CHeart: A Conditional Spatio-Temporal Generative Model for Cardiac Anatomy

Mengyun Qiao, Shuo Wang, Huaqi Qiu, Antonio de Marvao, Declan P. O'Regan, Daniel Rueckert, *IEEE Fellow* and Wenjia Bai

Abstract—Two key questions in cardiac image analysis are to assess the anatomy and motion of the heart from images; and to understand how they are associated with non-imaging clinical factors such as gender, age and diseases. While the first question can often be addressed by image segmentation and motion tracking algorithms, our capability to model and answer the second question is still limited. In this work, we propose a novel conditional generative model to describe the 4D spatio-temporal anatomy of the heart and its interaction with non-imaging clinical factors. The clinical factors are integrated as the conditions of the generative modelling, which allows us to investigate how these factors influence the cardiac anatomy. We evaluate the model performance in mainly two tasks, anatomical sequence completion and sequence generation. The model achieves high performance in anatomical sequence completion, comparable to or outperforming other state-of-the-art generative models. In terms of sequence generation, given clinical conditions, the model can generate realistic synthetic 4D sequential anatomies that share similar distributions with the real data. The code and the trained generative model are available at https://github.com/MengyunQ/CHeart.

Index Terms—Conditional generative model, synthetic data generation, cardiac image analysis, cardiac anatomy and motion

M. Qiao and W. Bai are jointly affiliated with Department of Computing, Department of Brain Sciences and Data Science Institute, Imperial College London, London, SW7 2AZ, United Kingdom. H. Qiu and D. Rueckert are with Biomedical Image Analysis Group (BioMedIA), Department of Computing, Imperial College London. D. Rueckert is also with Klinikum rechts der Isar, Technical University of Munich, Munich, Germany. E-mail: m.qiao21@imperial.ac.uk.

A. de Marvao and D.P. O'Regan are with the MRC Laboratory of Medical Sciences, Imperial College London, London W12 0HS, United Kingdom. A. de Marvao is also with the Department of Women and Children's Health, and British Heart Foundation Centre of Research Excellence, School of Cardiovascular and Metabolic Medicine and Sciences, King's College London, London, United Kingdom.

S. Wang is with Digital Medical Research Center, School of Basic Medical Sciences, Fudan University and Shanghai Key Laboratory of MICCAI, Shanghai, China.

I. INTRODUCTION

CARDIAC imaging plays an essential role in cardiovascular image diagnosis and management [10]. Imaging modalities such as cine cardiac magnetic resonance (CMR) or ultrasound scans reveals the anatomical structure of the heart as well as its contracting and relaxing patterns [26]. A classical but long-standing research problem is to explore the associations between the three-dimensional (3D) cardiac anatomy and other non-imaging clinical factors, such as age, gender, diseases [5]. Besides 3D anatomical information, the temporal dynamic motion of the heart also contains information that is useful for clinical diagnosis and therapy selection [20], [32], [47]. It is of particular interest to develop computational tools that can bridge between spatial-temporal imaging features and non-imaging clinical factors. In this work, we aim to improve our understanding of the spatial-temporal cardiac anatomy and clinical factors from a generative modelling perspective. We propose a conditional generative model to model the interaction between imaging features and clinical factors. Given clinical factors as conditions, the proposed model can generate corresponding 4D spatial-temporal cardiac anatomies. We demonstrate that the generated 4D anatomies are realistic and consistent with real data distribution.

Lately, the field of conditional generative modelling has made tremendous progress, greatly driven by deep learning methods such as conditional generative adversarial networks (GAN) [34], conditional variational autoencoders (VAEs) [29], [48], flow-based models [41] and diffusion models [36]. These approaches enable efficient approximation of the underlying conditional distributions and generation of high-quality samples. Improvements in conditional generative models have been characterised by numerous developments in different generation tasks: image-to-image translation [13], [24], [27], style and lyrics-to-music generation [16] and text-to-image synthesis [12].

Apart from generating static images [36], generative models have also been applied to sequential data, such as videos [46], [53] and music [16]. In these applications, it is important to learn a model that is able to capture the inner connection of temporal sequences. To this end, long short-term memory (LSTM) [28], [52] and transformers [56] have been explored to learn the sequential progression of the latent representations of the samples. Some work also introduces spatiotemporal convolution and attention layers to learn temporal world dynamics

from a collection of videos [46]. Sequential data contain both structural variations and motion information. Disentangled representation learning approaches such as DiSCVAE [59] have been proposed to separate representations of the motion features from the structural features.

In the field of medical imaging, several papers have explored incorporating non-imaging clinical factors into the image generation process. Dalca et al. [15] proposed a learning framework for building deformable brain image templates conditioned on age. Xia et al. [54] developed a model to generate synthetic brain images conditioned on age and the status of Alzheimer's disease. For cardiac images, Biffi et al. [7] presented LVAE for interpretable classification of anatomical shapes into different clinical conditions. Krebs et al. [30] proposed to learn a probabilistic motion model for spatio-temporal cardiac image registration. Reynaud et al. [40] proposed a causal generative model to generate synthetic 3D ultrasound videos conditioned on a given input image and an expected ejection fraction. Campello et al. [9] proposed a conditional generative model in cardiac imaging to extract longitudinal patterns related to aging. Duchateau et al. [17] built a scheme for synthesizing pathological cardiac sequences from real healthy sequences. Amirrajab et al. [1] developed a framework for simulating cardiac MRI with variable anatomical and imaging characteristics. For cardiac temporal modeling scheme, some work [57], [60], [61] showed dynamic cardiac data could be described by low-dimensional latent representations, i.e. a conditional autoencoder to capture latent representations of data [61] or temporal smoothness applied as a regularisation term in the reconstruction loss function [60], [61]. These works provide useful insights for conditional medical image generation. However, the generation of a sequence of spatial-temporal cardiac anatomies from multiple clinical factors has been less explored.

In this work, we propose a conditional generative model that can generate realistic cardiac anatomical sequences conditioned on non-imaging factors including age, gender, weight, height and blood pressure. We name the Conditional Heart generation model as CHeart. The model employs a variational autoencoder to learn the latent representations for cardiac anatomies and a condition encoder to embed the clinical conditions into a condition latent vector. Then, a *Temporal Module* is designed to generate the condition-related sequential latent space based on the anatomy latent representations and the condition latent vector. The proposed model demonstrates a high diversity and fidelity in the generation, evaluated using structural overlaps and surface distance metrics, as well as clinical measure (ventricular volume and mass) distributions. The main contributions in this work are summarised as follows:

- We propose a spatial-temporal generative model for 3D cardiac anatomy that accounts for both the spatial variations and the temporal variations i.e. motion during the cardiac cycle.
- We leverage both imaging data and non-imaging clinical data to train the model, which allows the model to generate cardiac anatomical sequences conditioned on multiple clinical factors.

- We introduce a temporal module into the latent space of cardiac anatomy and conditions to model the complex sequential patterns of a beating heart.
- We demonstrate that the model can generate highly realistic and diverse cardiac anatomical sequences that follow the real data distributions.

II. METHODS

The proposed generative model takes non-imaging clinical factors as input and generates a cardiac anatomical sequence. Fig. 1 illustrates the overall framework. The following sections provide more technical details. First, we introduce the conditional generative model. Then, we describe the temporal module for learning the sequential latent representations due to cardiac motion. Lastly, we demonstrate two applications of the generative model at the inference stage: *anatomical sequence completion* and *anatomical sequence generation*.

A. Conditional generative model

Assume that we observe a dynamic sequence of anatomies of a subject, $x_t(t = 0, 1, \cdots, T - 1)$, where x_t denotes the anatomical segmentation at the t-th frame and T denotes the total number of time frames in a sequence. We also observe some clinical conditions c for this subject, where c could include factors such as age, gender, weight, height, blood pressure etc. Our aim is to learn the probability distribution of the anatomy x conditioned on c with a chosen model, $p_\theta(x|c)$, where θ denotes the model parameters. We seek a model $p_\theta(x|c)$ which is sufficiently flexible to be able to describe the data x. Deep neural networks have often been used for this modelling due to its complex modelling capacity [21], [29], [48]. Without losing generality, we first attempt to learn the distribution of anatomy at the first time frame, $p_\theta(x_0|c)$, which is often the end-diastolic (ED) frame in cardiac imaging.

We adopt the conditional β-VAE model [21], [29], [48] to learn the data distribution. The condition c is embedded as a condition latent vector z_c by the MLP, which integrates multiple clinical factors and enables exploration across the conditional latent space. The model consists of a decoder $p_\theta(x_0|z_0, z_c)$ and an encoder $q_\phi(z_0|x_0, z_c)$. The decoder $p_\theta(x_0|z_0, z_c)$ with parameters θ maps the latent variables z_0, z_c to the anatomy x_0. We assume a prior distribution $p(z_0)$ over the latent variable z_0. The prior and the decoder together define a joint distribution, denoted as $p_\theta(x_0, z_0|z_c)$, which is parameterized by θ.

To turn the intractable posterior inference and learning problem into a tractable problem, we introduce a parametric encoder model $q_\phi(z_0|x_0, z_c)$ with ϕ as the variational parameters, which approximates the true but intractable posterior distribution $p_\theta(z_0|x_0, z_c)$ of the generative model, given an input x_0 and condition space z_c:

$$q_\phi(z_0|x_0, z_c) \approx p_\theta(z_0|x_0, z_c) \tag{1}$$

where $q_\phi(z_0|x_0, z_c)$ often adopts a simpler form, e.g. the Gaussian distribution. By introducing the approximate posterior $q_\phi(z_0|x_0, z_c)$, the log-likelihood of $p_\theta(x_0|z_c)$ can be
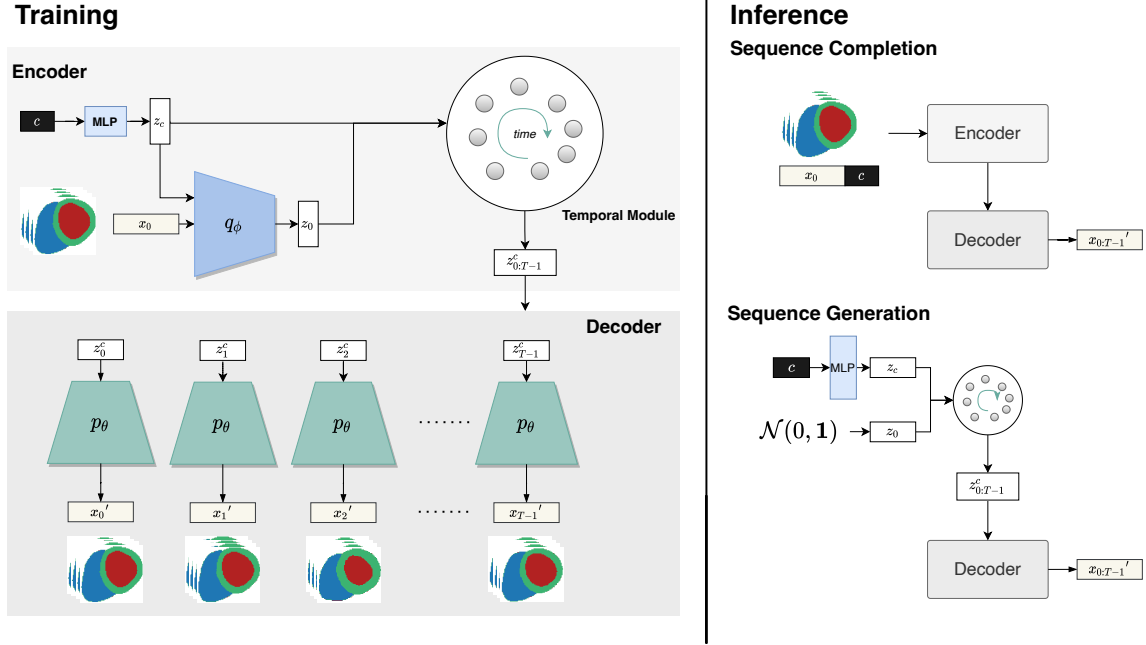
Fig. 1. Overview of the CHeart model including training and inference stages. During training, an encoder is applied to learn the latent representations z_c, z_0 for the clinical conditions c and anatomy at the first time frame x_0. A temporal module models the trajectory of $z^c_{0:T-1}$ in the latent space across the temporal dimension from the initial latent vectors z_c and z_0. The decoder then generates the 4D cardiac anatomy sequence $x_{0:T-1}$ from the latent vectors on the temporal trajectory. The training process enables two inference mechanisms at test time: sequence completion and sequence generation. In sequence completion, the model is given x_0 and c, and generates the remaining sequence of anatomies in the cardiac cycle. In sequence generation, a random latent code z_0 sampled from the prior distribution and c are given to the model and the temporal module to generate the latent vector sequence $z^c_{0:T-1}$, which are used to generate synthetic cardiac anatomical sequence $x'_{0:T-1}$.

formulated as:

$$
\begin{aligned}
\log p_\theta(x_0|z_c) &= \mathbb{E}_{z_0 \sim q_\phi(z_0|x_0,z_c)} \log\left[p_\theta(x_0|z_c)\right] \\
&= \mathbb{E}_{z_0 \sim q_\phi(z_0|x_0,z_c)} \log\left[\frac{p_\theta(x_0, z_0|z_c)}{q_\phi(z_0|x_0,z_c)}\right] \\
&\quad + \mathbb{E}_{z_0 \sim q_\phi(z_0|x_0,z_c)} \log\left[\frac{q_\phi(z_0|x_0,z_c)}{p_\theta(x_0|z_0,z_c)}\right]
\end{aligned} \tag{2}
$$

where the second term denotes the Kullback-Leibler (KL) divergence $D_{KL}(q_\phi \parallel p_\theta)$, between $q_\phi(z_0|x_0,z_c)$ and $p_\theta(z_0|x_0,z_c)$. It is non-negative and zero only if the approximate posterior $q_\phi(z_0|x_0,z_c)$ equals the true posterior distribution $p_\theta(z_0|x_0,z_c)$. Due to the non-negativity of the KL divergence, the first term in Eq. 2 is the lower bound of the evidence $\log[p_\theta(x_0|z_c)]$, known as the evidence lower bound (ELBO). Instead of optimising the evidence $\log[p_\theta(x_0|z_c)]$ which is often intractable, we optimise the ELBO:

$$
\max_{\theta,\phi} ELBO = \log[p_\theta(x_0|z_c)] - D_{KL} \tag{3}
$$

To better control the encoding representation capacity and encourage more efficient latent encoding, we adopt β-VAE by modifying VAE with an adjustable hyperparameter β [21]. As a result, the loss function of the generative model is formulated as:

$$
\begin{aligned}
\mathcal{L}_{\theta,\phi} = &- \mathbb{E}_{z_0 \sim q_\phi(z_0|x_0)} \log[p_\theta(x_0|z_0,c)] \\
&+ \beta \cdot D_{KL}[q_\phi(z_0|x_0,c) \parallel p_\theta(z_0)]
\end{aligned} \tag{4}
$$

where the sign is negated so as we can minimise the loss function.

In practice, we use the reconstruction loss for the first term., i.e. how accurate the generative model $p_\theta(x_0)$ can be

for reconstructing the anatomy x_0 from the latent vector z_0 using the decoder. The reparameterization trick is applied to replace the subscript of the expectation and express the random variable $z_0 \sim q_\phi(z_0|x_0,z_c)$ as some differentiable and invertible transformation of another random variable ϵ, so the expectation does not rely on q itself.

B. Motion modelling in the latent space
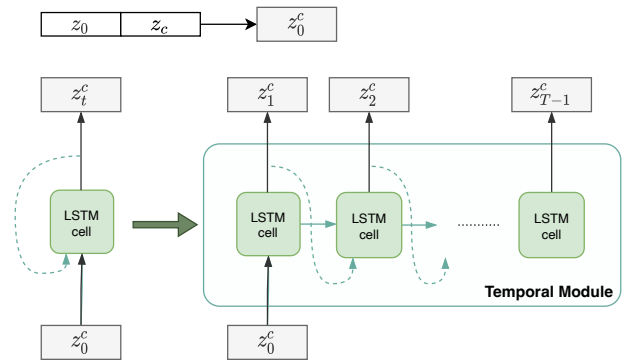
Fig. 2. The temporal module for generating the sequential latent codes $z_{0:T-1}$, constructed with a one-to-many long short-term memory (LSTM) structure.

In the previous section, we modelled $q_\phi(z_0|x_0,z_c)$ and $p_\theta(x_0|z_0,z_c)$ for the first frame x_0 in a sequence. Here, to model the whole anatomical sequence $x_0, x_1, ..., x_{T-1}$ on the clinical conditions c, we design a *Temporal Module* constructed using a one-to-many LSTM structure [49] with

parameters ω, which generates the condition-related sequential latent codes based on z_0 and z_c. The detailed structure of the temporal module is illustrated in Fig. 2.

LSTM [22] is a variant of recurrent neural networks that consists of gating mechanisms and cell memory blocks. The first LSTM cell of the module takes the concatenation of the anatomy latent representation z_0 and the condition latent representation z_c as input, which is denoted as z_0^c. With the hidden state h_0 and cell state $cell_0$ being initialised to zero, it infers the latent z_1^c at the next time frame. Each following LSTM cell, with shared weights, takes z_{t-1}^c as input, updates the hidden state h_t and cell state $cell_t$, and infers the latent z_t^c. All the LSTM cells have shared weights. Each latent code z_t^c contains information of both the anatomy at time t and the clinical conditions c. The cardiac anatomy of a dynamic sequence forms a temporal sequence z_t^c in the latent space, where $t = 0, 1, \ldots, T$. After the temporal module computes the latent codes $z_{0:T-1}^c$ across all the time frames, the decoder generates the anatomical sequence x_t' from z_t^c, illustrated in Fig. 1.

The overall loss function for modelling the anatomical sequence generation is formulated based on Eq.4:

$$\mathcal{L}_{\theta,\phi,\omega} = -\sum_{t=0}^{T-1} \mathbb{E}_{z_0 \sim q_\phi(z_0|x_0)}(\log(p_\theta(x_t|z_t, z_c))) \quad (5)$$
$$+ \beta D_{KL}(q_\phi(z_0|x_0, z_c) \parallel p_\theta(z_0))$$

The training loss function is composed of two parts: 1) the reconstruction accuracy at all time frames, where we use cross-entropy for evaluating the performance in reconstructing the segmentation maps; 2) the KL divergence term, penalising the discrepancy between the learned prior and posterior distributions. The whole training process is performed end-to-end, with the encoder, temporal module and decoder being trained together. The VAE enables the model to learn a low-dimensional latent space that captures the underlying anatomical variations. By incorporating the temporal module, the model can effectively model the temporal dynamics in the cardiac images, enabling the generation of anatomically consistent and coherent sequences over time.

C. Inference

To demonstrate the performance of the proposed generative model at the inference stage, we carry out two benchmark tasks, namely anatomical sequence completion and anatomical sequence generation, as shown in the right panel of Fig. 1.

In *anatomical sequence completion*, the model is given the anatomy at the first time frame x_0 and clinical conditions c. It is asked to generate the remaining sequence of anatomies across the cardiac cycle. The model maps x_0 and c to their latent representations z_0 and z_c, predicts the sequential latent codes $z_{0:T-1}^c$ through the temporal module and finally reconstructs the full sequence of cardiac anatomy $x_{0:T-1}'$ using the shared-weight decoders.

In *anatomical sequence generation*, the model is only conditioned on the clinical factors c and it does not require any anatomy as input. Since the model has learnt the distribution of anatomical latent variable p_{z_0}, we can draw samples z_0

in the latent space from a Gaussian distribution $\mathcal{N}(0, 1)$ and concatenate it with the clinical latent code z_c. We then provide the concatenated latent code z_0^c to the temporal module to predict $z_{0:T-1}^c$ and generate the full anatomical sequence $x_{0:T-1}'$ using the decoder.

D. Evaluation

To evaluate the conditional generative model, we use quantitative measures to assess the generated anatomy, as well as use clinical measures to assess the distribution similarity.

First, we employ the Dice coefficient, the Hausdorff distance (HD) and the average symmetric surface distance (ASSD) which compare the similarity of the generated cardiac anatomy to the ground truth anatomy associated with the same clinical conditions.

Second, we derive five imaging phenotypes including the left ventricular myocardial mass (LVM), LV end-diastolic volume (LVEDV), LV end-systolic volume (LVESV), right ventricular end-diastolic volume (RVEDV) and RV end-systolic volume (RVESV). We evaluate differences between generated data and real data with the same clinical conditions, denoted as $d_{\text{phenotype}}$. Furthermore, these phenotypes are closely associated with age and gender [5]. We calculate the distributions of the imaging phenotypes against age and gender, and compare the generated data to the real data. The comparison is illustrated qualitatively using density plots and quantitatively using the Kullback–Leibler (KL) divergence and Wasserstein distance (WD). The KL divergence [14] is an information-theoretic measurement of the similarity between two probability mass functions. Similarly, WD [2] measures the distance between two probability distributions and can be computed as:

$$\text{WD} = \inf_{\gamma \sim \prod(P,Q)} \mathbb{E}_{(u,v) \sim \gamma}[\|u - v\|] \quad (6)$$

where $\prod(P, Q)$ is the set of all joint distributions over u and v. WD can be seen as the minimum work needed to transform one distribution to another, where work is defined as the amount of mass that must be moved from u to v to transform P to Q and the distance to be moved.

III. EXPERIMENTS

A. Data sets

A short-axis 3D cardiac MR dataset of 1,383 subjects was used, acquired from Hammersmith Hospital, Imperial College London. Each cardiac cine image sequence comprises 20 time frames ($T = 20$) covering one complete cardiac cycle, with a spatial resolution of $1.25~mm \times 1.25~mm \times 2~mm$. The temporal resolution ranges from 0.041 to 0.048 seconds per frame, accommodating variations in the heart rate. The cardiac anatomy is described by the image segmentation map with four labels: background, the left ventricle (LV), myocardium (Myo) and the right ventricle (RV). Ground truth segmentation at end-diastolic (ED) and end-systolic (ES) frames was generated by using a multi-atlas segmentation method [3], then quality controlled and manually corrected by an experienced cardiologist using itkSNAP [58]. A state-of-the-art nnU-net model [23] was trained using the ED and ES segmentation and then deployed

to all time frames generating the 3D-t segmentation, followed by manual quality control. To eliminate the influence of image orientations in the generation, all 3D-t segmentation were rigidly aligned to a template space using MIRTK [42], [44] and cropped to a standard size of $128 \times 128 \times 64$. In this way, the generative model will focus on learning subject-specific variations of the anatomy instead of image orientations.

In terms of demographic information, all subjects were healthy volunteers, with 775 females and 608 males, aged between 18-73 years old, weighed between 33-131 kg, with height between 142-195 cm and systolic blood pressure (SBP) between 79-183 mm Hg. When incorporating the clinical information into the model, age was represented as a categorical factor with seven age groups with an interval of 10 years, from 10 to 80 years old. The dataset was randomly split into three subsets for training ($n = 968$), validation ($n = 138$) and test ($n = 277$).

B. Experimental setup

1) Implementation: The model was implemented in PyTorch [37]. The encoder q_ϕ consisted of four 3D convolution layers, one flatten layer and one bottleneck layer, outputting the latent code z_0. The condition mapping network was constructed using an MLP, outputting latent code z_c for input conditions c. A latent dimension of 32 was used for both z_0 and z_c, and 64 for the concatenated latent vector z_0^c. The decoder consisted of one flatten layer and four 3D transposed convolution layers. All convolution and transposed convolution layers in the encoder and the decoder used a kernel size of 4. The temporal module was built with one-layer LSTMCells. The regularisation weight β in β-VAE was set to 0.001. The model was trained using the Adam optimiser with a learning rate of $5 \cdot 10^{-4}$ and a batch size of 8. It was trained for 500 epochs and an early stopping criterion was used based on the validation set performance. The training took 17 hours on an NVIDIA RTX A6000 GPU.

2) Baseline methods: Currently, there is no other existing work for performing conditional generation of 3D-t cardiac anatomies. For comparison, we implemented the following baseline generation methods developed in other application domains, extending them from 2D image generation to 3D-t data generation:

CGAN: A conditional version of the generative adversarial network (GAN) originally developed for MNIST images [34]. Note that the model can only perform cardiac sequence generation, not sequence completion.

CVAE: The conditional generative model CVAE [48]. It was modified to adapt to this application. CVAE applied condition incorporation by concatenating conditions and anatomies in both the encoder and decoder.

CVAE-GAN: A conditional variational generative adversarial network proposed in [6]. It is a general learning framework that combines a VAE with a GAN for synthesizing natural images in fine-grained categories.

PCA: The principal component analysis (PCA) [25]. It is a classical method for dimensionality reduction, which aims to preserve as much of the variation in data as possible using

the principal components. Note that the PCA is only used for performing sequence completion, but not for sequence generation.

C. Sequence completion

	Dice (unit: 1)			
	LV	Myo	RV	Average
CVAE-GAN [6]	$0.845^*_{\pm 0.028}$	$0.697^*_{\pm 0.054}$	$0.832^*_{\pm 0.028}$	$0.791^*_{\pm 0.032}$
CVAE [48]	$0.900_{\pm 0.023}$	$0.800^*_{\pm 0.040}$	$0.894_{\pm 0.023}$	$0.864^*_{\pm 0.026}$
PCA [25]	$0.906_{\pm 0.022}$	$0.810_{\pm 0.038}$	$0.901_{\pm 0.023}$	$0.872_{\pm 0.025}$
Proposed	$\mathbf{0.908}_{\pm 0.023}$	$\mathbf{0.814}_{\pm 0.037}$	$\mathbf{0.902}_{\pm 0.021}$	$\mathbf{0.874}_{\pm 0.024}$
	HD (unit: mm)			
	LV	Myo	RV	Average
CVAE-GAN [6]	$10.361^*_{\pm 1.475}$	$9.571^*_{\pm 1.379}$	$14.070^*_{\pm 3.736}$	$11.334^*_{\pm 1.849}$
CVAE [48]	$5.920^*_{\pm 1.335}$	$5.891^*_{\pm 1.055}$	$6.525_{\pm 1.076}$	$6.112^*_{\pm 1.049}$
PCA [25]	$\mathbf{5.517}_{\pm 1.029}$	$5.710_{\pm 1.125}$	$\mathbf{6.165}_{\pm 1.072}$	$\mathbf{5.797}_{\pm 0.978}$
Proposed	$5.535_{\pm 1.180}$	$\mathbf{5.576}_{\pm 0.955}$	$6.445_{\pm 1.067}$	$5.842_{\pm 1.017}$
	ASSD (unit: mm)			
	LV	Myo	RV	Average
CVAE-GAN [6]	$2.120^*_{\pm 0.390}$	$1.670^*_{\pm 0.236}$	$2.244^*_{\pm 0.399}$	$1.983^*_{\pm 0.306}$
CVAE [48]	$1.657^*_{\pm 0.348}$	$1.376^*_{\pm 0.212}$	$1.622_{\pm 0.305}$	$1.461_{\pm 0.280}$
PCA [25]	$1.565_{\pm 0.324}$	$1.319^*_{\pm 0.221}$	$\mathbf{1.519}_{\pm 0.301}$	$1.490_{\pm 0.305}$
Proposed	$\mathbf{1.535}_{\pm 0.330}$	$\mathbf{1.298}_{\pm 0.208}$	$1.620_{\pm 0.323}$	$\mathbf{1.462}_{\pm 0.266}$

A well-known challenge to generative modelling is the difficulty in evaluation, as we normally do not have access to the ground truth data distribution, e.g. the distribution of all possible cardiac anatomies in our case. Therefore, we adopt anatomical sequence completion as a surrogate task for evaluating the model performance. The sequence completion experiments were conducted to assess the ability of capturing the sequential information given the first frame of a cardiac anatomy sequence. One example of sequence completion is shown in Fig. 3. It can be seen in the figure that the generated anatomies across time frames maintain the same heart structures as the ED frame and capture the temporal motion pattern through time, contracting first and then expanding.

The sequence completion accuracy is evaluated between the generated anatomy and ground truth across the whole sequence in terms of the Dice metric, HD and ASSD for three structures: LV, Myo and RV. Table I reports the sequence completion accuracy of the proposed model and compares it to other generative models including CVAE-GAN [6], CVAE [48] and PCA [25]. It shows that the proposed model achieves a good sequence completion accuracy with an average Dice metric of 0.874, HD of 5.842 mm and ASSD of 1.462 mm, which is comparable to or outperforms the other three generative models in most metrics. In addition, we conducted evaluations at the basal, mid-cavity, and apical slices. The proposed model
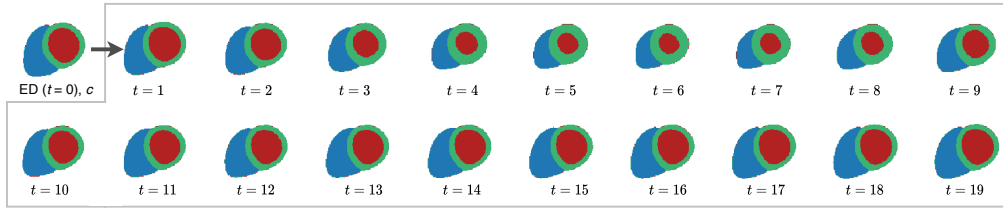
Fig. 3. An example of sequence completion, arranged in two rows with the left-to-right and top-to-bottom order. With the end-diastolic (ED) frame in time $t = 0$ and conditions c as input, the model generates the remaining anatomical sequence at time frame $t = 1-19$, shown within the gray box. The top row depicts anatomy images at time frame $t = 0 - 9$, and the bottom row depicts at time frame $t = 10 - 19$.

TABLE II

COMPARISON OF SEQUENCE GENERATION PERFORMANCE BETWEEN CGAN, CVAE, CVAE-GAN AND THE PROPOSED MODEL, IN TERMS OF MEAN AND BEST DICE METRIC AND CONTOUR DISTANCE METRICS FOR THE AVERAGE PERFORMANCE OVER LV, RV AND MYO. THE BEST VALUE ACROSS 20 SAMPLES FOR DICE METRIC (MAXIMUM), HD (MINIMUM) AND ASSD (MINIMUM) ARE REPORTED. ASTERISKS INDICATE STATISTICAL SIGNIFICANCE (*: $P \leq 0.05$) WHEN USING A PAIRED STUDENT'S t-TEST COMPARING THE PERFORMANCE OF THE PROPOSED METHOD TO OTHER METHODS.

Model	Dice (unit: 1)		HD (unit: mm)		ASSD (unit: mm)	
	mean	best/max	mean	best/min	mean	best/min
CGAN [34]	$\mathbf{0.713}_{\pm 0.061}$	$0.717^*_{\pm 0.061}$	$15.533^*_{\pm 2.258}$	$13.956^*_{\pm 2.326}$	$\mathbf{3.004}_{\pm 0.714}$	$2.862^*_{\pm 0.712}$
CVAE [48]	$0.694_{\pm 0.056}$	$0.789_{\pm 0.049}$	$11.461^*_{\pm 1.809}$	$8.321_{\pm 1.536}$	$3.380^*_{\pm 0.710}$	$2.317^*_{\pm 0.540}$
CVAE-GAN [6]	$0.645^*_{\pm 0.052}$	$0.774_{\pm 0.039}$	$16.844^*_{\pm 2.008}$	$12.105^*_{\pm 1.815}$	$3.693^*_{\pm 0.709}$	$2.185_{\pm 0.394}$
Proposed	$\mathbf{0.713}_{\pm 0.058}$	$\mathbf{0.793}_{\pm 0.052}$	$\mathbf{10.940}_{\pm 2.343}$	$\mathbf{8.166}_{\pm 1.621}$	$3.023_{\pm 0.757}$	$\mathbf{2.049}_{\pm 0.521}$

TABLE III

COMPARISON OF SEQUENCE GENERATION PERFORMANCE AMONG CGAN, CVAE, CVAE-GAN AND THE PROPOSED MODEL. THE CLINICAL MEASURES DERIVED FROM EACH REAL SAMPLE ARE COMPARED TO THOSE DERIVED FROM 20 SYNTHETIC SAMPLES OF EXACTLY THE SAME CONDITIONS. THE MEAN AND THE MINIMAL DIFFERENCES OF THE CLINICAL MEASURES ARE REPORTED HERE.

Model	d_{LVEDV} (mL)		d_{LVESV} (mL)		d_{RVEDV} (mL)		d_{RVESV} (mL)		d_{LVM} (g)	
	mean	best/min	mean	best/min	mean	best/min	mean	best/min	mean	best/min
CGAN [34]	$35.58_{\pm 20.33}$	$15.66_{\pm 16.67}$	$20.06_{\pm 9.71}$	$19.74_{\pm 9.72}$	$51.47_{\pm 25.25}$	$14.71_{\pm 17.12}$	$17.57_{\pm 12.19}$	$17.04_{\pm 12.18}$	$38.26_{\pm 19.15}$	$10.40_{\pm 11.23}$
CVAE [48]	$35.74_{\pm 16.99}$	$\mathbf{4.91}_{\pm 9.84}$	$13.92_{\pm 6.06}$	$1.87_{\pm 3.46}$	$44.97_{\pm 21.58}$	$\mathbf{6.46}_{\pm 12.92}$	$19.49_{\pm 9.21}$	$2.86_{\pm 5.74}$	$23.07_{\pm 9.96}$	$\mathbf{2.70}_{\pm 4.33}$
CVAE-GAN [6]	$51.32_{\pm 20.40}$	$6.33_{\pm 11.96}$	$19.80_{\pm 6.53}$	$\mathbf{1.69}_{\pm 2.57}$	$48.94_{\pm 28.66}$	$8.28_{\pm 17.52}$	$25.26_{\pm 10.99}$	$\mathbf{2.57}_{\pm 4.11}$	$51.03_{\pm 11.40}$	$8.29_{\pm 7.91}$
Proposed	$\mathbf{25.93}_{\pm 17.47}$	$6.87_{\pm 12.09}$	$\mathbf{11.74}_{\pm 8.41}$	$3.54_{\pm 6.25}$	$\mathbf{34.63}_{\pm 21.31}$	$6.88_{\pm 12.87}$	$\mathbf{15.54}_{\pm 11.33}$	$5.12_{\pm 9.19}$	$\mathbf{17.34}_{\pm 9.89}$	$2.95_{\pm 5.62}$

achieved an average Dice metric of 0.929, 0.927, and 0.878 for LV at the three locations, surpassing the corresponding metrics of the other three generative models.

We also performed paired student's t-tests between the results generated by our method and those of competing methods. The performance metrics of the proposed model marked with asterisk in Table I were significantly better than other methods at a p value smaller than 0.05. On a different cardiac MR dataset, [4] reports an average Dice metric of 0.94, 0.88, 0.90 for LV, myocardium and RV, respectively, for inter-observer variability in manual cardiac image segmentation (Table 3 of [4]). The Dice metric of the proposed generative model is close to this value, which indicates its high performance and capability for anatomical sequence completion.

D. Sequence generation

Apart from the sequence completion task, we also perform anatomical sequence generation and evaluate how close the generated anatomical sequences are to the real data. In this experiment, we generate new synthetic anatomies of the heart by providing the clinical conditions as the only input to the model. Given the stochastic nature of the VAE generation, for each set of input conditions, multiple anatomical sequences can be generated. We draw 20 random samples from the Gaussian distribution of the latent vector, and correspondingly generate 20 synthetic anatomical sequences for this input condition set.

We first compare the synthetic anatomies to the real anatomy with the same clinical conditions and evaluate the mean similarity and the best similarity across 20 samples, in terms of the Dice metric, HD, ASSD and differences of clinical measures. This is similar to the random average or random best evaluation in other recent generation works in computer vision [38]. Table II shows that the proposed model achieves a reasonably good sequence generation accuracy with a mean Dice metric of 0.713, HD of 10.940 mm and ASSD of 3.023 mm. We also reported the best value of each measurement, with a significantly improved maximum Dice of 0.793, minimum HD of 8.166 mm, and ASSD of 2.049 mm. This perhaps means the proposed method can capture a wide variation of anatomies and thus draw a sample that is close to the real sample. When we compare the differences of clinical phenotypes, Table III shows that our model achieved the lower measurement difference with a mean difference of 25.93 mL,
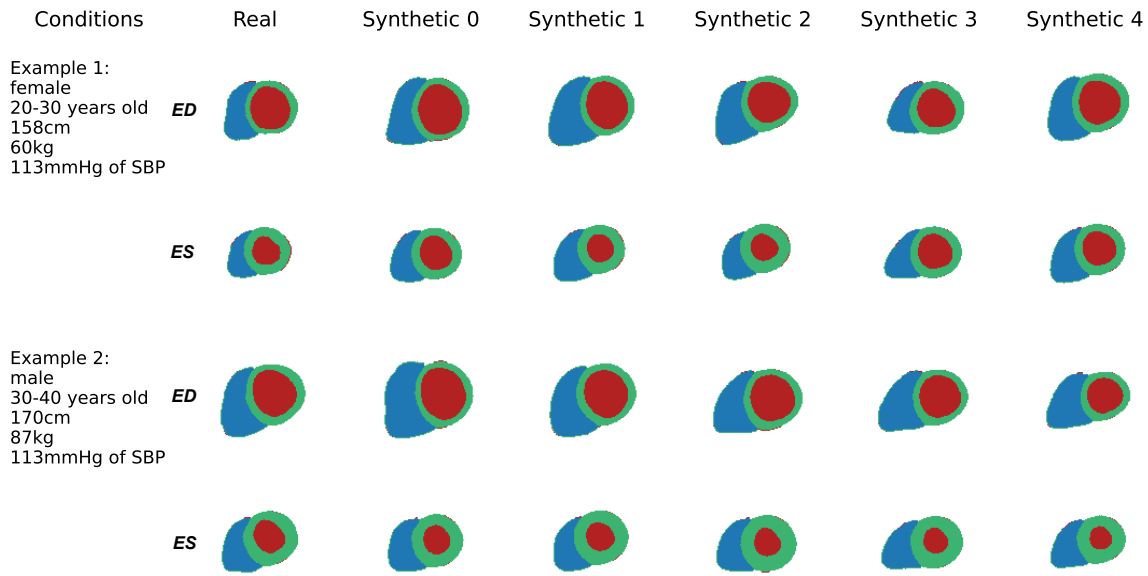
Fig. 4. Visualisation of synthetic anatomies (last five columns) generated by the model, compared to the real anatomy (first column) with the same clinical conditions (text annotation). The whole anatomical sequence is generated but only ED and ES frames are shown here. The first and second rows of each example show the ED and ES frames of the cardiac anatomical sequence.

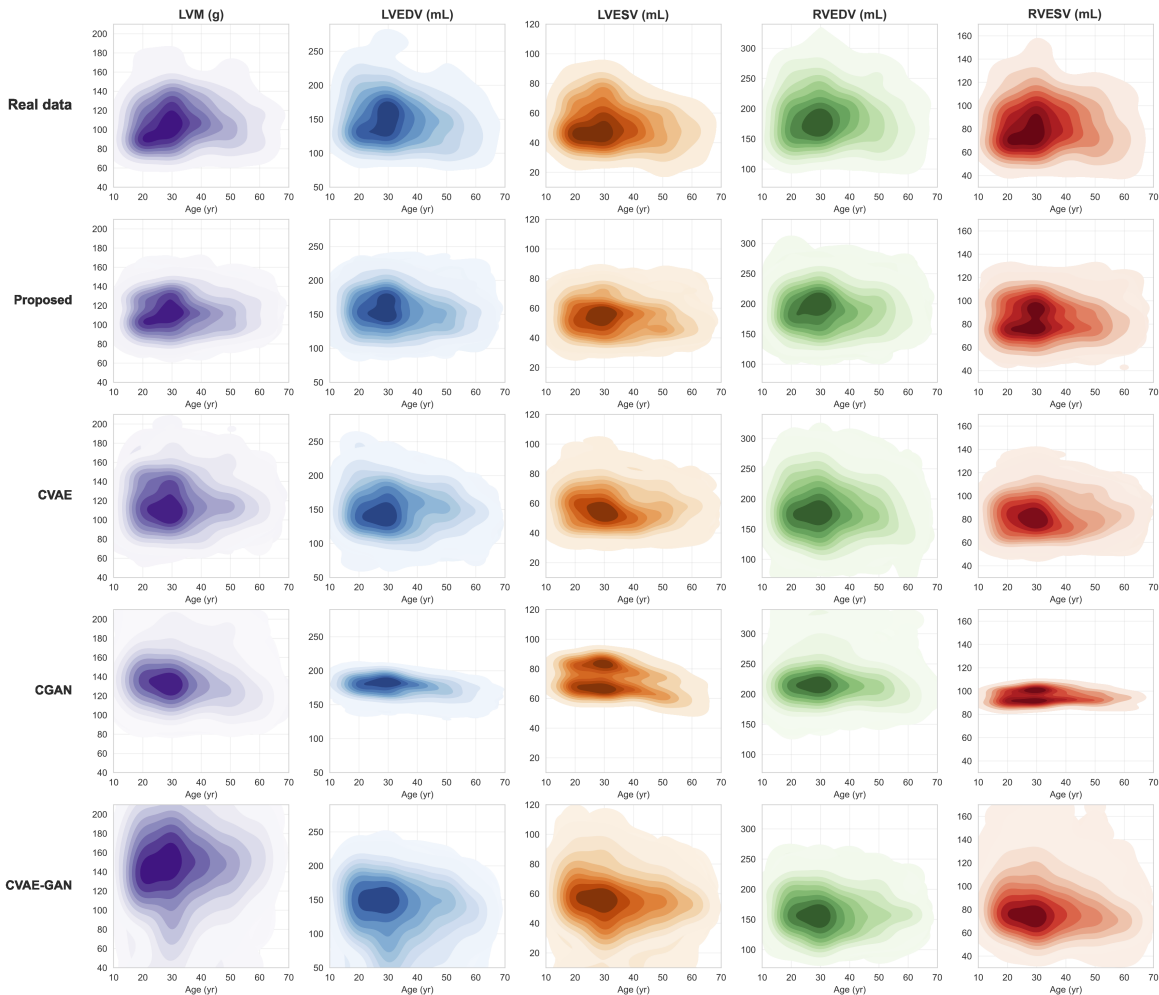Fig. 5. Distributions of clinical measures for real data and synthetic data. Each graph displays a kernel density plot of an imaging phenotype (LVM, LVEDV, LVESV, RVEDV, RVESV) against age. For each plot, the x-axis denotes age and the y-axis denotes the value of the imaging phenotype. Darker areas in the plot indicate the regions where the data is more concentrated. Lighter areas show the regions where the data is sparser.

TABLE IV

KL DIVERGENCE AND WASSERSTEIN DISTANCE BETWEEN SYNTHETIC DATA DISTRIBUTION AND REAL DATA DISTRIBUTION.

Distribution similarity	Kullback–Leibler (KL) divergence					Wasserstein Distance (WD)				
	LVEDV	LVESV	RVEDV	RVESV	LVM	LVEDV	LVESV	RVEDV	RVESV	LVM
CGAN [34]	$\mathbf{0.023}_{\pm 0.001}$	$\mathbf{0.019}_{\pm 0.001}$	$0.036_{\pm 0.004}$	$\mathbf{0.022}_{\pm 0.001}$	$0.050_{\pm 0.006}$	$33.687_{\pm 1.173}$	$19.982_{\pm 0.025}$	$41.643_{\pm 4.161}$	$17.434_{\pm 0.036}$	$35.395_{\pm 2.933}$
CVAE [48]	$0.039_{\pm 0.005}$	$0.041_{\pm 0.004}$	$0.042_{\pm 0.004}$	$0.034_{\pm 0.003}$	$0.030_{\pm 0.003}$	$\mathbf{11.929}_{\pm 2.116}$	$7.017_{\pm 0.964}$	$14.680_{\pm 2.869}$	$9.665_{\pm 1.051}$	$10.365_{\pm 1.703}$
CVAE-GAN [6]	$0.153_{\pm 0.025}$	$0.023_{\pm 0.003}$	$0.046_{\pm 0.006}$	$0.064_{\pm 0.008}$	$0.098_{\pm 0.019}$	$27.001_{\pm 2.809}$	$8.425_{\pm 1.771}$	$24.614_{\pm 3.202}$	$9.748_{\pm 2.675}$	$43.251_{\pm 4.566}$
Proposed	$0.034_{\pm 0.002}$	$0.043_{\pm 0.002}$	$\mathbf{0.034}_{\pm 0.002}$	$0.039_{\pm 0.002}$	$\mathbf{0.031}_{\pm 0.002}$	$15.053_{\pm 3.597}$	$\mathbf{5.773}_{\pm 1.358}$	$\mathbf{12.214}_{\pm 2.408}$	$\mathbf{9.182}_{\pm 2.145}$	$\mathbf{9.215}_{\pm 1.713}$

11.74 mL, 34.63 mL, 15.54 mL and 17.34 g and minimum difference of 6.87 mL, 3.54 mL, 6.88 mL, 5.12 mL and 2.95 g for LVEDV, LVESV, RVEDV, RVESV and LVM, respectively. The results of mean and best values indicate that our model achieves similar (Dice) or better sequence generation accuracy (HD, ASSD, difference in clinical measures) compared to other methods. The best values of the metrics indicate the high fidelity of the proposed generative model, which refers to the degree to which the generated samples resemble the real ones [35], [43]. It is important to acknowledge that in anatomical sequence generation, the model is not expected to replicate existing anatomies. But instead, the model generates a plausible anatomy that fulfils certain conditions, which is compared to a real anatomy with the same conditions.

Further, we visualised two examples of anatomical sequence generation in Fig. 4. For each example, we show five random synthetic samples which share the same clinical conditions as the real sample. It illustrates that the LV and RV structures look realistic and their shapes share a high similarity to the real anatomy. The contracting pattern of the ventricles and myocardium from ED to ES frame also looks realistic and similar to the real sample. This demonstrates our model can capture the overall anatomy and temporal dynamics of the heart during generation. The five samples with the same conditions also present certain degrees of variations, which demonstrates the diversity of synthetic data. This is due to the Gaussian sampling part of the generation process and reflects the individual differences between two hearts even if they are of the same gender and age, which can be caused by genetic, environmental, lifestyle and many other factors that are not easily accounted for by the model.

To further evaluate whether fidelity and diversity of the generated samples with respect to the real samples, we assess the distance between their distributions, conditioned on age, a common factor of interest in clinical research. In addition to quantitative assessments, we conducted qualitative comparisons by evaluating the distributions of five clinical measures for both real and synthetic anatomies against age, including LVM, LVEDV, LVEV, RVEDV, and RVEF, illustrated in Fig. 5. Compared to other methods, the synthetic data distributions from our model closely resemble the real distributions and cover the full variability of the real samples. Table IV reports the KL divergence and Wasserstein distance between synthetic and real data distributions. The proposed model achieves the best KL or WD metrics in most clinical measurements, with KL divergence values of 0.034, 0.043, 0.034, 0.039, 0.031,

and WD values of 15.053, 5.773, 12.214, 9.182, 9.215 for LVEDV, LVESV, RVEDV, RVESV, and LVM, respectively. These results demonstrate that the synthetic data generated by our model maintains a distribution against age that is similar to the real data.
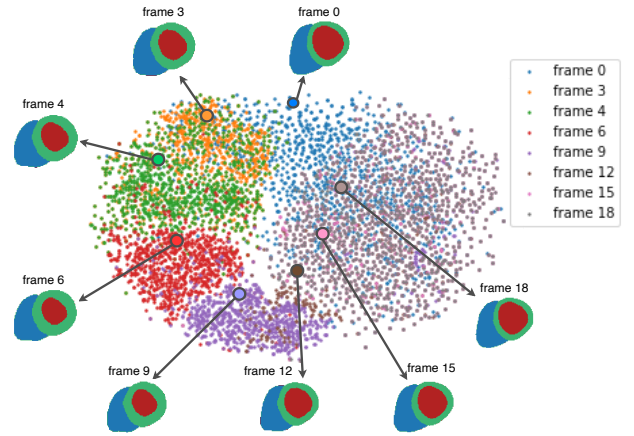
E. Temporal dynamics

Fig. 6. T-distributed stochastic neighbor embedding (t-SNE) visualization of latent space for generated anatomical sequences from frame 0 to frame 18. Each dot represents a single time frame of a sample, with colors indicating the frame index. A sequence of anatomies, decoded from a corresponding sequence of latent codes, belonging to one subject, is visualised in the figure.

The proposed model encoded the anatomical and clinical information z_0^c of the first frame (ED) and generated the latent vectors z_t^c for the following frames by the temporal module. We use the dimensionality reduction technique, t-distributed stochastic neighbor embedding (t-SNE) [31], to visualise the latent space z_t^c of the generated anatomical sequences, as shown in Fig. 6. The sequential latent codes $z_{0:T-1}^c$ start at ED ($t = 0$) and move along a cyclic path in the latent space. It shows that the generative model can capture the temporal dynamics of the anatomy during the heartbeat and form a cyclic pattern as a real heart [45]. More overlapped areas between frames 9 to 18 show that the variation of anatomies is smaller in the relaxation stage, which demonstrates the nonlinear trajectories of cardiac motion. We plotted one example of the anatomical sequence at time frame $0, 3, 4, 6, 9, 12, 15, 18$ in the figure. Through the time frames, the anatomies present first decreased and then increased LV volumes. The thickness of Myo has the opposite trend, which

is consistent with the contraction and relaxation pattern of the heart [19].

F. Condition manipulation

With the conditional generative model, we are able to simulate the change of anatomy when certain conditions (e.g. age) change. Fig. 7(a) shows a series of generated anatomies during ageing, when the condition age increases but all the other conditions as well as the latent vectors drawn from the Gaussian distribution are fixed. The difference map comparing the aged anatomy to the anatomy at 10-20 years old shows subtle changes to the LV and RV structures. We further generate 200 random samples of the synthetic ageing anatomies and derive the clinical measures. Fig. 7(b) illustrates the longitudinal evolution of these measures, stratified by gender. We observe a longitudinally increasing trend in LVM during ageing and a decreasing trend in LVEDV, consistent with findings in clinical literature [18] (Figure 3 of [18]). It demonstrates the potential of using this model for simulating anatomical data distributions. However, we need to be cautious in interpreting this result, as our training data is cross-sectional instead of longitudinal and also the mechanism of cardiac ageing is complex, confounded by more factors (genetics, lifestyle etc) than the five conditions we used in this work.

IV. Discussion

The proposed model is built upon a β-VAE for learning the latent space of the cardiac anatomy. It integrates a conditional branch to model the influence of multiple clinical factors on the generation process and uses a temporal module to model the temporal relationship of anatomical latent vectors during cardiac motion. The experiments demonstrate good performance in both anatomical sequence completion and sequence generation tasks, qualitatively and quantitatively. The model enables condition manipulation for demonstrating the impact of clinical factors on anatomical shape variation. When using the common clinical measures (ventricular volumes and mass) for evaluation, the distribution of generated anatomies is close to the real data distribution visually (Fig. 5) and quantitatively (Table IV), which indicate both the fidelity and diversity of the generation. While the model performs well in generating anatomically coherent structures, further improvement can be made in terms of achieving a closer similarity between the distribution of generated anatomies and real data distribution. There is also potential for further exploration of the relationship between cardiac motion and clinical conditions

We foresee there are several potential downstream tasks for the generative cardiac anatomy model, including discovering patterns in large datasets, facilitating out-of-distribution detection and generating synthetic data etc. First, by training a generative model on a large dataset of cardiac anatomies, the trained model can capture complex patterns and variations of the anatomy associated with different clinical factors. This knowledge can be valuable for understanding population-level characteristics, identifying risk factors and informing public health strategies. Second, by learning the distribution

of normal cardiac anatomy and dynamics, the proposed model can identify patterns of a given anatomy that deviate from the norm, indicating potential anomalies that require further investigation. More importantly, the proposed method is a conditional generative model, which means it can learn the norm specifically for certain conditions (e.g. a gender and age group) and evaluate the deviation from the norm in a personalised manner. Third, the trained generative model can provide a large amount of synthetic data for other tasks. Synthetic data can be used for performing data augmentation for training machine learning models [8], creating synthetic fair data to improve the fairness of prediction models [11], [50], or used as digital anatomies for performing in-silico trials [55]. Diverse and realistic synthetic data will alleviate the data scarcity issue in the medical field, where real data are often limited or not easy to share. This includes the creation of synthetic data for privacy-preserving research [39], [51].

There are a few limitations of this work. The first limitation is the high computational cost during training to learn the spatio-temporal patterns from 4D data, even after cropping the images to $128 \times 128 \times 64$ and using sequences of only 20 time frames. An interesting future direction is to reduce the computational complexity of high-dimensional and high-resolution medical imaging data. Second, here we use a segmentation map as a representation of the anatomy so that the generative model can focus on learning the variations of anatomy, instead of intensity image styles. Future explorations could be extended to the generation of intensity images for the heart [1] or using mesh as a representation for the anatomy [33], which may be computationally more efficient. Third, we use a cross-sectional imaging dataset of mainly healthy volunteers for training the generative model, due to the challenge of curating large-scale longitudinal datasets with high spatial resolution. It would be interesting to extend this to longitudinal and clinical imaging cohorts with cardiac diseases in the future.

V. Conclusion

In this work, we propose a novel conditional generative model that is able to synthesise spatial-temporal cardiac anatomies given clinical factors as input. It demonstrates the feasibility of generating highly realistic synthetic 3D-t anatomies for the heart that captures both the anatomical variations and motion of the heart. The work paves the way for further generative modelling research in cardiac imaging, such as incorporating disease types or representing anatomy as meshes. It also has the potential to be applied to downstream tasks, such as performing data augmentation based on various anatomies, building condition-specific atlases and performing biomechanical modelling of the heart etc.

References

[1] S. Amirrajab, Y. Al Khalil, C. Lorenz, J. Weese, J. Pluim, and M. Breeuwer. A framework for simulating cardiac MR images with varying anatomy and contrast. *IEEE Transactions on Medical Imaging*, 2022.

[2] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223. PMLR, 2017.
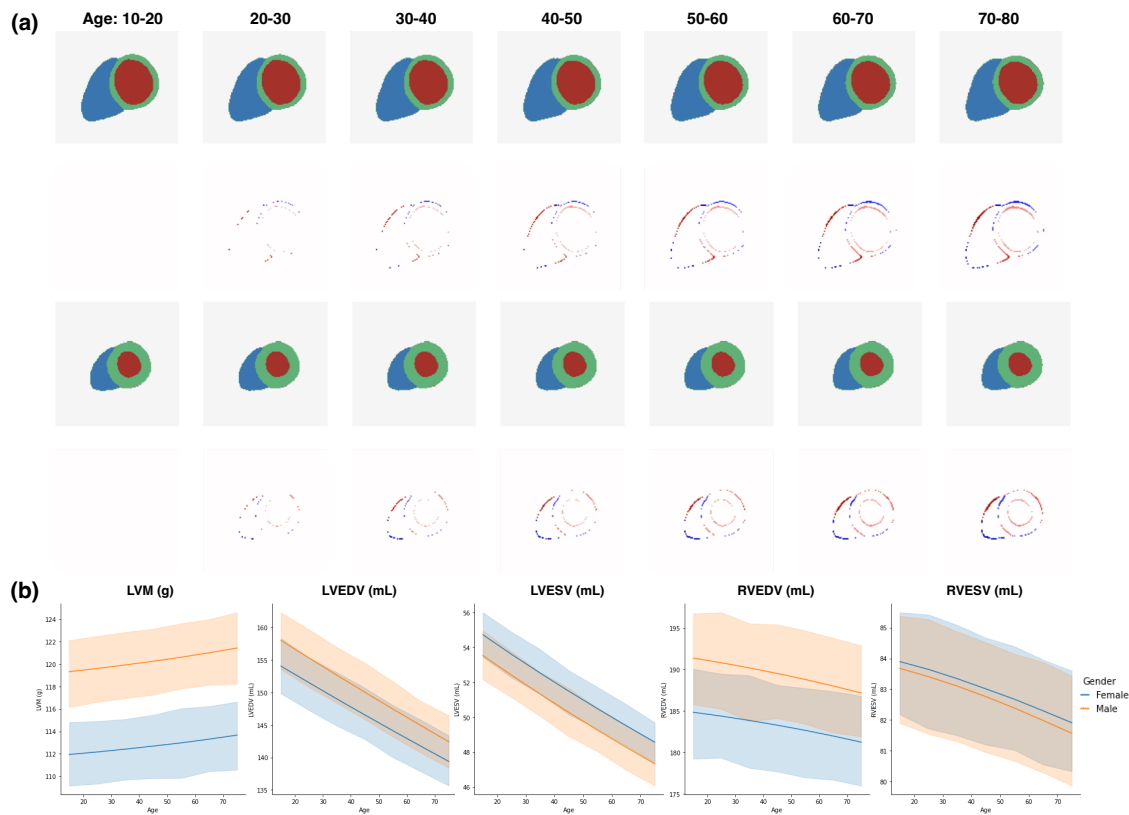
Fig. 7.   (a) An example of the synthetic cardiac anatomy during ageing. The first and third rows show the cardiac anatomies at end-diastolic (ED) and end-systolic (ES) frames. The second and fourth rows show the difference maps between the aged anatomy 20-80 years old and the anatomy at 10-20 years old. (b) The simulated evolution of clinical measures (LVM, LVEDV, LVESV, RVEDV, RVESV) by generating 200 samples of gender-specific ageing cardiac anatomy and plotting their mean measures with 95% confidence interval.

[3] W. Bai, W. Shi, D. P. O'regan, T. Tong, H. Wang, S. Jamil-Copley, N. S. Peters, and D. Rueckert.   A probabilistic patch-based label fusion model for multi-atlas segmentation with registration refinement: application to cardiac MR images.   *IEEE Transactions on Medical Imaging*, 32(7):1302–1315, 2013.

[4] W. Bai, M. Sinclair, G. Tarroni, O. Oktay, M. Rajchl, G. Vaillant, A. M. Lee, N. Aung, E. Lukaschuk, M. M. Sanghvi, et al. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. *Journal of Cardiovascular Magnetic Resonance*, 20(1):65, 2018.

[5] W. Bai, H. Suzuki, J. Huang, C. Francis, S. Wang, G. Tarroni, F. Guitton, N. Aung, K. Fung, S. E. Petersen, et al. A population-based phenome-wide association study of cardiac and aortic structure and function. *Nature Medicine*, 26(10):1654–1662, 2020.

[6] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua.   CVAE-GAN: fine-grained image generation through asymmetric training. In *International Conference on Computer Vision*, pages 2745–2754, 2017.

[7] C. Biffi, J. J. Cerrolaza, G. Tarroni, W. Bai, A. De Marvao, O. Oktay, C. Ledig, L. Le Folgoc, K. Kamnitsas, G. Doumou, et al. Explainable anatomical shape analysis through deep hierarchical generative models. *IEEE Transactions on Medical Imaging*, 39(6):2088–2099, 2020.

[8] B. Billot, D. N. Greve, O. Puonti, A. Thielscher, K. Van Leemput, B. Fischl, A. V. Dalca, J. E. Iglesias, et al. Synthseg: Segmentation of brain mri scans of any contrast and resolution without retraining. *Medical image analysis*, 86:102789, 2023.

[9] V. M. Campello, T. Xia, X. Liu, P. Sanchez, C. Martín-Isla, S. E. Petersen, S. Seguí, S. A. Tsaftaris, and K. Lekadir. Cardiac aging synthesis from cross-sectional data with conditional generative adversarial networks. *Frontiers in Cardiovascular Medicine*, 9, 2022.

[10] N. Cardim, M. Galderisi, T. Edvardsen, S. Plein, B. A. Popescu, A. d'Andrea, O. Bruder, B. Cosyns, L. Davin, E. Donal, et al. Role of multimodality cardiac imaging in the management of patients with hypertrophic cardiomyopathy: an expert consensus of the European Association of Cardiovascular Imaging Endorsed by the Saudi Heart Association.   *European Heart Journal - Cardiovascular Imaging*, 16(3):280–280, 2015.

[11] R. J. Chen, M. Y. Lu, T. Y. Chen, D. F. Williamson, and F. Mahmood. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, 5(6):493–497, 2021.

[12] Z. Chen, V. G. Kim, M. Fisher, N. Aigerman, H. Zhang, and S. Chaudhuri. Decor-gan: 3d shape detailization by conditional refinement. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 15740–15749, 2021.

[13] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8789–8797, 2018.

[14] T. M. Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[15] A. V. Dalca, M. Rakic, J. Guttag, and M. R. Sabuncu.   Learning conditional deformable templates with convolutional networks.   In *Neural Information Processing Systems*, pages 806–818, 2019.

[16] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*, 2020.

[17] N. Duchateau, M. Sermesant, H. Delingette, and N. Ayache. Model-based generation of large databases of cardiac images: Synthesis of pathological cine mr sequences from real healthy cases. *IEEE Transactions on Medical Imaging*, 37(3):755–766, 2018.

[18] J. Eng, R. L. McClelland, A. S. Gomes, W. G. Hundley, S. Cheng, C. O. Wu, J. J. Carr, S. Shea, D. A. Bluemke, and J. A. Lima. Adverse left ventricular remodeling and age assessed with cardiac MR imaging: the multi-ethnic study of atherosclerosis. *Radiology*, 278(3):714–722, 2016.

[19] H. Fukuta and W. C. Little. The cardiac cycle and the physiologic basis of left ventricular contraction, ejection, relaxation, and filling. *Heart Failure Clinics*, 4(1):1–11, 2008.

[20] K. Gilbert, C. Mauger, A. A. Young, and A. Suinesiaputra. Artificial intelligence in cardiac imaging with statistical atlases of cardiac anatomy. *Frontiers in Cardiovascular Medicine*, 7, 2020.

[21] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick,

S. Mohamed, and A. Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

[22] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[23] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.

[24] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017.

[25] I. T. Jolliffe. *Principal component analysis for special types of data*. Springer, 2002.

[26] T. D. Karamitsos, J. M. Francis, S. Myerson, J. B. Selvanayagam, and S. Neubauer. The role of cardiovascular magnetic resonance imaging in heart failure. *Journal of the American College of Cardiology*, 54(15):1407–1424, 2009.

[27] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[28] A. Khamparia, B. Pandey, S. Tiwari, D. Gupta, A. Khanna, and J. J. P. C. Rodrigues. An Integrated Hybrid CNN–RNN Model for Visual Description and Generation of Captions. *Circuits, Systems, and Signal Processing*, 39(2):776–788, 2020.

[29] D. P. Kingma and M. Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

[30] J. Krebs, H. Delingette, N. Ayache, and T. Mansi. Learning a generative motion model from image sequences based on a latent motion matrix. *IEEE Transactions on Medical Imaging*, 40(5):1405–1416, 2021.

[31] L. v. d. Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[32] C. A. Mauger, S. Govil, R. Chabiniok, K. Gilbert, S. Hegde, T. Hussain, A. D. McCulloch, C. J. Occleshaw, J. Omens, J. C. m. v. o. Perry, K. Pushparajah, A. Suinesiaputra, L. Zhong, and A. A. Young. Right-left ventricular shape variations in tetralogy of fallot: associations with pulmonary regurgitation. *Journal of Cardiovascular Magnetic Resonance*, 23(1):1–14, 2021.

[33] Q. Meng, W. Bai, T. Liu, D. P. O'Regan, and D. Rueckert. Mesh-based 3D motion tracking in cardiac MRI using deep learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2022.

[34] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[35] M. F. Naeem, S. J. Oh, Y. Uh, Y. Choi, and J. Yoo. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning*, volume 119, pages 7176–7185. PMLR, 13–18 Jul 2020.

[36] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, 2022.

[37] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Neural Information Processing Systems*, 32:8026–8037, 2019.

[38] M. Petrovich, M. J. Black, and G. Varol. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision*, 2022.

[39] Z. Qian, T. Callender, B. Cebere, S. M. Janes, N. Navani, and M. van der Schaar. Synthetic data for privacy-preserving clinical risk prediction. *medRxiv*, pages 2023–05, 2023.

[40] H. Reynaud, A. Vlontzos, M. Dombrowski, C. Lee, A. Beqiri, P. Leeson, and B. Kainz. D'artagnan: Counterfactual video generation. In *Medical Image Computing and Computer Assisted Intervention*, pages 599–609, 2022.

[41] D. Rezende and S. Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning*, pages 1530–1538. PMLR, 2015.

[42] D. Rueckert, L. I. Sonoda, E. R. Denton, S. Rankin, C. Hayes, M. O. Leach, D. L. Hill, and D. J. Hawkes. Comparison and evaluation of rigid and nonrigid registration of breast mr images. In *Medical Imaging 1999: Image Processing*, volume 3661, pages 78–88. International Society for Optics and Photonics, 1999.

[43] M. S. Sajjadi, O. Bachem, M. Lucic, O. Bousquet, and S. Gelly. Assessing generative models via precision and recall. *Advances in Neural Information Processing Systems*, 31, 2018.

[44] A. Schuh, A. Makropoulos, E. C. Robinson, L. Cordero-Grande, E. Hughes, J. Hutter, A. N. Price, M. Murgasova, R. P. A. Teixeira, N. Tusor, et al. Unbiased construction of a temporally consistent morphological atlas of neonatal brain development. *bioRxiv*, page 251512, 2018.

[45] A. D. Scott, J. Keegan, and D. N. Firmin. Motion in cardiovascular MR imaging. *Radiology*, 250(2):331–351, 2009.

[46] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

[47] O. A. Smiseth, H. Torp, A. Opdahl, K. H. Haugaa, and S. Urheim. Myocardial strain imaging: how useful is it in clinical decision making? *European Heart Journal*, 37(15):1196–1207, 2016.

[48] K. Sohn, H. Lee, and X. Yan. Learning structured output representation using deep conditional generative models. *Neural Information Processing Systems*, 28:3483–3491, 2015.

[49] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Neural Information Processing Systems*, pages 3104–3112, 2014.

[50] B. van Breugel, T. Kyono, J. Berrevoets, and M. van der Schaar. Decaf: Generating fair synthetic data using causally-aware generative networks. *Advances in Neural Information Processing Systems*, 34:22221–22233, 2021.

[51] B. van Breugel and M. van der Schaar. Beyond privacy: Navigating the opportunities and challenges of synthetic data. *arXiv preprint arXiv:2304.03722*, 2023.

[52] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2015.

[53] J. Walker, A. Razavi, and A. v. d. Oord. Predicting video with vqvae. *arXiv preprint arXiv:2103.01950*, 2021.

[54] T. Xia, A. Chartsias, C. Wang, S. A. Tsaftaris, A. D. N. Initiative, et al. Learning to synthesise the ageing brain without longitudinal data. *Medical Image Analysis*, 73:102169, 2021.

[55] Y. Xia, N. Ravikumar, T. Lassila, and A. F. Frangi. Virtual high-resolution mr angiography from non-angiographic multi-contrast mris: synthetic vascular model populations for in-silico trials. *Medical Image Analysis*, 87:102814, 2023.

[56] W. Yan, Y. Zhang, P. Abbeel, and A. Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021.

[57] J. Yoo, K. H. Jin, H. Gupta, J. Yerly, M. Stuber, and M. Unser. Time-dependent deep image prior for dynamic mri. *IEEE Transactions on Medical Imaging*, 40(12):3337–3348, 2021.

[58] P. A. Yushkevich, J. Piven, H. C. Hazlett, R. G. Smith, S. Ho, J. C. Gee, and G. Gerig. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *Neuroimage*, 31(3):1116–1128, 2006.

[59] M. Zolotas and Y. Demiris. Disentangled sequence clustering for human intention inference. In *International Conference on Intelligent Robots and Systems*, 2022.

[60] Q. Zou, A. H. Ahmed, P. Nagpal, S. Priya, R. F. Schulte, and M. Jacob. Variational manifold learning from incomplete data: application to multi-slice dynamic mri. *IEEE transactions on medical imaging*, 41(12):3552–3561, 2022.

[61] Q. Zou, S. Priya, P. Nagpal, and M. Jacob. Joint cardiac t 1 mapping and cardiac cine using manifold modeling. *Bioengineering*, 10(3):345, 2023.