Training-Free Condition Video Diffusion Models for single frame Spatial-Semantic Echocardiogram Synthesis

Van Phi Nguyen[1], Tri Nhan Luong Ha[2], Huy Hieu Pham[2], and Quoc Long Tran [†1] (✉)

[1] Institute for Artificial Intelligence, University of Engineering and Technology, Vietnam National University, Hanoi, Vietnam tqlong@vnu.edu.vn
[2] VinUni-Illinois Smart Health Center, VinUni, Hanoi, Vietnam

Abstract. Condition video diffusion models (CDM) have shown promising results for video synthesis, potentially enabling the generation of realistic echocardiograms to address the problem of data scarcity. However, current CDMs require a paired segmentation map and echocardiogram dataset. We present a new method called Free-Echo for generating realistic echocardiograms from a single end-diastolic segmentation map without additional training data. Our method is based on the 3D-Unet with Temporal Attention Layers model and is conditioned on the segmentation map using a training-free conditioning method based on SDEdit. We evaluate our model on two public echocardiogram datasets, CAMUS and EchoNet-Dynamic. We show that our model can generate plausible echocardiograms that are spatially aligned with the input segmentation map, achieving performance comparable to training-based CDMs. Our work opens up new possibilities for generating echocardiograms from a single segmentation map, which can be used for data augmentation, domain adaptation, and other applications in medical imaging. Our code is available at https://github.com/gungui98/echo-free

Keywords: Deep Learning · Ultrasound · Cardiac · Generative · Video · Diffusion

1 Introduction

Echocardiogram is a widely used imaging modality for assessing cardiac function and structure. It is a non-invasive, cost-effective, and widely available imaging modality that provides real-time information about the structure and function of the heart. Meanwhile, the interpretation of echocardiograms highly depends on the operator's experience and the quality of the images. Automated analysis of echocardiograms has the potential to improve the accuracy and efficiency of the diagnosis and treatment of cardiovascular diseases [20]. However, the development of machine learning models for echocardiogram analysis is challenging

---

[†] Corresponding authors.

due to the scarcity of labeled data, the complexity of echocardiograms, and the variability of echocardiograms across different patients and imaging systems [7].

Synthetic medical imaging therefore has the potential to address the problem of data scarcity in medical image processing. In the domain of echocardiogram, synthetic data is often generated from physics-based simulators and data-driven deep generative models. Physics-based simulators create synthetic echocardiograms by solving the wave equation, which mimics the physical process of ultrasound imaging [14,2,6]. From a segmentation map, the simulator could generate an echocardiogram by simulating the ultrasound wave propagation through the tissue, typically with the help of a scattering model. However, physics-based simulators are computationally expensive and require expert knowledge to tune the parameters, such as the speed of sound, the attenuation coefficient, and the scattering coefficient, to generate realistic echocardiograms. In addition, obtaining a tissue scatter map from the segmentation map is non-trivial and often causes unrealistic textures in the generated echocardiograms. In addition to physics-based simulators, data-driven deep-generative models, such as Generative Adversarial Networks (GANs), have been proposed to generate realistic echocardiograms [23,27,1,3]. Despite promising results, the quality of the generated echocardiograms is often limited due to the mode collapse problem. Diffusion models (DMs) have recently emerged as a promising alternative to GANs for generating more realistic videos and conditions [8,9,11,5,29], due to easier training procedures and better sample quality. However, controlling the generation of current DMs requires a paired dataset of segmentation maps and echocardiogram, which is not always available in the domain of medical imaging. In this work, we propose a training-free condition video diffusion model (CDM) for echocardiogram synthesis that allows the generation of realistic echocardiograms from a single end-diastolic segmentation map without the need for any additional training data.

Contribution: In this paper, we propose a new method for CDM for echocardiogram synthesis. Our model is training-free and requires no additional paired dataset of segmentation map and echocardiogram data, while still generating realistic echocardiograms from a single end-diastolic segmentation map. Based on SDEdit [18], we propose, Free-Echo, which start the reverse denoising process with a noisy version of a pseudo-video obtained from the segmentation map instead of pure Gaussian noise. We demonstrate the effectiveness of our model on two public echocardiogram datasets, CAMUS and EchoNet-Dynamic, showing comparable performance to training-based CDMs.

Related work: Echocardiogram synthesis: Several methods have been proposed to generate realistic echocardiograms. Salehi *et al.* [23] used a physic simulator to generate patient-specific echocardiograms. Liang [17] proposed a GAN-based model to generate ultrasound from sketch image. Tomar *et al.* [27] also used Cycle-GAN to generate content-preserving echocardiograms with unpaired data. While GAN-based methods have shown promising results for echocardiogram synthesis, the quality of generated echocardiograms is often limited due to the mode collapse problem and the difficulty of training GANs. DMs have been

proposed to generate realistic echocardiograms [22,21,25]. Reynaud *et al.* [22] developed a cascaded diffusion model, conditioned on an End Diastolic (ED) frame, to produce ultrasound images with varying left ventricle ejection fractions (LVEFs). Stojanovski *et al.* [25] utilized Denoising Diffusion Probabilistic Model (DDPM) to generate synthetic echoechocardiograms with conditions made from semantic label maps. Another DDPM-based method by Phi *et al.* [21] applied dynamic semantic label mapping of diastolic frame in a multi-scale decoder to produce realistic echocardiography sequences with diverse anatomical structures. Diffusion-based methods have shown success, however current methods still require a paired dataset of segmentation map and echocardiogram data. In this work, we propose a training-free CDM for echocardiogram synthesis to address this problem.

Conditional video synthesis: CDM aims to synthesise realistic sample given some condition, from semantic description from text to spatial layout such as bouding box, segmentation map. Especially in the domain of medical imaging, CDM enables synthetic data generation with different anatomy, pathology, and acquisition parameters. With paired dataset of condition and the sample, classifier-guided [4] and classifier-free conditioning [13] have been proposed to the CDM on spatial layout. In some domains, such as medical imaging, where a paired dataset of conditions is not always available, researchers search for a training-free conditioning method, such as SDEdit [18]. However, SDEdit which requires a colorized version of the spatial layout, suffered from content leakage. In this work, we propose a training-free conditioning method based on SDEdit, that replaces the colorized version of the spatial layout with the pseudo-video obtained from the spatial layout using optimal transport.
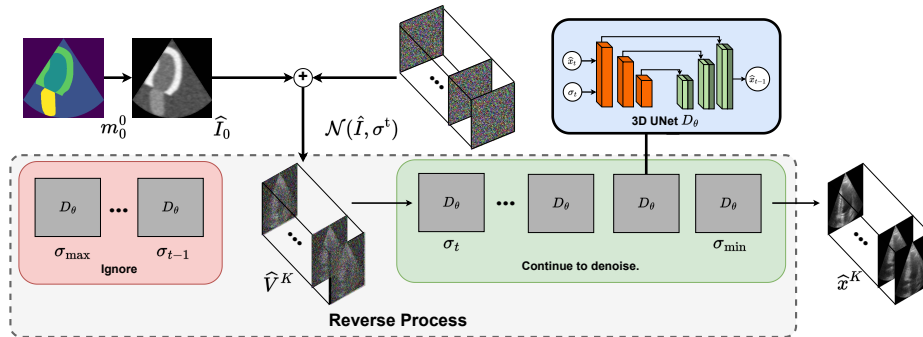
2   Method

Fig. 1. Illustration of our method. Given a single end-diastolic segmentation map m_0, we first solve the optimal transport problem to obtain the pseudo-image $\hat{I}_0$. We then start the reverse process of the DM from the diffusion step t with the noisy version of the pseudo video $\hat{V}^K$, obtained by adding Gaussian noise to the pseudo image. The reverse process is continued until the diffusion step $t = 0$, at which point we obtain the generated echocardiogram $\hat{x}^K$.

Video Diffusion Model: DMs [24,12] are a class of generative models that learn to transform pure Gaussian noise into data distribution by gradually removing the noise. DMs consist of two processes: the forward process and the reverse sampling process. Given the data distribution, $p(x_0^K)$ has a standard deviation of σ_{data}, where x_0^K is a video volume that contains K frames. The forward process creates a noisy version s_0^K of the video x_0^K by adding Gaussian noise of standard deviation σ to the video volume $s_0^K = x_0^K + \sigma n$, where $n \sim \mathcal{N}(0, I)$ is a pure Gaussian noise. When the standard deviation or noise level σ is substantially larger than σ_{data}: $\sigma_{\text{max}} \gg \sigma_{\text{data}}$ and the noisy volume s_0^K is almost indistinguishable from pure Gaussian noise, the σ_{max} is the maximum noise level.

The reverse process, used to synthesize a new video $\hat{x}_0^K$, randomly initializes the noisy video s_0^K by sampling from pure Gaussian noise $\mathcal{N}(0, \sigma_{\text{max}}^2 I)$, and then gradually removes the noise from the noisy video s_0^K from high noise level σ_{max} until reach the zero noise level $\sigma = 0$, at which point we obtain the generated echocardiogram $\hat{x}_0^K$. To estimate the noise added to the noisy video s_0^K at the noise level σ, the reverse process trains a denoiser $D_\theta(\cdot, \sigma)$ by minimizing the following loss function:

$$\mathcal{L}(\theta) = \mathbb{E}_{\sigma \sim p(\sigma), x_0^K \sim p(x_0^K)} \left[w(\sigma) \| D_\theta(x_0^K + \sigma n, \sigma) - x_0^K \|_2^2 \right] , \tag{1}$$

where $w(\sigma)$ is the weight function, used to balance the contribution of the different noise levels to the loss function, and $p(\sigma)$ is the noise level distribution. To sample the DMs, starting from the noisy version of the video x at noise level σ_{max}, the reverse process is continued until noise level σ_{min} by solving the following ordinary differential equation (ODE):

$$\frac{d\hat{x}_t}{d\sigma} = -\sigma \Delta_x log(p(\hat{x}_t, \sigma)) = \frac{\hat{x}_t - D_\theta(\hat{x}_t, \sigma)}{\sigma} , \tag{2}$$

for $t \in [0, T]$ and $\sigma \in [\sigma_{\text{max}}, \sigma_{\text{min}}]$ and T is the denoising step. The solution of the ODE is obtained by the ODE solver, such as the Euler-Maruyama or 2nd order Heun's method.

Conditioning: Conventionally, the denoiser $D_\theta(\cdot, \sigma)$ is modified to take condition c as input, for example, the end-diastolic segmentation map m_0^0. The denoiser is then trained via classifier-free [13] or classifier-guidance [4]. In both cases, the conditioning requires a pair dataset of segmentation maps and the video. To address those challenges, we propose a training-free conditioning method based on SDEdit [18]. The idea is that instead of denoising pure Gaussian noise $s_0^K \sim \mathcal{N}(0, \sigma_{\text{max}}^2 I)$ to σ_{min}, at diffusion step t_i, we replace the noisy version of video s_0^K with the noisy version of a pseudo-video obtained from the segmentation map m_0^0 and continue the reverse process until σ_{min}. This allows us to generate spatially coherent echocardiograms from the segmentation map while still being able to fill realistic textures with pseudo-video. In the original SDEdit, the pseudo-video is a colorized version of segmentation, which makes it harder for denoiser to map from the noisy version of the pseudo-video to the clean version of the pseudo-video. While recent research on unpaired domain-to-domain

translation has shown that the pseudo-video can be generated from the segmentation map using optimal transport [26], current works require extensive training and data from both source and target domains. In our method, we replace the colorized version of the pseudo-video with an adaptive pseudo-video generation.

Pseudo Video Generation: Given the segmentation map m_0^0, we first map each label to a unique intensity value to obtain an intensity image $\hat{I}_0$. We noticed that simple mapping may cause a mismatch in the intensity histogram between the pseudo-image and the video data. To address this issue, we solve the optimal transport problem between the pseudo-image $\hat{I}_0$ and the video data x_0^K to obtain the pseudo-image $\hat{I}_0$. Formally, we solve the following optimal transport problem:

$$\hat{I}_0 = \arg\min_{\hat{I}_0} \int_{\hat{I}_0} \int_{x_0^K} c(x, \hat{x})\mu(x)\mu(\hat{x})dxd\hat{x}\,, \tag{3}$$

where $c(j, \hat{j})$ is the cost function, and $\mu(x)$ and $\mu(\hat{x})$ are the probability distribution of the video data x_0^K and the pseudo video $\hat{x}_0^K$, respectively. We then use the Sinkhorn algorithm to solve the optimal transport problem with the cost function $c(x, \hat{x}) = \|x - \hat{x}\|_2^2$ and a regularization parameter of 10^{-3}. Similar to the original SDEdit, we add Gaussian noise at step t_i to the pseudo image $\hat{I}_0$ to obtain the noisy version of the pseudo video $\hat{V}^K = \hat{I}_0 + \sigma(i)n$ with $\sigma(i)$ being the noise level at diffusion step t_i. We then continue the reverse process of the DM in equation 2 from the diffusion step t_i with the noisy version of the pseudo-video $\hat{V}^K$ until we obtain the generated ultrasound video.

Denoiser Formulation: We follow the parametrization from Elucidated Diffusion Model (EDM) [15]. The noise level σ follows the log-normal distribution $\log \sigma \sim \mathcal{N}(P_{\text{mean}}$ where $P_{\text{mean}} = -1.2, P_{\text{std}} = 1.2)$.

The input and output of the denoiser $D_\theta(\cdot, \sigma)$ are scaled as:

$$\hat{D}_\theta(s, \sigma) = \frac{\sigma_{\text{data}}}{\sigma^*}s + \frac{\sigma \cdot \sigma_{\text{data}}}{\sigma^*}F_\theta\left(\frac{s}{\sigma^*}, \frac{\ln(\sigma)}{4}\right)\,, \tag{4}$$

where $\sigma_{\text{data}} = 0.5$ and $\sigma^* = \sqrt{\sigma^2 + \sigma_{\text{data}}^2}$. F_θ is a neural network. The weight function $w(\sigma)$ is set to $w(\sigma) = (\sigma^*/(\sigma \cdot \sigma_{\text{data}}))^2$ that cancels the weight of F_θ.

3 Experiment Settings

Data: We evaluate our model on two datasets, CAMUS [16] and EchoNet-Dynamic [19]. The CAMUS public dataset consists of 500 patient records with 2 chamber-view echocardiograms. It contains segmentation map for the left ventricle endocardium, myocardium, and left atrial endocardium. The segmentation annotation of CAMUS is available for the end-diastolic and end-systolic frames. We split the dataset into 400 videos for training, 50 videos for validation, and 50 videos for testing. The EchoNet-Dynamic dataset consists of 10,030 4 chamber view ultrasound videos. Different from CAMUS, it contains the segmentation map only for the left ventricle at the end-diastolic frame and the end-systolic

frame. We split the dataset into 8,024 videos for training, 1,003 videos for validation, and 1,003 videos for testing. For both datasets, we sample the first 24 frames of the video, with condition on the end-diastolic segmentation map.

Implementation: We adopt the 3D-Unet from Ho et.al. [11] as our denoiser. The model is trained on 24 frames of 128×128 pixels. For the training of the denoiser, we use the Adam optimizer with a learning rate of 10^{-3}, and a batch size of 16 for 100,000 iterations. We use the same noise level distribution as EDM [15]. We start the reverse process from the diffusion step $t_i = 15$ in a total of 64 diffusion steps. All experiments are conducted on a 3x NVIDIA H100 GPU with 80GB of memory.

Evaluation: We evaluate our model on the following metrics: (1) Structural Similarity Index (SSIM) [30], (2) Peak Signal-to-Noise Ratio (PSNR), (3) Fréchet Inception Distance (FID) [10], (4) Fréchet Video Distance (FVD) [28] between the generated and the ground truth echocardiograms. For quantitative evaluation, we generated 10 samples for each segmentation in the test set, for FID and FVD, we use the InceptionV3 and the R(2+1)D networks, respectively. The non-deep learning metrics, such as SSIM, PSNR, are computed based on the average of the metrics between the generated and ground truth echocardiograms given the same segmentation map. We compare our model with SDEdit method [18] and the CDM using classifier-free conditioning [13], with a classifier-free guidance factor of 7.0. All models are trained on the same datasets, denoiser hyperparameters, and noise level distribution.

4 Experiment Results

Dataset	Method	Step. t	SSIM ↑	PSNR ↑	FID ↓	FVD ↓
CAMUS [16]	SDEdit [18]	-	0.23	19.54	32.45	213.50
	Cls-Free [13]	-	0.52	22.49	22.50	150.80
	Free-Echo (Ours)	15	0.48	20.43	26.31	195.41
	Free-Echo (Ours)	35	0.27	16.54	53.71	312.12
	Free-Echo (Ours)	55	0.12	13.45	67.15	412.55
Echonet-Dynamic [20]	SDEdit [18]	-	0.34	21.52	25.49	250.45
	Cls-Free [13]	-	0.56	20.04	18.56	130.85
	Free-Echo (Ours)	15	0.51	20.78	24.22	180.39
	Free-Echo (Ours)	35	0.32	18.72	44.67	332.12
	Free-Echo (Ours)	55	0.21	16.66	59.85	395.73

Table 1. Quantitative evaluation of our model on the CAMUS and EchoNet-Dynamic datasets.

Table 1 shows the quantitative evaluation of our model on the CAMUS and EchoNet-Dynamic datasets. We observe that our model achieves comparable per-

formance to the CDM using classifier-free conditioning, with around 10% drop in SSIM, PSNR, and L2 distance, and 20% increase in FID and FVD. Our model outperforms SDEdit, showing the benefit of pseudo-video generation. Since there is trade-off between the diffusion step when we start the reverse process from the diffusion step t_i, we observe that the best performance is achieved at the diffusion step $t_i = 15$ for both datasets, and the performance decrease significantly when we start the reverse process from the diffusion step $t_i = 55$. This trade-off suggests that the high-level anatomy and motion of the heart can be generated at very early diffusion step, and the texture and fine details of the heart can be generated at later diffusion step. The performance of our model is also consistent across the two datasets. In Figure 2, we present a visual comparison of our model against the CDM, using ground truth data from the CAMUS and EchoNet-Dynamic datasets. For the CAMUS dataset, our model shows an ability to integrate motion that aligns with the segmentation map, even when generating from a static pseudo-video. However, the fidelity of motion diminishes in later diffusion steps, emphasizing the importance of selecting the right t_i for realistic echocardiogram generation. On the EchoNet-Dynamic dataset, our model effectively synthesizes echocardiograms with realistic LV region motion, adhering to the segmentation map's anotomy. We noticed that the CDM struggles with consistent ultrasound structure, particularly in areas with artifacts and deformities like the cone area. This is likely due to the label map containing only LV region, making it difficult for the CDM to generate uniform representations. Our use of pseudo-labels in these areas helps overcome this issue, ensuring consistency in the generated ultrasound shapes. This approach shows promise in handling the inherent variability and complexity of echocardiogram data, as seen in EchoNet-Dynamic's diverse shapes.

Some limitations should be considered. The generated echocardiograms are imperfect, with low resolution and short duration. The performance of our model is also sensitive to the diffusion step t_i, and the optimal diffusion step may vary across different datasets. Downstream applications of our model still require further evaluation since the motion of the heart is not always consistent with the segmentation map. Future will focus on a more robust method for generating echocardiograms from a single segmentation map, and exploring the potential of our model for data augmentation, domain adaptation, and other applications in medical imaging.
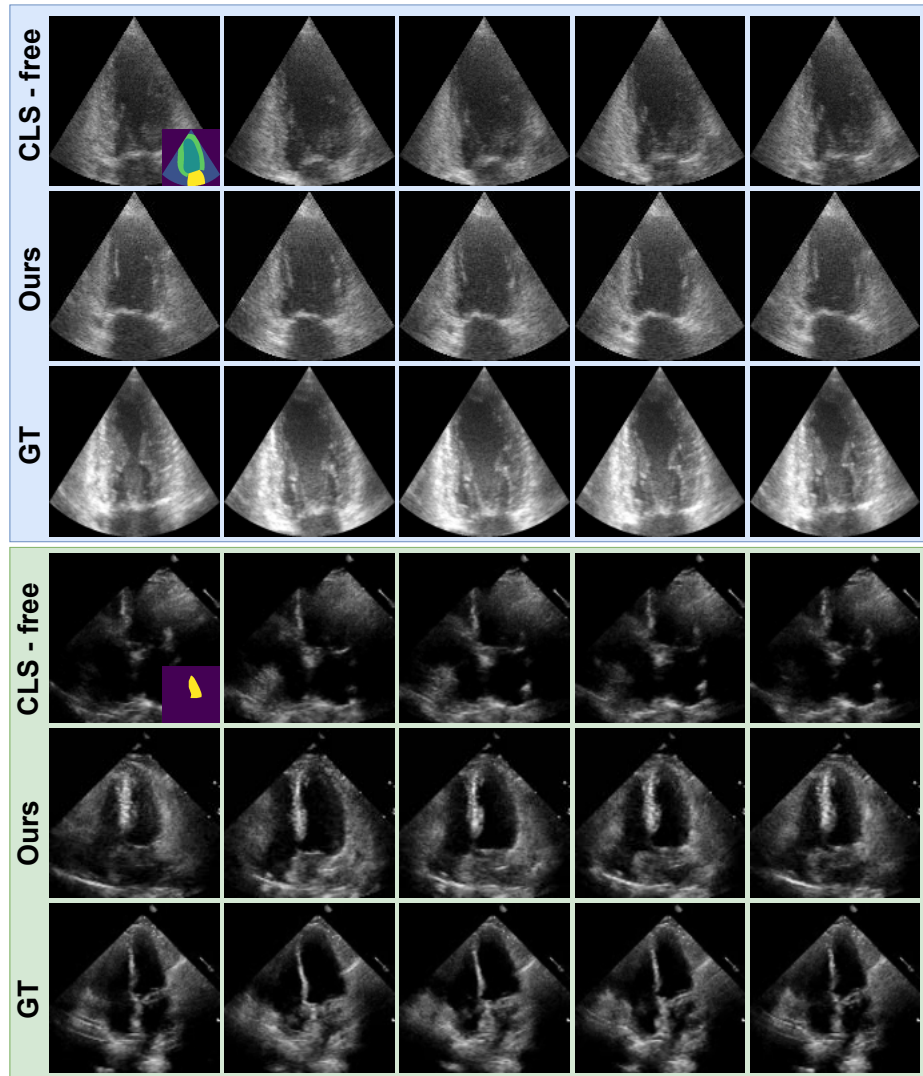
Fig. 2. Visual comparison of our model with the CDM and the ground truth echocardiograms trained on the CAMUS (blue) and EchoNet-Dynamic (green) datasets.

5   Conclusion

We propose a training-free CDM for echocardiogram synthesis. Our model can generate realistic echocardiograms from a single end-diastolic segmentation map. We demonstrate the effectiveness of our model on two public echocardiogram datasets CAMUS and EchoNet-Dynamic, and our in-house dataset. We further conduct human evaluation to assess the visual quality of the generated echocardiograms. Through extensive experiments, we show that our model can generate plausible echocardiograms with that both temporal coherent and spatially align with the input segmentation map without additional training data. Our work can be used for data augmentation, domain adaptation, and other applications in medical imaging.

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

References

1. Bargsten, L., Schlaefer, A.: SpeckleGAN: a generative adversarial network with an adaptive speckle layer to augment limited training data for ultrasound image processing. IJCARS **15**, 1427–1436 (2020)
2. Burger, B., Bettinghausen, S., Radle, M., Hesser, J.: Real-Time GPU-Based Ultrasound Simulation Using Deformable Mesh Models. IEEE Transaction on Medical Imaging **32**, 609–618 (2013)
3. Cronin, N.J., Finni, T., Seynnes, O.: Using deep learning to generate synthetic B-mode musculoskeletal ultrasound images. Computer Methods and Programs in Biomedicine **196**, 105583 (2020)
4. Dhariwal, P., Nichol, A.Q.: Diffusion models beat gans on image synthesis. In: Ranzato, M., Beygelzimer, A., Dauphin, Y.N., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual. pp. 8780–8794 (2021)
5. Esser, P., Chiu, J., Atighehchian, P., Granskog, J., Germanidis, A.: Structure and content-guided video synthesis with diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7346–7356 (2023)
6. Garcia, D.: Simus: an open-source simulator for medical ultrasound imaging. part i: theory & examples. Computer Methods and Programs in Biomedicine **218**, 106726 (2022)
7. Ghorbani, A., Ouyang, D., Abid, A., He, B., Chen, J.H., Harrington, R.A., Liang, D.H., Ashley, E.A., Zou, J.Y.: Deep learning interpretation of echocardiograms. NPJ digital medicine **3**(1), 10 (2020)
8. Gupta, A., Tian, S., Zhang, Y., Wu, J., Martín-Martín, R., Fei-Fei, L.: MaskViT: Masked Visual Pre-Training for Video Prediction (2022)
9. Harvey, W., Naderiparizi, S., Masrani, V., Weilbach, C., Wood, F.: Flexible Diffusion Modeling of Long Videos (2022)
10. 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: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. pp. 6626–6637 (2017)
11. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, Alexey, e.a.: Imagen Video: High Definition Video Generation with Diffusion Models (2022)
12. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual (2020)
13. Ho, J., Salimans, T.: Classifier-Free Diffusion Guidance (2022)
14. Jensen, J.: Simulation of advanced ultrasound systems using Field II. In: 2004 2nd IEEE International Symposium on Biomedical Imaging: Nano to Macro (IEEE Cat No. 04EX821). pp. 636–639 Vol. 1 (2004)
15. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. Advances in Neural Information Processing Systems **35**, 26565–26577 (2022)
16. Leclerc, S., Smistad, E., Pedrosa, J., Østvik, A., Cervenansky, F., Espinosa, F., Espeland, T., Berg, E.A.R., Jodoin, P.M., Grenier, T., et al.: Deep learning for

segmentation using an open large-scale dataset in 2d echocardiography. IEEE TMI **38**(9), 2198–2210 (2019)

17. Liang, J., Yang, X., Huang, Y., Li, H., He, S., Hu, Xindi, e.a.: Sketch guided and progressive growing GAN for realistic and editable ultrasound image synthesis. Medical Image Analysis **79**, 102461 (2022)

18. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. In: Proc. of ICLR (2022)

19. Ouyang, D., He, B., Ghorbani, A., Yuan, N., Ebinger, J., Langlotz, C.P., Heidenreich, P.A., Harrington, R.A., Liang, D.H., Ashley, E.A., et al.: Video-based ai for beat-to-beat assessment of cardiac function. Nature **580**(7802), 252–256 (2020)

20. Ouyang, D., He, B., Ghorbani, A., Yuan, N., Ebinger, J., Langlotz, Curtis P., e.a.: Video-based AI for beat-to-beat assessment of cardiac function. Nature **580**, 252–256 (2020)

21. Phi, N.V., Duc, T.M., Hieu, P.H., Long, T.Q.: Echocardiography video synthesis from end diastolic semantic map via diffusion model. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2024)

22. Reynaud, H., Qiao, M., Dombrowski, M., Day, T., Razavi, R., Gomez, A., Leeson, P., Kainz, B.: Feature-Conditioned Cascaded Video Diffusion Models for Precise Echocardiogram Synthesis, p. 142–152 (2023)

23. Salehi, M., Ahmadi, S.A., Prevost, R., Navab, N., Wein, W.: Patient-specific 3D Ultrasound Simulation Based on Convolutional Ray-tracing and Appearance Optimization. In: MICCAI. pp. 510–518 (2015)

24. Sohl-Dickstein, J., Weiss, E.A., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: Bach, F.R., Blei, D.M. (eds.) Proc. of ICML. JMLR Workshop and Conference Proceedings, vol. 37, pp. 2256–2265 (2015)

25. Stojanovski, D., Hermida, U., Lamata, P., Beqiri, A., Gomez, A.: Echo from noise: synthetic ultrasound image generation using diffusion models for real image segmentation (2023)

26. Su, X., Song, J., Meng, C., Ermon, S.: Dual diffusion implicit bridges for image-to-image translation. In: International Conference on Learning Representations (2023)

27. Tomar, D., Zhang, L., Portenier, T., Goksel, O.: Content-Preserving Unpaired Translation from Simulated to Realistic Ultrasound Images. In: MICCAI. pp. 659–669 (2021)

28. Unterthiner, T., Steenkiste, S.v., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: FVD: A new Metric for Video Generation. ICLR 2022 Workshop: Deep Generative Models for Highly Structured Data (2019)

29. Villegas, R., Babaeizadeh, M., Kindermans, P.J., Moraldo, H., Zhang, H., Saffar, Mohammad Taghi, e.a.: Phenaki: Variable Length Video Generation From Open Domain Textual Description (2022)

30. 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)