
LiteVAE-L	36	Latent dim (n_z)	64	(1, 2, 4)

F   Implementation details

All models were trained with a batch size of 16 on two GPUs until the autoencoder could produce high-quality reconstructions. The training duration was 200k steps for the ImageNet 128×128 models, and 100k for the ImageNet 256×256 and FFHQ models. We use Adam optimizer [34] with a learning rate of 10^{-4} and $(\beta_1, \beta_2) = (0.5, 0.9)$. The details of the model architecture for feature-extraction and feature-aggregation modules are given in Tables 23 and 24. Our implementation of the UNet used for feature extraction and aggregation closely follows the ADM model [10] without spatial down/upsampling layers. The decoder in LiteVAE exactly follows the implementation of the decoder from Stable Diffusion VAE [55], except for the SMC experiment. For training the latent diffusion and the standard VAE models, we closely follow Rombach et al. [55] to ensure a fair comparison.

20

G Pseudocode for different LiteVAE blocks

In this section, we present additional pseudocode for various LiteVAE components. The core element of LiteVAE is the Haar wavelet transform, which can be implemented in PyTorch as shown below:

```python
class HaarTransform(nn.Module):
  def __init__(self, level=3, mode="symmetric", with_grad=False) -> None:
    super().__init__()
    self.wavelet = pywt.Wavelet("haar")
    self.level = level
    self.mode = mode
    self.with_grad = with_grad

  def dwt(self, x, level=None):
    with torch.set_grad_enabled(self.with_grad):
      level = level or self.level
      x_low, *x_high = ptwt.wavedec2(
        x.float(),
        wavelet=self.wavelet,
        level=level,
        mode=self.mode,
      )
      x_combined = torch.cat(
        [x_low, x_high[0][0], x_high[0][1], x_high[0][2]], dim=1
      )
      return x_combined

  def idwt(self, x):
    with torch.set_grad_enabled(self.with_grad):
      x_low, x_high = x[:, :3], x[:, 3:]
      x_high = torch.chunk(x_high, 3, dim=1)
      x_recon = ptwt.waverec2([x_low.float(), x_high.float()],
        ↪  wavelet=self.wavelet)
      return x_recon

  def forward(self, x, inverse=False):
    if inverse:
      return self.idwt(x)
    return self.dwt(x)
```

The PyTorch implementation of the self-modulated convolution block introduced in Section 4.2 is provided below:

```python
class SMC(nn.Module):
  def __init__(
    self,
    in_channels: int,
    out_channels: int = None,
    kernel_size: int = 3,
```

21

```
7      stride: int = None,
8      padding: int = None,
9      bias: bool = True,
10 ):
11     super().__init__()
12
13     # setting the default values
14     out_channels = out_channels or in_channels
15     padding_ = int(kernel_size // 2) if padding is None else padding
16     stride_ = 1 if stride is None else stride
17
18     self.padding = padding_
19
20     self.conv = nn.Conv2d(
21       in_channels,
22       out_channels,
23       kernel_size=kernel_size,
24       padding=padding_,
25       stride=stride_,
26       bias=bias,
27     )
28
29     self.gain = nn.Parameter(torch.ones(1))
30     self.scales = nn.Parameter(torch.ones(in_channels))
31
32   def forward(self, x: torch.Tensor) -> torch.Tensor:
33     scales = self.scales.expand(x.shape[0], -1)
34     out = modulated_conv2d(
35       x=x,
36       w=self.conv.weight,
37       s=scales,
38       padding=self.padding,
39       input_gain=self.gain,
40     )
41     if self.conv.bias is not None:
42       out = out + self.conv.bias.view(1, -1, 1, 1)
43     return out
```

Next, we present the code for the residual blocks utilized in the LiteVAE UNet networks:

```
1 class ResBlock(nn.Module):
2   def __init__(
3     self,
4     in_channels: int,
5     dropout: float = 0.0,
6     out_channels: int = None,
7     use_conv: bool = False,
8     activation: str = "swish",
```

22

```python
      norm_num_groups: int = 32,
      scale_factor: float = 1,
   ):

   super().__init__()
   self.in_channels = in_channels
   self.out_channels = out_channels or in_channels

   self.norm_in = GroupNorm(in_channels, norm_num_groups)
   self.act_in = SiLU()
   self.conv_in = ConvLayer2D(in_channels, out_channels, 3)
   self.norm_out = GroupNorm(out_channels, norm_num_groups)
   self.act_out = SiLU()
   self.dropout = Dropout(dropout)
   self.conv_out = ConvLayer2D(out_channels, 3)

   if self.out_channels == in_channels:
      self.skip_connection = Identity()
   elif use_conv:
      self.skip_connection = ConvLayer2D(in_channels, out_channels, 3)
   else:
      self.skip_connection = ConvLayer2D(in_channels, out_channels, 1)
   self.scale_factor = scale_factor

def forward(self, x):
   # input layers
   h = self.norm_in(x)
   h = self.act_in(h)
   h = self.conv_in(h)
   # output layers
   h = self.norm_out(h)
   h = self.act_out(h)
   h = self.dropout(h)
   h = self.conv_out(h)
   return (self.skip_connection(x) + h) / self.scale_factor

class ResBlockWithSMC(nn.Module):
   def __init__(
      self,
      in_channels: int,
      dropout: float = 0.0,
      out_channels: int = None,
      use_conv: bool = False,
      activation: str = "swish",
      norm_num_groups: int = 32,
      scale_factor: float = 1,
   ):
```

23

```python
57
58      super().__init__()
59      self.in_channels = in_channels
60      self.out_channels = out_channels or in_channels
61
62      self.act_in = SiLU()
63      self.conv_in = SMC(in_channels, out_channels, 3)
64      self.act_out = SiLU()
65      self.dropout = Dropout(dropout)
66      self.conv_out = SMC(out_channels, 3)
67
68      if self.out_channels == in_channels:
69        self.skip_connection = Identity()
70      elif use_conv:
71        self.skip_connection = ConvLayer2D(in_channels, out_channels, 3)
72      else:
73        self.skip_connection = ConvLayer2D(in_channels, out_channels, 1)
74      self.scale_factor = scale_factor
75
76    def forward(self, x):
77      # input layers
78      h = self.act_in(x)
79      h = self.conv_in(h)
80      # output layers
81      h = self.act_out(h)
82      h = self.dropout(h)
83      h = self.conv_out(h)
84      return (self.skip_connection(x) + h) / self.scale_factor
85
86
87  class MidBlock2D(nn.Module):
88    def __init__(
89      self,
90      in_channels: int,
91      out_channels: int,
92      dropout: float = 0.0,
93      use_smc: bool = True,
94    ) -> None:
95      super().__init__()
96      resblock_class = ResBlockWithSMC if use_smc else ResBlock
97      self.res0 = resblock_class(
98        in_channels=in_channels,
99        out_channels=out_channels,
100       dropout=dropout,
101     )
102     self.res1 = resblock_class(
103       in_channels=out_channels,
104       out_channels=out_channels,
```

24

```
105        dropout=dropout,
106      )
107    def forward(self, x):
108      x = self.res0(x)
109      x = self.res1(x)
110      return x
```

Additionally, the feature-extraction and feature-aggregation UNets can be implemented as follows:

```
1  class LiteVAEUNetBlock(nn.Module):
2    def __init__(
3      self,
4      in_channels: int,
5      out_channels: int,
6      model_channels: int,
7      ch_multiplies: list[int] = [1, 2, 4],
8      num_res_blocks: int = 2,
9      use_smc: bool = False,
10   ):
11     super().__init__()
12     self.in_layer = ConvLayer2D(in_channels, model_channels, 3)
13     self.out_layer = ConvLayer2D(model_channels, out_channels, 3)
14
15     resblock_class = ResBlockWithSMC if use_smc else ResBlock
16
17     # ----------------------------------------------------------------------
18     # UNet encoder path
19     # ----------------------------------------------------------------------
20     channel = model_channels
21     in_channel_list = [model_channels]
22     self.encoder_blocks = []
23     for level, ch_mult in enumerate(ch_multiplies):
24       for i in range(num_res_blocks):
25         self.encoder_blocks.append(
26           resblock_class(
27             in_channels=channel,
28             out_channels=model_channels * ch_mult
29           )
30         )
31         channel = model_channels * ch_mult
32         in_channel_list.append(channel)
33     self.encoder_blocks = nn.ModuleList(self.encoder_blocks)
34     # ----------------------------------------------------------------------
35     # UNet middle block
36     # ----------------------------------------------------------------------
37     self.mid_block = MidBlock2D(
38       in_channels=channel,
39       out_channels=channel,
```

25

```python
40        embed_channels=0,
41        legacy=legacy
42      )
43      # --------------------------------------------------------------------
44      # UNet decoder path
45      # --------------------------------------------------------------------
46      self.decoder_blocks = []
47      for level, ch_mult in reversed(list(enumerate(ch_multiplies))):
48        for i in range(num_res_blocks):
49          self.decoder_blocks.append(
50            resblock_class(
51              in_channels=channel + in_channel_list.pop(),
52              out_channels=model_channels * ch_mult
53            )
54          )
55          channel = model_channels * ch_mult
56      self.decoder_blocks = nn.ModuleList(self.decoder_blocks)

58  def forward(self, x):
59      x = self.in_layer(x)
60      skip_features = [x]
61      # the encoder path
62      for enc_block in self.encoder_blocks:
63        x = enc_block(x)
64        skip_features.append(x)
65      # the middle block
66      x = self.mid_block(x)
67      # the decoder path
68      for dec_block in self.decoder_blocks:
69        x_cat = torch.cat([x, skip_features.pop()], dim=1)
70        x = dec_block(x_cat)
71      return self.out_layer(x)
```

The LiteVAE encoder can be implemented as shown below:

```python
1  class LiteVAEEncoder(nn.Module):
2    def __init__(
3      self,
4      in_channels: int,
5      out_channels: int,
6      wavelet_fn: HaarTransform,
7      feature_extractor_params: dict,
8      feature_aggregator_params: dict,
9    ):
10     super().__init__()
11     self.wavelet_fn = wavelet_fn
12     self.feature_extractor_L1 = LiteVAEUNetBlock(
13       in_channels, in_channels, **feature_extractor_params
```

26

```
14        )
15      self.feature_extractors_L2 = LiteVAEUNetBlock(
16          in_channels, in_channels, **feature_extractor_params
17      )
18      self.feature_extractor_L3 = LiteVAEUNetBlock(
19          in_channels, in_channels, **feature_extractor_params
20      )
21      out_channels = out_channels * 2 # for VAE mean and log_var
22      aggregated_channels = in_channels * 3
23      self.feature_aggregator = LiteVAEUNetBlock(
24          aggregated_channels, out_channels, **feature_aggregator_params
25      )
26      self.downsample_block_L1 = Downsample2D(in_channels, scale_factor=4)
27      self.downsample_block_L2 = Downsample2D(in_channels, scale_factor=2)
28
29   def forward(self, image):
30      dwt_L1 = self.wavelet_fn.dwt(image, level=1) / 2
31      dwt_L2 = self.wavelet_fn.dwt(image, level=2) / 4
32      dwt_L3 = self.wavelet_fn.dwt(image, level=3) / 8
33      features_L1 = self.downsample_block_L1(
34          self.feature_extractor_L1(dwt_L1)
35      )
36      features_L2 = self.downsample_block_L2(
37          self.feature_extractor_L1(dwt_L2)
38      )
39      features_L3 = self.feature_extractor_L3(dwt_L3)
40      dwt_features = [features_L1, features_L2, features_L3]
41      latent = self.feature_aggregator(torch.cat(features, dim=1))
42      return latent
```

Finally, the code for LiteVAE is also provided below.

```
1  class LiteVAE(nn.Module):
2    def __init__(
3      self,
4      encoder: LiteVAEEncoder,
5      decoder: SDVAEDecoder,
6      config: DictConfig,
7      output_type: str = "image",
8    ):
9      super().__init__()
10     assert output_type in ["image", "wavelet"]
11     self.encoder = encoder
12     self.decoder = decoder
13     self.wavelet_fn = encoder.wavelet_fn
14     self.output_type = output_type
15
16     pre_channels = config.latent_dim * 2 # for VAE mean and log_var
```


27


```python
    post_channels = config.latent_dim
    if config.get("use_1x1_conv", False):
      self.pre_conv = nn.Conv2d(pre_channels, pre_channels, 1)
      self.post_conv = nn.Conv2d(post_channels, post_channels, 1)
    else:
      self.pre_conv = nn.Identity()
      self.post_conv = nn.Identity()

  def encode(self, image):
    return self.pre_conv(self.encoder(image))

  def decode(self, latent):
  latent = self.post_conv(latent)
    if self.output_type == "image":
      image_recon = self.decoder(latent)
      wavelet_recon = self.wavelet_fn.dwt(image_recon, level=1) / 2
    elif self.output_type == "wavelet":
      wavelet_recon = self.decoder(latent)
      image_recon = self.wavelet_fn.idwt(wavelet_recon, level=1) * 2
    return image_recon, wavelet_recon

  def forward(self, image, sample=True):
    latent = self.encode(image)
    latent_dist = DiagonalGaussianDistribution(latent)
    latent = latent_dist.sample() if sample else latent_dist.mode()
    kl_reg = latent_dist.kl().mean()
    image_recon, wavelet_recon = self.decode(latent)
    return Dict(
      {
        "sample": image_recon,
        "wavelet": wavelet_recon,
        "latent": latent,
        "kl_reg": kl_reg,
        "latent_dist": latent_dist,
      }
    )
```

28

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims are supported via detailed empirical analysis.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are discussed in Section 7.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper does not include theoretical results.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The hyperparameters, algorithms, and implementation details are provided in the appendix for proper reproducibility of our work.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: While we do not provide open access to our codebase, the hyperparameters, algorithms, and implementation details are provided in the appendix to ensure reproducibility.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide the details of the experiments in the main text and the appendix.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Computing the error bar in the context of our work is too computationally demanding.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide hardware details whenever appropriate.

29

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The paper follows the NeurIPS Code of Ethics as stated on the website.

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The broader impact statements are discussed in Appendix A.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our work does not release such models or datasets.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We use well-established datasets and models with proper academic licenses.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

30