HeartBeat: Towards Controllable Echocardiography Video Synthesis with Multimodal Conditions-Guided Diffusion Models

Xinrui Zhou[1,2,3⋆], Yuhao Huang[1,2,3⋆], Wufeng Xue[1,2,3], Haoran Dou[4], Jun Cheng[1,2,3], Han Zhou[1,2,3], and Dong Ni[1,2,3(✉)]

[1]National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, School of Biomedical Engineering, Medical School, Shenzhen University, China
nidong@szu.edu.cn
[2]Medical Ultrasound Image Computing (MUSIC) Lab, Shenzhen University, China
[3]Marshall Laboratory of Biomedical Engineering, Shenzhen University, China
[4]Centre for Computational Imaging and Simulation Technologies in Biomedicine (CISTIB), University of Leeds, UK

Abstract. Echocardiography (ECHO) video is widely used for cardiac examination. In clinical, this procedure heavily relies on operator experience, which needs years of training and maybe the assistance of deep learning-based systems for enhanced accuracy and efficiency. However, it is challenging since acquiring sufficient customized data (e.g., abnormal cases) for novice training and deep model development is clinically unrealistic. Hence, controllable ECHO video synthesis is highly desirable. In this paper, we propose a novel diffusion-based framework named HeartBeat towards controllable and high-fidelity ECHO video synthesis. Our highlight is three-fold. First, HeartBeat serves as a unified framework that enables perceiving multimodal conditions simultaneously to guide controllable generation. Second, we factorize the multimodal conditions into local and global ones, with two insertion strategies separately provided fine- and coarse-grained controls in a composable and flexible manner. In this way, users can synthesize ECHO videos that conform to their mental imagery by combining multimodal control signals. Third, we propose to decouple the visual concepts and temporal dynamics learning using a two-stage training scheme for simplifying the model training. One more interesting thing is that HeartBeat can easily generalize to mask-guided cardiac MRI synthesis in a few shots, showcasing its scalability to broader applications. Extensive experiments on two public datasets show the efficacy of the proposed HeartBeat.

Keywords: Echocardiography · Controllable synthesis · Diffusion model

1 Introduction

Cardiac ultrasound (US), i.e., echocardiography (ECHO), is widely utilized to evaluate cardiac function and diagnose cardiovascular diseases. Compared to

⋆ Xinrui Zhou and Yuhao Huang contribute equally to this work.

other modalities, US has the advantages of real-time imaging, radiation-free, and affordability in clinical practice [26]. ECHO is a dynamic examination that heavily relies on operator experience. It usually takes more than 10 years to train qualified sonographers. Recent deep learning-based systems have shown efficacy in automatic cardiac diagnosis, which can assist sonographers in clinical analysis [13]. Collecting abundant ECHO videos tailored to specific anatomy composition (e.g., plane types, specific mitral valve (MV) motion directions, etc.) or abnormal cases has the potential to accelerate novice training and intelligent system development. However, collecting adequate customized ECHO sequences is impractical in clinical scenarios [12]. Hence, building a framework for controllable ECHO video synthesis is greatly desired to solve the data scarcity issue.

ECHO video synthesis is challenging due to the US speckle noise, complex motion trajectories (e.g., mitral valve motion), and varying sizes of anatomical structures. Recently, driven by the significant progress of the Denoising Diffusion Probabilistic Models (DDPMs) [7], several DDPM-based approaches have been proposed to synthesize ECHO videos. Stojanovski et al. [20] synthesized 2D ECHO videos with the guidance of semantic masks. Though generating realistic images, it focused on static image synthesis instead of dynamic sequences, limiting its clinical practice. Reynaud et al. [15] proposed a cascaded video diffusion model to generate ECHO videos conditioned on the end-diastolic frame and left ventricle ejection fractions. Van et al. [23] introduced a single semantic mask condition to compose ECHO videos. However, most of these methods show poor controllability and flexibility due to restricted control signals. Moreover, they directly utilized 3D diffusion architectures, resulting in substantial computational cost and training difficulty. Thus, they may not be suitable for clinical practice.

Most recently, studies have been proposed to achieve controllable video synthesis. They can be coarsely classified into two types. **(1) Text-to-Video (T2V) Synthesis.** Most approaches [1, 6, 18, 32] took the random Gaussian noise and text prompt as inputs and learned both visual concepts and temporal dynamics. However, these methods typically fell short in controlling the visual appearance and geometric structure of the generated videos due to the lack of fine-grained conditional controls. **(2) Multiple Conditions-guided T2V Synthesis.** Compared to pure T2V synthesis, these methods enabled controllable and precise synthesis driven by multiple conditions. For facilitating the global controls regarding visual appearance, AnimateDiff [4] and MoonShot [28] conditioned the synthesis on both image and text inputs simultaneously. However, these methods overlooked fine-grained controls, limiting their applicability in the medical field. For local controls related to geometric structure, ControlVideo [30] reused the pretrained ControlNet [29] that was designed for conditional image generation to achieve controllable video synthesis. Wang et al. [24] proposed VideoComposer to achieve multiple conditions-guided T2V synthesis in a composable fashion. However, it required high training costs by introducing numerous temporal layers to model the time-series knowledge of conditions.

In this study, we propose a novel framework called HeartBeat to achieve controllable and high-fidelity ECHO video synthesis. We believe that this is the

first exploration of highly customized US video synthesis based on the guidance of multimodal conditions. Our contributions are three-fold. First, we introduce HeartBeat, a uniform framework that enables perceiving versatile conditions simultaneously to guide controllable video generation. Second, we factorize the multimodal conditions into local (i.e., structural cues like sketches, hand-crafted motion direction of MV, etc.) and global (i.e., plane types and image priors) parts. For instance, sketches allow for flexible edits in anatomical structures and hand-crafted strokes of MV make its motion controllable. Moreover, generating abnormal cases becomes feasible due to the transfer of visual patterns in image priors. Two corresponding insertion methods are further proposed to provide fine- and coarse-grained controls, respectively. Note that all conditions can be manipulated in a composable and flexible manner. In this case, users can seamlessly combine multimodal conditions to produce ECHO videos. Third, we propose a two-stage training scheme that decouples the visual concepts and temporal dynamics learning to ease the model training. Furthermore, we validate the generalization ability of HeartBeat by transferring ECHO video synthesis into 3D cardiac MRI (CMR) generation using few-shot learning. Extensive experimental results prove the effectiveness of HeartBeat.

2 Methodology

Fig. 1 shows the pipeline of our proposed HeartBeat. It supports composable and customized ECHO video synthesis via multimodal conditions. HeartBeat involves a two-stage training scheme design. In the pretraining stage, HeartBeat focuses on high-quality *visual concepts learning* for controllable text-to-image (T2I) generation. While in the finetuning stage, the domain-specific knowledge gained from the pretext task is reused to facilitate *temporal dynamics modeling* for customized text-to-video (T2V) generation. During inference, given a group of multimodal prompts as conditional inputs, HeartBeat enables customizable ECHO video synthesis from Gaussian noise.

2.1 Preliminaries of T2V DDPMs.

T2V DDPMs are generative models developed upon traditional DDPMs [7, 19]. They are trained to learn video distribution by iteratively recovering noisy inputs with the guidance of text prompts. To reduce the computational burden caused by pixel-space training, latent diffusion models (LDMs) that operate in the latent space are introduced in the video domain for perceptual video compression. In this way, the optimized objective is adapted to $\min_\theta \mathbb{E}_{z_0,\epsilon,\mathbf{c},t}[||\epsilon - \epsilon_\theta(z_t, \mathbf{c}, t)||_2]$, where z_0 is the latent representation of the training video from a pretrained variational autoencoder (VAE) [3]. t stands for the time step. ϵ_θ and ϵ represent the predicted and target noise, respectively. For our model, $\mathbf{c}$ refers to multimodal conditions including local and global conditions.

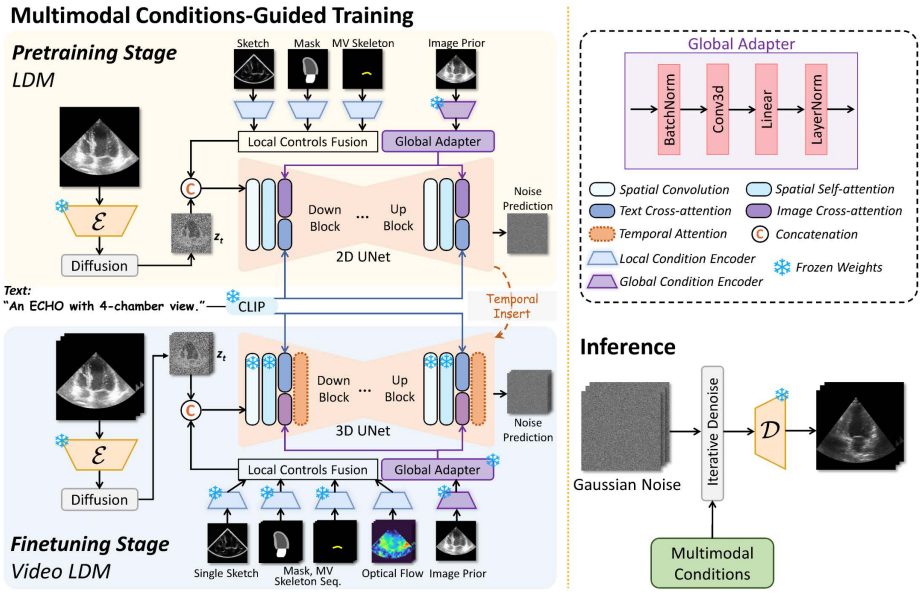

Fig. 1. Pipeline of HeartBeat. $\mathcal{E}, \mathcal{D}$ denote pretrained encoder and decoder in VAE, respectively. z_t refers to latent features. LDM, latent diffusion model. MV, mitral valve.

2.2 HeartBeat

Decoupled learning of visual concepts and temporal dynamics. Directly employing video LDMs (VLDMs) is an intuitive way to achieve ECHO video synthesis. However, it is challenging to simultaneously model visual appearance and temporal variation driven by scarce medical data, while producing high training costs. Thus, in HeartBeat, we first pretrain an LDM and extend it to a VLDM for controllable video generation. Note that the VLDM is an inflation of the 2D counterpart over the spatial-temporal dimension. Leveraging the decoupling training scheme, HeartBeat can achieve high-fidelity and temporal-coherent ECHO video synthesis, while reducing training difficulty and computational burden.

2D-to-3D Model Inflation for Video Synthesis. In the pretraining stage, following LDM [16], we employ a 2D UNet [17] as the denoising network for controllable image generation. Typical UNet in LDMs is composed of stacked blocks, where each contains a spatial convolution layer, a spatial self-attention layer, and a cross-attention layer that controls the synthesis by texts. Motivated by [8,27], we extend the 2D UNet to the 3D counterpart using a simple inflation strategy. Specifically, we inflate all the spatial convolution layers at the temporal dimension with *t=1*. To model the temporal dynamics, we insert a temporal self-attention layer following the cross-attention layer in each block. This design

ensures the feature distribution of the spatial layers will not be altered significantly [28]. Thus, HeartBeat can reuse the rich visual concepts regarding ECHO video patterns preserved in LDM and focus on temporal features integration.

Controllable Video Synthesis with Multimodal Conditions. It is well acknowledged that pure T2I/T2V models suffer from limited control of spatial composition [29]. Generating customized ECHO videos of normal and abnormal cases that precisely match their mental imagery is crucial for junior sonographers to improve their reading and diagnostic skills. To achieve this, our proposed HeartBeat introduces multimodal conditions to guide customizable ECHO video synthesis in a flexible and composable manner.

As shown in Fig. 1, the overall generation process is comprehensively controlled by employing multimodal control signals. In this case, HeartBeat allows users to flexibly select/edit any available single condition or their combination to guide high-fidelity, temporal-coherent, and customized ECHO video synthesis. In our study, we involve six multimodal conditions and factorize them into four local conditions and two global conditions for fine-grained and coarse-grained controls, respectively. To fully perceive multimodal conditions at local and global dimensions, we separately design condition injection methods for two conditions. Meanwhile, HeartBeat assures the composability of different conditions by adopting such injection strategies. Note that LDM and VLDM in the proposed HeartBeat share identical conditions injection methods.

1) Local Conditions Guidance for Fine-grained Control. To achieve fine-grained geometric structure guidance, we apply *sketch*, *semantic mask*, and *skeleton of mitral valve* as three local spatial conditions to perform controllable image generation. To improve the temporal consistency of generated videos, in VLDM, we extend the above local spatial conditions to time-series ones except for *sketch*. Optical flow is also employed as an additional local condition to explicitly indicate the pixel-wise movements between adjacent frames (see Fig. 1). We propose a **local condition injection method**, which consists of a lightweight encoder [3] for embedding each local condition and a fusion operation between conditions and noisy latents z_t. Specifically, each condition is first encoded in parallel, allowing HeartBeat to capture local spatial knowledge. The obtained local conditional features are then fused by element-wise addition. Last, such features are concatenated with z_t along channel dimension to form local control signals. Note that each encoder shares the same architecture.

2) Global Conditions Perception for Coarse-grained Control. Towards highly-customized video synthesis solely guided by local conditions has a limitation, that is, the synthetic results lack diversity due to the various fine-grained conditional inputs. However, acquiring sufficient diverse ECHO videos is desirable for novices undergoing clinical training or for deep learning systems development. Thus, we also consider two global conditions, i.e., *text* and *image prior* to balance the trade-off between highly-customized and diverse generations (see Fig. 1). The text offers an intuitive indication in terms of coarse-grained visual content, while the latter provides richer information beyond text

Table 1. Quantitative results of methods. Conditional controls involve text (T), sketch (S), image prior (I), mask (M), the skeleton of the mitral valve (MV), and optical flow (O). Note that "HeartBeat" and "VideoComposer" use the joint training strategy, while other models merely use selected conditions for training.

Methods	Controls						A2C			A4C		
	T	S	I	M	MV	O	FID↓	FVD↓	SSIM↑	FID↓	FVD↓	SSIM↑
MoonShot [28]	✓		✓				48.44	12.65	0.63	61.57	18.13	0.62
VideoComposer [24]	✓	✓	✓	✓	✓	✓	37.68	10.31	0.60	35.04	11.27	0.61
HeartBeat-3D	✓			✓			54.63	13.64	0.60	41.07	14.01	0.60
HeartBeat-Base	✓			✓			23.57	6.97	0.65	31.08	10.80	0.64
HeartBeat	✓						107.66	19.07	0.53	76.46	23.83	0.53
	✓		✓				36.34	9.60	0.61	36.00	12.58	0.61
	✓	✓	✓				27.38	7.04	0.66	35.42	13.48	0.64
	✓	✓	✓	✓			26.30	6.74	0.66	33.77	12.59	0.64
	✓	✓	✓	✓	✓		25.98	6.94	0.66	33.98	12.00	0.64
	✓	✓	✓	✓	✓	✓	25.23	6.08	0.66	31.99	9.96	0.65

(e.g., visual patterns regarding abnormal disease). We develop a **global condition injection method** to align the global image prior embeddings f_i extracted from the pretrained MedSAM image encoder [10] with the text embeddings f_t derived from CLIP text encoder [14]. For the global conditions integration, an intuitive way is to directly input concatenated features of both into the frozen cross-attention layers. However, it will hinder the model from capturing visual patterns from the image prior. To this end, we replace the original text-guided cross-attention layer in each UNet block with a factorized cross-attention layer for handling texts and image priors. In this layer, two separate Query-Key-Value (QKV) projections are added and optimized for both conditions, respectively. Take the proposed VLDM as an example, this attention process can be defined as: $\mathit{CrossAttention}(\mathbf{Q}^T, \mathbf{K}^T, \mathbf{V}^T) + \mathit{CrossAttention}(\mathbf{Q}^I, \mathbf{K}^I, \mathbf{V}^I)$, where $Q^T, Q^I \in \mathbb{R}^{BF\times H\times W\times C}$, $K^T, K^I, V^T, V^I \in \mathbb{R}^{BF\times N\times C}$, with B the batchsize, F the length of frames, H the height, W the width, N the token numbers of text (T) and image prior condition (I), and C the hidden dimension. Besides, we propose a learnable global adapter to align f_i and f_t.

3 Experimental Results

Dataset and Implementations. We validated HeartBeat on the public CAMUS dataset [11]. It consists of 900 ECHO videos collected from 450 patients. Videos labeled with poor quality in [11] were excluded for reliability. Finally, 884 videos with 431 apical two-chamber (A2C) and 453 apical four-chamber (A4C) views were included. The dataset was split randomly into 793 and 91 videos with 16 frames for training and testing at the patient level. We set text prompts (i.e., "An ECHO with 2/4-chamber view.") for all videos according to the actual view.

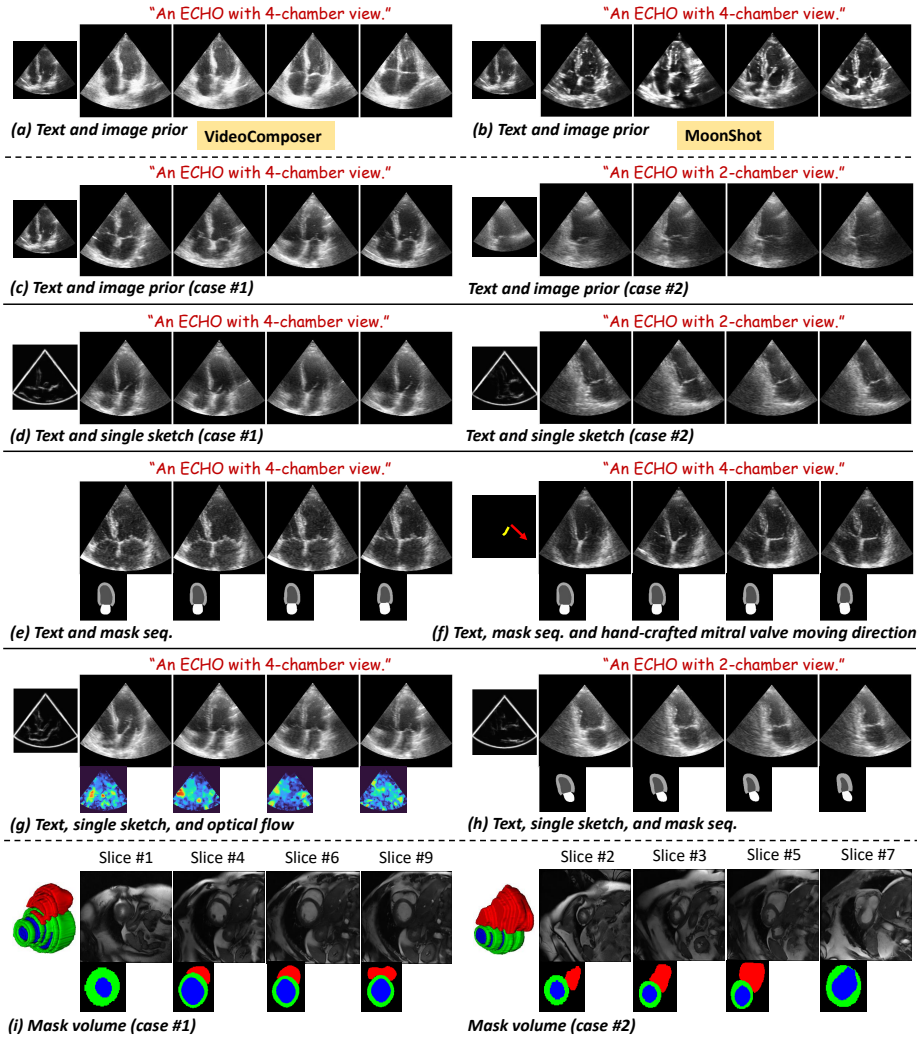

Fig. 2. Typical results of HeartBeat and two most related approaches (first row).

Sketches were extracted by PiDiNet [21]. For few-shot generalization validation, we employed 50 CMR volumes from M&Ms Challenge [2] as the training set.

In this study, we implemented HeartBeat in *Pytorch*, using four NVIDIA A6000 GPUs. All frames/slices were resized to 256×256. During pretraining, our LDM was developed upon Stable Diffusion [16] and initialized using the public pretrained weights[1]. During finetuning, we froze the spatial weights except for the newly added optical flow encoder and kept the temporal layers trainable.

[1] https://huggingface.co/CompVis/stable-diffusion-v1-4

We updated the query projection in cross-attention layers to refine the text- and image prior-video alignment. We set the batch size as 64 and 16 for the first and second stage training. During the whole training, we adopted the all conditions joint training strategy [9]. In this way, HeartBeat was not required to be finetuned for each unique combination of multimodal conditions every time and enables flexibly dropping several conditions during inference. The CMR synthesizer was merely conditioned on mask volumes, and initialized with US-trained HeartBeat. The learning rate was initialized to 1e-4 after 500 steps of warm-up strategy and decayed by a cosine annealing scheduler. All models were trained using the Adam optimizer for 200 epochs. We chose models of the last epoch to work with HeartBeat.

Quantitative and Qualitative Analysis. We validated the generative performance of HeartBeat on achieving various tasks in a controllable manner. Conditions can be tailored to meet the specific needs of users by leveraging the composable control capabilities of HeartBeat. In this study, three metrics were used to evaluate the performance, including 1) Fréchet Inception Distance (FID) [5] for image quality evaluation in feature level, 2) Fréchet Video Distance (FVD) [22] for visual quality and temporal consistency assessment in video level, and 3) Structure Similarity Index (SSIM) score [25] to assess the controllability [31].

Fig. 2 shows the qualitative results: **1) Image prior-controlled ECHO video synthesis.** Fig. 2(c) shows synthesizing realistic videos driven by global knowledge (i.e., coarse-grained visual patterns) of image priors. Given priors with normal (*left*) and left ventricular hypertrophy (*right*), HeartBeat generates videos with the same disease status accordingly, allowing the normal/abnormal transfer abilities of the model. **2) Sketch-controlled ECHO video synthesis.** HeartBeat is able to animate static sketch for generating realistic videos (Fig. 2(d)). **3) Mitral valve motion-controlled ECHO video synthesis.** Observation of MV motion is clinically important. With the simple MV skeleton, masks, and hand-crafted strokes indicating the motion direction, i.e., red arrow in Fig. 2(f), HeartBeat enables precise motion control compared to the same case with only mask condition (Fig. 2(e)). Such features endow users with ease of use, flexibility, and high controllability when using HeartBeat, showcasing the potential for clinical applications. **4) Various conditions-controlled ECHO video synthesis.** Fig. 2(g-h) demonstrates the superior controllability and generative quality of HeartBeat. **5) Generalize to 3D CMR synthesis.** Thanks to the flexibility of control signals manipulation, HeartBeat enables controllable CMR synthesis with few-shot tuning solely guided by the 3D mask (Fig. 2(i)). It shows that the generated anatomy areas are highly consistent with the masks. Besides, HeartBeat produces sharper frames with high fidelity and coherence than others (Fig. 2(a-c)). The quantitative results in Table 1 are in line with the qualitative ones. Notably, the "HeartBeat-Base" model with a two-stage training scheme performs better than that with pure 3D one-stage training using VLDM (i.e., HeartBeat-3D), showcasing the efficacy of our training scheme. Ablation studies (last 6 rows) show that adding more conditions enhances generative performance.

4 Conclusion

In this study, we propose a novel diffusion-based HeartBeat framework for controllable and flexible ECHO video synthesis. Specifically, HeartBeat is driven by multimodal control signals in a composable fashion, including local and global conditions to separately provide fine- and coarse-grained guidance. Extensive experiments on two datasets validate the powerful controllability and generality of HeartBeat, showing its clinical practicality. In the future, we will extend HeartBeat to more challenging datasets for further generality validation.

Acknowledgments. This work was supported by the grant from National Natural Science Foundation of China (Nos. 12326619, 62171290), Science and Technology Planning Project of Guangdong Province (No. 2023A0505020002), Shenzhen Science and Technology Program (No. SGDX20201103095613036), Natural Science Foundation of Guangdong Province (No. 2024A1515030143), and Shenzhen Science and Technology Program (No. 20220810145705001).

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

References

1. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., et al.: Align your latents: High-resolution video synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22563–22575 (2023)
2. Campello, V.M., Gkontra, P., Izquierdo, C., Martin-Isla, C., Sojoudi, A., Full, P.M., et al.: Multi-centre, multi-vendor and multi-disease cardiac segmentation: the m&ms challenge. IEEE TMI **40**(12), 3543–3554 (2021)
3. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021)
4. Guo, Y., Yang, C., Rao, A., Wang, Y., Qiao, Y., Lin, D., et al.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725 (2023)
5. 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. Advances in neural information processing systems **30** (2017)
6. Ho, J., Chan, W., Saharia, C., Whang, J., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022)
7. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
8. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. NeurIPS **35**, 8633–8646 (2022)
9. Huang, L., Chen, D., Liu, Y., et al.: Composer: Creative and controllable image synthesis with composable conditions. arXiv preprint arXiv:2302.09778 (2023)
10. Huang, Y., Yang, X., Liu, L., Zhou, H., Chang, A., Zhou, X., et al.: Segment anything model for medical images? Medical Image Analysis **92**, 103061 (2024)

11. Leclerc, S., Smistad, E., Pedrosa, J., Østvik, A., Cervenansky, F., Espinosa, F., et al.: Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. IEEE TMI **38**(9), 2198–2210 (2019)
12. Liang, J., Yang, X., Huang, Y., Liu, K., Zhou, X., Hu, X., et al.: Weakly-supervised high-fidelity ultrasound video synthesis with feature decoupling. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 310–319. Springer (2022)
13. Ouyang, D., He, B., Ghorbani, A., Yuan, N., Ebinger, J., et al.: Video-based ai for beat-to-beat assessment of cardiac function. Nature **580**(7802), 252–256 (2020)
14. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., , et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763. PMLR (2021)
15. Reynaud, H., Qiao, M., Dombrowski, M., Day, T., Razavi, R., Gomez, A., et al.: Feature-conditioned cascaded video diffusion models for precise echocardiogram synthesis. In: MICCAI. pp. 142–152. Springer (2023)
16. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
17. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
18. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., et al.: Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792 (2022)
19. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. PMLR (2015)
20. Stojanovski, D., Hermida, U., Lamata, P., Beqiri, A., Gomez, A.: Echo from noise: synthetic ultrasound image generation using diffusion models for real image segmentation. In: International Workshop on Advances in Simplifying Medical Ultrasound. pp. 34–43. Springer (2023)
21. Su, Z., Liu, W., Yu, Z., Hu, D., Liao, Q., Tian, Q., et al.: Pixel difference networks for efficient edge detection. In: ICCV. pp. 5117–5127 (2021)
22. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges. arXiv preprint arXiv:1812.01717 (2018)
23. Van Phi, N., 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). pp. 13461–13465. IEEE (2024)
24. Wang, X., Yuan, H., Zhang, S., Chen, D., Wang, J., Zhang, Y., et al.: Videocomposer: Compositional video synthesis with motion controllability. Advances in Neural Information Processing Systems **36** (2024)
25. 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)
26. Wei, H., Cao, H., Cao, Y., Zhou, Y., Xue, W., Ni, D., et al.: Temporal-consistent segmentation of echocardiography with co-learning from appearance and shape. In: MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23. pp. 623–632. Springer (2020)
27. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., et al.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In: ICCV. pp. 7623–7633 (2023)

28. Zhang, D.J., Li, D., Le, H., Shou, M.Z., Xiong, C., Sahoo, D.: Moonshot: Towards controllable video generation and editing with multimodal conditions. arXiv preprint arXiv:2401.01827 (2024)
29. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV. pp. 3836–3847 (2023)
30. Zhang, Y., Wei, Y., Jiang, D., Zhang, X., et al.: Controlvideo: Training-free controllable text-to-video generation. arXiv preprint arXiv:2305.13077 (2023)
31. Zhao, S., Chen, D., Chen, Y.C., Bao, J., Hao, S., Yuan, L., et al.: Uni-controlnet: All-in-one control to text-to-image diffusion models. Advances in Neural Information Processing Systems **36** (2024)
32. Zhou, D., Wang, W., Yan, H., Lv, W., Zhu, Y., Feng, J.: Magicvideo: Efficient video generation with latent diffusion models. arXiv preprint arXiv:2211.11018 (2022)