FA-MoE: Improving Medical Image Generation through Frequency-Aware Mixture of Experts

Supplementary Material

Overview of Supplementary Materials This supplementary material provides additional details and experimental results to support our main paper. It is organised as follows:

- Section A presents detailed pseudo-code for our DCT-based frequency-aware tokenisation scheme in Section 4.1.
- Section B reports further experimental details of our configurations.
- Section C presents the model efficiency analysis of our experiments.
- Section D presents additional visualisation examples that illustrate the generation quality of our method and all competitors.

A. Pseudo Code

The Algorithm 1 follows the same three stages introduced in the main text: *block-wise DCT transform*, *zigzag ordering and frequency truncation*, and *macro-block formation and frequency-band grouping*. Each step in the algorithm matches the operations explained in the method description and shows how the DCT coefficients are grouped into frequency bands and combined to form the final set of tokens for each image X.

Algorithm 1 DCT-Based Frequency-Aware Tokenisation

Input: Greyscale image $X \in \mathbb{R}^{H\times W}$; block size b; truncation length m; macro-block grid (r_x, r_y); number of bands k (with $m = k\ell$).
Output: Token set $\mathcal{T}$.
Step 1: *Split the image into blocks*
Divide X into $N = \frac{H\times W}{b^2}$ non-overlapping blocks $A_i \in \mathbb{R}^{b\times b}$.
Step 2: *DCT, zigzag, truncation, band split*
for $i = 1$ to N **do**
    Compute the 2D DCT of A_i to obtain $D_i(u, v)$.
    Apply zigzag ordering to form $c_i \in \mathbb{R}^{b^2}$.
    Keep the first m entries, giving $\tilde{c}_i \in \mathbb{R}^m$.
    Split $\tilde{\mathbf{c}}_i$ into k segments $\tilde{\mathbf{c}}_i^{(j)} \in \mathbb{R}^\ell$, $j = 1, \ldots, k$, with $\ell = m/k$.
end for
Step 3: *Macro-blocks and tokens*
Group every $r_x \times r_y$ neighbouring blocks into one macro-block (each has $r = r_x r_y$ blocks).
Initialise $\mathcal{T} \leftarrow \emptyset$.
for each macro-block p **do**
    Let the r blocks in this macro-block have local indices $q = 1, \ldots, r$.
    for $j = 1$ to k **do**
        Form the token for band j:

$$t_p^{(j)} = [\,\tilde{\mathbf{c}}_{p,1}^{(j)}, \ldots, \tilde{\mathbf{c}}_{p,r}^{(j)}\,] \in \mathbb{R}^{r\ell},$$

        where $\tilde{\mathbf{c}}_{p,q}^{(j)}$ is the j-th segment of the q-th block in macro-block p.
        Add $t_p^{(j)}$ to $\mathcal{T}$.
    end for
end for
return $\mathcal{T}$

B. Implementation Details

We currently employ the U-ViT architecture as backbone network for our proposed **FA-MoE**. It is worth noting that FA-MoE is equally applicable to other transformer-based network such as DiT and conditional generation scenarios. For the diffusion process, we employ a continuous-time diffusion framework based on score functions with DPM-Solver(ODE) for sampling, setting the number of function evaluations (NFEs) to 50. Table 5 contains the important details of our experiment configurations.

Our U-ViT base model consists of 16 transformer blocks, and we employ the FA-MoE blocks exclusively on the first and last transformer blocks. Intuitively, it is straightforward to substitute all transformer blocks with FA-MoE blocks, but we have discovered that such an implementation causes unacceptable computational overhead in our current condition. Our experiment results in Section 5 have proven that only 2 FA-MoE blocks (the first and last) are capable enough to improve the model performance without compromising training efficiency. For all FA-MoE configurations, we fix the segment number $k = 2$, which means we divide the remaining m DCT coefficients from a single block into 2 segments, one for low frequency and one for high frequency. All experiments were conducted on a Linux platform with 2× NVIDIA A100-80G GPUs. 500,000 training steps and a batch size of 512 are set for each configuration.

C. Model Efficiency Comparison

To better evaluate the computational efficiency of the proposed method **FA-MoE**, we calculated the activated parameter count and GFLOPs during our experiments in Table 6. Note that we used the same configurations as in Table 3.

Table 5. Summary of experiment configurations.

Method	Backbone	Tokenization / VAE	Sampling	NFEs/Timesteps
FA-MoE	U-ViT (16 Blocks)	DCT (Block Size: 4)	DPM-Solver(ODE)	50 NFEs
LDM	U-Net (4 Scales)	FLUX.1-dev VAE	DDPM	1000(Train) / 256(Sample)
DCTdiff	U-ViT (16 Blocks)	DCT (Block Size: 4)	DPM-Solver(ODE)	50 NFEs
Med-D3CG	U-Net (4 Scales)	WT(db4)	DDPM	1000(Train) / 256(Sample)

Table 6. Model activated parameters and GFLOPs comparison.

Dataset	Model	# Parameters	GFLOPs
ACDC	**FA-MoE**	163M	58.89
	LDM	213M (83M + 130M)	83.35
	DCTdiff	145M	76.14
	Med-D3CG	21M	2140.581
EchoNet	**FA-MoE**	163M	45.30
	LDM	213M (83M + 130M)	112.79
	DCTdiff	145M	76.69
	Med-D3CG	21M	2913.57

As shown in Table 6, the 83M active parameters introduced by the VAE impose a significant computational burden on LDM, leading to suboptimal GFLOPs and parameter efficiency compared to FA-MoE and DCTdiff. As evidenced by the higher active parameter counts for FA-MoE and DCTdiff compared to the CNN-based LDM and Med-D3CG, we found that although transformer-based backbones are typically more parameter-intensive, this trend reverses regarding GFLOPs. Sharing the same hidden dimension as DCTdiff and having an high parameter count versus LDM, FA-MoE achieves the lowest GFLOPs by utilizing our frequency-aware tokenization and omitting the VAE. Conversely, the high computational cost of Med-D3CG is attributed to the Wavelet Transform (WT), which is notoriously complex and highly sensitive to the choice of the base wavelet.

D. Visualisation

We provide more samples from our experiments as well as samples from the ACDC dataset here. For each figure, we select 16 samples randomly.

In Figure 8 we present random samples from the preprocessed ACDC dataset and in Figure 9 we present random samples from the preprocessed EchoNet dataset. In Figure 10 and Figure 11 we present random samples drawn from checkpoints corresponding to the lowest FID score of **FA-MoE** in ACDC dataset and EchoNet dataset, respectively.

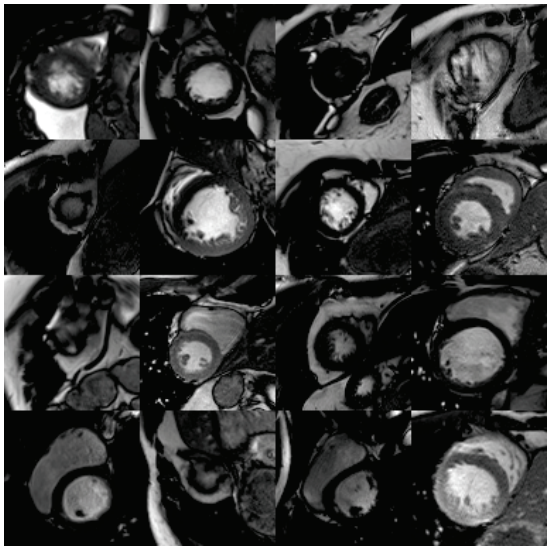

Figure 8. Random samples from preprocessed ACDC dataset.

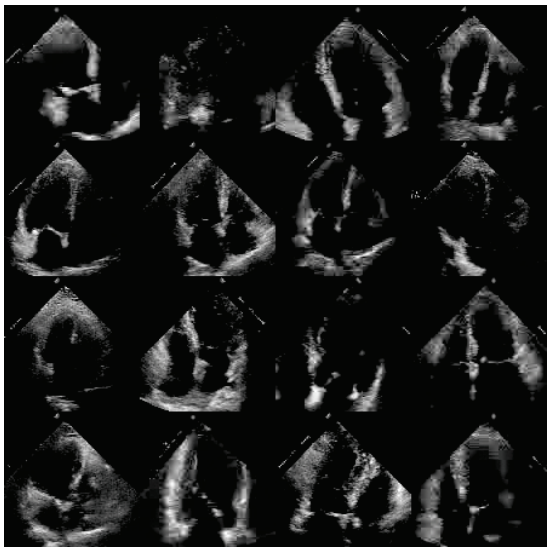

Figure 9. Random samples from preprocessed EchoNet dataset.

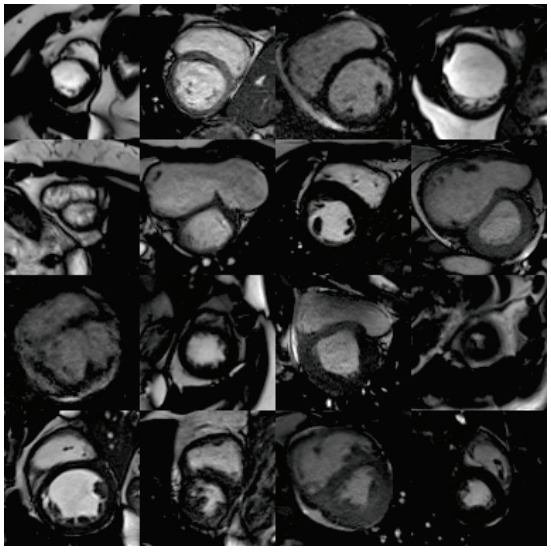

Figure 10. Random samples from **FA-MoE** experiment.

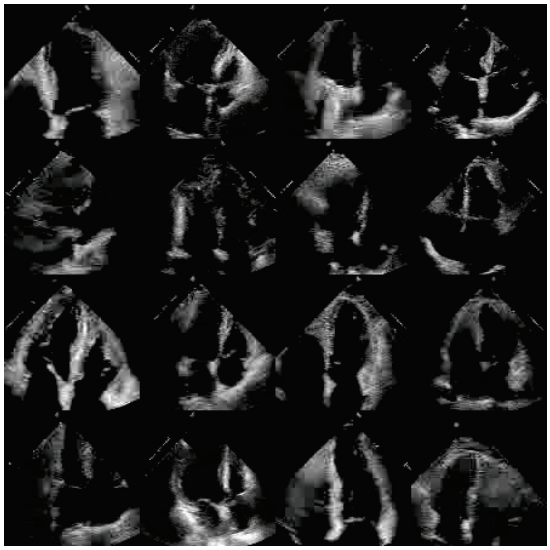

Figure 11. Random samples from **FA-MoE** experiment.