FB-Diff: Fourier Basis-guided Diffusion for Temporal Interpolation of 4D Medical Imaging

Supplementary Material

6. Dataset Settings

ACDC. The ACDC dataset contains 80% pathological cardiac cases, including pathologies with myocardial infarction, cardiomyopathy. All MRI volumes are resampled with a voxel space of $1.5 \times 1.5 \times 3.12mm^3$. Besides, all cardiac scans have been cropped with a centered patch. The patch size is set as $128 \times 128 \times 32$. The frame number N shows a range of $[6, 16]$. Min-max scaling at $[0, 1]$ is applied to ensure consistent scaling across all scans.

4D Lung. In the case of the 4D-lung dataset, the models are trained to predict the four intermediate frames $(10\%, 20\%, 30\%, 40\%)$ between the end-inspiratory (0%) and end-expiratory (50%) phases. Only CT images captured using kilovoltage energy are included in the study due to their superior image quality. The data preprocessing strategy is the same as that in [30].

7. Implementation Details

Network Details. For the first stage, the VAE is not to regulate the whole pipeline, but to utilize a MedNeXt [42] structure for encoding temporal features and learning Fourier bases. The VAE maps the image space into the downsampled latent space with a ratio of $1/8$. Specifically, the core component for MedNeXt is the MedNeXtBlock. For more details of the VAE, please refer to the source codes released here. For the latent diffusion UNet, we select a more lightweight MedNeXt as the baseline, with the downsampling scale equal to $1/4$. The diffusion timestep is set as 1000. L_2 norm is chosen as the loss function for the diffusion process.

Training Details. All models are trained using AdamW optimizer with the linear warm-up strategy. For the taining of VAE, the initial learning rate is set as $3e$-4 with a cosine learning rate decay scheduler, and weight decay is set as $1e$-5. While for the training of the diffusion model, the learning rate is set as $1e$-4. The batch size is set as 2. Experiments are implemented based on Pytorch and 2 NVIDIA RTX 4090 GPUs.

8. Model Efficiency

We have added the model efficiency metrics. Table 5 reports the training time, FLOPs, and per-case inference speed for models. Overall, FB-Diff offers a good trade-off in performance and model efficiency.

Table 5. (a) Generalization on cardiac ultrasound in EchoNet-Dynamic [40]. (b) Model efficiency on training time, computational costs, and per-case inference speed.

Model	(a) Cardiac ultrasound			(b) Model efficiency		
	PSNR (dB) ↑	LPIPS ↓	FVD ↓	Training time (h)	FLOPs (T)	Inference (s)
Voxelmorph [2]	28.40	2.492	295.3	5.6	0.49	1.09
IFRNet [31]	29.95	2.017	261.8	21.3	1.92	1.27
UVI-Net [30]	**30.87**	<u>1.818</u>	<u>243.7</u>	18.5	1.27	0.63
Conditional diff [16]	26.67	2.578	337.2	12.4	2.37	37.80
FB-Diff	<u>30.51</u>	**1.654**	**227.0**	18.0	1.58	29.50

9. Generalization to other modalities

We tested FB-Diff on a different imaging modality to confirm generality. Using the cardiac ultrasound dataset proposed by [40], FB-Diff achieves comparable or better performance than benchmark methods. As revealed in Table 5, FB-Diff achieves better temporal consistency for interpolated videos while maintaining promising reconstruction metrics.

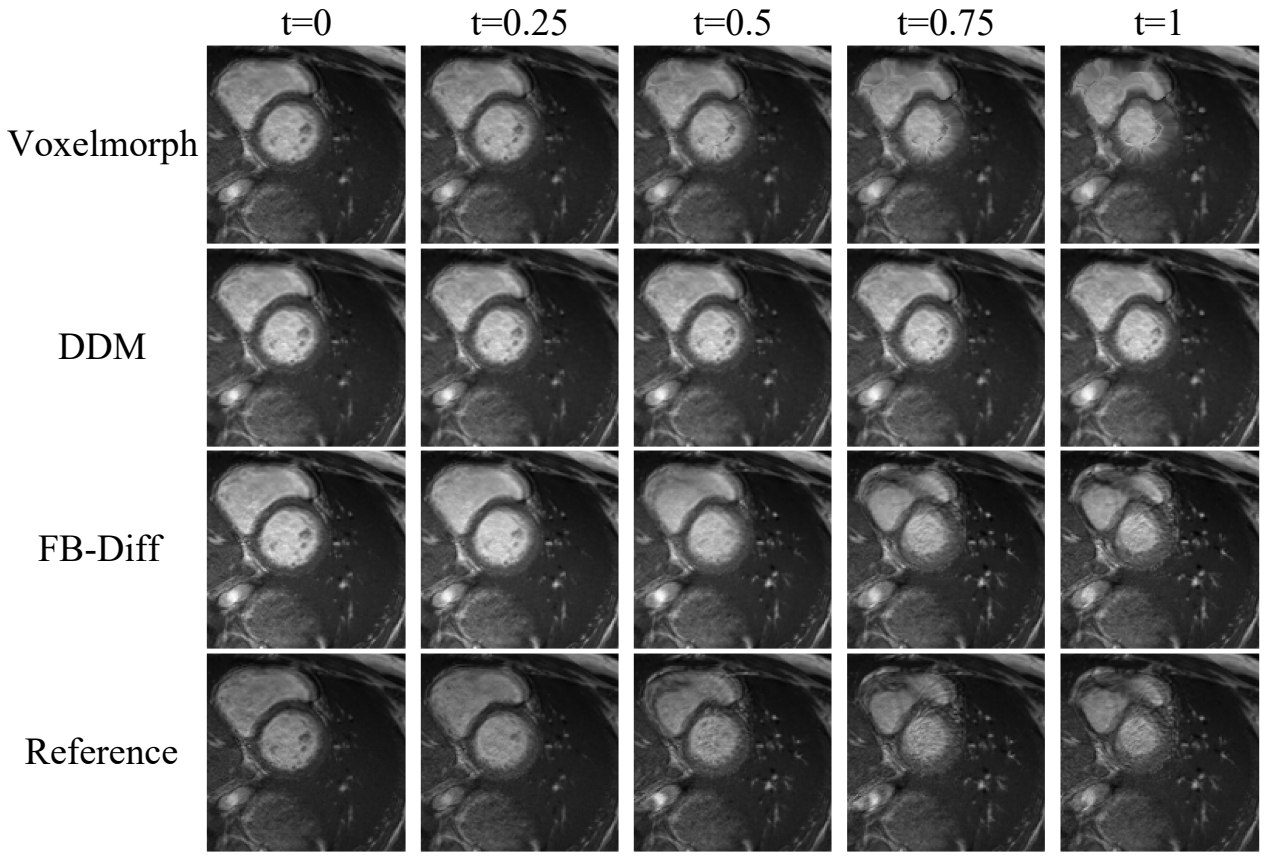

Figure 8. Temporal variation comparison between FB-Diff and existing methods with the linear motion hypothesis.

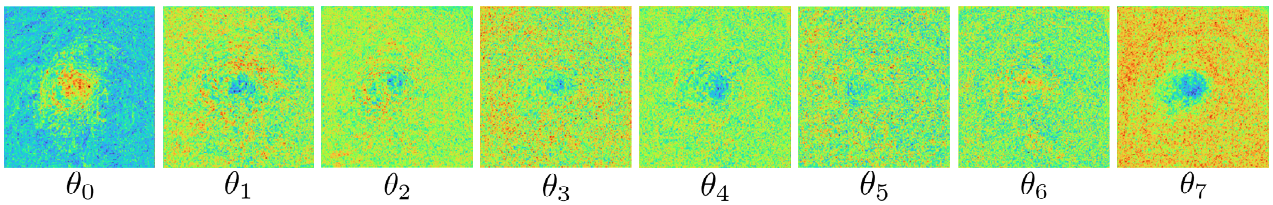

Figure 9. The spectral intensity visualizations of the first eight well-learned physiology motion priors on ACDC.