CardioDiT: Latent Diffusion Transformers for 4D Cardiac MRI Synthesis

Marvin Seyfarth[1,2,3*], Sarah Kaye Müller[1,2,3], Arman Ghanaat[1], Isabelle Ayx[4], Fabian Fastenrath[5], Philipp Wild[6], Alexander Hertel[4], Theano Papavassiliu[3,7], Salman Ul Hassan Dar[1,2,3], and Sandy Engelhardt[1,2,3]

[1] Institute for Artificial Intelligence in Cardiovascular Medicine, Medical Faculty of Heidelberg University, Heidelberg University, Germany
[2] Department of Cardiology, Angiology, Pneumology, Heidelberg University Hospital, Heidelberg, Germany
[3] DZHK (German Centre for Cardiovascular Research), Partner Site Heidelberg/Mannheim, Heidelberg, Germany
[4] Department of Radiology and Nuclear Medicine, University Medical Center Mannheim, Germany
[5] Section for Invasive Cardiology and Electrophysiology, I. Medical Department, Cardiology, Hemostasiology and Intensive Care, University Medical Centre Mannheim, Mannheim, Germany
[6] University Medical Center of the Johannes Gutenberg-University Mainz, Germany
[7] Department of Cardiology, Angiology, Hemostasis, and Medical Intensive Care, University Medical Centre Mannheim, Medical Faculty Mannheim, University of Heidelberg, Germany
[*]Corresponding author: Marvin.Seyfarth@med.uni-heidelberg.de

Abstract. Latent diffusion models (LDMs) have recently achieved strong performance in 3D medical image synthesis. However, modalities like cine cardiac MRI (CMR), representing a temporally synchronized 3D volume across the cardiac cycle, add an additional dimension that most generative approaches do not model directly. Instead, they factorize space and time or enforce temporal consistency through auxiliary mechanisms such as anatomical masks. Such strategies introduce structural biases that may limit global context integration and lead to subtle spatiotemporal discontinuities or physiologically inconsistent cardiac dynamics. We investigate whether a unified 4D generative model can learn continuous cardiac dynamics without architectural factorization. We propose CardioDiT, a 3D+t latent diffusion framework for short-axis cine CMR synthesis based on diffusion transformers. A spatiotemporal VQ-VAE encodes 2D+t slices into compact latents, which a diffusion transformer then models jointly as complete 3D+t volumes, coupling space and time throughout the generative process. We evaluate CardioDiT on public CMR datasets and a larger private cohort, comparing it to baselines with progressively stronger spatiotemporal coupling. Results show improved inter-slice consistency, temporally coherent motion, and realistic cardiac function distributions, suggesting that explicit 4D modeling with a diffusion transformer provides a principled foundation for spatiotemporal cardiac image synthesis. Code, configurations and models trained on public data are available at https://github.com/Cardio-AI/cardiodit.

Keywords: Diffusion transformers · CMR synthesis · 4D synthesis

1 Introduction

Generative models, particularly latent diffusion models (LDMs), have recently shown strong performance in synthesizing high-quality 3D medical images [8,5,19], enabling applications such as data augmentation or privacy-preserving data sharing where annotated datasets are scarce. Some clinically relevant modalities, like short-axis cardiac MRI (CMR), include a temporal component, meaning both spatial anatomy and temporal dynamics are critical for capturing cardiac function. Most existing generative approaches do not model the full 3D+t distribution directly. Instead, they factorize space and time, e.g., generating 2D slices with temporal modules [10], modeling selected cardiac phases such as end-diastole and end-systole [18], or enforcing temporal consistency via auxiliary components [1]. Preliminary work [15] took a first step toward single-pass $3D{+}t$ CMR synthesis using the same slice-wise spatiotemporal latent encoding adopted here. In that formulation, the latent representations of the $2D{+}t$ slices were concatenated along the channel dimension and jointly denoised by a 3D U-Net operating over the in-plane spatial and temporal dimensions. While this approach enables joint volume generation, it merges through-plane depth into the channel dimension rather than retaining it as an explicit modeling axis. In this work, we investigate whether explicitly representing the complete latent cine volume as four-dimensional spatiotemporal patches and modeling it with global self-attention provides an effective alternative to factorized and channel-merged formulations. To this end, we introduce **CardioDiT**, a latent diffusion transformer for joint $3D{+}t$ CMR synthesis. CardioDiT first encodes 2D+t slices into a compact latent representation using a vector-quantized VAE, and then models the full 3D+t volume in latent space with a diffusion transformer operating on the complete 4D tensor [13]. Global self-attention thereby enables direct long-range interactions across slices and cardiac phases without requiring auxiliary temporal-consistency constraints. We compare CardioDiT with models exhibiting progressively stronger forms of spatiotemporal coupling: (i) a factorized $2D{+}t$ LDM conditioned on slice position and (ii) the channel-merged $3D{+}t$ U-Net [15]. Our contributions are threefold: (1) a latent diffusion transformer that models complete cine CMR volumes; (2) a systematic comparison with factorized $2D{+}t$ and channel-merged $3D{+}t$ U-Net formulations, characterizing their respective fidelity–consistency trade-offs; and (3) a comprehensive evaluation of distributional fidelity, through-plane anatomy, temporal dynamics, and cardiac function across public and private CMR cohorts.

2 Methodology

CardioDiT extends our concurrently developed 3D Diffusion Transformer formulation [17] from static volumetric imaging to dynamic cine CMR. A spatiotemporal VQ-GAN first encodes individual $2D{+}t$ slices, whose latent representations

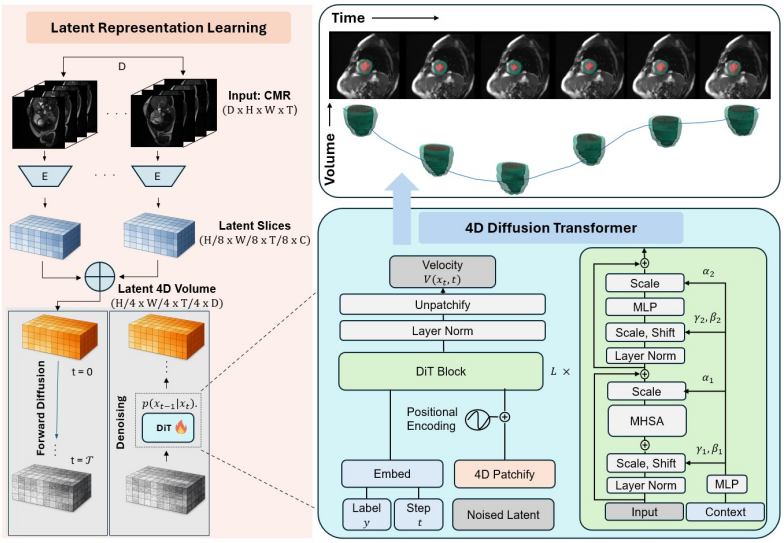

Fig. 1. CardioDiT framework for cine CMR synthesis. A spatiotemporal VQ-GAN encodes $2D{+}t$ slices, whose latents are stacked along depth into a 4D latent volume. A diffusion transformer denoises the noisy latent using spatiotemporal patches and global self-attention. The denoised latent is split along depth, decoded slice-wise, and concatenated to reconstruct the full cine CMR volume.

are stacked along the depth axis to form a complete latent $3D{+}t$ volume. This volume is then partitioned into patches along depth, height, width, and time and jointly modeled by a diffusion transformer using 4D positional encodings and global self-attention. Throughout this work, "4D" refers specifically to the diffusion model operating jointly on the complete latent $3D{+}t$ cine volume. The autoencoder itself operates slice-wise on $2D{+}t$ inputs, reflecting the anisotropic through-plane resolution of short-axis CMR. We focus on unconditional CMR synthesis in this study (Fig. 1).

2.1 Spatiotemporal Latent Autoencoder

We first train a spatiotemporal VQ-GAN to learn a compact representation of $2D{+}t$ cine slices. Let $\mathbf{v} \in \mathbb{R}^{c\times d\times h\times w\times t}$ denote a cine CMR volume, where c, d, h, w, and t correspond to image channels, depth, height, width, and temporal frames, respectively. The VQ-GAN is implemented using 3D convolutions over the (h, w, t) dimensions. During training, a depth position s is randomly sampled from each volume, and the VQ-GAN learns to reconstruct the corresponding $2D{+}t$ slice. After training, all slices of a volume are encoded independently using the encoder. For a slice $\mathbf{v}_s \in \mathbb{R}^{c\times h\times w\times t}$, the encoder produces a latent representation $\mathbf{z}_s \in \mathbb{R}^{c_z\times h_z\times w_z\times t_z}$, where c_z is the latent channel dimension and $(h_z, w_z, t_z) = (h/f, w/f, t/f)$ denote the downsampled latent dimensions. The slice latents are then stacked along the depth axis to form $\mathbf{z} \in \mathbb{R}^{c_z\times d\times h_z\times w_z\times t_z}$.

The resulting tensor provides the complete latent $3D{+}t$ representation modeled by CardioDiT.

2.2 4D Diffusion Transformer

Next, we train a diffusion transformer to model the 4D latent CMR representation. The latent volumes are partitioned into 4D patches of size $(p_d, p_h, p_w, p_t) = (1, 4, 4, 2)$ and linearly projected into token embeddings. Given a latent volume $\mathbf{z} \in \mathbb{R}^{c_z \times d \times h_z \times w_z \times t_z}$, where c_z denotes the latent channel dimension, the patchification and embedding process produces tokens $\mathbf{x} \in \mathbb{R}^{N \times e}$, where N is the number of 4D patches and e is the token embedding dimension. The tokens are augmented with 4D sine–cosine positional encodings to preserve their depth, height, width, and temporal positions. The complete token sequence is then processed jointly using global self-attention, enabling direct interactions across spatial locations, slices, and temporal frames. To enable memory-efficient training with long token sequences, we employ FlashAttention [4]. After denoising, the predicted 4D latent volume is split along the depth axis into $2D{+}t$ latent slices. Each slice is decoded by the VQ-GAN decoder, and the decoded slices are concatenated along the depth axis to reconstruct the full $3D{+}t$ CMR volume.

2.3 Datasets

Public: Cine short-axis CMR from Automated Cardiac Diagnosis Challenge (ACDC) [2] and Multi-Centre, Multi-View, Multi-Vendor & Multi-Disease Cardiac Image Segmentation Challenge (M&Ms-2) [3,11] were used. For ACDC, 100 cases were used for training and 50 for validation; for M&Ms-2, 289 for training and 58 for validation. All volumes were resampled to the median spatial resolution, center-cropped, or padded to $6 \times 256 \times 256$ (d, h, w) while preserving cardiac context. The temporal dimension was standardized to 32 frames via cyclic repetition. The training cases were merged into a single public training pool, and evaluation was performed jointly on the pooled public validation data.
Private: A private CMR dataset was collected at University Clinic Mannheim (*ethic no. 2020-808R*) using three different scanners. 813 samples were used for training and 102 for validation. All volumes were resampled to the median spatial resolution of corresponding scanners, cropped or padded to a fixed resolution of $10 \times 224 \times 224 \times 32$, matching the preprocessing applied to the public datasets.

2.4 Experimental Setup

To compare formulations with progressively more joint spatiotemporal processing, we evaluate CardioDiT against two representative latent diffusion baselines: **Factorized** $2D{+}t$ **LDM:** A U-Net-based $2D{+}t$ latent diffusion model conditioned on slice position s, modeling each short-axis slice independently as $p(X) = \prod_{s=1}^{D} p(x_{s,1:T} \mid s)$, where $x_{s,1:T}$ denotes the temporal sequence at slice position s.

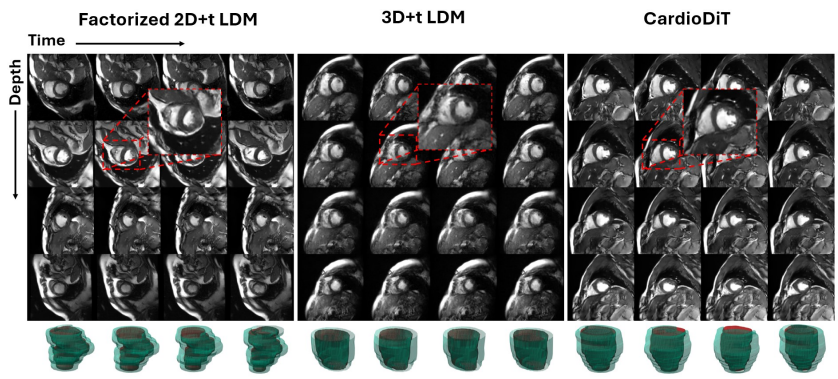

Fig. 2. Slices across time and depth for one sample generated by each model trained on the public data, together with the corresponding LV meshes. The factorized $2D+t$ LDM exhibits inter-slice discontinuities arising from independent slice synthesis, whereas the $3D+t$ U-Net produces smoother but blurrier anatomical boundaries. CardioDiT preserves both through-plane anatomical consistency and fine image detail.

Channel-merged $3D+t$ LDM: Following [15], the latent representations of all slices are concatenated along the channel dimension and jointly denoised by a 3D U-Net operating over the (h, w, t) dimensions. Thus, all slices are processed in one network pass, but depth is represented implicitly through channels rather than as an explicit geometric axis. Both U-Net architectures follow the implementation of [16].

CardioDiT (ours): CardioDiT retains depth, height, width, and time as explicit axes of the latent representation and jointly processes the resulting spatiotemporal patches using global self-attention.

To evaluate physiological plausibility, we analyze left ventricular (LV) volume curves across time. For each generated and real sample, we compute LV cavity volumes using a segmentation model trained with the nnU-Net framework [7] on theM&Ms-2 [3,11] training cohort. To evaluate through-plane anatomical consistency, we compute the *d-SSIM*, defined as the structural similarity between consecutive slices along the short-axis (d) direction for each time point [20]. For each volume, SSIM values are aggregated across both spatial (d) and temporal (t) dimensions, and mean across all samples is reported. In addition, we quantify geometric consistency of the left ventricle at end-diastole (ED). For each slice, the center of mass of the LV segmentation is computed, and a linear regression line is fitted through these points along the through-plane direction. The mean absolute residual AR_{ED} from this fitted axis is then reported as a measure of inter-slice alignment. Analogous to d-SSIM, performance is assessed relative to the real reference distribution, where values closer to the reference indicate better anatomical alignment. For distributional similarity, we compute the Fréchet Inception Distance (FID) using the pretrained CMR foundation model *CineMA* [6]. CineMA features are extracted for all temporal frames, after which FID, preci-

sion, and recall are computed [9]. This provides denser feature-space sampling than one feature vector per volume. The neighborhood parameter k for precision and recall is selected such that a real–real comparison yields scores greater than 0.95, ensuring a stable and meaningful operating point. All comparisons are performed against the complete set of real validation samples for $N = 50$ synthetic samples. To assess temporal motion modeling, one complete cardiac cycle was extracted using key frame detection [12]. We report the mean and standard deviation of the ejection fraction (EF), defined as $\mathrm{EF} = \frac{V_{\mathrm{ED}} - V_{\mathrm{ES}}}{V_{\mathrm{ED}}}$, where V_{ED} and V_{ES} denote the end-diastolic and end-systolic volumes, respectively, and quantify distributional differences in cardiac function by computing the Wasserstein distance (W_2) between the EF distributions of synthetic and real validation samples [14]. Additionally, we analyze the scalability of CardioDiT by evaluating three CardioDiT variants with increasing numbers of DiT blocks $(S, B, L) = (8, 12, 16)$ on the private dataset. The autoencoder was trained for 500 epochs with compression factor f along the encoded (h, w, t) dimensions, codebook size $c' = 4096$ and embedding dimension $emb = 8$. CardioDiT was trained for 600,000 iterations using a cosine noise schedule with $T_{\mathrm{diff}} = 300$ timesteps, AdamW optimizer and a fixed learning rate of $LR = 10e-4$. CardioDiT can run on 24GB VRAM.

3 Results

3.1 Visual Assessment of 4D Consistency

We qualitatively assess spatial fidelity and temporal coherence across models. Fig. 2 shows a representative CMR volume per model spanning the cardiac cycle. The factorized 2D+t LDM captures temporally plausible contraction and

Table 1. Quantitative comparison of CMR generation models using distributional (FID, Precision, Recall), structural (d-SSIM, $\mathrm{AR}_{\mathrm{ED}}$) and clinical EF distribution statistics. Lower is better ↓, higher is better ↑, closest to real is better ∼.

Model	FID ↓	Precision ↑	Recall ↑	d-SSIM ∼	$\mathrm{AR}_{\mathrm{ED}}$ ∼	$\overline{EF}$ (Std) ∼	W_2 ↓
Public Dataset							
Real	52.8	0.95	0.96	0.55 (0.10)	2.1 (1.0)	47.6 (17.1)	3.09
Fact. 2D+t LDM	100.1	0.89	0.17	0.39 (0.04)	5.1 (1.1)	45.3 (8.0)*	8.90
3D+t U-Net	221.5	0.59	0.64	0.75 (0.10)	1.0 (0.3)	44.5 (10.0)	7.14
CardioDiT	**71.9**	**0.97**	**0.70**	**0.69 (0.10)**	**1.3 (0.5)**	**49.3 (15.1)**	**2.67**
Private Dataset							
Real	11.4	0.92	0.92	0.63 (0.12)	1.9 (0.7)	44.6 (7.3)	2.11
Fact. 2D+t LDM	32.9	0.58	0.36	0.50 (0.04)	5.0 (1.0)	41.1 (6.5)*	3.51
3D+t U-Net	51.3	0.43	0.23	0.73 (0.03)	1.1(0.3)	38.1 (5.1)	7.23
CardioDiT	**21.2**	**0.82**	**0.54**	**0.69 (0.09)**	**1.7 (0.4)**	**42.3 (6.3)**	**3.05**

* EF is a clinically meaningful volumetric measurement. Inter-slice discontinuities in the 2D+t LDM make the computed EF physiologically uninterpretable.

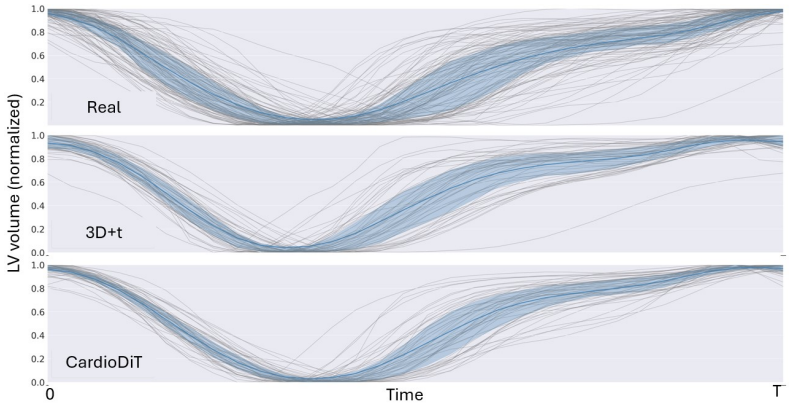

Fig. 3. Normalized left ventricular volume–time curves over one detected cardiac cycle, defined from peak to peak and resampled to a common length T. The mean curve, interquartile range (IQR), and individual subject curves are shown for the real validation data and the evaluated volumetric generative models.

relaxation patterns within individual slices. Across depth, however, there are strong inter-slice discontinuities, leading to inconsistent ventricular structures despite roughly correct global positioning of the left ventricle. The 3D+t U-Net improves volumetric consistency, producing smoother structures across slices while maintaining reasonable temporal dynamics. Nevertheless, image fidelity is reduced, with blurred anatomical boundaries and diminished high-frequency detail. In contrast, CardioDiT generates anatomically consistent ventricular structures across slices, maintains smooth and physiologically plausible motion across time, and preserves fine anatomical detail.

3.2 Quantitative Evaluation

As shown in Tab. 1 CardioDiT achieves the best FID and precision on both datasets, while substantially improving recall compared to the baselines. This indicates that our model generates high-fidelity samples and improved coverage in the CineMA feature space. CardioDiT achieves d-SSIM and AR_{ED} closest to the real images among the evaluated models, indicating superior anatomical consistency across the through-plane direction, which is consistent with the qualitative results in Fig. 2. For clinical EF statistics, CardioDiT achieves the closest match to the real EF distribution, reflected in both mean/standard deviation and the lowest W_2. While the factorized 2D+t LDM yields EF statistics within a comparable range, the computation of EF does not explicitly enforce anatomical consistency in the through-plane direction. EF is a clinically meaningful volumetric measurement, and inter-slice discontinuities in the 2D+t LDM make the computed values physiologically uninterpretable. Additionally, Fig. 3 shows the mean normalized LV volume curves across the cardiac cycle for the volumetric models. Both CardioDiT and the 3D+t LDM capture the overall temporal

trend, with mean curves and interquartile ranges closely following the real data. Differences become apparent at the individual curve level, particularly during the transition from end-systole to early/mid-diastole, where CardioDiT exhibits variability more closely aligned with the real distribution. Volume curves for the factorized $2D{+}t$ LDM are not reported because the inter-slice discontinuities observed for this baseline make its volumetric measurements physiologically unreliable. Overall, both volumetric models reproduce the principal cardiac motion pattern, while CardioDiT provides the best overall balance of fidelity, distributional alignment, through-plane consistency, and temporal variability among the evaluated methods.

Architecture Ablation Study As shown in Tab. 2 increasing model capacity consistently improves generative quality, as reflected by the monotonic decrease in FID from CardioDiT-S to CardioDiT-L. Slice consistency (d-SSIM) remains stable across scales. Notably, CardioDiT-S, despite having substantially fewer parameters, achieves better scores than all baseline methods in Tab. 1, demonstrating the efficiency of the proposed architecture. Clinical EF statistics remain comparable across model sizes, suggesting that scaling primarily enhances distributional fidelity rather than altering cardiac motion properties.

4 Discussion

In this work, we investigated whether joint 4D transformer-based modeling can be used for cine CMR synthesis that may support data augmentation and privacy-aware data sharing. By modeling the complete latent $3D{+}t$ representation with a diffusion transformer backbone, CardioDiT achieves coherent cardiac anatomy and plausible motion dynamics using a conceptually simple, globally coupled architecture. Results indicate improved through-plane consistency compared to slice-wise and channel-merged baselines, while temporal analysis shows smooth contraction–relaxation dynamics and realistic ejection fraction distributions. These findings suggest that the combined use of 4D latent representations and a transformer backbone is an effective approach for spatiotemporally coherent cine CMR synthesis. A key design principle is architectural simplicity: rather than introducing modality-specific temporal mechanisms, we rely on a unified transformer operating on 4D latent tokens. The ablation study shows that increasing model capacity improves distributional fidelity within the evaluated configurations, suggesting that scaling the joint transformer representation can improve sample quality without additional structural constraints [13]. Although this study focused on unconditional synthesis, the framework naturally supports conditional generation via adaptive layer normalization or lightweight token adapters, enabling integration of demographic, pathological, or spatial priors without modifying the core architecture [17]. Beyond cine CMR, the proposed joint 4D modeling framework may be applicable to other dynamic imaging domains, including settings with a substantially higher number of d-slices. Limitations include temporal standardization via cyclic repetition, and the limited

Table 2. Scalability of CardioDiT. Inference time T, VRAM and number of parameters were measured for a single sample on a NVIDIA RTX 4090 GPU.

Model	FID ↓	d-SSIM ∼	$\overline{EF}$ (Std) ∼	W_2 ↓	#Par (M)	T(s) ↓	VRAM(GB)↓
CardioDiT-S	30.5	0.68(0.06)	41.8 (7.7)	3.56	92.23	30	**1.2**
CardioDiT-B	25.9	0.69(0.06)	41.4 (4.7)	4.13	134.74	45	1.3
CardioDiT-L	**21.2**	**0.69(0.09)**	**42.3 (6.3)**	**3.05**	177.24	60	1.5
Fact. 2D+t	32.9	0.50(0.04)	41.1 (6.5)	3.51	166	20	1.5
3D+t	51.3	0.73(0.03)	38.1 (5.1)	7.23	578	**15**	3.8

number of evaluated samples. Future work should explore variable-length formulations, larger evaluation sizes, and comparisons with factorized modules.

Acknowledgments. This work was supported by Heidelberg Faculty of Medicine at Heidelberg University, by the Multi-DimensionAI project of the Carl Zeiss Foundation (P2022-08-010), by the EU-Horizon Project *DVPS* (101213369), by the Klaus Tschira Foundation, and by the DZHK (ref. number 81X3500151, proj. number 2098). The authors acknowledge *1-* the data storage service SDS@hd supported by the Ministry of Science, Research and the Arts Baden-Württemberg (MWK) and the German Research Foundation (DFG) through grant INST 35/1314-1 FUGG and INST 35/1503-1 FUGG, *2-* the state of Baden-Württemberg through bwHPC and the DFG through grant INST 35/1597-1 FUGG.

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

References

1. Amirrajab, S., Al Khalil, Y., Lorenz, C., Weese, J., Pluim, J., Breeuwer, M.: Label-informed cardiac magnetic resonance image synthesis through conditional generative adversarial networks. Computerized Medical Imaging and Graphics **101**, 102123 (2022). https://doi.org/https://doi.org/10.1016/j.compmedimag.2022.102123, https://www.sciencedirect.com/science/article/pii/S0895611122000933
2. Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Gonzalez Ballester, M.A., Sanroma, G., Napel, S., Petersen, S., Tziritas, G., Grinias, E., Khened, M., Kollerathu, V.A., Krishnamurthi, G., Rohé, M.M., Pennec, X., Sermesant, M., Isensee, F., Jäger, P., Maier-Hein, K.H., Full, P.M., Wolf, I., Engelhardt, S., Baumgartner, C.F., Koch, L.M., Wolterink, J.M., Išgum, I., Jang, Y., Hong, Y., Patravali, J., Jain, S., Humbert, O., Jodoin, P.M.: Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved? IEEE Transactions on Medical Imaging **37**(11), 2514–2525 (2018). https://doi.org/10.1109/TMI.2018.2837502
3. Campello, V.M., Gkontra, P., Izquierdo, C., Martín-Isla, C., Sojoudi, A., Full, P.M., Maier-Hein, K., Zhang, Y., He, Z., Ma, J., Parreño, M., Albiol, A., Kong, F., Shadden, S.C., Acero, J.C., Sundaresan, V., Saber, M., Elattar, M., Li, H., Menze, B., Khader, F., Haarburger, C., Scannell, C.M., Veta, M., Carscadden, A.,

Punithakumar, K., Liu, X., Tsaftaris, S.A., Huang, X., Yang, X., Li, L., Zhuang, X., Viladés, D., Descalzo, M.L., Guala, A., Mura, L.L., Friedrich, M.G., Garg, R., Lebel, J., Henriques, F., Karakas, M., Çavuş, E., Petersen, S.E., Escalera, S., Seguí, S., Rodríguez-Palomares, J.F., Lekadir, K.: Multi-centre, multi-vendor and multi-disease cardiac segmentation: The m&ms challenge. IEEE Transactions on Medical Imaging **40**(12), 3543–3554 (2021). https://doi.org/10.1109/TMI.2021.3090082

4. Dao, T., Fu, D.Y., Ermon, S., Rudra, A., Ré, C.: Flashattention: Fast and memory-efficient exact attention with io-awareness (2022), https://arxiv.org/abs/2205.14135
5. Friedrich, P., Wolleb, J., Bieder, F., Durrer, A., Cattin, P.C.: Wdm: 3d wavelet diffusion models for high-resolution medical image synthesis. In: Mukhopadhyay, A., Oksuz, I., Engelhardt, S., Mehrof, D., Yuan, Y. (eds.) Deep Generative Models. pp. 11–21. Springer Nature Switzerland, Cham (2025)
6. Fu, Y., Bai, W., Yi, W., Manisty, C., Bhuva, A.N., Treibel, T.A., Moon, J.C., Clarkson, M.J., Davies, R.H., Hu, Y.: A versatile foundation model for cine cardiac magnetic resonance image analysis tasks (2025), https://arxiv.org/abs/2506.00679
7. Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, Klaus, Jäger, P.: nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. arXiv preprint (2024)
8. Khader, F., Müller-Franzes, G., Tayebi Arasteh, S., Han, T., Haarburger, C., Schulze-Hagen, M., Schad, P., Engelhardt, S., Baeßler, B., Foersch, S., Stegmaier, J., Kuhl, C., Nebelung, S., Kather, J.N., Truhn, D.: Denoising diffusion probabilistic models for 3d medical image generation. Scientific Reports **13**(1), 7303 (May 2023). https://doi.org/10.1038/s41598-023-34341-2, https://doi.org/10.1038/s41598-023-34341-2
9. Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., Aila, T.: Improved precision and recall metric for assessing generative models. Curran Associates Inc., Red Hook, NY, USA (2019)
10. Liu, C., Yuan, X., Yu, Z., Wang, Y.: Texdc: Text-driven disease-aware 4d cardiac cine mri images generation. In: Cho, M., Laptev, I., Tran, D., Yao, A., Zha, H. (eds.) Computer Vision – ACCV 2024. pp. 191–208. Springer Nature Singapore, Singapore (2025)
11. Martín-Isla, C., Campello, V.M., Izquierdo, C., Kushibar, K., Sendra-Balcells, C., Gkontra, P., Sojoudi, A., Fulton, M.J., Arega, T.W., Punithakumar, K., Li, L., Sun, X., Al Khalil, Y., Liu, D., Jabbar, S., Queirós, S., Galati, F., Mazher, M., Gao, Z., Beetz, M., Tautz, L., Galazis, C., Varela, M., Hüllebrand, M., Grau, V., Zhuang, X., Puig, D., Zuluaga, M.A., Mohy-ud Din, H., Metaxas, D., Breeuwer, M., van der Geest, R.J., Noga, M., Bricq, S., Rentschler, M.E., Guala, A., Petersen, S.E., Escalera, S., Palomares, J.F.R., Lekadir, K.: Deep learning segmentation of the right ventricle in cardiac mri: The m&ms challenge. IEEE Journal of Biomedical and Health Informatics **27**(7), 3302–3313 (2023). https://doi.org/10.1109/JBHI.2023.3267857
12. Mueller, S.K., Koehler, S., Kiekenap, J., Greil, G., Hussain, T., Sarikouch, S., Andre, F., Frey, N., Engelhardt, S.: Deformable image registration for self-supervised cardiac phase detection in cardiac magnetic resonance images of patients with various diseases. Medical Image Analysis p. 104142 (2026). https://doi.org/https://doi.org/10.1016/j.media.2026.104142, https://www.sciencedirect.com/science/article/pii/S1361841526002112

13. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4172–4182 (2023). https://doi.org/10.1109/ICCV51070.2023.00387
14. Ramdas, A., Garcia, N., Cuturi, M.: On wasserstein two sample testing and related families of nonparametric tests (2015), https://arxiv.org/abs/1509.02237
15. Seyfarth, M., Mueller, S.K., Dar, S.U.H., Engelhardt, S.: Latent diffusion for single-pass full 4d cine cardiac magnetic resonance imaging synthesis with spatiotemporal autoencoding. European Heart Journal - Digital Health **7**, ztaf143.028 (01 2026). https://doi.org/10.1093/ehjdh/ztaf143.028
16. Seyfarth, M., Dar, S.U.H., Ayx, I., Fink, M.A., Schoenberg, S.O., Kauczor, H.U., Engelhardt, S.: Medlord: A medical low-resource diffusion model for high-resolution 3d ct image synthesis. In: Fernandez, V., Wiesner, D., Zuo, L., Casamitjana, A., Remedios, S.W. (eds.) Simulation and Synthesis in Medical Imaging. pp. 1–12. Springer Nature Switzerland, Cham (2026)
17. Seyfarth, M., Dar, S.U.H., Frisch, Y., Wild, P., Frey, N., André, F., Engelhardt, S.: Voldit: Controllable volumetric medical image synthesis with diffusion transformers (2026), https://arxiv.org/abs/2603.25181
18. Skorupko, G., Osuala, R., Szafranowska, Z., Kushibar, K., Dang, V.N., Aung, N., Petersen, S.E., Lekadir, K., Gkontra, P.: Fairness-aware data augmentation for cardiac mri using text-conditioned diffusion models. In: Puyol-Antón, E., Ferrante, E., Feragen, A., King, A., Cheplygina, V., Ganz-Benjaminsen, M., Glocker, B., Petersen, E., Lee, H. (eds.) Fairness of AI in Medical Imaging. pp. 63–73. Springer Nature Switzerland, Cham (2026)
19. Wang, H., Liu, Z., Sun, K., Wang, X., Shen, D., Cui, Z.: 3d meddiffusion: A 3d medical latent diffusion model for controllable and high-quality medical image generation. IEEE Transactions on Medical Imaging **44**(12), 4960–4972 (2025). https://doi.org/10.1109/TMI.2025.3585372
20. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004). https://doi.org/10.1109/TIP.2003.819861