Cardiac Ultrasound Video Generation Using a Diffusion Model with Temporal Transformer

Wenbin Wang and Ping Lu

School of Computing, University of Leeds, Leeds, United Kingdom
cnml0744@leeds.ac.uk and p.lu@leeds.ac.uk

Abstract. Cardiac ultrasound is widely used for the diagnosis and monitoring of cardiovascular diseases due to its noninvasive nature, real-time imaging capability, and low cost. However, its clinical utility is often limited by noise sensitivity and acquisition variability, which adversely affect automated interpretation and sequence consistency. To overcome these limitations, this paper presents a multimodal deep learning framework that combines a denoising diffusion model with a Temporal Transformer to generate high-quality cardiac ultrasound videos. A unified preprocessing pipeline with intensity normalisation and standardisation is employed to reduce intersample variation and enhance anatomical structures. Spatial features are first extracted from individual frames, followed by temporal modelling across sequences using the Temporal Transformer. These features guide the latent-space denoising process, optionally augmented by ControlNet for structure-aware generation. The experimental results demonstrate that the proposed method achieves robust performance, with an FID of 43.50, an FVD of 274.52, and an inception score of 8.62. Ablation studies further verify the critical contributions of ControlNet and composite loss design, highlighting the effectiveness of the framework in ensuring both spatial fidelity and temporal coherence.

Keywords: Cardiac Ultrasound · Diffusion Model · Temporal Transformer · ControlNet · Multimodal Generation.

1 Introduction

Cardiac ultrasound plays a critical role in the clinical diagnosis and research of cardiovascular diseases due to its noninvasiveness, high resolution, and excellent contrast. However, current cardiac ultrasound acquisition often faces significant noise interference, unstable image quality, and data format variability caused by diverse medical equipment and imaging protocols, greatly restricting the efficiency and precision of subsequent clinical analyses [1]. Furthermore, traditional imaging techniques often lack effective control over frame-to-frame continuity when handling temporal sequence data, leading to temporal inconsistency that compromises the reliability of diagnostic results [2].

In recent years, diffusion models, known for their powerful generative capabilities, have gradually been applied in medical imaging reconstruction and synthesis, outperforming traditional generative models such as Generative Adversarial Networks (GANs) in certain tasks [38]. Simultaneously, Transformers have become prominent in vision tasks due to their exceptional global information-capturing abilities, progressively replacing conventional Convolutional Neural Networks (CNNs) [22]. However, existing studies mostly concentrate on spatial feature extraction, rarely addressing precise temporal feature modelling, especially lacking joint spatial-temporal modelling tailored to the specific characteristics of medical imaging data.

To address the above issues, we first propose a deep learning framework combining diffusion models and temporal transformers. Through unified data preprocessing strategies with intensity normalisation and standardisation, noise suppression and key structural information extraction are effectively enhanced. Furthermore, a comprehensive loss function integrating the Structural Similarity Index Measure (SSIM), Temporal Mean Squared Error (Temporal MSE), which computes the mean squared difference between adjacent frames to reflect temporal consistency, is adopted as one of the loss components [44], and pixel-level error (L1) is designed to significantly enhance the temporal stability and image quality of cardiac ultrasound videos generated.

The primary innovative contributions of this paper are summarised as follows:

1. A flexible architecture is proposed, utilising a Temporal Transformer module to model temporal relationships in cardiac ultrasound sequences.
2. ControlNet is optionally integrated into the diffusion model, using original cardiac ultrasound images as structural guidance conditions to significantly enhance the structural fidelity of the generated images.
3. A comprehensive loss function based on SSIM, Temporal MSE, and L1 loss is designed to balance image quality and temporal continuity effectively.

2 Related Works

2.1 Overview of Medical Image Generation Techniques

In recent years, deep learning-based generative models have been widely applied to medical image reconstruction, enhancement, and synthesis. Among them, **Generative Adversarial Networks (GANs)** and **Variational Autoencoders (VAEs)** have emerged as two of the most prevalent frameworks. GANs are known for generating visually realistic images, whilst VAEs offer a principled latent space for data representation. However, both methods face limitations that hinder their broad adoption in clinical scenarios. GANs suffer from training instability and mode collapse, which often leads to the generation of limited or repetitive patterns [17]. VAEs, on the other hand, tend to produce blurry outputs due to the nature of their probabilistic reconstruction.

To address these issues, **Flow-based** and **Score-based** models have been proposed as alternatives. These methods offer advantages in terms of generation stability and likelihood estimation, enabling better modelling of complex medical data distributions [18]. Nevertheless, their adoption remains limited due to high computational demands and difficulty in scaling to high-resolution 3D medical data.

Recent work in multi-modal medical generation has also seen progress in report generation tasks. For example, a multinodal method for chest radiology report synthesis [35] leverages visual-textual alignment to generate semantically rich findings. While our framework focuses purely on visual generation, the concept of incorporating semantic priors into medical generative models remains a promising direction.

2.2 Diffusion Models and Their Applications in Medical Imaging

In recent years, **Transformer architectures** have garnered significant attention in the field of medical imaging, particularly in tasks that require modelling **long-range dependencies** and **temporal dynamics**. Unlike convolutional neural networks (CNNs), which are primarily effective at extracting local features, Transformers leverage self-attention mechanisms to capture **global contextual relationships**, making them especially suitable for modelling complex **temporal interactions**.

An increasing number of studies have explored the use of Transformers for **dynamic medical imaging**, such as cardiac ultrasound, cine MRI, and functional brain imaging, where data is represented as temporal sequences. For instance, some works introduce **temporal attention mechanisms** or stack frame-wise features across the time axis to learn organ motion patterns more effectively, leading to improved recognition of cardiac cycles and physiological rhythms [25] [28] [22].

However, most existing Transformer-based models are still designed with a **spatial modelling focus**, and **temporal continuity**—which is critical in medical image sequences—has not been adequately addressed. In particular, for generative tasks involving dynamic image synthesis, current approaches often fail to maintain **inter-frame structural consistency** and **motion coherence**, resulting in artefacts such as flickering, anatomical distortion, or loss of periodic motion. Therefore, designing Transformer modules that can **jointly capture spatial structures and temporal evolution** has become a critical challenge in the domain of dynamic medical image generation [23] [24].

Recent studies have also explored counterfactual video generation as a means to model alternative outcomes or plausible trajectories. For instance, D‘ARTAGNAN [41] proposes a generative architecture that conditions video synthesis on hypothetical interventions, demonstrating promising results in generating temporally coherent counterfactual sequences. While our method does not explicitly model causality, future extensions could integrate such mechanisms for interpretability in clinical contexts.

3 Method

3.1 Model Architecture and Feature Extraction

We propose *HeartDiffusionModel*,a modular deep generative framework tailored for cardiac ultrasound sequence generation. The model integrates transformer-based temporal encoding [25], diffusion-based generation in latent space [37], and a structure-aware ControlNet module [27]. The overall design aims to ensure both **spatial fidelity** and **temporal coherence**, two essential factors for clinical-quality video synthesis.

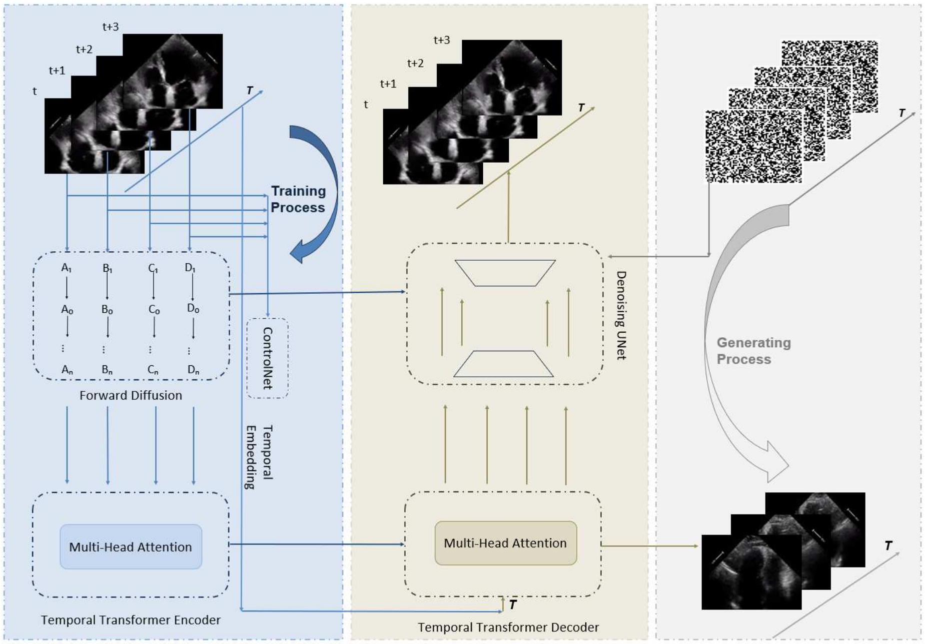

Fig. 1: Our model is structured as a modular architecture consisting of ControlNet, a Temporal Transformer, and a Diffusion model.

Temporal Feature Encoding Given an input sequence of frames $\mathbf{x}_i \in \mathbb{R}^{1\times H\times W}$, a U-Net encoder is employed to extract spatial features $\mathbf{f}_i^{unet}$ for each frame. These features are then stacked along the temporal dimension to form a temporal feature sequence:

$$\mathbf{F}^{unet} = [\mathbf{f}_1^{unet}, \mathbf{f}_2^{unet}, \ldots, \mathbf{f}_T^{unet}] \in \mathbb{R}^{T\times C\times H'\times W'} \tag{1}$$

To capture long-range temporal dependencies, we adopt a **Temporal Transformer** [28], which operates along the time axis. With positional encoding and multi-head self-attention, the temporally enriched features are defined as:

$$\mathbf{F}^{temp} = \text{TemporalTransformer}(\mathbf{F}^{cnn} + \mathbf{E}_{pos}) \tag{2}$$

where $\mathbf{E}_{pos}$ denotes the learnable positional encoding. This module models long-range motion patterns in the cardiac cycle and improves rhythm consistency in the generated sequences.

Latent Diffusion and Feature Projection The output features are projected to match the latent space of the diffusion backbone:

$$\mathbf{F}^{proj} = \mathbf{W}_{proj} \cdot \mathbf{F}^{temp} + \mathbf{b}_{proj} \tag{3}$$

A denoising diffusion probabilistic model (DDPM) [37] operates in latent space. Given the initial latent code $\mathbf{z}_0$, Gaussian noise is gradually added to generate corrupted latents $\mathbf{z}_t$:

$$\mathbf{z}_t = \sqrt{\alpha_t} \cdot \mathbf{z}_0 + \sqrt{1 - \alpha_t} \cdot \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}) \tag{4}$$

The denoising network predicts the noise component:

$$\hat{\epsilon} = \text{UNet}(\mathbf{z}_t,\, t,\, \mathbf{F}^{proj},\, \mathbf{R}_{down},\, \mathbf{R}_{mid}) \tag{5}$$

Structure Guidance via ControlNet To enhance anatomical consistency, we integrate a ControlNet module [27] that runs parallel to the UNet backbone. Given the conditional input $\mathbf{x}$ and projected features, ControlNet produces residual conditions:

$$\{\mathbf{R}_{down}, \mathbf{R}_{mid}\} = \text{ControlNet}(\mathbf{z}_t,\, t,\, \mathbf{F}^{proj},\, \mathbf{x}) \tag{6}$$

These residuals are injected into the UNet's corresponding blocks to guide structural generation, particularly effective for preserving cardiac anatomical features. The module is switchable for ablation studies.

3.2 Loss Function Design

To jointly optimise for spatial detail and temporal smoothness, we define a composite loss function [29] [30] [31]:

$$\mathcal{L}_{\text{total}} = \alpha \cdot \mathcal{L}_{\text{SSIM}} + \beta \cdot \mathcal{L}_{\text{Temporal}} + \gamma \cdot \mathcal{L}_{\text{L1}} \tag{7}$$

Structural Similarity Loss (SSIM): This loss promotes high-level structural similarity between the predicted and ground truth frames:

$$\mathcal{L}_{\text{SSIM}} = \frac{1}{B \cdot T} \sum_{b=1}^{B} \sum_{t=1}^{T} \left[\frac{1 - \text{SSIM}(\mathbf{X}_{b,t}, \hat{\mathbf{X}}_{b,t})}{2} \right] \tag{8}$$

Temporal Consistency Loss: To enforce smooth motion between frames:

$$\mathcal{L}_{\text{Temporal}} = \frac{1}{B \cdot (T-1)} \sum_{b=1}^{B} \sum_{t=1}^{T-1} \text{MSE}(\hat{\mathbf{X}}_{b,t}, \hat{\mathbf{X}}_{b,t+1}) \tag{9}$$

Pixel-wise L1 Loss: To maintain pixel-level fidelity:

$$\mathcal{L}_{\text{L1}} = \frac{1}{B \cdot T \cdot C \cdot H \cdot W} \sum_{b,t,c,h,w} \left| \mathbf{X}_{b,t,c,h,w} - \hat{\mathbf{X}}_{b,t,c,h,w} \right| \tag{10}$$

However, most existing studies primarily focus on **spatial aspects** of medical images, with limited attention to **temporal modelling**—a crucial factor in dynamic imaging modalities such as cardiac ultrasound or cine MRI. Moreover, whilst diffusion models offer strong performance in static image generation, their adaptation to **temporal consistency** and **sequence-level coherence** remains an open research challenge.

4 Experiments

4.1 Data Acquisition

We conducted our experiments on the publicly available EchoNet-Dynamic dataset provided by Stanford University [40]. The EchoNet-Dynamic dataset consists of 10,030 echocardiographic videos collected from patients undergoing echocardiography examinations. Each video captures cardiac cycles and includes important cardiac function metrics such as ejection fraction (EF), which are crucial to assessing cardiac functionality. To evaluate the performance of the proposed model, we randomly selected 1,500 frames from the EchoNet-Dynamic dataset. These frames were split into training, validation, and testing sets(8:1:1). This dataset has also been extended in EchoNet-Synthetic [33], which demonstrates the value of privacy-preserving video generation for secure and ethical sharing of medical imaging data.

4.2 Setup

Training: The experiments were conducted using four H20-NVLink GPUs, each equipped with 20 cores and 96GB memory. During training, distributed data parallelism was employed to efficiently utilise computational resources and reduce training time. To investigate the contribution of each key component in our proposed architecture, we conducted comprehensive ablation studies by selectively removing individual modules, including the Temporal Transformer, ControlNet, MISS loss, and Temporal loss.

Table 1: Configurations of model variants for ablation studies

Model Configuration	**Temporal Transformer**	**ControlNet**	**MISS Loss**	**Temporal Loss**
Ours (full components)	✓	✓	✓	✓
w/o Temporal Transformer	×	✓	✓	✓
w/o ControlNet	✓	×	✓	✓
w/o MISS Loss	✓	✓	×	✓
w/o Temporal Loss	✓	✓	✓	×

Note: Each ablation variant (denoted as "w/o", short for "without") removes one core component from the baseline model to evaluate its unique contribution to spatial fidelity and temporal coherence in the generated ultrasound sequences.

4.3 Comparative Study Design

Ablation Study Design Five model variants were created by removing key modules from the full architecture to isolate the effects of each component. Table 1 details these variants:

All configurations were trained under the same conditions and assessed using the same metrics to ensure reproducible comparisons. Each variant was trained for 200 epochs with early stopping based on validation performance, keeping learning rate (1e−5), batch size, and optimiser settings consistent across experiments.

Comparison with Representative Models We further validated our approach by comparing it against another widely known video generation models:

- **TATS (Temporally-Aware Token Synthesis)** [36]: A Transformer-driven video generation technique that synthesises frame tokens in an autoregressive manner whilst modelling temporal features explicitly. Its progressive generation process is comparable to ours in its ability to recover structured video content from noise.

We implemented these baselines following their official specifications, adapting only the minimum domain-specific elements for cardiac ultrasound data. All models were trained under identical computational constraints for impartial evaluation.

4.4 Evaluation Metrics

To quantitatively assess the quality and fidelity of our generated videos, we employ three well-established metrics that evaluate both perceptual quality and statistical similarity:

- **FID (Fréchet Inception Distance)** [12]:

$$\mathrm{FID} = \|\mu_r - \mu_g\|^2 + \mathrm{Tr}\left(\Sigma_r + \Sigma_g - 2\left(\Sigma_r \Sigma_g\right)^{1/2}\right), \tag{11}$$

  where μ_r, μ_g and Σ_r, Σ_g represent the mean and covariance matrices of real and generated image feature distributions, respectively. FID measures the distributional similarity between real and generated images through feature representations extracted from InceptionV3 network.

- **FVD (Fréchet Video Distance)** [13]: An extension of FID to the video domain that captures temporal dynamics by utilising features from a pre-trained 3D convolutional network. FVD is defined analogously to FID but operates on spatio-temporal features:

$$\mathrm{FVD} = \|\mu_r^v - \mu_g^v\|^2 + \mathrm{Tr}\left(\Sigma_r^v + \Sigma_g^v - 2\left(\Sigma_r^v \Sigma_g^v\right)^{1/2}\right), \tag{12}$$

  where the superscript v indicates features extracted from video sequences.

- **IS (Inception Score)** [16]:

$$\mathrm{IS} = \exp\left(\mathbb{E}_x\left[D_{KL}(p(y|x)\|p(y))\right]\right), \tag{13}$$

  where $p(y|x)$ represents the conditional class distribution for a generated sample x as predicted by the Inception model, and $p(y)$ is the marginal class distribution. IS jointly quantises quality and diversity by measuring how distinctive and recognisable the generated samples are.

For all metrics, we calculate the scores across multiple samples to ensure robust evaluation. Lower FID and FVD values indicate better quality and temporal consistency, with values closer to zero representing perfect alignment with the real data distribution. Conversely, higher IS values signify improved quality and diversity. Through this complementary set of metrics, we comprehensively evaluate both the spatial fidelity and temporal coherence of our generated cardiac ultrasound sequences.

5 Results

5.1 Quantitative Results

Table 2: Quantitative comparison of the proposed method, its ablation variants, and TATS on ultrasound video generation.

Model Configuration	**FID ↓**	**FVD ↓**	**IS ↑**
TATS	41.40	**174.30**	7.06
w/o Temporal Transformer	**41.30**	–	5.32
w/o ControlNet	72.26	310.72	5.29
w/o MISS Loss	53.32	297.20	6.62
w/o Temporal Loss	48.10	291.74	7.78
Ours (Full Model)	43.50	274.52	**8.62**

As shown in Table 2, we conduct a comprehensive comparison between the proposed full model, its ablation variants, and the TATS baseline. While TATS achieves the best FID (41.40) and FVD (174.30), indicating strong image-level fidelity and temporal coherence, it lags behind in perceptual quality, with an IS score of 7.06. In contrast, our full model achieves the highest IS (8.62), suggesting superior perceptual sharpness and diversity, while maintaining competitive FID (43.50) and FVD (274.52), which reflects a good balance between spatial quality and temporal consistency.

Among the ablation variants, removing the ControlNet module leads to the most significant performance degradation, with FID and FVD increasing to 72.26 and 310.72 respectively, and IS dropping to 5.29. This underscores the importance of ControlNet in preserving spatial structures during generation. The MISS loss also plays a key role, as its removal causes FID to rise to 53.32 and FVD to 297.20, indicating a loss of fine-grained anatomical consistency.

Interestingly, removing the Temporal Loss results in the highest IS score (7.78), yet degrades FVD to 291.74, suggesting a trade-off where improved perceptual clarity comes at the cost of motion coherence. The variant without the Temporal Transformer yields the lowest FID (41.30), outperforming TATS in this metric. However, it lacks FVD evaluation due to unstable video generation, indicating compromised temporal modeling despite strong spatial accuracy.

Overall, our full model delivers a robust and well-rounded performance, effectively integrating spatial fidelity, perceptual realism, and temporal coherence, validating the superiority of our proposed architecture for high-quality cardiac ultrasound video synthesis.

5.2 Representative Examples

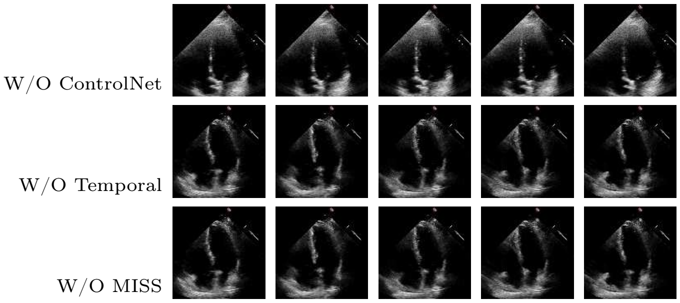

Fig. 2: Qualitative comparison of five internal frames for ablation model variants. Frames are sampled at every 12-frame interval from the generated videos.

5.3 Discussion

The experimental findings underscore the effectiveness of our proposed model in generating high-quality cardiac ultrasound sequences that preserve both spatial fidelity and temporal consistency. Compared to the TATS baseline [36], our

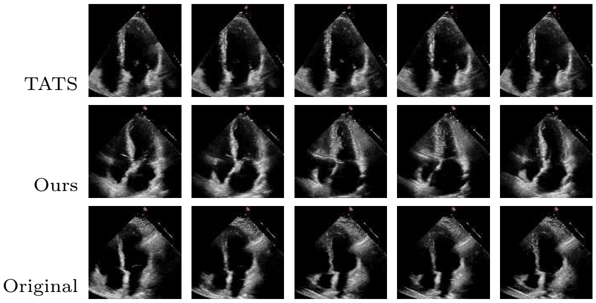

Fig. 3: Qualitative comparison of five consecutive frames generated by TATS and our proposed model, alongside the corresponding ground truth (Original). The frames are uniformly sampled every 12 frames from each video sequence.

model achieves a higher Inception Score, indicating enhanced perceptual realism and diversity, while maintaining competitive FID and FVD metrics. This performance highlights the advantage of combining diffusion-based generation [37, 38] with dedicated temporal modeling.

Ablation experiments further validate the significance of individual modules. The removal of ControlNet leads to severe degradation across all metrics, confirming its role in preserving anatomical structure. Likewise, the MISS Loss contributes to multi-scale structural consistency, and its absence results in decreased spatial fidelity. Interestingly, eliminating the Temporal Loss yields sharper individual frames (higher IS), but leads to unstable motion patterns, emphasizing the trade-off between perceptual quality and temporal coherence.

Despite its demonstrated robustness, the proposed model exhibits limitations in accurately capturing fine-grained cardiac motion and managing noisy real-world ultrasound inputs. In future work, we will will explore more advanced spatio-temporal architectures, including the incorporation of recurrent units such as LSTM [39] or GRU, to better model long-term temporal dependencies. Moreover, adaptive attention mechanisms and context-aware diffusion control strategies may further enhance generation fidelity in clinically complex scenarios. In addition, we plan to extend our evaluation across multiple datasets—such as the CAMUS [42] and EchoNet-LVH datasets [43]—to improve the generalizability and reliability of the results under diverse imaging protocols and patient populations. As a potential clinical application, we also intend to incorporate ejection fraction (EF) estimation as a downstream task, enabling quantitative assessment of cardiac function from the generated ultrasound sequences.

6 Conclusion

In this work, we present a novel multimodal framework for cardiac ultrasound video generation by integrating diffusion models [37, 38] with temporal transformers. The architecture leverages ControlNet for spatial conditioning, MISS Loss for structural consistency, and a Temporal Loss for maintaining motion coherence. A unified preprocessing pipeline also ensures normalization across highly variable ultrasound inputs.

Quantitative experiments on the EchoNet-Dynamic dataset [40] and qualitative comparisons with the TATS model [36] demonstrate that our approach generates perceptually realistic and temporally smooth sequences. Ablation results confirm that each component contributes meaningfully to the overall performance, particularly in balancing visual quality and anatomical correctness.

In the future, we aim to explore LSTM-based temporal modeling [39] and spatio-temporal attention mechanisms to further improve long-range motion continuity and fine detail reconstruction. These improvements could enhance the utility of generative models in clinical simulation, diagnostic support, and privacy-preserving data augmentation.

References

1. Bernard, O. et al.: 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)
2. Qin, C. et al.: Convolutional recurrent neural networks for dynamic MR image reconstruction. *IEEE Transactions on Medical Imaging*, **38**(1), 280–290 (2018)
3. Wolleb, J. et al.: Diffusion models for medical anomaly detection. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 35–45 (2022)
4. Chen, J. et al.: TransUNet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306* (2021)
5. Ouyang, D. et al.: Video-based AI for beat-to-beat assessment of cardiac function. *Nature*, **580**(7802), 252–256 (2020)
6. Vaswani, A. et al.: Attention is all you need. In: *Advances in Neural Information Processing Systems*, pp. 5998–6008 (2017)
7. Zhang, L. et al.: Adding Conditional Control to Text-to-Image Diffusion Models. In: *Proc. of ICCV*, pp. 18436–18446 (2023)
8. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: *Advances in Neural Information Processing Systems*, **33**, 6840–6851 (2020)
9. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, **13**(4), 600–612 (2004)
10. Huang, M. et al.: Learning Temporal Coherence via Self-Supervision for GAN-based Video Generation. *ACM Transactions on Graphics (TOG)*, **39**(4), 1–12 (2020)
11. Zhao, H., Gallo, O., Frosio, I., Kautz, J.: Loss functions for image restoration with neural networks. *IEEE Transactions on Computational Imaging*, **3**(1), 47–57 (2016)
12. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: *Advances in Neural Information Processing Systems*, pp. 6626–6637 (2017)
13. Unterthiner, T. et al.: FVD: A new metric for video generation. In: *ICLR Workshop on Deep Generative Models for Highly Structured Data* (2019)
14. Voleti, V. et al.: MCVD: Masked Conditional Video Diffusion for Prediction, Generation, and Interpolation. *arXiv preprint arXiv:2205.09853* (2022)

15. Ge, S. et al.: Long video generation with time-agnostic VQGAN and time-sensitive Transformer. In: *European Conference on Computer Vision*, pp. 370–386. Springer (2022)
16. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved Techniques for Training GANs. In: *Advances in Neural Information Processing Systems*, pp. 2234–2242 (2016)
17. Yi, X., Walia, E., & Babyn, P. (2019). Generative adversarial network in medical imaging: A review. *Medical Image Analysis*, 58, 101552.
18. Song, Y., & Ermon, S. (2019). Generative modelling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems (NeurIPS)*.
19. Wolleb, J., Van Gool, L., & Lüthi, M. (2022). Diffusion Models for Medical Image Analysis: A Comprehensive Survey. *arXiv preprint arXiv:2211.07804*.
20. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
21. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., & Schmid, C. (2021). ViViT: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6836–6846.
22. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., & Zhou, Y. (2021). TransUNet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*.
23. Yan, K., Wang, C., Yu, H., & Liang, D. (2023). Neural motion fields for dynamic MRI reconstruction. *IEEE Transactions on Medical Imaging*, 42(2), 427–438.
24. Shi, X., Li, Z., Zhao, Y., Chen, M., Liu, S., Wang, M., & Huang, G. (2022). Video diffusion models. *arXiv preprint arXiv:2204.03458*.
25. Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.
26. Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33, 6840–6851.
27. Zhang, L., Chen, H., Zhang, J., et al. (2023). Adding conditional control to text-to-image diffusion models. *CVPR 2023*, pp. 14065–14075.
28. Arnab, A., Dehghani, M., Heigold, G., et al. (2021). ViViT: A video vision transformer. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6836–6846.
29. Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4), 600–612.
30. Huang, Y., Wu, W., et al. (2020). Learning of temporal consistency for video super-resolution. *IEEE Transactions on Image Processing*, 29, 9140–9153.
31. Zhao, H., Gallo, O., Frosio, I., & Kautz, J. (2016). Loss functions for image restoration with neural networks. *IEEE Transactions on Computational Imaging*, 3(1), 47–57.
32. Author(s). *D'ARTAGNAN: Counterfactual Video Generation*. Proceedings of [Conference Name], 2023.
33. Author(s). *EchoNet-Synthetic: Privacy-preserving Video Generation for Safe Medical Data Sharing*. Proceedings of [Conference Name], 2023.
34. Author(s). *Generating Chest Radiology Report Findings Using a Multinodal Method*. [Journal or Conference], 2023.
35. Author(s). *Generating Chest Radiology Report Findings Using a Multinodal Method*. [Journal or Conference], 2023.

36. Ge, Y., Dai, B., Wu, J., Torralba, A., Freeman, W. T., and Liu, C.: Long-term Video Synthesis with Time-Agnostic VQGAN and Time-Aware Transformer. In: *Advances in Neural Information Processing Systems* (NeurIPS), (2022).
37. Ho, J., Jain, A., and Abbeel, P.: Denoising Diffusion Probabilistic Models. In: *Advances in Neural Information Processing Systems* (NeurIPS), (2020).
38. Wolleb, J., Sandküehler, R., Karlen, W., Czempiel, T., and Rieke, N.: Diffusion Models for Medical Anomaly Detection. In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 35–45, (2022).
39. Hochreiter, S., and Schmidhuber, J.: Long Short-Term Memory. *Neural Computation*, **9**(8), 1735–1780 (1997).
40. Ouyang, D., He, B., Ghorbani, A., Yuan, N., Ebinger, J., Langlotz, C., Heidenreich, P., Harrington, R., Liang, D., and Zou, J.: Video-based AI for beat-to-beat assessment of cardiac function. *Nature*, **580**(7802), 252–256 (2020).
41. Yuan L., Hu Y., Du Y., Jiang L., Wang W., Lin D., Qian C. D'ARTAGNAN: Counterfactual Video Generation with Latent Diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20187–20197.
42. S. Leclerc, E. Smistad, J. Pedrosa, A. Ostvik, et al., "Deep Learning for Segmentation using an Open Large-Scale Dataset in 2D Echocardiography," *IEEE Transactions on Medical Imaging*, vol. 38, no. 9, pp. 2198–2210, Sept. 2019. doi: https://doi.org/10.1109/TMI.2019.290051610.1109/TMI.2019.2900516.
43. A. Y. Hannun, P. Rajpurkar, M. S. M. Saeed, et al., "A Large-Scale Benchmark for Automated Cardiac View Classification and Disease Detection," in *NeurIPS 2019 Machine Learning for Health Workshop*, 2019. Available: https://echonet.github.io/dataset/download.html
44. R. Villegas, J. Yang, S. Hong, X. Lin, and H. Lee, "Decomposing Motion and Content for Natural Video Sequence Prediction," in *International Conference on Learning Representations (ICLR)*, 2017.