DCT-MoE: Frequency-Domain Medical Image Generation with Mixture-of-Experts

Yifan Sun YXS443@STUDENT.BHAM.AC.UK
Qingjie Meng M.QINGJIE@BHAM.AC.UK
Wayne-Anthony Ezechukwu WXE381@STUDENT.BHAM.AC.UK
Tao Chen T.CHEN@BHAM.AC.UK
Huiping Chen H.CHEN.13@BHAM.AC.UK
University of Birmingham, United Kingdom

Editors: Under Review for MIDL 2026

Abstract

Medical image synthesis is a critical solution to data scarcity, yet standard Latent Diffusion Models (LDMs) are often limited by their reliance on Variational Autoencoders (VAEs) pre-trained on RGB images. Such reliance introduces domain shift and channel mismatch between the training domain and grayscale medical scans, which degrades fine anatomical detail and amplifies reconstruction artefacts. To address these limitations, we introduce DCT-MoE, a diffusion model that adopts a deterministic block-wise Discrete Cosine Transform (DCT) representation instead of a learnable VAE latent space. In detail, the proposed method maps grayscale images to a compact block-wise DCT representation that acts as a fixed, low-dimensional space. On top of this representation, a Mixture-of-Experts (MoE) backbone is integrated into the Diffusion model, providing scalable expressivity without a proportional increase in computational cost. Extensive experiments on cardiac MRI and echocardiography generation demonstrate that DCT-MoE achieves high image quality and inference efficiency compared to the state-of-the-art spatial-domain LDMs and frequency-domain generation methods.

Keywords: Diffusion model, medical image generation, frequency domain, MoE.

1. Introduction

Artificial Intelligence (AI) has rapidly transformed societal sectors, including the medical field, where it is driving innovation. Notably, computer-vision models show substantial potential for disease classification and lesion segmentation. Real world applications have saved many patients' lives (Cao et al., 2023; Kondepudi et al., 2025). High-quality data remains the cornerstone of AI training, yet medical data presents significant acquisition challenges due to high costs, privacy concerns and legal constraints (e.g., the General Data Protection Regulation (GDPR) in the EU).

Consequently, addressing data scarcity has become a critical priority in medical AI research. In the domain of medical imaging, researchers have increasingly investigated the potential of generative models (Kazerouni et al., 2023; Yi et al., 2019). These architectures, most notably diffusion models, are designed to reconstruct coherent data samples from random noise (Ho et al., 2020). This capability offers a promising workflow: a limited dataset of high-quality, real-world images can be used to train a diffusion model, which can

subsequently generate an extensive volume of high-fidelity synthetic samples to augment the original data (Kazerouni et al., 2022; Khader et al., 2023).

Despite the remarkable success of diffusion models in natural image generation (Rombach et al., 2022; Dhariwal and Nichol, 2021), directly adapting these pipelines to medical imaging presents significant challenges. A primary obstacle lies in the Variational-Autoencoder (VAE) component of Latent Diffusion Models (LDMs). VAEs are essential for encoding high-dimensional pixel data into a compressed latent space and subsequently decoding the latent representation back into pixel space (Kingma and Welling, 2013). This compression significantly enhances computational efficiency, making LDMs the predominant architecture in contemporary diffusion modeling (Rombach et al., 2022). Current research usually relies on pretrained VAEs derived from large-scale foundation models, such as Stable Diffusion and SDXL, to construct medical image LDM pipelines (Pinaya et al., 2022; Khader et al., 2023; Reynaud et al., 2024). However, we contend that using VAE-based LDM in medical image diffusion is suboptimal. Here are the main reasons:

First, a *channel mismatch* exists between natural and medical images. Unlike natural scenes, medical images are predominantly grayscale, rendering standard RGB-based VAEs suboptimal. Replicating the single channel to fit 3-channel input introduces memory redundancy and creates the risk of chromatic artifacts during the decoding process. Second, the potential *domain shift* between natural and medical distributions imposes challenges for standard VAEs, which are typically pre-trained on datasets like ImageNet (Raghu et al., 2019). Non-negligible distribution shifts also exist within the medical domain itself, such as the disparity between brain and cardiac MRI. Third, addressing these limitations by training or fine-tuning a medical-specific VAE from scratch incurs an inevitable *computational overhead*.

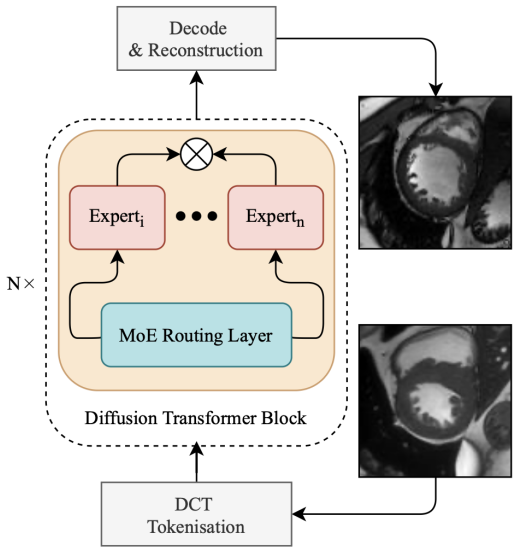

Figure 1: Overview of DCT-MoE.

To address the limitations of current medical image diffusion, we propose **DCT-MoE** (as illustrated in Figure 1), a novel framework for grayscale medical image generation that integrates the Discrete Cosine Transform (DCT) and a Mixture-of-Experts (MoE) architecture. Diverging from standard LDMs that rely on learned latent spaces (VAEs), our approach shifts the generative process to the frequency domain. Specifically, we utilize DCT to transform input grayscale images into spectral coefficients, where we achieve efficient compression and tokenization directly within the DCT space. A Diffusion Transformer is employed to model the distribution of these coefficients, incorporating an MoE design to enhance scalability and performance without sacrificing training and sampling efficiency. Finally, high-fidelity images are reconstructed from the sampled coefficients via the Inverse DCT (IDCT).

We summarise our main contributions as follows:

- We employ a frequency-domain diffusion framework for medical image generation that operates directly on block-wise DCT coefficients, replacing the common VAE latent space with a simple DCT module. This moves grayscale medical images into the frequency domain without adding extra channels and reduces the domain shift and reconstruction artifacts associated with VAEs trained on natural images.

- We propose an MoE-based diffusion transformer pipeline that operates directly on DCT coefficients. MoE is capable of scaling our diffusion model without compromising training and sampling efficiency.

- We evaluate our method on two medical image modalities: MRI and Ultrasound. Experiments demonstrate that the proposed method achieves high image fidelity and computational efficiency, outperforming the state-of-the-art methods.

2. Related Work

2.1. Medical image generation

Initially, Generative Adversarial Networks (GANs) were the standard for tasks such as cross-modality translation (e.g., MRI-to-CT) and super-resolution (Yi et al., 2019; Armanious et al., 2020). However, their applicability is frequently questioned by training instability and mode collapse. While VAEs provide a more stable probabilistic alternative, their pixel-wise optimization objectives often result in over-smoothed reconstructions, ruining fine-grained diagnostic details (Pinaya et al., 2022). Recently, (latent) diffusion models, especially Denoising Diffusion Probabilistic Models (DDPMs) have established a new state-of-the-art in medical synthesis, delivering superior textural fidelity and mode coverage (Kazerouni et al., 2023; Reynaud et al., 2024; Takagi and Nishimoto, 2023). Nevertheless, while LDMs lower the computational costs of high-resolution medical data, they typically rely on perceptual compression optimized for natural RGB scenes. This approach remains suboptimal for grayscale medical data generation.

2.2. Frequency domain approaches

Recent deep learning approaches in computer vision have increasingly integrated frequency-aware mechanisms to enhance visual quality across image generation (Yang et al., 2023; Phung et al., 2023), compression (Li et al., 2023), and enhancement (He et al., 2024). Notable examples include Wavelet Diffusion (Phung et al., 2023), which utilizes multi-scale sub-bands for efficient sampling, and Spectral Diffusion (Yang et al., 2023), which leverages wavelet gating to preserve high-frequency details.

Meanwhile, some researchers have shifted towards modeling directly within frequency domains rather than pixel space, such as DCT, Fourier, or Wavelet. For instance, Transformers have been applied to sparse DCT sequences for autoregressive generation (Nash et al., 2021) and to Fourier coefficients for tomographic reconstruction (Buchholz and Jug, 2022). Wavelet-based tokenization has similarly demonstrated improvements in throughput and scalability for both generative and discriminative tasks (Mattar et al., 2024; Zhu and Soricut, 2024). Most recently, pure frequency-space diffusion models, such as FDDM (Ziashahabi et al., 2024) and DCTdiff (Ning et al., 2025), have emerged. By performing

diffusion directly on DCT coefficients, these methods achieve superior training efficiency and generative fidelity compared to their pixel-space/latent-space counterparts.

2.3. MoE in vision and medical imaging

Originally proposed in natural language processing, MoE architectures have recently demonstrated remarkable scalability in general computer vision as well. Fundamental approaches, such as V-MoE (Riquelme et al., 2021) established that replacing dense feed-forward networks with sparse MoE layers allows for massive parameter expansion with minor computational overhead. By conditionally activating only a subset of experts per token, these models achieve superior performance on large-scale benchmarks like ImageNet. Following this trajectory, the medical image community has increasingly adopted MoE strategies to address the high across-class variance and data heterogeneity inherent in medical image datasets. Recent applications utilize MoE for multi-modal fusion and dynamic region-of-interest processing in segmentation tasks (Płotka et al., 2024), effectively decoupling model capacity from inference cost to handle complex anatomical variations.

3. Background

Diffusion models (Ho et al., 2020; Song et al., 2020) are deep generative frameworks that gradually corrupt data with noise (the forward process) and then learn to invert this process to synthesise new samples (the backward process). Let the distribution of the original data be denoted as $p_{data}(\mathbf{x})$ with a standard deviation σ_{data}, and $\mathbf{x}_0$ be a random sample drawn from $p_{data}(\mathbf{x})$. In the discrete-time diffusion model, the forward process is achieved via adding i.i.d. Gaussian noise over $T \in \mathbb{N}$ timesteps via a Markov chain $q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1})$, so that $\mathbf{x}_T$ is nearly isotropic Gaussian noise. The backward process then learns the reverse transitions $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$, often achieved by approximating the Gaussian noise added during the forward process through a neural network. Notably, discrete-time diffusion requires timestep to be discrete and can not be overridden. In continuous-time diffusion model, the timestep $t \in [0, 1]$. The forward process is also considered as injecting i.i.d. Gaussian noise into the clean data samples. The Gaussian noise is controlled and measured by its standard deviation $\sigma(t)$ over time t. The noisy sample distribution is therefore noted as $p(\mathbf{x}; \sigma(t))$. As the noise level $\sigma_{\max} \gg \sigma_{data}$, the original information of $p_{data}(\mathbf{x})$ is considered to be destroyed by noise injected over steps, leaving a final state $p(\mathbf{x}; \sigma_{\max}) \approx \mathcal{N}(0, \sigma_{\max}^2 \mathbf{I}^2)$. During the forward process, the score function $\nabla_{\mathbf{x}} \log(p(\mathbf{x}; \sigma(t))$ is recorded as the trajectory of distribution shift from $p_{data}(\mathbf{x})$ to $\mathcal{N}(0, \sigma_{\max}^2 \mathbf{I}^2)$. The learning objective of the score model $s_\theta(\mathbf{x}_t, t)$ is to approximate the score function $\nabla_{\mathbf{x}} \log(p(\mathbf{x}; \sigma(t))$ with model parameters θ via denoising score-matching. To sample clean images (backward process), the score function learnt during forward process is used to reverse the distribution $p(\mathbf{x}; \sigma)$ from Gaussian noise $(0, \sigma_{\max}^2 \mathbf{I}^2)$ to the original data $p_{data}(\mathbf{x})$. The reverse process is described by an ordinary differential equation (ODE) or a stochastic differential equation (SDE) (Song et al., 2021):

$$\begin{aligned} \text{ODE} : \mathrm{d}\mathbf{x} &= -\dot{\sigma}(t)\sigma(t)\nabla_{\mathbf{x}} \log(p(\mathbf{x}; \sigma(t))\mathrm{d}t \\ \text{SDE} : \mathrm{d}\mathbf{x} &= -2\dot{\sigma}(t)\sigma(t)\nabla_{\mathbf{x}} \log(p(\mathbf{x}; \sigma(t))\mathrm{d}t \\ &\quad + \sqrt{2\dot{\sigma}(t)\sigma(t)}\mathrm{d}W_t \end{aligned} \tag{1}$$

where $\sigma(t)$ is the noise level given time t, and W_t is the standard Wiener process. Again, by replacing the score function $\nabla_{\mathbf{x}} \log(p(\mathbf{x}; \sigma(t))$ with $s_\theta(\mathbf{x}_t, t)$, any numerical solver can be applied to obtain a generated data sample $\hat{\mathbf{x}}_0$ (Lu et al., 2022).

4. Methodology

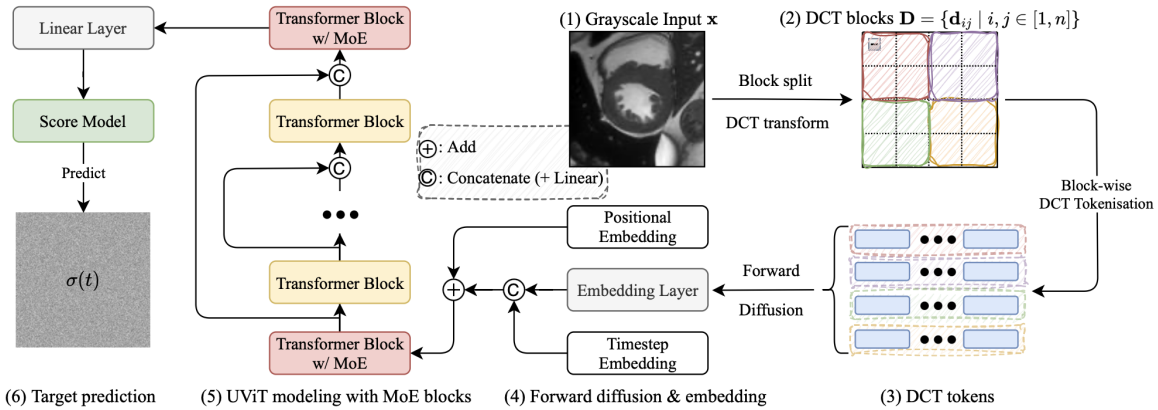

Figure 2: The training process of proposed **DCT-MoE**. We use a U-ViT transformer backbone with long skip connection and continuous-time diffusion as the diffusion model. We set noise as the prediction target. We employ the MoE transformer blocks at the first and last transformer blocks.

In this section, we will deliver the methodology of our proposed method DCT-MoE. Figure 2 presents the training pipeline of DCT-MoE. We will initially start with DCT tokenization of grayscal images; then followed by the design of MoE blocks in our diffusion transformers. Note that a DCT-MoE achieves scaling by employing a lightweight MoE transformer architecture, i.e., we only set the first and last transformer block as MoE blocks and leaving other transformer blocks unchanged. Ablation study on Section 5 proves that we achieve a balance between efficiency and generation quality.

4.1. DCT tokenization

In this section, we detail the tokenization process within our framework based on DCT[1]. DCT has a well-known foundational role in signal processing, notably in the JPEG image compression standard (Wallace, 1991). Recently, it has also emerged as a potential pre-processing tool in generative modeling, offering a deterministic alternative to the VAEs used in standard LDMs (Ning et al., 2025).

However, prior research has predominantly focused on natural RGB images, which necessitate conversion to the YCbCr color space and subsequent chroma sub-sampling. Conversely, the application of DCT is uniquely favorable and reasonable in our medical images scenario. As most medical images are inherently grayscale (equivalent to the luminance Y channel), they eliminate the need for color space transformation or channel down-sampling,

1. Unless otherwise stated, references to DCT imply the standard Type-II DCT.

thereby streamlining the tokenization pipeline and eliminating potential information loss due to chroma down-sampling.

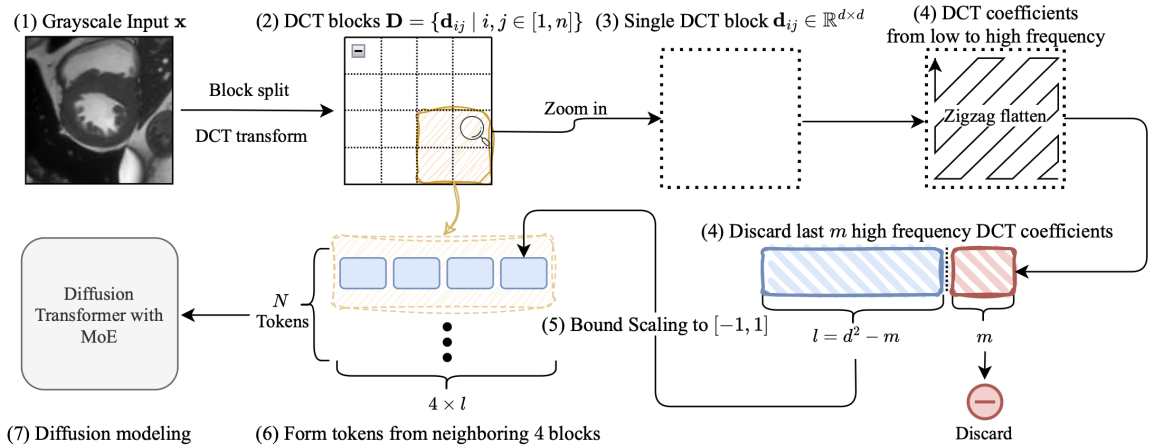

Figure 3: DCT tokenisation process. A grayscale image input is transformed to a sequence of tokens consists of DCT coefficients.

We adopt our DCT tokenisation method mainly from DCTdiff (Ning et al., 2025), but we modify and adjust it to suit grayscale images as input only. Figure 3 presents the pipeline of DCT tokenisation in our proposed method. We first split the input grayscale image into small, non-overlapping blocks before applying the block-wise DCT transform. Given a grayscale input image $\mathbf{x} \in \mathbb{R}^{H \times W}$, we first partition X into a grid of $b \times b$ blocks $A_i(x, y)$, with i being the block index and $0 \leq x,\ y \leq b$ the intra-block coordinates. Each block is transformed via a 2D DCT (Ahmed et al., 1974). The DCT converts the spatial block into a block of frequency-based coefficients $D_i(u, v)$ of the same size. For all $u, v \in [0, b]$, the transform is defined as:

$$D_i(u,v) = \alpha(u)\,\alpha(v) \sum_{x=0}^{b-1} \sum_{y=0}^{b-1} A_i(x,y)\, \cos\left[\frac{(2x+1)u\pi}{2b}\right] \cos\left[\frac{(2y+1)v\pi}{2b}\right], \tag{2}$$

$$\text{where} \quad \alpha(u) = \begin{cases} \sqrt{\frac{1}{B}}, & u = 0 \\ \sqrt{\frac{2}{B}}, & u \neq 0 \end{cases}$$

Here, $D_i(0, 0)$ represents the DC (zero-frequency) component, which corresponds to the average intensity of the pixels, while larger indices (u, v) denote increasingly higher spatial frequencies. This structure reflects a fundamental property of the DCT: *energy compaction*. Typically, the signal energy is concentrated in the lower frequencies, which are prioritized via a zigzag order starting from the DC component up to the left. Psychovisual research has long established that the Human Visual System (HVS) is highly sensitive to low-frequency information, such as smooth luminance variations, but exhibits reduced sensitivity to high-frequency details like fine-grained texture or noise (Campbell and Robson, 1968). Consequently, effective signal compression can be achieved by truncating high-frequency coefficients without compromising perceptual quality. This principle not only

underpins the JPEG standard (Wallace, 1991), but has also proven feasible in recent DCT-based diffusion modeling (Ning et al., 2025).

As illustrated in Figure 3, let m denote the number of high-frequency DCT coefficients to be discarded. Following a standard zigzag order, we truncate the final m coefficients and retain the remaining components as a flattened sequence. To construct a single feature token, we concatenate the flattened sequences from four spatially adjacent blocks. This process is repeated across non-overlapping regions of the input image $\mathbf{x}$, yielding a total of N tokens. Formally, for each token we apply a global bound normalization defined as $x_{\text{norm}} = \frac{x}{C_{\max}}$, where $C_{\max}$ represents the maximum absolute value of DC component computed across the entire training dataset.

Thanks to its nature in orthogonality and strong energy compaction, the DCT effectively concentrates signal energy into low-frequency components. This property enables efficient data compression via high-frequency truncation, yielding a compact representation without the computational cost of training distribution-specific VAEs. Furthermore, given the theoretical alignment between pixel-space diffusion and DCT autoregression (Ning et al., 2025), the DCT serves as a viable, deterministic alternative to learned latent spaces for diffusion modeling.

4.2. Diffusion Transformer with MoE

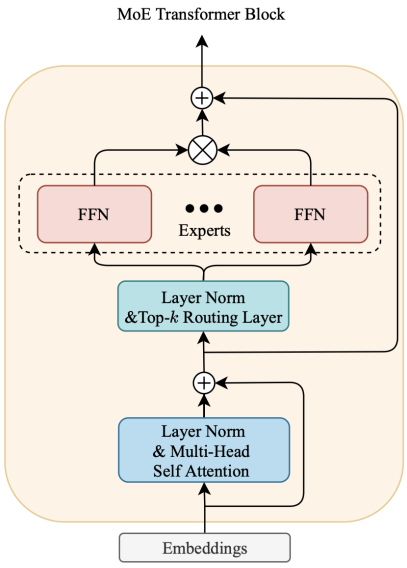

Figure 4: MoE Transformer block.

The **Mixture-of-Experts (MoE)** architecture is designed to scale model capacity with minor computational cost. An MoE layer comprises two primary components: a **gating network (router)** and a set of **expert networks**. The experts are independent neural networks that contribute the majority of the model's parameters. The gating network is responsible for calculating importance weights for each input token and conditionally routing it to a small, activated subset of experts. The final layer output is then computed as the weighted sum of the selected experts' outputs, thereby ensuring sparse activation (Shazeer et al., 2017).

MoE has demonstrated significant efficiency in generative modeling, particularly within Transformer-based diffusion frameworks (Sun et al., 2025). As illustrated in Figure 4, in our proposed method, we integrate an MoE mechanism by replacing the standard Feed-Forward Network (FFN) in the Transformer block with a set of parallel expert networks. And we employ a top-k gating strategy to route tokens. Let the input sequence after the second layer-norm be denoted as $\mathbf{x} \in \mathbb{R}^{d \times s}$, where s represents the sequence length and d the hidden dimension. A learnable projection maps $\mathbf{x}$ to the routing logits $\mathbf{E} \in \mathbb{R}^{e \times s}$, where e is the total number of experts. The routing probabilities $\mathbf{P}$ are subsequently obtained via a Softmax operation: $\mathbf{P} = \text{Softmax}(\mathbf{E})$. We then determine the indices $\mathbf{I}$ of

the experts by applying a top-k operation along the expert dimension. Consequently, each token is routed to its selected experts, and the final layer output $\mathbf{x}_{out}$ is computed as the probability-weighted sum of the experts' outputs:

$$\mathbf{x}_{out} = \sum_{i \in \mathbf{I}} P_i(\mathbf{x}) \cdot E_i(\mathbf{x}) \tag{3}$$

where E_i represents the i-th expert FFN, and P_i the corresponding routing probability.

Meanwhile, to keep the distribution of tokens balance, it is crucial to set up an auxiliary loss that encourages an average proportion of experts selected by each token. The design of auxiliary loss will be elaborated in Appendix A.

5. Experiments

5.1. Dataset

To evaluate our proposed method properly, we employ two diverse medical datasets: the **Automated Cardiac Diagnosis Challenge (ACDC)** (Bernard et al., 2018), comprising cardiac MRI scans; and **EchoNet-Dynamic (EchoNet)** (Ouyang et al., 2020), a large-scale cardiac ultrasound dataset. Both of the datasets are consists of grayscale data. The objective of our experiments is unconditional image generation. We extract 2D slices from the original volumes and apply center cropping to standardize the spatial dimensions. The details of the preprocessed datasets are listed below in Table 1.

Table 1: Datasets overview

Dataset(Preprocessed)	Height×Width	Image Number	Field
ACDC	96×96	25,022	MRI Scan
EchoNet	112×112	358,259	Ultrasound

5.2. Experiment Setup

To evaluate **DCT-MoE**, we benchmark it against three targeted baselines. First, to isolate the efficacy of the MoE, we compare it against **DCT-G**, a dense variant of our model without MoE design. Second, to compare our frequency-domain generation with latent space, we employ **LDM-C**, which retains the continuous-time U-ViT backbone but substitutes our DCT blocks with a standard pre-trained VAE for latent diffusion. Finally, we include **LDM-D**, a classic U-Net-based LDM with discrete DDPM sampling, as the prevailing standard in medical image synthesis. In summary, we employ 4 different experiment configurations[2]. Table 4 details our configurations.

5.3. Experiment Results

To quantitatively assess generative quality, we employ the Fréchet Inception Distance (FID) (Heusel et al., 2017) and Learned Perceptual Image Patch Similarity (LPIPS) (Zhang

2. Our code will be available after the review.

et al., 2018) metrics, both derived from a pre-trained InceptionV3 backbone.[3] For the primary comparative analysis, we fix the number of experts at $e = 4$ and the number of truncated high-frequency coefficients at $m = 0$ for both **DCT-MoE** and DCT-G. The specific impact of varying these hyperparameters is examined in the later ablation studies.

Table 2: FID and LPIPS evaluations. The best relative result is highlighted in **bold**, and the second best is underlined.

	ACDC		EchoNet	
	FID↓	LPIPS↓	FID↓	LPIPS↓
DCT-G	9.57	0.5100	7.17	0.4583
LDM-D	19.58	**0.4965**	7.33	**0.4039**
LDM-C	**6.64**	0.5253	6.99	0.4481
DCT-MoE	7.05	0.5093	**6.24**	0.4472

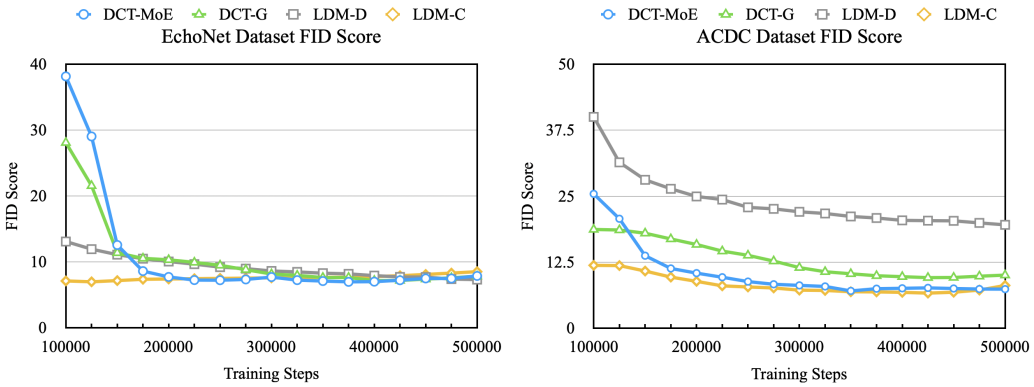

Figure 5: FID score over training steps.

Table 3: Model activated parameters and average training time per step comparison.

Dataset	Model	# Parameters	Avg. Time (S) / Step
ACDC	DCT-MoE	163M	2.1252
	DCT-G	135M	2.0475
EchoNet	DCT-MoE	163M	2.8842
	DCT-G	135M	2.6102

Table 2 and Figure 5 present the quantitative evaluation results based on FID and LPIPS metrics. Given the marginal variance in LPIPS scores, we only present visualization of the FID trajectories. The results demonstrate that DCT-MoE outperforms the dense baseline DCT-G, effectively validating the scaling capability of our design. Furthermore, DCT-MoE

3. For FID we use 50,000 generated images and for LPIPS we use 2,000 generated images.

exhibits superior generative fidelity compared to LDM-D while maintaining competitive performance level with LDM-C. These findings suggest that DCT-MoE serves as a robust and effective alternative to VAE-based latent representations for medical image synthesis.

Table 3 demonstrates the scalability of the proposed **DCT-MoE**. By integrating MoE blocks exclusively into the first and last transformer blocks, we increased the model capacity from 135M to 163M parameters. Crucially, the expansion improved generative quality while incurring negligible computational overhead in terms of time per training step.

5.4. Ablation studies

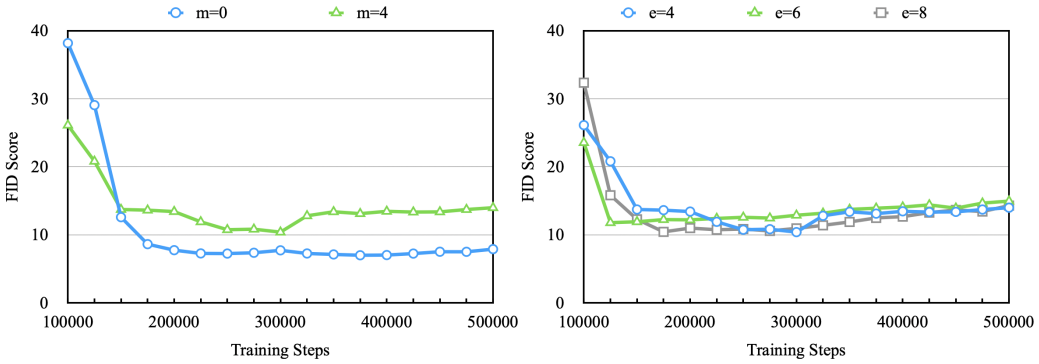

Figure 6: FID score of ACDC dataset. Left: Impact of truncation parameter m. Right: Impact of expert count e.

Impact of truncated high-frequency coefficients m. We first analyze the impact of high-frequency truncation (m) on the ACDC dataset. As shown in the left side of Figure 6, the configuration with no truncation ($m = 0$) consistently outperforms the truncated setting ($m = 4$) from the early training stages. This result underscores the fundamental trade-off between compression efficiency and information preservation.

Impact of number of experts e. We subsequently analyze the influence of the expert count, e. Due to efficiency and lightweight concerns, we limit our evaluation to a compact range of configurations, specifically $e \in \{4, 6, 8\}$, while holding all other hyperparameters constant ($m = 4$). As illustrated in the right panel of Figure 6, increasing the number of experts accelerates the convergence, effectively demonstrating the efficiency of MoE scaling.

6. Conclusion

This study addresses the dual challenges of data scarcity and inconvenience of traditional LDM in grayscale medical image synthesis. By introducing **DCT-MoE**, we have demonstrated that integrating deterministic frequency-domain tokenization with sparse MoE scaling offers a robust, lightweight and domain-friendly alternative to standard LDMs without compromising generation quality. The future work of this study mainly focuses on exploring more possibility of frequency domain diffusion in medical image generation, such as conditional generation.

References

N. Ahmed, T. Natarajan, and K.R. Rao. Discrete cosine transform. *IEEE Transactions on Computers*, C-23(1):90–93, 1974. doi: 10.1109/T-C.1974.223784.

Karim Armanious, Chenming Jiang, Marc Fischer, Thomas Küstner, Tobias Hepp, Konstantin Nikolaou, Sergios Gatidis, and Bin Yang. Medgan: Medical image translation using gans. *Computerized medical imaging and graphics*, 79:101684, 2020.

Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11): 2514–2525, 2018.

Black Forest Labs. FLUX.1 dev. https://github.com/black-forest-labs/flux, 2024.

Tim-Oliver Buchholz and Florian Jug. Fourier image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1846–1854, 2022.

Fergus W Campbell and John G Robson. Application of fourier analysis to the visibility of gratings. *The Journal of physiology*, 197(3):551, 1968.

Kai Cao, Yingda Xia, Jiawen Yao, Xu Han, Lukas Lambert, Tingting Zhang, Wei Tang, Gang Jin, Hui Jiang, Xu Fang, et al. Large-scale pancreatic cancer detection via non-contrast ct and deep learning. *Nature medicine*, 29(12):3033–3043, 2023.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Xuanhua He, Keyu Yan, Rui Li, Chengjun Xie, Jie Zhang, and Man Zhou. Frequency-adaptive pan-sharpening with mixture of experts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2121–2129, 2024.

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. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. Diffusion models for medical image analysis: A comprehensive survey. *arXiv preprint arXiv:2211.07804*, 2022.

Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. Diffusion models in medical imaging: A comprehensive survey. *Medical image analysis*, 88:102846, 2023.

Firas Khader, Gustav Müller-Franzes, Soroosh Tayebi Arasteh, Tianyu Han, Christoph Haarburger, Maximilian Schulze-Hagen, Philipp Schad, Sandy Engelhardt, Bettina Baeßler, Sebastian Foersch, et al. Denoising diffusion probabilistic models for 3d medical image generation. *Scientific Reports*, 13(1):7303, 2023.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Akhil Kondepudi, Melike Pekmezci, Xinhai Hou, Katie Scotford, Cheng Jiang, Akshay Rao, Edward S Harake, Asadur Chowdury, Wajd Al-Holou, Lin Wang, et al. Foundation models for fast, label-free detection of glioma infiltration. *Nature*, 637(8045):439–445, 2025.

Han Li, Shaohui Li, Wenrui Dai, Chenglin Li, Junni Zou, and Hongkai Xiong. Frequency-aware transformer for learned image compression. *arXiv preprint arXiv:2310.16387*, 2023.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in neural information processing systems*, 35:5775–5787, 2022.

Wael Mattar, Idan Levy, Nir Sharon, and Shai Dekel. Wavelets are all you need for autoregressive image generation. *arXiv preprint arXiv:2406.19997*, 2024.

Charlie Nash, Jacob Menick, Sander Dieleman, and Peter W Battaglia. Generating images with sparse representations. *arXiv preprint arXiv:2103.03841*, 2021.

Mang Ning, Mingxiao Li, Jianlin Su, Haozhe Jia, Lanmiao Liu, Martin Beneš, Wenshuo Chen, Albert Ali Salah, and Itir Onal Ertugrul. DCTdiff: Intriguing Properties of Image Generative Modeling in the DCT Space, May 2025. URL http://arxiv.org/abs/2412.15032. arXiv:2412.15032 [cs].

David Ouyang, Bryan He, Amirata Ghorbani, Neal Yuan, Joseph Ebinger, Curtis P Langlotz, Paul A Heidenreich, Robert A Harrington, David H Liang, Euan A Ashley, et al. Video-based ai for beat-to-beat assessment of cardiac function. *Nature*, 580(7802):252–256, 2020.

Hao Phung, Quan Dao, and Anh Tran. Wavelet diffusion models are fast and scalable image generators. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10199–10208, 2023.

Walter HL Pinaya, Petru-Daniel Tudosiu, Jessica Dafflon, Pedro F Da Costa, Virginia Fernandez, Parashkev Nachev, Sebastien Ourselin, and M Jorge Cardoso. Brain imaging generation with latent diffusion models. In *MICCAI workshop on deep generative models*, pages 117–126. Springer, 2022.

Szymon Płotka, Maciej Chrabaszcz, and Przemyslaw Biecek. Swin smt: Global sequential modeling for enhancing 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 689–698. Springer, 2024.

Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. *Advances in neural information processing systems*, 32, 2019.

Hadrien Reynaud, Qingjie Meng, Mischa Dombrowski, Arijit Ghosh, Thomas Day, Alberto Gomez, Paul Leeson, and Bernhard Kainz. Echonet-synthetic: Privacy-preserving video generation for safe medical data sharing. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 285–295. Springer, 2024.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34:8583–8595, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *Advances in neural information processing systems*, 34: 1415–1428, 2021.

Haotian Sun, Tao Lei, Bowen Zhang, Yanghao Li, Haoshuo Huang, Ruoming Pang, Bo Dai, and Nan Du. EC-DIT: Scaling diffusion transformers with adaptive expert-choice routing, 2025. URL http://arxiv.org/abs/2410.02098.

Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14453–14463, 2023.

Gregory K Wallace. The jpeg still picture compression standard. *Communications of the ACM*, 34(4):30–44, 1991.

Xingyi Yang, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Diffusion probabilistic model made slim. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 22552–22562, 2023.

Xin Yi, Ekta Walia, and Paul Babyn. Generative adversarial network in medical imaging: A review. *Medical image analysis*, 58:101552, 2019.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

Zhenhai Zhu and Radu Soricut. Wavelet-based image tokenizer for vision transformers. *arXiv preprint arXiv:2405.18616*, 2024.

Amir Ziashahabi, Baturalp Buyukates, Artan Sheshmani, Yi-Zhuang You, and Salman Avestimehr. Frequency domain diffusion model with scale-dependent noise schedule. In *2024 IEEE International Symposium on Information Theory (ISIT)*, pages 19–24. IEEE, 2024.

Appendix A. Design of Auxiliary Load Balancing Loss

To prevent *expert collapse*, a phenomenon where the gating network converges to utilizing only a small subset of experts while leaving others idle, we incorporate an auxiliary load balancing loss during training. Following the design in (Shazeer et al., 2017), this loss encourages both an equal distribution of router confidence (soft probabilities) and actual token assignments (hard selection) across all experts. Let N denote the number of experts and $\mathbf{x} \in \mathbb{R}^{B\times L\times D}$ be the input batch of tokens with total count $T = B \times L$. For each expert i, we compute two statistics: the *importance* P_i, defined as the average routing probability assigned to expert i across the batch, and the *load* f_i, representing the fraction of tokens actually dispatched to expert i. The auxiliary loss $\mathcal{L}_{aux}$ is formulated as the scaled dot product of these distributions:

$$\mathcal{L}_{aux} = N \sum_{i=1}^{N} f_i \cdot P_i \tag{4}$$

Minimizing this objective promotes a uniform distribution for both f and P, thereby ensuring equitable expert utilization. In our implementation, we employ a Straight-Through Estimator (STE) to approximate the gradients for the discrete routing decisions involved in calculating f_i.

Algorithm 1: Top-K Routing with Auxiliary Load Balancing

Require: Input tokens $\mathbf{X} \in \mathbb{R}^{T\times D}$, Router weights $\mathbf{W}_g$, Noise scale ϵ
Ensure: Routed outputs, Auxiliary loss $\mathcal{L}_{aux}$

Forward Pass:
$\mathbf{H} \leftarrow \mathbf{X}\mathbf{W}_g + \mathcal{N}(0, \epsilon)$ ▷ Compute noisy logits
$\mathbf{P} \leftarrow \text{Softmax}(\mathbf{H})$ ▷ Routing probabilities
$\text{val}, \text{idx} \leftarrow \text{TopK}(\mathbf{P}, k)$ ▷ Select top-k experts
$\mathbf{M}_{hard} \leftarrow \text{OneHot}(\text{idx})$ ▷ Create binary selection mask
$\mathbf{W}_{ste} \leftarrow \mathbf{M}_{hard} - \text{detach}(\mathbf{P}) + \mathbf{P}$ ▷ Apply Straight-Through Estimator
Load Balancing:
$P_i \leftarrow \frac{1}{T}\sum_{t=1}^{T} \mathbf{P}_{t,i}$ ▷ Average probability for expert i
$f_i \leftarrow \frac{1}{T}\sum_{t=1}^{T} \mathbf{W}_{ste,t,i}$ ▷ Actual selection fraction for expert i
$\mathcal{L}_{aux} \leftarrow N \sum_{i=1}^{N} f_i \cdot P_i$ ▷ Compute auxiliary loss
return $\text{Dispatch}(\mathbf{X}, \mathbf{W}_{ste})$, $\mathcal{L}_{aux}$

Appendix B. Details of Experiment Eetup

Here, we detail the configuration of our experiments in Section 5.2. Table 4 shows the specific setting of each configuration. Common settings across all models include **500k** training steps, Batch Size **512**, and dual NVIDIA **A100** GPUs. Score Model sampling uses DPM-Solver (SDE) (Lu et al., 2022). VAE is from FLUX.1-dev (Black Forest Labs, 2024).

Table 4: Summary of experimental configurations.

Model	Backbone	Tokenization / VAE	Sampling / Training Scheme	Specifics
DCT-MoE	U-ViT (16 Blocks)	DCT (Block Size: 4)	100 NFE (SDE)	**First & Last blocks: MoE**
DCT-G	U-ViT (16 Blocks)	DCT (Block Size: 4)	100 NFE (SDE)	All blocks: Dense
LDM-C	U-ViT (16 Blocks)	FLUX.1-dev VAE	100 NFE (SDE)	All blocks: Dense
LDM-D	U-Net (4 Scales)	FLUX.1-dev VAE	128 Steps (Sample) / 256 (Train)	Discrete Timesteps

Appendix C. Visualization of experiments results

To facilitate qualitative evaluation, we present randomly selected samples ($N = 16$) from the ground truth dataset, the dense baseline DCT-G, and the proposed **DCT-MoE**. Visual comparisons with LDM-C and LDM-D are omitted due to space constraints. The original and generated samples for the ACDC and EchoNet datasets are illustrated in Figure 7 and Figure 8, respectively.

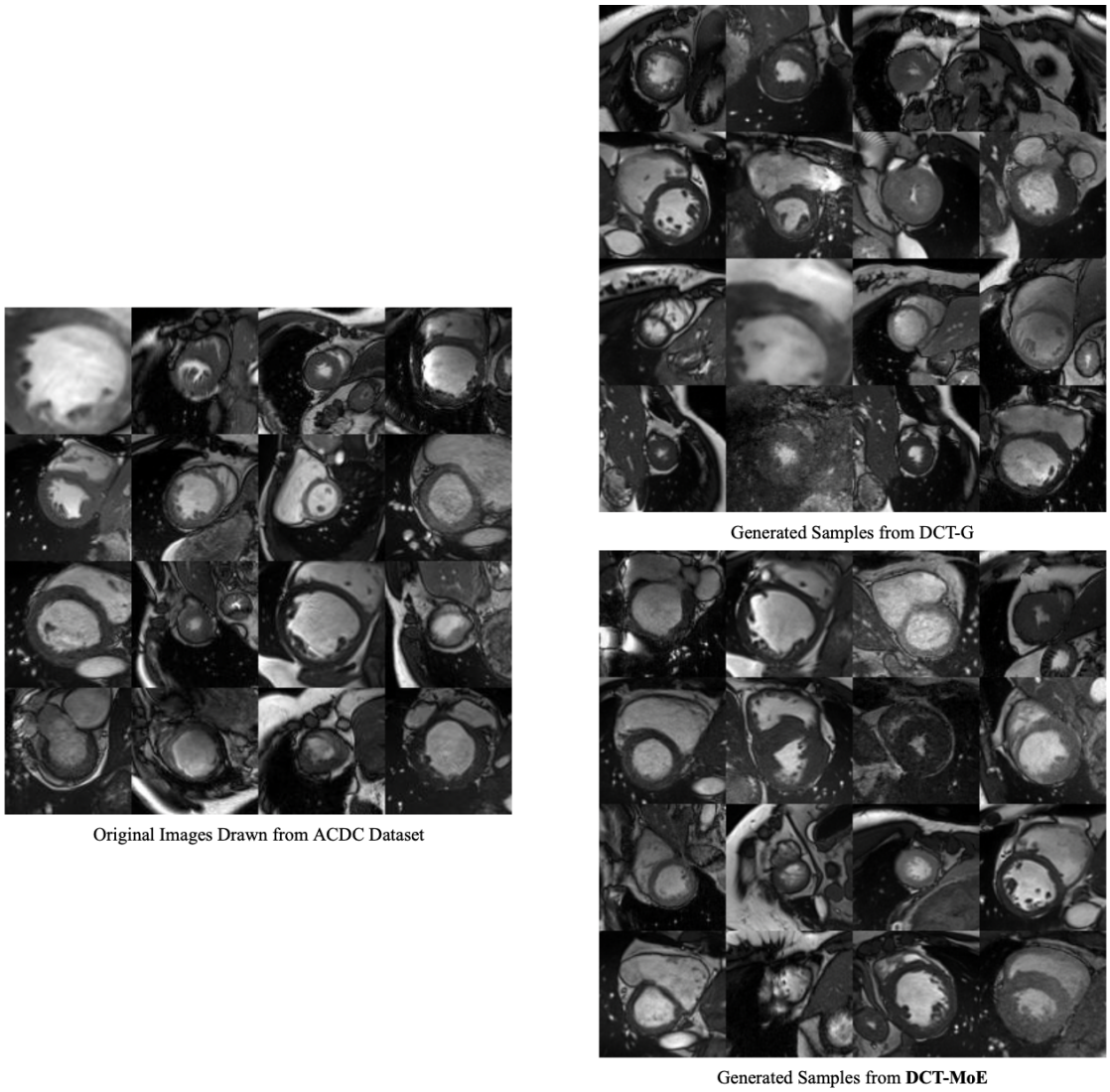

Figure 7: Visualization of DCT-G and **DCT-MoE** on ACDC dataset.

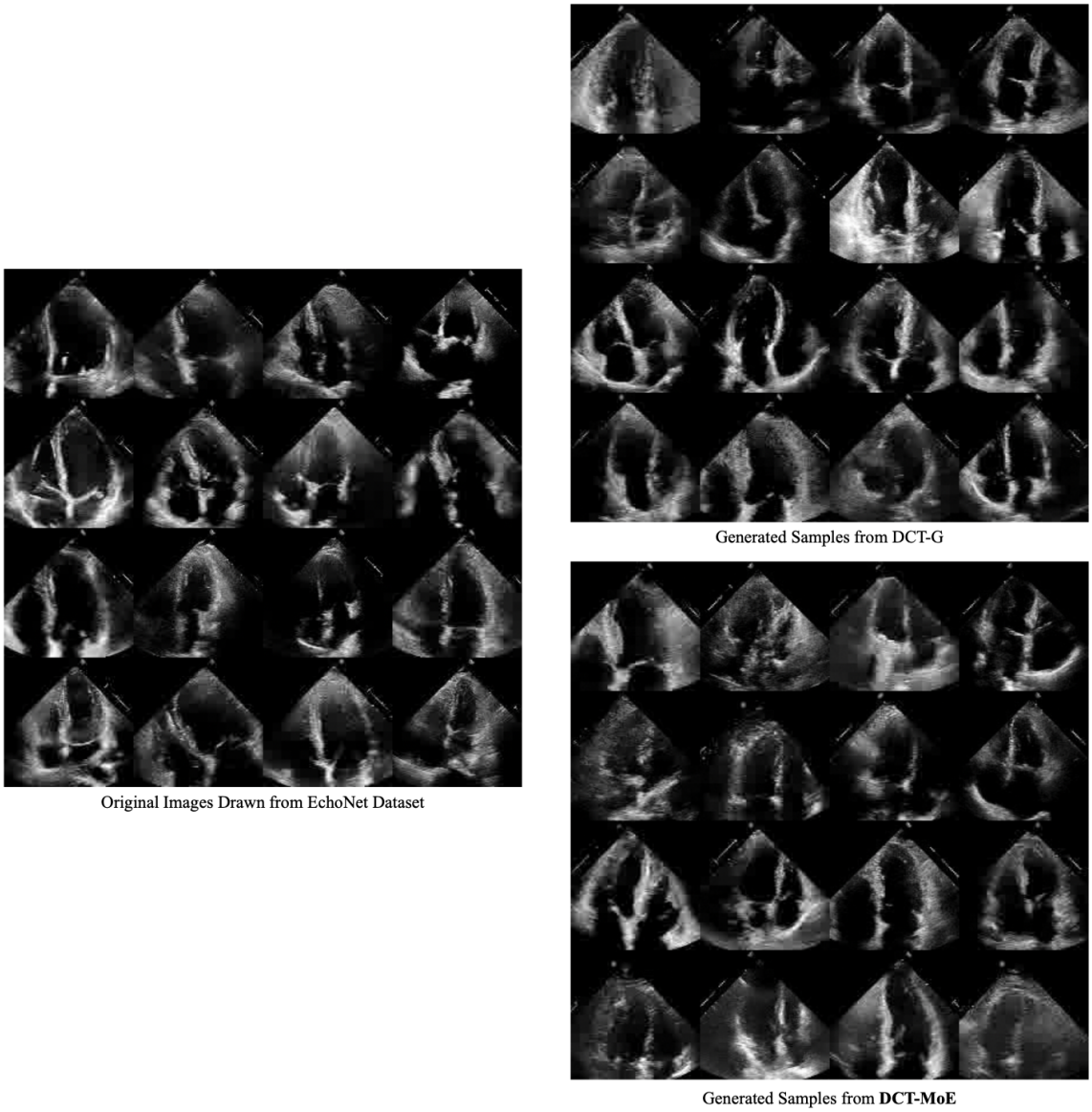

Figure 8: Visualization of DCT-G and **DCT-MoE** on EchoNet dataset.