Multi-Architecture Multi-Expert Diffusion Models

Yunsung Lee[2*], JinYoung Kim[3*], Hyojun Go[3*], Myeongho Jeong[4], Shinhyeok Oh[1], Seungtaek Choi[4]

[1]Riiid AI Research
[2]Wrtn Technologies
[3]Twelvelabs
[4]Yanolja

sung@wrtn.io, seago0828@gmail.com, gohyojun15@gmail.com
myeongho.jeong@yanolja.com, shinhyeok.oh@riiid.co seungtaek.choi@yanolja.com

Abstract

In this paper, we address the performance degradation of efficient diffusion models by introducing Multi-architecturE Multi-Expert diffusion models (MEME). We identify the need for tailored operations at different time-steps in diffusion processes and leverage this insight to create compact yet high-performing models. MEME assigns distinct architectures to different time-step intervals, balancing convolution and self-attention operations based on observed frequency characteristics. We also introduce a soft interval assignment strategy for comprehensive training. Empirically, MEME operates 3.3 times faster than baselines while improving image generation quality (FID scores) by 0.62 (FFHQ) and 0.37 (CelebA). Though we validate the effectiveness of assigning more optimal architecture per time-step, where efficient models outperform the larger models, we argue that MEME opens a new design choice for diffusion models that can be easily applied in other scenarios, such as large multi-expert models.

Introduction

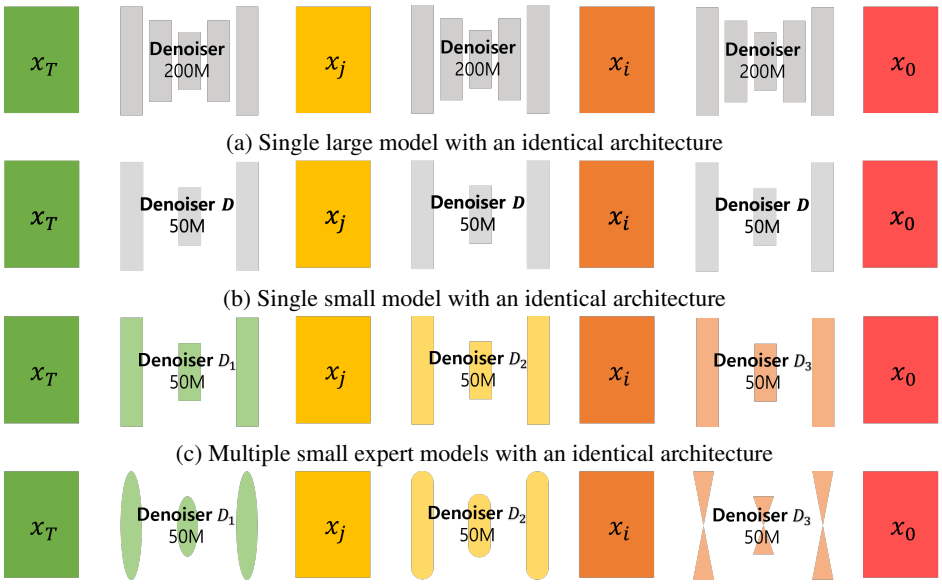
Diffusion models (Sohl-Dickstein et al. 2015; Song and Ermon 2019; Ho, Jain, and Abbeel 2020) are a promising approach for generative modeling, and they are likely to play an increasingly important role in diverse domains, including image (Dhariwal and Nichol 2021; Rombach et al. 2022; Balaji et al. 2022), audio (Kong et al. 2021; Kim, Kim, and Yoon 2022), video (Ho et al. 2022b,a; Zhou et al. 2022), and 3D generation (Poole et al. 2022; Seo et al. 2023). However, despite their impressive performance, diffusion models suffer from high computation costs, which stem from the following two orthogonal factors: (i) the lengthy iterative denoising process, and (ii) the cumbersome denoiser networks. Though there have been several efforts to overcome such limitations (Song, Meng, and Ermon 2021; Bao et al. 2022; Lu et al. 2022; Meng et al. 2022; Song et al. 2023), most of these efforts have focused only on resolving the first factor, such that the cumbersome denoisers still limit their applicability to real-world scenarios. A few efforts reduce the size of the denoisers based on post-training low-bit quantization (Shang et al. 2022) and distillation (Yang et al. 2022), as we illustrated in Fig. 1b, but they usually achieve such efficiency by compromising on accuracy.

In this paper, we thus aim to build a diffusion model that is compact yet comparable in performance to the large models. For this purpose, we first ask a research question "*why the traditional diffusion models require such massive parameters?*". (Choi et al. 2022; Go et al. 2023) suggests that the difficulty of learning diffusion models is that they have to learn all the different tasks at many different time-steps. One step further, from a frequency perspective, (Yang et al. 2022) theoretical explains that diffusion models should learn too many different features in varying time-steps, where diffusion models tend to initially form low-frequency components (e.g., overall image contour) and subsequently fill in high-frequency components (e.g., detailed textures). However, as they assume the denoiser network to be a linear filter, which is not practical, we aim to investigate empirical evidence to support this claim. Specifically, we analyze the per-layer Fourier spectrum for the input x_t at each diffusion time-step t, finding that there are significant and consistent variations in the relative log amplitudes of the Fourier-transformed feature maps as t progresses. This finding indicates that the costly training of large models indeed involves learning to adapt to the different frequency characteristics at each time-step t.

One way to leverage this finding is to assign distinct time-step intervals to multiple diffusion models (Go et al. 2022; Balaji et al. 2022), referred to as the *multi-expert* strategy, in order for models to be specialized in the assigned time-step intervals as shown in Fig. 1c. However, since (Go et al. 2022) utilized the multi-expert strategy for the conditioned generation with guidance and (Balaji et al. 2022) focused on high performance, the architecture efficiency is not considered. Most importantly, to the best of our knowledge, previous works have constructed diffusion models with a single architecture, missing the possibility of optimal architectures that better solve the task at each time-step of diffusion.

To this end, we propose to assign different models with **different architectures** for each different time-step interval, whose base operations vary according to their respective frequency ranges, which we dub **Multi-architecturE Multi-Expert diffusion models (MEME)** (Fig. 1d). Specifi-

*These authors contributed equally.



(a) Single large model with an identical architecture

(b) Single small model with an identical architecture

(c) Multiple small expert models with an identical architecture

(d) Multiple small expert models with multiple architectures (MEME)

Figure 1: Comparative illustration of single/multiple-expert models with single/multiple architectures. Figure 1a depict the standard diffusion models approach, which employs a single large denoiser. To reduce the cost due to the large-scale of the model, a diffusion model with a small denoiser is designed with post-training low-bit or distillation as illustrated in Fig. 1b. In Fig. 1c, to alleviate the performance drop, we consider a configuration with multiple small expert models having identical architectures. Finally, our proposed method, the Multi-architecturE Multi-Expert diffusion models (MEME), constructs small expert models with unique optimal architectures for their respective assigned time-step intervals, as visualized in Fig. 1d.

cally, we leverage the property that convolutions are advantageous for handling high-frequency components ($t \sim 0$), while multi-head self-attention (MHSA) excels in processing low-frequency components ($t \sim T$) (d'Ascoli et al. 2021; Dai et al. 2021; Park and Kim 2022; Si et al. 2022). However, a naive hard-shuffling of convolution and MHSA at different time intervals would be suboptimal because the features are inherently a combination of high- and low-frequency components (d'Ascoli et al. 2021; Si et al. 2022).

In order to better adapt to such frequency-specific components, we propose a more flexible denoiser architecture called **iU-Net**, which incorporates an iFormer (Si et al. 2022) block that allows for adjusting the channel-wise balance ratio between the convolution operations and MHSA operations. We take advantage of the characteristic of diffusion models we discovered that first recover low-frequency components during the denoising process and gradually add high-frequency features. Consequently, we configure each architecture to have a different proportion of MHSA, effectively tailoring each architecture to suit the distinct requirements at different time-step intervals of the diffusion process.

We further explore methods for effectively assigning focus on specific time-step intervals to our flexible iU-Net. Specifically, we identify a soft interval assignment strategy for the multi-expert models that prefers a soft division over a hard segmentation. This strategy allows the experts assigned to intervals closer to T to have more chance to be trained with the entire time-step, which prevents excessive exposure to meaningless noises at the time-step $t \sim T$.

Empirically, our MEME diffusion models effectively perform more specialized processing for each time-step interval, resulting in improved performance compared to the baselines. MEME, with LDM (Rombach et al. 2022) as the baseline, has managed to reduce the computation cost by 3.3 times while training on FFHQ (Karras, Laine, and Aila 2019) and CelebA-HQ (Karras et al. 2018) datasets from scratch and has simultaneously improved image generation performance by 0.62 and 0.37 in FID scores, respectively. By comparing the Fourier-transformed feature maps of MEME and multi-expert with identical architecture, we have confirmed that MEME's multi-architecture approach allows for distinct frequency characteristics suitable for each interval. Furthermore, MEME not only improves performance when combined with the LDM baseline but also demonstrates successful performance enhancements when integrated with the other diffusion model, DDPM (Ho, Jain, and Abbeel 2020).

Our main contributions are summarized as follows:

- As far as we know, we are the first to identify and address the limitation of diffusion models where vastly different functionalities at each time-step in diffusion processes yield sub-optimal performances.
- We propose MEME, a novel diffusion models framework composed of multi-architecture multi-expert denoisers that can balance operations for low- and high-frequency, performing distinct operations for each time-step interval.
- MEME surpasses its large counterparts in terms of generation quality while providing more efficient inference.

Trained from scratch on the FFHQ and CelebA datasets, MEME operates **3.3 times faster** than baselines while **improving FID** scores by **0.62** and **0.37**, respectively.

Related Work

Diffusion Models

Diffusion models (Sohl-Dickstein et al. 2015; Song and Ermon 2019), a subclass of generative models, generate data through an iterative denoising process. Trained by denoising score-matching objectives, these models demonstrate impressive performance and versatility in various domains, including image (Dhariwal and Nichol 2021; Rombach et al. 2022; Balaji et al. 2022), audio (Kong et al. 2021; Popov et al. 2021; Kim, Kim, and Yoon 2022), video (Ho et al. 2022b; Zhou et al. 2022), and 3D (Poole et al. 2022; Zeng et al. 2022; Seo et al. 2023) generation. However, diffusion models suffer from significant drawbacks, such as high memory and computation time costs (Kong and Ping 2021; Lu et al. 2022). These issues primarily stem from two factors: the lengthy iterative denoising process and the substantial number of parameters in the denoiser model. A majority of studies addressing the computation cost issues of diffusion models have focused on accelerating the iteration process. Among these, (Watson et al. 2022, 2021; Dockhorn, Vahdat, and Kreis 2022) have employed more efficient differential equation solvers. Other studies have sought to reduce the lengthy iterations by using truncated diffusion (Lyu et al. 2022; Zheng et al. 2022) or knowledge distillation (Luhman and Luhman 2021; Salimans and Ho 2022; Song et al. 2023).

In contrast, (Shang et al. 2022) and (Yang et al. 2022) focus on reducing the size of diffusion models. (Shang et al. 2022) proposes a post-training low-bit quantization specifically tailored for diffusion models. (Yang et al. 2022) analyzes diffusion models based on frequency, enabling small models to effectively handle high-frequency dynamics by applying wavelet gating and spectrum-aware distillation. However, these attempts at lightweight models usually fail to match the performance of large models and rely on resource-intensive training, which assumes the availability of a pretrained diffusion model. To overcome such dependency, multi-expert strategies have been explored to increase the model capacity (Balaji et al. 2022) while keeping the inference cost at each time step. In this work, our distinction is to enhance the multi-expert strategy, by focusing more on the fact that diffusion models have to learn very different functionality at each time-step, while the existing diffusion models leverages a single operation across the diffusion processes.

Combination of Convolutions and Self-attentions

Since the advent of the Vision Transformer (Dosovitskiy et al. 2021), there has been active research into why self-attention works effectively in the image domain and how it differs from convolution operations. (Park and Kim 2022; Si et al. 2022; Wang et al. 2022; Bai et al. 2022) explain empirically or theoretically that this is because self-attention operations better capture global features and act as low-pass filters. There have been efforts (Dai et al. 2021; d'Ascoli et al. 2021; Park and Kim 2022; Si et al. 2022; Bu, Huang, and Cui 2023) aiming to design optimal architectures that better combine the advantages of self-attention and convolution. (Park and Kim 2022; Dai et al. 2021) suggest structuring networks with convolution-focused front layers, which are advantageous for high-pass filtering, and self-attention-focused rear layers, which are advantageous for low-pass filtering. (d'Ascoli et al. 2021; Si et al. 2022) propose new blocks that perform operations intermediate between convolution and self-attention. Notably, (Si et al. 2022) proposes the iFormer block that allows for adjustable ratios between convolution and self-attention operations. While previous efforts have been focused on exploring better operations and architectures for image recognition, we are the first to explore the same questions in diffusion models, revealing the need for diffusion-specific strategies.

Background

Spectrum Evolution over Diffusion Process

Diffusion models (Sohl-Dickstein et al. 2015; Song and Ermon 2019) work by inverting a stepwise noise process using latent variables. Data points $\mathbf{x}_0$ from the true distribution are perturbed by Gaussian noise with zero mean and β_t variance across T steps, eventually reaching Gaussian white noise. As in (Ho, Jain, and Abbeel 2020), efficiently sampling from the noise-altered distribution $q(\mathbf{x}_t)$ is achieved through a closed-form expression to generate arbitrary time-step $\mathbf{x}_t$:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \tag{1}$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, $\alpha_t = 1 - \beta_t$, and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$.

The denoiser, a time-conditioned denoising neural network $\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}, t)$ with trainable parameters θ, is trained to reverse the diffusion process by minimizing re-weighted evidence lower bound (Song and Ermon 2019), as follows:

$$\mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\left[||\nabla \mathbf{x}_t \log p(\mathbf{x}_t|\mathbf{x}_0) - \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)||_2^2\right] \tag{2}$$

In essence, the denoiser learns to recover the gradient that optimizes the data log-likelihood. The previous step data x_{t-1} is generated by inverting the Markov chain:

$$\mathbf{x}_{t-1} \leftarrow \frac{1}{\sqrt{1-\beta_t}}(\mathbf{x}_t + \beta_t \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)) + \sqrt{\beta_t}\boldsymbol{\epsilon}_t \tag{3}$$

In this reverse process, the insight that diffusion models evolve from rough to detailed was gained through several empirical observations (Ho, Jain, and Abbeel 2020; Rombach et al. 2022). Beyond them, (Yang et al. 2022) provides a numerical explanation of this insight from a frequency perspective by considering the network as a linear filter. In this case, the optimal filter, known as the Wiener filter (Wiener et al. 1949), can be expressed in terms of its spectrum response at every time-step. Under the widely accepted assumption that the power spectra $\mathbb{E}[|X_0(f)|^2] = A_s(\theta)/f^{\alpha_S(\theta)}$ of natural images x_0 follows a power law (Burton and Moorhead 1987; Field 1987; Tolhurst, Tadmor, and Chao 1992) the frequency response of the signal reconstruction filter is determined by the amplitude scaling factor $A_s(\theta)$ and the frequency exponent $\alpha_S(\theta)$. As the reverse denoising process progresses from $t = T$ to $t = 0$, and $\bar{\alpha}$ increases from 0 to 1, diffusion models,

as analyzed by (Yang et al. 2022), exhibit spectrum-varying behavior over time. Initially, a narrow-banded filter restores only low-frequency components responsible for rough structures. As t decreases and $\bar{\alpha}$ increases, more high-frequency components, such as human hair, wrinkles, and pores, are gradually restored in the images.

Inception Transformer

The limitation of transformers in the field of vision is well-known as they tend to capture low-frequency features that convey global information but are less proficient at capturing high-frequency features that correspond to local information (Dosovitskiy et al. 2021; Si et al. 2022). To address this shortcoming, (Si et al. 2022) introduced the Inception Transformer, which combines a convolution layer with a transformer, utilizing the Inception module (Szegedy et al. 2015, 2016, 2017). To elaborate, the input feature $\mathbf{Z} \in \mathbb{R}^{N \times d}$ is first separated into $\mathbf{Z}_h \in \mathbb{R}^{n \times d_h}$ and $\mathbf{Z}_l \in \mathbb{R}^{n \times d_l}$ along the channel dimension, where $d = d_h + d_l$. The iFormer block then applies a high-frequency mixer to $\mathbf{Z}_h$ and a low-frequency mixer to $\mathbf{Z}_l$. Specifically, $\mathbf{Z}_h$ is further split into $\mathbf{Z}_{h_1}$ and $\mathbf{Z}_{h_2}$ along the channel dimension as follows:

$$\mathbf{Y}_{h_1} = \text{FC}(\text{MP}(\mathbf{Z}_{h_1})), \tag{4}$$

$$\mathbf{Y}_{h_2} = \text{D-Conv}(\text{FC}(\mathbf{Z}_{h_2})), \tag{5}$$

where $\mathbf{Y}$ denotes the outputs of high-frequency mixer, FC is fully-connected layer, MP represents max pooling layer, and D-Conv is depth-wise convolutional layer.

In the low-frequency mixer, MHSA is utilized to acquire a comprehensive and cohesive representation, as shown in Eq. 6. This global representation is then combined with the output from the high-frequency mixer as in Eq. 7. However, due to the potential oversmoothing effect of the upsample operation in Eq. 6, a fusion module is introduced to counteract this issue and produce the final output, outlined in Eq. 8:

$$\mathbf{Y}_l = \text{Up}(\text{MHSA}(\text{AP}(\mathbf{Z}_{h_2}))), \tag{6}$$

$$\mathbf{Y}_c = \text{Concat}(\mathbf{Y}_{h_1}, \mathbf{Y}_{h_2}, \mathbf{Y}_l), \tag{7}$$

$$\mathbf{Y} = \text{FC}(\mathbf{Y}_c + \text{D-Conv}(\mathbf{Y}_c)), \tag{8}$$

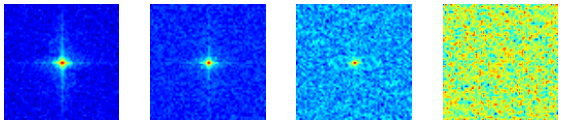
where Up denotes upsampling, AP is average pooling, and Concat represents concatenation.

Frequency Analysis for Diffusion Models

It is useful to design the architecture with distinct blocks capturing appropriate frequency according to the depth of the block (d'Ascoli et al. 2021; Touvron et al. 2021). In this section, we analyze the frequency-based properties of latents and extracted features according to time-step.

Frequency Component from Latents.

From the fact that a Gaussian filter prioritizes the filtering out of high-frequency (Gonzalez and Woods 2002), it is evident that the training data fed into diffusion models progressively lose their high-frequency spectrum as t increases. In Fig. 2, we visualize the Fourier spectrum of input latent, output of autoencoder, for training LDM (Rombach et al. 2022) with FFHQ (Karras, Laine, and Aila 2019) dataset. By illustrating the Fourier coefficients of the periodic function against the corresponding frequency, the training data gradually lose their high-frequency spectrum as t increases. It suggests designing a diffusion model that filters different frequency components according to the time-step for dealing with the corresponding features.



(a) $t = 125$ (b) $t = 375$ (c) $t = 625$ (d) $t = 875$

Figure 2: Visualization of the Fourier spectrum of the inputs used in training diffusion models. As t increases from 0 to T, the high-frequency feature spectrum, initially concentrated towards the center, gradually disappears.

Frequency Component Focused by Model.

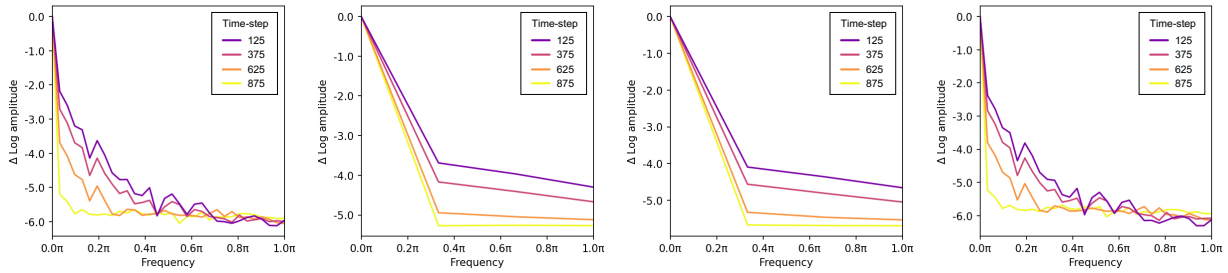
Here, we first introduce the analysis on the frequency for each layer with the distinct depth as (d'Ascoli et al. 2021) did. We examine the relative log amplitudes of Fourier-transformed feature maps obtained from the pre-trained latent diffusion model (LDM). (Park and Kim 2022) reveals that image recognition deep neural networks primarily perform high-pass filtering in earlier layers and low-pass filtering in later layers. Additionally, in this paper, we further analyze the frequency components focused by the diffusion model with respect to the diffusion time-step. The captured feature with frequency perspective is illustrated in each subfigure of Fig. 3, indicating that diffusion models tend to attenuate low-frequency signals more prominently as t increases. These findings align with the well-established characteristics of a Gaussian filter, known for its tendency to suppress high-frequency components primarily (Gonzalez and Woods 2002).

Multi-Architecture Multi-Expert

Based on our frequency analysis for diffusion models, we propose the following significant hypothesis: *By structuring the denoiser model with operations that vary according to each time-step interval, it could potentially enhance the efficiency of the diffusion model's learning process.* To validate this hypothesis, two key elements are needed: i) a denoiser architecture with the capacity to adjust the degree of its specialization towards either high or low frequencies, and ii) a strategy for varying the application of this tailored architecture throughout the diffusion process.

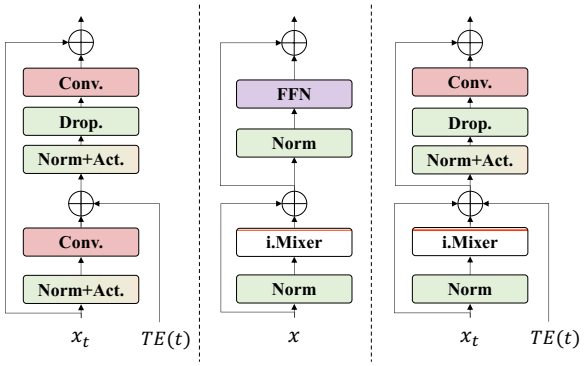
iU-Net Architecture

We propose the iU-Net architecture, a variant of U-Net (Ronneberger, Fischer, and Brox 2015) that allows for adjusting the ratio of operations favorable to high and low frequencies. We utilize a block referred to as the inception transformer (iFormer) (Si et al. 2022), which intertwines convolution operations, suitable for high-pass filtering, and Multi-Head Self-Attention (MHSA) operations, suitable for low-pass filtering, with an inception (Szegedy et al. 2015) mixer. Figure 4 illustrates the manner in which we have adapted the



(a) The former layer in the encoder (b) The later layer in the encoder (c) The former layer in the decoder (d) The later layer in the decoder

Figure 3: Visualization of relative log amplitudes of Fourier transformed feature map obtained from the pre-trained large LDM. The ΔLog amplitude of high-frequency signals is a difference with log amplitudes at the frequency of 0.0π and π. We compute it with 10K image samples at each time-step $t \in \{125, 375, 625, 875\}$. It shows the tendency of ΔLog amplitude to be interpolated as t is changed. In particular, as $t \to T$, the Fourier transformed features from the model are rapidly changed after 0.0π.



(a) U-Net (b) iFormer (c) Ours (iU-Net)

Figure 4: Comparative illustration of the block in the diffusion models. The $\oplus$ denotes element-wise addition and TE denotes the time-embedding lookup table.

iFormer block to fit the denoiser architecture of diffusion. This setup allows the iFormer block to regulate the ratio between the convolution-heavy high-frequency mixer and the MHSA-heavy low-frequency mixer in the architecture's composition. Following (Park and Kim 2022; Dai et al. 2021; Wu et al. 2021; d'Ascoli et al. 2021; Si et al. 2022) that tried to combine convolution and MHSA, we set the iU-Net encoder to perform more MHSA operations in the later layers. We discuss it more technically in a later section. Furthermore, as in (Cao et al. 2022), rather than completely replacing the block architecture from the U-Net block to the iFormer block, asymmetrically merging the two is effective in constructing an architecture for diffusion model that exploits iFormer.

Multi-Architecture Multi-Expert Strategy

Architecture Design for Experts To facilitate the construction of structures capable of accommodating diverse architectures, we employ a multi-expert strategy (Go et al. 2022; Balaji et al. 2022), but assign different architectures to each expert according to the frequency component. In each architecture, the ratio of dimension sizes for high and low channels is defined by two factors: layer depth and diffusion time-step. The former is well-known to enable the frequency dynamic feature extraction by focusing on lower frequency as a deeper layer (Park and Kim 2022; Dai et al. 2021). For more technical derivation, let d^k be the channel size in the k-th layer, d_h^k be the dimension size for the high mixer, and d_l^k for the low mixer, satisfying $d^k = d_h^k + d_l^k$. Based on the analysis in Fig. 3, the ratio in each iFormer block is defined for dealing with appropriate frequency components according to the depth; d_h^k/d_l^k decreases as a deeper block. On the other hand, the latter (diffusion time-step) can be associated with the frequency components based on the observation we found; as time-step t increases, the lower frequency components are focused. Therefore, we configure the iU-Net architecture such that the ratio of d_h^k/d_l^k decreases faster for the denoiser taking charge of the expert on the larger t.

Soft Expert Strategy. As suggested in (Go et al. 2022), one of N experts Θ_n is trained on the uniform and disjoint interval $\mathbb{I}_n = \left\{ t \middle| t \in \left(\frac{(n-1)}{N}T, \frac{n}{N}T \right] \right\}$ for $n = 1, ..., N$. However, for the large n, expert Θ_n takes as noised input images by near Gaussian noise $\epsilon_n \sim \mathcal{N}(\sqrt{\bar{\alpha}_n}x_0, (1-\bar{\alpha}_n)\mathbf{I})$, which makes it challenging for meaningful learning to take place with Θ_n. To address this, we propose a *soft expert strategy*, where each Θ_n learns on the interval $\mathbb{I}_n$ with a probability of p_n denoted as the expertization probability[1]. Otherwise, it learns on the entire interval $\bigcup_{n=1}^{N} \mathbb{I}_n$ with the remaining probability of $(1 - p_n)$.

Since it is evident that Θ_n for large n takes more noised images, larger p_n as $n \to N$ is a more flexible strategy for training multi-expert, yielding $p_1 \geq \cdots \geq p_N$.

Experiments

In this section, we demonstrate the capability of MEME to enhance the efficiency of diffusion models. Firstly, we showcase how our model can achieve superior performance over the baseline models, despite being executed with less computation. Secondly, we verify if our MEME model, as hy-

[1]Note that When $p_n = 1$ regardless of n, it can be denoted as *hard expert strategy* (Go et al. 2022).

FFHQ 256×256 (DDIM-200)					
Model	#Param↓	MACs↓	FID↓	Prec.↑	Recall ↑
LDM-L** (635K iter) (Rombach et al. 2022)	274.1M	288.2G	<u>9.03</u>	**0.72**	<u>0.49</u>
Lite-LDM† (Yang et al. 2022)	22.4M	23.6G	17.3	-	-
SD (with Distill.) (Yang et al. 2022)	21.1M	-	10.5	-	-
LDM-L* (540K iter)	274.1M	288.2G	9.14	**0.72**	0.48
LDM-S*	89.5M(3.1×)	94.2G(3.1×)	11.41(−2.27)	0.66(−0.06)	0.44(−0.04)
iU-LDM-S*	**82.6M**(3.3×)	<u>90.5G</u>(3.2×)	11.64(−2.50)	0.65(−0.07)	0.45(−0.03)
Multi-Expert* (w/o Soft)	89.5M×4(3.1×)	94.2G(3.1×)	10.42(−1.28)	0.69(−0.03)	0.46(−0.02)
Multi-Expert*	89.5M×4(3.1×)	94.2G(3.1×)	9.58(−0.44)	<u>0.70</u>(−0.02)	0.46(−0.02)
MEME*(w/o Soft)	<u>82.9M</u>‡×4(3.3×)	**90.4G**‡(3.3×)	9.20(−0.06)	<u>0.70</u>(−0.02)	0.48(+0.00)
MEME*	<u>82.9M</u>‡×4(3.3×)	**90.4G**‡(3.3×)	**8.52**(+0.62)	**0.72**(+0.00)	**0.50**(+0.02)
CelebA-HQ 256×256 (DDIM-50)					
Model	#Param↓	MACs↓	FID↓	Prec.↑	Recall ↑
LDM-L** (410K iter) (Rombach et al. 2022)	274.1M	288.2G	<u>5.92</u>	<u>0.71</u>	**0.49**
Lite-LDM† (Yang et al. 2022)	22.4M	23.6G	14.3	-	-
SD† (with Distill.) (Yang et al. 2022)	21.1M	-	9.3	-	-
LDM-S*	89.5M(3.1×)	94.2G(3.1×)	9.11(−3.19)	0.61(−0.10)	0.45(−0.04)
iU-LDM-S*	**82.6M**(3.3×)	<u>90.5G</u>(3.2×)	9.06(−3.14)	0.60(−0.11)	0.47(−0.02)
Multi-Expert*	89.5M×4(3.1×)	94.2G(3.1×)	7.00(−1.08)	0.67(−0.04)	<u>0.48</u>(−0.01)
MEME*	<u>82.9M</u>‡×4(3.3×)	**90.4G**‡(3.2×)	**5.55**(+0.37)	**0.73**(+0.02)	**0.49**(+0.00)

Table 1: Overall Results of Unconditional Generation on FFHQ and CelebA-HQ We use the Clean-FID implementation to ensure reproducibility. We sample 200 steps using DDIM on the FFHQ, and 50 steps on the CelebA-HQ. Even with N models trained using Multi-Expert and MEME, the total training cost was equivalent to that of a large model. SD is trained through knowledge distillation, which is dependent on having a large pretrained model already, but we can build an efficient model from scratch. The symbols denote †: values reported in the original source; ‡: average value across four architectures; $*$: calculated using checkpoints from our training; $**$: recalculated using pretrained checkpoints from the official repository.

pothesized, indeed incorporates appropriate Fourier features for each time-step interval input.

We evaluated the unconditional generation of models on two datasets, FFHQ (Karras, Laine, and Aila 2019) and CelebA-HQ (Karras et al. 2018). We construct models based on the LDM framework. All pre-trained auto-encoders for LDM were obtained from the official repository[2].

MEME employs a multi-expert structure composed of multiple small models, each of which has its channel dimension reduced from 224 to 128. The use of these smaller models is denoted by appending an 'S' to the model name, such as in LDM-S and iU-LDM-S. We set the number of experts N to 4 for all multi-expert settings, including MEME.

All experiments were conducted on a single NVIDIA A100 GPU. We primarily utilize the AdamW optimizer (Loshchilov and Hutter 2017). The base learning rate is set according to the oigianl LDM (Rombach et al. 2022). Notably, our smaller models employ a setting that doubles the batch size, which is not feasible with the original LDM on a single A100. Correspondingly, the base learning rate for our smaller models is also doubled compared to the standard settings.

We assess the quality of our generated models using the FID score (Heusel et al. 2017). As the FID score can be challenging to replicate due to the settings of the reference set, we calculate it using the publicly available Clean-FID (Parmar, Zhang, and Zhu 2022) implementation[3]. Particularly for the FFHQ dataset, the availability of a fixed reference set allows for a fair comparison of generation quality across all evaluated generative models on Clean-FID. To verify the efficiency of our trained model, we compare its model size and computational cost using the number of parameters and Multiply-Add cumulation (MACs)[4] as metrics. We provide detailed configurations and hyperparameters regarding the models in the Appendix.

Image Generation Results

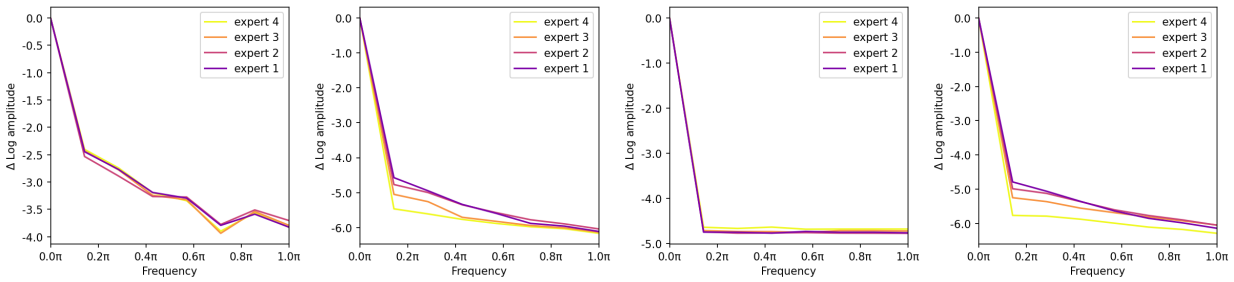
The results of our model trained on FFHQ (Karras, Laine, and Aila 2019) and CelebA-HQ (Karras et al. 2018) datasets are shown in Table 1. Despite requiring only 3.3 times less computation cost (MACs), our model demonstrates an improvement in performance (FID). Specifically, we observe a gain of 0.62 in FID for FFHQ and 0.4 in FID for CelebA. In the case of MEME and Multi-Expert, they require N models to be loaded into system memory for inference. However, in large-scale sample inference scenarios, it is possible to load only one expert into system memory while storing intermediate outputs on the disk, yielding less cost to the inferring process. Our approach allows for an improvement of 3.3 times in memory cost, which is equivalent to that of a single denoiser. In our experimental setting with $N = 4$, even if all experts are loaded into system memory for inference, it only requires an additional 20.9% of memory compared to the single large model. Further details regarding these two inference

[2]https://github.com/CompVis/latent-diffusion

[3]https://github.com/GaParmar/clean-fid

[4]https://github.com/sovrasov/flops-counter.pytorch

Figure 5: Samples from baseline LDM-L and MEME trained on FFHQ. The baseline often generates unnatural aspects in images, whereas our approach MEME shows fewer such cases.



(a) Multi-Expert, $t = 125$ (b) MEME, $t = 125$ (c) Multi-Expert, $t = 875$ (d) MEME, $t = 875$

Figure 6: Fourier Analysis Comparison between Multi-Expert and MEME Even with the same input t, we can confirm that MEME exhibits different characteristics for each expert. MEME demonstrates a similar trend as the pre-trained large model shown in Fig. 3, where experts responsible for intervals close to $t = T$ rapidly reduce high frequencies. In contrast, Multi-Expert composed of the same architecture shows that the frequency characteristics of features for the same time-step input are not significantly distinguished from each other.

scenarios can be found in the Appendix. It is also worth noting that in our experiments, the four experts of Multi-Expert and MEME incurred less than 30% of the computation time compared to LDM-L based on A100 GPU. Therefore, the overall training cost requires less than an additional 20% of resources. The qualitative results illustrated in Fig. 5 show that the generated images by our methods are superior to those by baseline. Further experiments on other domains (i.e., ImageNet) are in Appendix.

Module Ablation

Table 1 provides ablation study for various methods on FFHQ dataset. Firstly, when training with the baseline LDM-S, which involves standard diffusion training, performance drop (FID -2.27, -2.50 for LDM-S, iU-LDM-S, respectively) occurs. Although the incorporation of Multi-Expert mitigates the performance drop to some extent, there is still a degradation (FID -1.28) compared to the baseline LDM-L. In contrast, MEME not only reduces computational cost through the use of smaller-sized denoisers but also achieves performance improvement (FID +0.62).

Additionally, we found that the soft-expert strategy is more efficient than the hard-expert strategy, where each expert focuses on its designated region. We empirically discovered that a strategy for training the multi-expert with the expertization probability denoted as p_n is beneficial. We configured the probabilities: $p_1 = 0.8$, $p_2 = 0.4$, $p_3 = 0.2$, and $p_4 = 0.1$. Different configurations for p_n are provided in the Appendix.

Fourier Analysis of MEME

In this section, unlike the analysis shown in Fig. 3, we investigate whether the experts in MEME possess the ability to capture the corresponding frequency characteristics that are advantageous for their respective intervals as illustrated in Fig. 6. MEME, composed of various architectures, exhibits different characteristics for each expert; experts responsible for intervals closer to $t = T$ quickly reduce high frequencies. In contrast, Multi-Expert, composed of the same architecture, fails to significantly differentiate the frequency characteristics of features when the same time-step input is provided. Particularly for $t = 875$, which requires the ability to capture low-frequency components, it is difficult to distinguish the features of all experts.

MEME on Top of the Other Diffusion Baseline

In order to explore the generalizability of MEME, we adopted the experimental setup used for architecture validation in (Choi et al. 2022). We trained a lightweight version of ADM (Dhariwal and Nichol 2021) (referred to as ADM-S) on the CelebA-64 dataset with batch size 8 and 200K

CelebA 64×64		
Model	#Param↓	FID↓
ADM-S	90M	49.56
iU-ADM-S	**82M**($1.1\times$)	50.08(-0.52)
Multi-Expert	90M $\times 4$	47.29($+2.27$)
MEME	**82M** $\times 4$($1.1\times$)	**43.09**($+6.47$)

Table 2: Results when applied to ADM baseline MEME outperforms ADM-S baseline (Choi et al. 2022).

iterations. The FID measurement was conducted from 10K samples from DDIM (Song, Meng, and Ermon 2021) with 50 steps. The results demonstrate that our MEME exhibits effective performance (FID +6.47) even in the context of ADM. Furthermore, the consistent trend is in line with the results observed in the LDM experiments.

Conclusion

In this paper, we studied the problem of improving efficient diffusion models, with the distinction of adopting multiple architectures to suit the specific frequency requirements at different time-step intervals. By incorporating the iU-Net architecture, we provide a more flexible and efficient solution for handling the complex distribution of frequency-specific components across time-steps. Our experiments validated that the proposed method, named **MEME**, outperforms existing baselines in terms of both generation performance and computational efficiency, making it a more practical solution for real-world applications. While we confirm that assigning optimal architectures per time-step results in efficient models outperforming larger ones, we believe MEME offers a new design choice for diffusion models.

Acknowledgements

This research was done while all authors were working at Riiid. We would like to thank our colleagues at Riiid AI Research for their enthusiastic discussions and valuable advice.

References

Bai, J.; Yuan, L.; Xia, S.-T.; Yan, S.; Li, Z.; and Liu, W. 2022. Improving vision transformers by revisiting high-frequency components. In *European Conference on Computer Vision*, 1–18. Springer.

Balaji, Y.; Nah, S.; Huang, X.; Vahdat, A.; Song, J.; Kreis, K.; Aittala, M.; Aila, T.; Laine, S.; Catanzaro, B.; et al. 2022. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*.

Bao, F.; Li, C.; Zhu, J.; and Zhang, B. 2022. Analytic-DPM: an Analytic Estimate of the Optimal Reverse Variance in Diffusion Probabilistic Models. In *International Conference on Learning Representations*.

Bu, Q.; Huang, D.; and Cui, H. 2023. Towards Building More Robust Models with Frequency Bias. *arXiv preprint arXiv:2307.09763*.

Burton, G. J.; and Moorhead, I. R. 1987. Color and spatial structure in natural scenes. *Applied optics*, 26(1): 157–170.

Cao, H.; Wang, J.; Ren, T.; Qi, X.; Chen, Y.; Yao, Y.; and Zhang, L. 2022. Exploring Vision Transformers as Diffusion Learners. *arXiv preprint arXiv:2212.13771*.

Choi, J.; Lee, J.; Shin, C.; Kim, S.; Kim, H.; and Yoon, S. 2022. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11472–11481.

Dai, Z.; Liu, H.; Le, Q. V.; and Tan, M. 2021. Coatnet: Marrying convolution and attention for all data sizes. *Advances in Neural Information Processing Systems*, 34: 3965–3977.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34: 8780–8794.

Dockhorn, T.; Vahdat, A.; and Kreis, K. 2022. GENIE: Higher-order denoising diffusion solvers. *arXiv preprint arXiv:2210.05475*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

d'Ascoli, S.; Touvron, H.; Leavitt, M. L.; Morcos, A. S.; Biroli, G.; and Sagun, L. 2021. Convit: Improving vision transformers with soft convolutional inductive biases. In *International Conference on Machine Learning*, 2286–2296. PMLR.

Field, D. J. 1987. Relations between the statistics of natural images and the response properties of cortical cells. *Josa a*, 4(12): 2379–2394.

Go, H.; Kim, J.; Lee, Y.; Lee, S.; Oh, S.; Moon, H.; and Choi, S. 2023. Addressing Negative Transfer in Diffusion Models. *arXiv preprint arXiv:2306.00354*.

Go, H.; Lee, Y.; Kim, J.-Y.; Lee, S.; Jeong, M.; Lee, H. S.; and Choi, S. 2022. Towards Practical Plug-and-Play Diffusion Models. *arXiv preprint arXiv:2212.05973*.

Gonzalez, R. C.; and Woods, R. E. 2002. Digital image processing. upper saddle River. *J.: Prentice Hall*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Ho, J.; Chan, W.; Saharia, C.; Whang, J.; Gao, R.; Gritsenko, A.; Kingma, D. P.; Poole, B.; Norouzi, M.; Fleet, D. J.; et al. 2022a. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

Ho, J.; Salimans, T.; Gritsenko, A. A.; Chan, W.; Norouzi, M.; and Fleet, D. J. 2022b. Video Diffusion Models. In Oh, A. H.; Agarwal, A.; Belgrave, D.; and Cho, K., eds., *Advances in Neural Information Processing Systems*.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2018. Progressive Growing of GANs for Improved Quality, Stability, and

Variation. In *International Conference on Learning Representations*.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4401–4410.

Kim, H.; Kim, S.; and Yoon, S. 2022. Guided-tts: A diffusion model for text-to-speech via classifier guidance. In *International Conference on Machine Learning*, 11119–11133. PMLR.

Kong, Z.; and Ping, W. 2021. On fast sampling of diffusion probabilistic models. *arXiv preprint arXiv:2106.00132*.

Kong, Z.; Ping, W.; Huang, J.; Zhao, K.; and Catanzaro, B. 2021. DiffWave: A Versatile Diffusion Model for Audio Synthesis. In *International Conference on Learning Representations*.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lu, C.; Zhou, Y.; Bao, F.; Chen, J.; Li, C.; and Zhu, J. 2022. DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps. In Oh, A. H.; Agarwal, A.; Belgrave, D.; and Cho, K., eds., *Advances in Neural Information Processing Systems*.

Luhman, E.; and Luhman, T. 2021. Knowledge distillation in iterative generative models for improved sampling speed. *arXiv preprint arXiv:2101.02388*.

Lyu, Z.; Xu, X.; Yang, C.; Lin, D.; and Dai, B. 2022. Accelerating diffusion models via early stop of the diffusion process. *arXiv preprint arXiv:2205.12524*.

Meng, C.; Gao, R.; Kingma, D. P.; Ermon, S.; Ho, J.; and Salimans, T. 2022. On distillation of guided diffusion models. *arXiv preprint arXiv:2210.03142*.

Park, N.; and Kim, S. 2022. How Do Vision Transformers Work? In *International Conference on Learning Representations*.

Parmar, G.; Zhang, R.; and Zhu, J.-Y. 2022. On aliased resizing and surprising subtleties in gan evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11410–11420.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*.

Popov, V.; Vovk, I.; Gogoryan, V.; Sadekova, T.; and Kudinov, M. 2021. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*, 8599–8608. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 234–241. Springer.

Salimans, T.; and Ho, J. 2022. Progressive Distillation for Fast Sampling of Diffusion Models. In *International Conference on Learning Representations*.

Seo, J.; Jang, W.; Kwak, M.-S.; Ko, J.; Kim, H.; Kim, J.; Kim, J.-H.; Lee, J.; and Kim, S. 2023. Let 2D Diffusion Model Know 3D-Consistency for Robust Text-to-3D Generation. *arXiv preprint arXiv:2303.07937*.

Shang, Y.; Yuan, Z.; Xie, B.; Wu, B.; and Yan, Y. 2022. Post-training Quantization on Diffusion Models. *arXiv preprint arXiv:2211.15736*.

Si, C.; Yu, W.; Zhou, P.; Zhou, Y.; Wang, X.; and YAN, S. 2022. Inception Transformer. In Oh, A. H.; Agarwal, A.; Belgrave, D.; and Cho, K., eds., *Advances in Neural Information Processing Systems*.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, 2256–2265. PMLR.

Song, J.; Meng, C.; and Ermon, S. 2021. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.

Song, Y.; Dhariwal, P.; Chen, M.; and Sutskever, I. 2023. Consistency models. *arXiv preprint arXiv:2303.01469*.

Song, Y.; and Ermon, S. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32.

Szegedy, C.; Ioffe, S.; Vanhoucke, V.; and Alemi, A. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–9.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Tolhurst, D. J.; Tadmor, Y.; and Chao, T. 1992. Amplitude spectra of natural images. *Ophthalmic and Physiological Optics*, 12(2): 229–232.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, 10347–10357. PMLR.

Wang, P.; Zheng, W.; Chen, T.; and Wang, Z. 2022. Anti-Oversmoothing in Deep Vision Transformers via the Fourier Domain Analysis: From Theory to Practice. In *International Conference on Learning Representations*.

Watson, D.; Chan, W.; Ho, J.; and Norouzi, M. 2022. Learning fast samplers for diffusion models by differentiating through sample quality. In *International Conference on Learning Representations*.

Watson, D.; Ho, J.; Norouzi, M.; and Chan, W. 2021. Learning to efficiently sample from diffusion probabilistic models. *arXiv preprint arXiv:2106.03802*.

Wiener, N.; Wiener, N.; Mathematician, C.; Wiener, N.; Wiener, N.; and Mathématicien, C. 1949. *Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications*, volume 113. MIT press Cambridge, MA.

Wu, H.; Xiao, B.; Codella, N.; Liu, M.; Dai, X.; Yuan, L.; and Zhang, L. 2021. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22–31.

Yang, X.; Zhou, D.; Feng, J.; and Wang, X. 2022. Diffusion Probabilistic Model Made Slim. *arXiv preprint arXiv:2211.17106*.

Zeng, X.; Vahdat, A.; Williams, F.; Gojcic, Z.; Litany, O.; Fidler, S.; and Kreis, K. 2022. LION: Latent Point Diffusion Models for 3D Shape Generation. In Oh, A. H.; Agarwal, A.; Belgrave, D.; and Cho, K., eds., *Advances in Neural Information Processing Systems*.

Zheng, H.; He, P.; Chen, W.; and Zhou, M. 2022. Truncated diffusion probabilistic models. *stat*, 1050: 7.

Zhou, D.; Wang, W.; Yan, H.; Lv, W.; Zhu, Y.; and Feng, J. 2022. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*.