MICCAI

Temporal Differential Fields for 4D Motion Modeling via Image-to-Video Synthesis

Xin You[1,2,3], Minghui Zhang[1], Hanxiao Zhang[1],
Jie Yang[1(✉)], and Nassir Navab[2(✉)]

[1] Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai, China
[2] Computer Aided Medical Procedures, Technical University of Munich, Munich, Germany
[3] Munich Center for Machine Learning, Munich, Germany
sjtu_youxin@sjtu.edu.cn, nassir.navab@tum.de

Abstract. Temporal modeling on regular respiration-induced motions is crucial to image-guided clinical applications. Existing methods cannot simulate temporal motions unless high-dose imaging scans including starting and ending frames exist simultaneously. However, in the preoperative data acquisition stage, the slight movement of patients may result in dynamic backgrounds between the first and last frames in a respiratory period. This additional deviation can hardly be removed by image registration, thus affecting the temporal modeling. To address that limitation, we pioneeringly simulate the regular motion process via the image-to-video (I2V) synthesis framework, which animates with the first frame to forecast future frames of a given length. Besides, to promote the temporal consistency of animated videos, we devise the Temporal Differential Diffusion Model to generate temporal differential fields, which measure the relative differential representations between adjacent frames. The prompt attention layer is devised for fine-grained differential fields, and the field augmented layer is adopted to better interact these fields with the I2V framework, promoting more accurate temporal variation of synthesized videos. Extensive results on ACDC cardiac and 4D Lung datasets reveal that our approach simulates 4D videos along the intrinsic motion trajectory, rivaling other competitive methods on perceptual similarity and temporal consistency. Codes are available at https://github.com/AlexYouXin/Mo-Diff

Keywords: Motion Modeling · 4D · Diffusion · Temporal Consistency.

1 Introduction

4D Temporal modeling on breathing-induced motions, is significant to image-guided clinical applications [8, 15, 26], such as disease diagnosis and therapy planning. Particularly, 4D cardiac Magnetic Resonance Imaging (MRI), monitoring the anatomical variation of 3D cardiac structures [1, 30], is frequently used for cardiac disease analysis and surgical intervention. Besides, 4D pulmonary Computed Tomography (CT) is adopted to model the respiratory process, which

is beneficial to the intraoperative puncture of pulmonary nodules by assisting clinicians to determine the optimal puncture path [18].

Existing methods utilize flow-based interpolation models [9, 17] to deduce the regular motion process with the predefined starting and ending frames. And some researchers aim to synthesize the whole videos by means of diffusion models conditioning on these two prompting frames [16, 11]. Basically speaking, all these methods could barely simulate breathing-induced motions only if two prompting frames exist simultaneously in the test-time evaluation. However, in the preoperative data acquisition stage, the slight movement or unstable breathing of patients may result in dynamic backgrounds between the first and last frames in a respiratory period. This additional bias cannot be thoroughly removed through image registration, thereby affecting the temporal motion modeling. Moreover, clinical research indicates that two high-dose CT or MRI scans will cause prolonged radiation exposure, potentially impairing patients' health [21, 31].

To address these limitations, we pioneeringly simulate regular respiration-induced motions via the image-to-video (I2V) synthesis framework, which animates with the first frame to forecast future frames of a given length. Owing to the compatibility of conditional diffusion models across different inputs [34], the image-to-video paradigm can be seamlessly plugged into diffusion models. Herein, the diffusion model implicitly learns temporal correlations of 4D data with regular motion discipline, under the conditional guidance of the starting frame and frame number. Besides, to promote the temporal consistency of animated videos, we propose the Temporal Differential Diffusion Model (TDDM) to generate temporal differential fields, which measure the relative differential representations between adjacent frames. The prompt attention layer (PAL) is devised for fine-grained differential fields, and the field augmented layer (FAL) is adopted to better interact these fields with the I2V framework, boosting precise temporal consistency of synthesized videos. Our two-stage pipeline is aimed at the **Mo**tion simulation via conditional **Diff**usion models, termed **Mo-Diff**. Extensive results on ACDC cardiac and 4D Lung datasets reveal that Mo-Diff forecasts accurate volumes along the intrinsic motion trajectory, rivaling other competitive methods on perceptual similarity and temporal consistency, which are more clinically significant than pixel-centric reconstruction metrics.

Contributions: 1) We pioneeringly introduce the image-to-video (I2V) paradigm to simulate regular breathing-induced motions, and propose the two-stage pipeline termed Mo-Diff. 2) The temporal differential diffusion model is devised to yield temporal differential fields as conditional guidance, amplifying the temporal consistency of synthesized volumes by I2V. 3) The prompt attention layer is aimed at fine-grained differential fields, and the field augmented layer is adopted to effectively interact these fields with the I2V framework, promoting more accurate temporal consistency. 4) Experiments demonstrate that Mo-Diff can synthesize precise and realistic videos with better perceptual quality and temporal consistency, showcasing potentials of simulating regular cardiac and pulmonary motions.
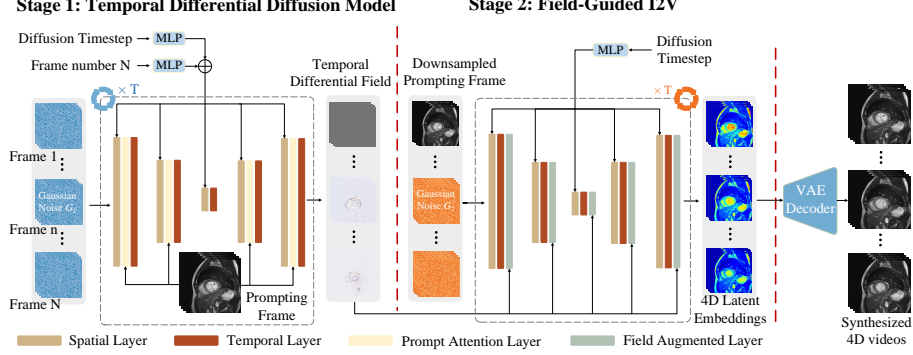
Fig. 1. The whole Mo-Diff pipeline. Stage 1: TDDM yields temporal differential fields conditioning on the prompting frame and frame number. Stage 2: the conditional diffusion model will synthesize 4D latent embeddings with the guidance of temporal fields. These latent embeddings are then transformed into 4D videos with regular motions.

Related work: Breathing-induced motion modeling has been a critical research area within medical image computing and computer-assisted intervention [27, 32]. Existing methods on this topic can be categorized into two groups. The first group of approaches try to interpolate intermediate frames with the starting and ending frames in a breathing period, based on the perspective of optical flow estimation. Specifically, SVIN [9] estimates forward and backward deformation fields, which will yield precise intermediate frames by linearly combining bidirectional information. UVI-Net [17] utilizes the flow calculation model with the time-domain cycle-consistency constraint and linear motion hypothesis, to realize motion modeling in an unsupervised style.

Recently, denoising diffusion probabilistic models [10, 11, 19] show promising performance in generating realistic images or videos, by transforming random Gaussian distributions into target data distributions. The second category of methods maximize the potential of diffusion models to resolve video interpolation tasks in a generative manner. Essentially, these approaches implicitly learn the temporal motions from the starting frame to the ending frame [14, 2, 6]. Furthermore, a recent work [16] devised a diffusion deformable model (DDM), which can learn spatial deformation maps between the starting and ending volumes and provide a latent code for generating intermediate frames along a geodesic path.

2   Methodology

2.1   Image-to-Video Conditional Diffusion Models

To eliminate the potential misalignment caused by patients' movements in clinical practice, we intend to avoid the data acquisition of the ending CT or MRI frame in a respiratory cycle, and only collect the starting frame for the motion simulation of future frames. Thus, we select the diffusion model conditioning on the first volume frame to simulate regular temporal motions. That generative

model can also be viewed as the image-to-video (I2V) framework. For saving GPU memory, the image space is transformed into latent space with a well-trained Variational Autoencoder (VAE). And our framework draws lessons from the Latent Video Diffusion Model (LVDM) [5, 23]. It conducts the denoising process in the latent space. Before training, the input video x_0 is first encoded into a latent embedding $z_0 = \mathcal{E}(x_0)$ with VAE encoder $\mathcal{E}(\cdot)$, and z_0 can be restored into x_0 via VAE decoder $\mathcal{D}(\cdot)$. Then the latent code z_0 is perturbed as:

$$z_t = \sqrt{\overline{\alpha}_t}\, z_0 + \sqrt{1 - \overline{\alpha}_t}\, \epsilon, \epsilon \sim \boldsymbol{N}(0, 1) \tag{1}$$

where $\overline{\alpha}_t = \prod_{i=1}^{t}(1 - \beta_t)$ with β_t is the noise strength coefficient at time step t, and t is uniformly sampled from the timestep index set $1, ..., T$. This process can be regarded as a Markov chain, which incrementally adds Gaussian noise to the latent code z_0. The denoising model ϵ_θ receives z_t as input and is optimized to learn the latent space distribution with the objective function:

$$\mathcal{L}_\epsilon = E_{z_t, \epsilon \sim \mathcal{N}(0, 1)}\, ||\epsilon - \epsilon_\theta(z_t, t, c)||^2 \tag{2}$$

where c represents the condition, and ϵ_θ is implemented as a light-weight U-Net. Using the devised LVDM, we aim to synthesize the latent embeddings showcasing motion dynamics of anatomical structures.

2.2   Temporal Differential Diffusion Model

The I2V model is aimed at modeling the motion discipline of the breathing process. However, as demonstrated by [19, 24], synthesized videos are prone to visualizing sequences with poor temporal consistency if no extra condition is injected into diffusion models. Thus, we devise the temporal differential diffusion model (TDDM) to generate temporal differential fields as an additional prompt. This conditional input can boost a temporally consistent synthesis of latent embeddings by the above-mentioned I2V framework, with smoother variation on anatomical shape and texture details.

Temporal differential fields are defined to measure the relative variance between two adjacent frames. Specifically, the ith frame of temporal fields $\mathcal{F}_i$ is calculated with the subtraction between image frame $\mathcal{I}_i$ and $\mathcal{I}_{i-1}$, with i ranging from 1 to frame number N. The detailed process is noted as follows:

$$\mathcal{F}_i = \begin{cases} \mathcal{I}_i \ominus \mathcal{I}_{i-1}, & \text{if } i \geq 2 \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

$\mathcal{F}_1$ is set as a zero mask to improve the convergence efficiency. After the score-based training, TDDM can transform the Gaussian noise $\mathcal{G}_1 \sim \mathcal{N}(0, 1)$ into synthesized differential fields $\hat{\mathcal{F}}$ corresponding to the prompting image $\mathcal{I}_1$ and frame number N, formulated as $\hat{\mathcal{F}} = TDDM(\mathcal{G}_1, t; \mathcal{I}_1, N)$. Here t refers to the diffusion time step. As revealed in Fig. 1, $\hat{\mathcal{F}}$ will suppress dynamically varying backgrounds, and highlight the foreground regions in the respiratory process. Thus, $\hat{\mathcal{F}}$ can enormously boost the reconstruction performance of future frames
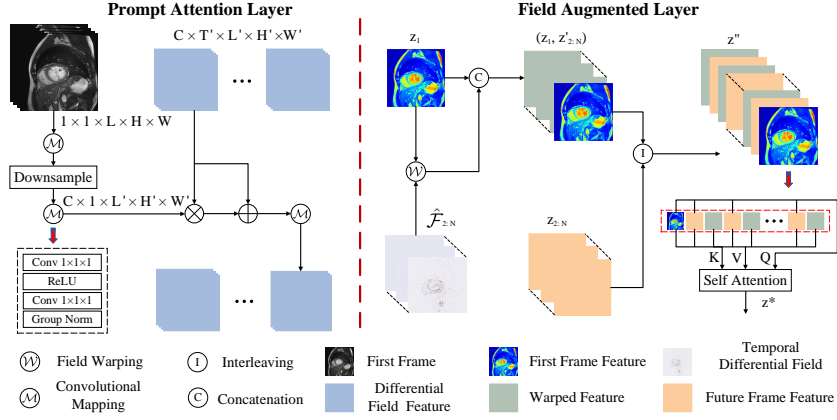
Fig. 2. The detailed structure of prompt attention layer and field augmented layer.

by the I2V network. Besides, temporal fields representing regular motions will significantly improve the temporal consistency of synthesized videos.

It is worth mentioning that frame number N is a significant conditional variable, serving as a compensation for the absence of the ending frame. Due to a fixed samping duration between frames, the specific N corresponds to the specific breathing period, acquired by electrocardiogram signals, thus influencing the rate of temporal motion variations. As shown in Fig. 1, N serves as an explicit constraint on the motion velocity of respiration, promoting TDDM to generate appropriate temporal differential fields, further improving the video synthesis with regular motions.

2.3   Network Details of Mo-Diff

Spatial & Temporal Layer: Mo-Diff is designed for yielding four-dimensional videos, which pose a huge challenge to the joint modeling on spatial and temporal domains. Inspired by [22], we try to extract spatial and temporal information in a separate manner. Given the noisy input $\mathcal{I}_{\mathcal{G}}$ with a dimension of $B \times C \times N \times L \times H \times W$ (corresponding to batch size, channel, frame number, length, height, and width individually), we respectively implement 3D spatial convolutions and 1D temporal convolutions on $\mathcal{I}_{\mathcal{G}}$ as visualized by Fig. 1(b).

Prompt Attention Layer (PAL): To integrate the prompt information into the first-stage TDDM, we do not simply concatenate the first frame with noisy differential fields for training as what previous works did [28, 20, 33]. Instead, a feature-level fusion is adopted to synthesize temporal differential fields with fine-grained textures and shapes, which hold importance for the synthesis performance of future frames. Thus, we propose the PAL to boost the representation of TDDM. Specifically, given the initial frame $\mathcal{I}_1$, we define the convolutional mapping $\mathcal{M}$ which can transform the image space into feature space. As visualized in Fig. 2, $\mathcal{M}$ is a combination of convolution, ReLU, and group normalization operations. Then the warped feature of the prompting frame is utilized to en-

Table 1. Baseline comparison with various models. Double prompting frame: starting and ending frames; Single prompting frame: only the starting frame (Blue, red values represent best evaluation metrics corresponding to two types of models respectively).

Prompting Frame	Model	ACDC Cardiac			4D Lung		
		PSNR (dB) ↑	LPIPS ↓	FVD ↓	PSNR (dB) ↑	LPIPS ↓	FVD ↓
Double	SVIN [9]	$31.43_{\pm0.421}$	$1.563_{\pm0.206}$	93.6	$30.49_{\pm0.304}$	$2.650_{\pm0.245}$	125.6
	Voxelmorph [3]	$30.77_{\pm0.502}$	$1.969_{\pm0.197}$	102.1	$29.90_{\pm0.373}$	$2.815_{\pm0.260}$	149.0
	UVI-Net [17]	$32.16_{\pm0.402}$	$1.662_{\pm0.245}$	94.2	$31.57_{\pm0.311}$	$2.211_{\pm0.216}$	121.7
	LDMVFI [7]	$27.11_{\pm0.460}$	$2.943_{\pm0.410}$	99.2	$26.31_{\pm0.453}$	$3.659_{\pm0.341}$	142.7
	DDM [16]	$29.79_{\pm0.504}$	$2.689_{\pm0.352}$	110.3	$29.67_{\pm0.420}$	$2.905_{\pm0.330}$	165.5
Single	LDDM [6]	$24.53_{\pm0.481}$	$2.634_{\pm0.308}$	105.7	$25.19_{\pm0.273}$	$2.914_{\pm0.287}$	146.3
	Condi-Diffusion [11]	$26.59_{\pm0.545}$	$2.460_{\pm0.357}$	95.7	$25.95_{\pm0.391}$	$3.294_{\pm0.375}$	133.0
	Mo-Diff	$30.79_{\pm0.409}$	$1.317_{\pm0.189}$	86.1	$29.86_{\pm0.282}$	$2.137_{\pm0.225}$	115.8

hance features corresponding to temporal differential fields via the element-wise multiplication and addition.

Field Augmented Layer (FAL): To synthesize 4D videos with temporal consistency, we propose the FAL to interact temporal differential fields with frame features. Specifically, z_1 denotes the feature of the first frame, and $z_{2:N}$ means the subsequent frame features. As revealed in Fig. 2, via the warping transform $\mathcal{W}$, differential fields of future frames $\hat{\mathcal{F}}_{2:N}$ are warped into frame features $z'_{2:N}$. The detailed process can be formulated as Eq (4).

$$z'_i = \mathcal{W}(z_1, \hat{\mathcal{F}}_{1\to i}), \ i = 2, ..., N \qquad (4)$$

z'_i shows similar spatial-temporal patterns to z_1. Then warped features are interleaved with future frame features $z_{2:N}$ in the temporal dimension. Specifically, two-branch features with the same frame index are arranged side by side, and the merged result is noted as augmented features $z'' = [z_1, z_2, z'_2, ..., z_N, z'_N]$, which is then projected as key and value vectors. And the query vector arises from z. After the self-attention operation, augmented frame features z can be attained to synthesize temporal-consistent latent embeddings.

3   Experiment

3.1   Experimental Settings

Dataset. To evaluate the efficacy of Mo-Diff, we conduct experiments on public ACDC cardiac [4] and 4D-Lung datasets [13]. For ACDC with the MRI modality, volume sequences from the diastolic to systolic phases are extracted as 4D data. Of all the 150 4D sequences, cases with identity 1-100, 101-120, and 121-150 serve as the training, validation, and testing sets. Besides, all MRI volumes are resampled with a voxel space of $1.5 \times 1.5 \times 3.12mm^3$ and cropped to $128 \times 128 \times 32$ [29]. And the frame number N shows a range of $[6, 16]$. For 4D-Lung CT images, the end-inspiratory and end-expiratory scans are set as the initial and final images. We collect 125 4D videos from [13], with resolutions equal to $10 \times 128 \times 128 \times 128$.
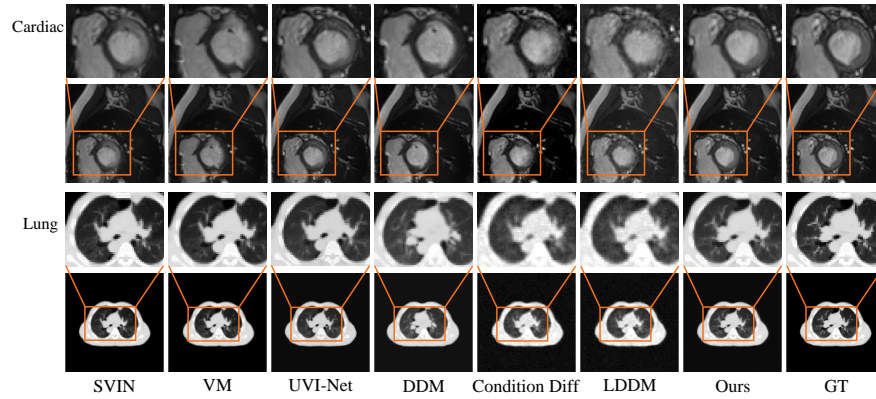
Fig. 3. Qualitative comparison of intermediate frames between different models.

They are split as 80/15/30 cases for training, validation, and inference. The intensity value is scaled as $[0, 1]$. We choose quantitative evaluation metrics including PSNR [12], Learned Perceptual Image Patch Similarity (LPIPS) [35], and Fréchet Video Distance (FVD) [25]. PSNR is the traditional pixel-centric reconstruction metric, which cannot well assess the quality of synthesized videos in clinical practice [14]. Instead, LPIPS and FVD are more reasonable metrics representing human visions. Specifically, LPIPS reveals perceptual variance. And FVD evaluates the feature distribution bias between ground truths and synthesized videos, measuring temporal consistency and continuity.

Implementation Details. The VAE maps the image space into the downsampled latent space with a ratio of $1/4$. For the proposed TDDM, the temporal channel of inputs is set as 16, but we only calculate the score-based loss $\mathcal{L}_\epsilon$ with the first N channels for efficient training. And the synthesized 4D latent embeddings by field-guided I2V are further transformed into temporal sequences via the latent decoder of VAE. All models are trained using AdamW optimizer with the linear warm-up strategy. The initial learning rate is set as $1e$-4 with a cosine learning rate decay scheduler, and weight decay is set as $1e$-5. The batch size, diffusion step T, and training epoch are equal to 2, 1000, and 500. Experiments are implemented based on Pytorch and 2 NVIDIA RTX 4090 GPUs.

3.2   Experimental Results

Baseline Comparison. For performance comparisons on the simulation of 4D temporal motions, we list out two categories of methods, classified as "Double" (conditioning on both starting and ending frames) and "Single" (conditioning on only the starting frame). Obviously, models with two prompting frames can achieve better pixel-wise reconstruction performance, with higher PSNR values. However, when evaluating the feature-level perceptual similarity with human judgments, our proposed model reveals a good LPIPS metric on ACDC, with a 0.246 ↓ lower value than supervision-based model SVIN [9] as illustrated by Table 2. Also, Mo-Diff synthesizes 4D videos with more consistent temporal

Table 2. Ablation study on key components. w/o PAL: replaced with a channel concatenation operation; w/o FAL: concatenate channels between frame features and differential fields, then conduct 1D temporal self-attention.

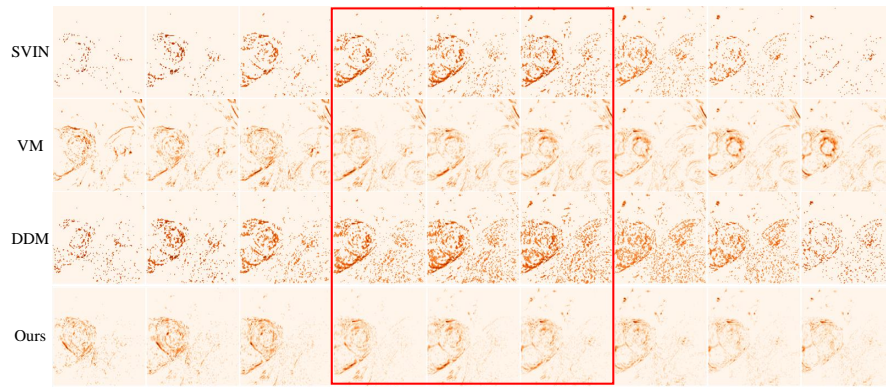| Frame Number	PAL	FAL	ACDC Cardiac			4D Lung		
			PSNR (dB) ↑	LPIPS ↓	FVD ↓	PSNR (dB) ↑	LPIPS ↓	FVD ↓
✗	✔	✔	28.07	2.206	109.7	27.44	2.732	142.9
✔	✗	✗	28.60	1.878	111.5	27.93	2.560	131.3
✔	✔	✗	29.86	1.683	94.7	28.46	2.351	119.4
✔	✗	✔	29.97	1.532	100.9	29.12	2.215	128.0
✔	✔	✔	**30.79**	**1.317**	**86.1**	**29.86**	**2.137**	**115.8**

Fig. 4. Temporal error maps between different models. All frame predictions are visualized except starting and ending frames (Red box: intermediate frames).

distributions with ground truth videos, outperforming UVI-Net [17] with a 5.9 ↓ FVD value. That convincingly validates the efficacy of the TDDM.

Qualitative Results. As illustrated by Fig. 3, our proposed model can forecast precise enough frames, rivaling other competitive flow-based models. Specifically, Mo-Diff reveals a promising shape synthesis of arteries for lung data and texture synthesis of myocardium for cardiac data. Also, Fig. 4 depicts temporal error maps for a qualitative evaluation on temporal consistency. Our results show more precise simulation for three intermediate frames, which are tricky to forecast due to a farther distance to prompting frames. In contrast, SVIN and Voxelmorph reveal better reconstruction performance for bilateral frames.

3.3   Ablation Study

We carried out the ablation study on Mo-Diff. 1) **Frame number** N: as shown in Table 2, removing the frame number off Mo-Diff will largely degrade the reconstruction and synthesis performance, with $2.42dB$ ↓ PSNR and 0.889 ↑ LPIPS on ACDC. Without N, the model cannot perceive the implicit period information of a respiratory process, affecting the motion modeling. 2) **PAL**: PAL aims to fully leverage the first frame by extracting prompting frame features.

A coarse channel concatenation fails to effectively inject prompting information into TDDM, resulting in unsatisfactory temporal differential fields. 3) **FAL**: replacing FAL with a channel fusion between these fields and frame features, will lead to modality confusion, which further affects synthesis performance with $1.40dB \downarrow$ PSNR and $0.214 \uparrow$ LPIPS.

4   Conclusion

We pioneeringly simulate regular respiration motions via the proposed Mo-Diff framework, which animates with the first frame to forecast future frames with a given length. Besides, the temporal differential diffusion model can generate temporal differential fields to promote the temporal consistency of animated videos. Mo-Diff can model relatively regular temporal motions via the prompting frame. However, it requires more clinical guidance, including electrocardiogram signals, to simulate highly unstable breathing, which is promising in future work.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

[1]   Endawoke Amsalu et al. "Spatial-temporal analysis of cause-specific cardiovascular hospital admission in Beijing, China". In: *International Journal of Environmental Health Research* 31.6 (2021), pp. 595–606.

[2]   Juyoung Bae, Elizabeth Tong, and Hao Chen. "Conditional Diffusion Model for Versatile Temporal Inpainting in 4D Cerebral CT Perfusion Imaging". In: *MICCAI*. Springer. 2024, pp. 67–77.

[3]   Guha Balakrishnan et al. "Voxelmorph: a learning framework for deformable medical image registration". In: *IEEE TMI* 38.8 (2019), pp. 1788–1800.

[4]   Olivier Bernard et al. "Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved?" In: *IEEE transactions on medical imaging* 37.11 (2018), pp. 2514–2525.

[5]   Andreas Blattmann et al. "Align your latents: High-resolution video synthesis with latent diffusion models". In: *CVPR*. 2023, pp. 22563–22575.

[6]   Tingxiu Chen et al. "Ultrasound Image-to-Video Synthesis via Latent Dynamic Diffusion Models". In: *MICCAI*. Springer. 2024, pp. 764–774.

[7]   Duolikun Danier, Fan Zhang, and David Bull. "Ldmvfi: Video frame interpolation with latent diffusion models". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 2. 2024, pp. 1472–1480.

[8]   Jan Ehrhardt, Cristian Lorenz, et al. *4D modeling and estimation of respiratory motion for radiation therapy*. Vol. 10. Springer, 2013.

[9]   Yuyu Guo et al. "A spatiotemporal volumetric interpolation network for 4d dynamic medical image". In: *CVPR*. 2020, pp. 4726–4735.

[10]   Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models". In: *NeurIPS* 33 (2020), pp. 6840–6851.

[11] Jonathan Ho et al. "Video diffusion models". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 8633–8646.

[12] Alain Hore and Djemel Ziou. "Image quality metrics: PSNR vs. SSIM". In: *2010 20th international conference on pattern recognition*. IEEE. 2010, pp. 2366–2369.

[13] Geoffrey D Hugo et al. "Data from 4D lung imaging of NSCLC patients". In: *(No Title)* (2016).

[14] Siddhant Jain et al. "Video interpolation with diffusion models". In: *CVPR*. 2024, pp. 7341–7351.

[15] Mi-Young Jeung et al. "Myocardial tagging with MR imaging: overview of normal and pathologic findings". In: *Radiographics* 32.5 (2012), pp. 1381–1398.

[16] Boah Kim and Jong Chul Ye. "Diffusion deformable model for 4D temporal medical image generation". In: *MICCAI*. Springer. 2022, pp. 539–548.

[17] JungEun Kim et al. "Data-Efficient Unsupervised Interpolation Without Any Intermediate Frame for 4D Medical Images". In: *CVPR*. 2024, pp. 11353–11364.

[18] Dongyuan Li et al. "A robotic system for transthoracic puncture of pulmonary nodules based on gated respiratory compensation". In: *Computer Methods and Programs in Biomedicine* 244 (2024), p. 107995.

[19] Yixin Liu et al. "Sora: A review on background, technology, limitations, and opportunities of large vision models". In: *arXiv preprint arXiv:2402.17177* (2024).

[20] Wei Peng et al. "Generating realistic brain mris via a conditional diffusion probabilistic model". In: *MICCAI*. Springer. 2023, pp. 14–24.

[21] Stephen P Power et al. "Computed tomography and patient risk: Facts, perceptions and uncertainties". In: *World journal of radiology* 8.12 (2016), p. 902.

[22] Zhaofan Qiu et al. "Learning spatio-temporal representation with pseudo-3d residual networks". In: *ICCV*. 2017, pp. 5533–5541.

[23] Robin Rombach et al. "High-resolution image synthesis with latent diffusion models". In: *CVPR*. 2022, pp. 10684–10695.

[24] Xiaoyu Shi et al. "Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling". In: *SIGGRAPH*. 2024, pp. 1–11.

[25] Thomas Unterthiner et al. "Towards accurate generative models of video: A new metric & challenges". In: *arXiv preprint arXiv:1812.01717* (2018).

[26] Lu Wang et al. "Dosimetric comparison of stereotactic body radiotherapy using 4D CT and multiphase CT images for treatment planning of lung cancer: evaluation of the impact on daily dose coverage". In: *Radiotherapy and Oncology* 91.3 (2009), pp. 314–324.

[27] Tzu-Ti Wei et al. "Mpvf: 4d medical image inpainting by multi-pyramid voxel flows". In: *IEEE Journal of Biomedical and Health Informatics* (2023).

[28] Zhaohu Xing et al. "Cross-conditioned diffusion model for medical image to image translation". In: *MICCAI*. Springer. 2024, pp. 201–211.

[29]   Xin You et al. "Learning with explicit shape priors for medical image segmentation". In: *IEEE Transactions on Medical Imaging* (2024).

[30]   Xin You et al. "Semantic difference guidance for the uncertain boundary segmentation of CT left atrial appendage". In: *MICCAI*. Springer. 2023, pp. 121–131.

[31]   Xin You et al. "SLoRD: Structural Low-Rank Descriptors for Shape Consistency in Vertebrae Segmentation". In: *arXiv preprint arXiv:2407.08555* (2024).

[32]   Xiaohan Yuan et al. "4D myocardium reconstruction with decoupled motion and shape model". In: *ICCV*. 2023, pp. 21252–21262.

[33]   Chuyan Zhang et al. "Pass: test-time prompting to adapt styles and semantic shapes in medical image segmentation". In: *IEEE TMI* (2024).

[34]   Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 3836–3847.

[35]   Richard Zhang et al. "The unreasonable effectiveness of deep features as a perceptual metric". In: *CVPR*. 2018, pp. 586–595.