Synthesis of Pathological Dual-Channel Color Doppler Echocardiograms for Equitable Diagnosis of Heart Diseases

Pooneh Roshanitabrizi[1][0000-0002-8853-2786]*, Pengfei Guo[2], Artur Arturi Aharonyan[1], Kelsey Brown[3], Taylor Gloria Broudy[1], Abhijeet Parida[1], Austin Tapp[1], Zhifan Jiang[1], Alison Tompsett[3], Joselyn Rwebembera[4], Emmy Okello[4], Andrea Beaton[5], Holger R. Roth[2], Daguang Xu[2], Syed Muhammad Anwar[1], Craig A. Sable[3,6], Marius George Linguraru[1,7]**

[1] Sheikh Zayed Institute for Pediatric Surgical Innovation, Children's National Hospital, Washington, DC, USA
[2] NVIDIA Corporation, Santa Clara, CA, USA
[3] Division of Cardiology, Children's National Hospital, Washington, DC, USA
[4] Uganda Heart Institute, Kampala, Uganda
[5] Department of Pediatric Cardiology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA
[6] Ochsner Children's Hospital, New Orleans, LA, USA
[7] Departments of Radiology and Pediatrics, George Washington University, School of Medicine and Health, Washington, DC, USA
*proshnani2@childrensnational.org,
**mlingura@childrensnational.org

Abstract. Rheumatic heart disease (RHD) is the leading global cardiac condition, affecting over 54 million people, predominantly in resource-constrained countries. Early detection via color Doppler echocardiography is crucial but often inaccessible due to reliance on specialized cardiologists. Consequently, such data from patients diagnosed with RHD are scarce. To address data limitations in developing robust RHD detection methods, we propose a novel AI-driven approach to synthesize color Doppler echocardiograms with matched B-mode ultrasound using a multi-factor conditioned diffusion model. To our knowledge, this is the first generative AI design for dual-channel color Doppler synthesis. Our model enhances realism by incorporating temporal information for motion consistency and class label for targeted synthesis. We use B-mode ultrasound to visualize anatomical structures and the Doppler-mode fields of view to define blood flow regions across key echocardiographic views (e.g., parasternal and apical). We synthesize one echocardiographic mode from another using cross-view translation to augment data and improve diversity. We evaluated our approach using synthetic data generated from echocardiograms of 589 Ugandan cases and the public CAMUS dataset. Our model outperformed state-of-the-art generative methods in fidelity and structural similarity. We trained and tested an RHD classifier on limited data from different devices. Training with synthetic data significantly improved detection performance compared to a model trained only on real data. These findings highlight the potential of diffusion-based synthetic data to democratize the diagnosis of heart diseases in marginalized populations and low-resource settings. Our approach is scalable, promotes health equity, and contributes to RHD prevention and reduced mortality.

Keywords: Classification, Color Doppler Echocardiogram, Conditional Diffusion Model, Dual-Channel Synthesis, Health Equity, Rheumatic Heart Disease.

1 Introduction

Rheumatic heart disease (RHD) is a major global health concern, affecting over 54 million people and causing more than 300,000 deaths annually—primarily in low-resource regions [1]. RHD is frequently diagnosed late, requiring complex surgical intervention unavailable in endemic regions [2]. Color Doppler echocardiography is essential for early RHD identification, as it visualizes blood flow direction with RGB overlays on B-mode (grayscale) ultrasound. B-mode provides detailed cardiac anatomy, highlighting structures such as ventricular walls, valves, chamber size, wall motion, and structural abnormalities. In screening sonography, RHD often presents as mitral regurgitation (MR), characterized by retrograde blood flow into the left atrium during ventricular systole, with a distinct blue jet. Nevertheless, this valuable diagnostic tool faces challenges in resource-constrained settings where skilled physicians are rarely available to interpret the data [3]. RHD develops during childhood and can easily go undetected. There are also significant challenges in developing technology to automatically detect RHD, including difficulties in curating data from vulnerable patients and from low-income areas, where the disease is endemic and access to medical technology and healthcare is limited. As a result, echocardiograms, including color Doppler imaging from children with RHD are rarely available, constraining the development of machine learning methods to aid automated analyses and detection.

Generative AI can enhance small datasets for training machine learning models, but no existing model adequately synthesizes color Doppler overlaid with B-mode ultrasound, as needed for RHD detection. Techniques like generative adversarial networks (GANs) and diffusion models (DMs) have been used for ultrasound synthesis [4–8], primarily for B-mode images, except for [4], which focuses solely on color Doppler. GANs face challenges like hallucinated anatomical structures and mode collapse [9, 10], whereas DMs can generate synthetic images/videos with domain-specific conditioning [11, 12]. Recent diffusion-based echocardiographic synthesis has explored semantic label-guided generation [5], privacy-preserving video diffusion [6], and dual-conditioned fetal ultrasound synthesis [7]. Zhou et al. [8] introduced a multimodal diffusion framework that integrates local (e.g., mitral valve motion) and global (e.g., image priors) conditions to enhance fine- and coarse-grained control and temporal coherence. However, this computationally intensive approach does not synthesize color Doppler. Sun et al. [13] proposed a numerical framework that combines patient-specific computational fluid dynamics with an ultrasound simulation environment to generate synthetic B-mode and color Doppler images. While effective at simulating Doppler artifacts (e.g., clutter, aliasing), it does not model pathological cases.

We present the first AI-driven method for generating clinically realistic, pathological Doppler–B-mode image pairs, leveraging multi-factor conditioning to enhance realism and diagnostic utility. Key contributions include: (1) AI-driven synthesis of color Doppler echocardiography paired with B-mode ultrasound; unlike the physics-based method in [13] or B-mode-only approaches such as [8], our data-driven method synthesizes both B-mode and color Doppler images—either individually or as paired sets—with or without pathology; (2) a dual-channel conditional DM to address data scarcity (e.g., RHD); (3) a 3.5D-DM framework with multi-factor conditioning, where "3.5D" refers to a hybrid temporal model for ultrasound videos that combines 2.5D modeling for Doppler and 1D for B-mode to capture modality-specific spatiotemporal dynamics; (4) cross-view translation to enhance data diversity and generalization by transforming echocardiographic views (e.g., apical to parasternal); and (5) validation on a downstream RHD detection task, demonstrating potential for equitable diagnosis in low-resource settings. This approach addresses data limitations, advancing AI-assisted cardiac diagnostics and accessibility.

2 Datasets

We trained our approach on a private dataset (D1) and evaluated it on D1, an independent private dataset (D2), and the publicly available CAMUS (CA) dataset [14]. Both D1 and D2 were IRB-approved by Children's National Hospital (#00010408). (1) D1 consists of 462 pediatric echocardiographic cases (ages 5–17, mean: 12 ± 3 years) from Uganda, where RHD is endemic. Data were acquired using GE Vivid Q/IQ ultrasound machines (5 MHz transducer) in apical 4-chamber (A4CC, 451 videos) and parasternal long-axis (PLAXC, 462 videos) views, both with color Doppler. Each clip has an average resolution of 592×817 pixels and 40 ± 18 frames. Expert cardiologists classified 207 cases as normal and 255 as RHD-positive (195 borderline, 44 definite, 16 severe) based on MR presence. Fig. 1 shows paired A4CC and PLAXC views for normal and borderline RHD cases. (2) D2 is an independent dataset of 127 Ugandan cases (ages 15–75), including 188 PLAXC and 117 A4CC videos. Of these, 79 cases were RHD-positive with MR, and 48 were normal. Data were collected using a Terason ultrasound machine (Terason Inc., Burlington, MA, USA), with an average resolution of 763×1033 pixels and 70 ± 21 frames per clip. Expert cardiologists confirmed all diagnoses. (3) CA is a publicly available dataset containing 500 echocardiographic cases collected using a GE Vivid E95 ultrasound system. Each case includes apical 4-chamber and 2-chamber B-mode views, with manual annotations of cardiac structures at end-diastole and end-systole. We used the apical 4-chamber data to evaluate synthetic color Doppler echocardiograms in the A4CC view and translate them to the PLAXC view.

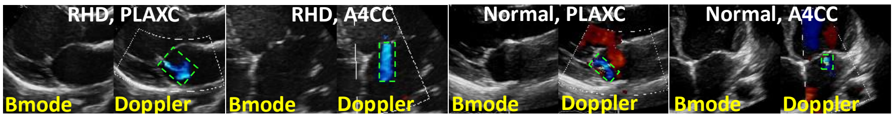


Fig. 1. Paired A4CC and PLAXC views in B-mode and color Doppler during ventricular systole, showing the MR jet (rectangle) in a borderline RHD case (left) and a normal case (right).

Preprocessing: Matched B-mode and color Doppler echocardiograms were temporally aligned using maximum cross-correlation and spatially registered via affine transformation using the ANTs toolbox [15]. Cropping was performed with square bounding boxes designed to retain essential flow features in the color Doppler images. These cropping regions were defined based on manually identified Doppler-mode fields of view (FOVs) and applied uniformly across all frames in each video to ensure spatial and temporal coherence.

3 Methods

Our proposed framework, illustrated in Fig. 2, synthesized data using a 3.5D-conditional DM, followed by sampling and quality checking during inference. The high-quality synthesized samples were then used in a downstream task for RHD detection.

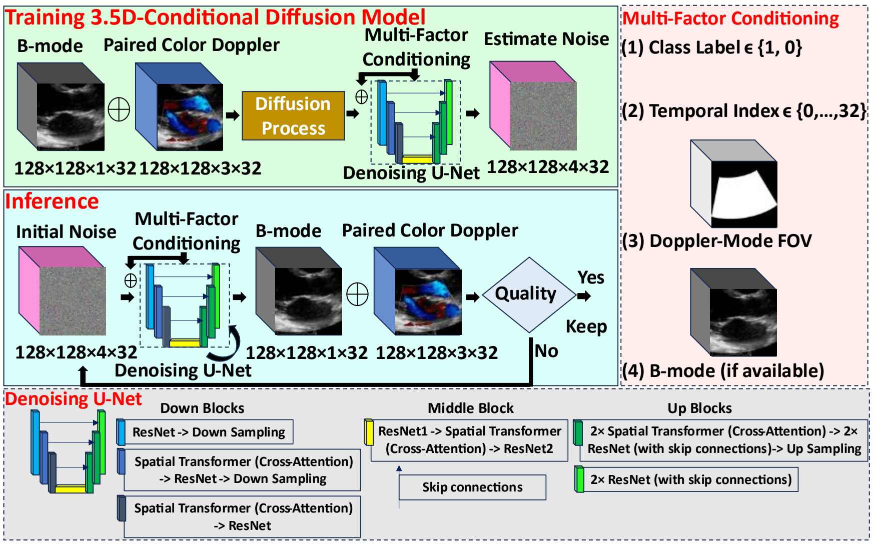

Training 3.5D-Conditional Diffusion Model
B-mode
Paired Color Doppler
Multi-Factor Conditioning
Estimate Noise
Diffusion Process
128×128×1×32
128×128×3×32
Denoising U-Net
128×128×4×32
Inference
Initial Noise
Multi-Factor Conditioning
B-mode
Paired Color Doppler
Quality
Yes
Keep
No
128×128×4×32
Denoising U-Net
128×128×1×32
128×128×3×32
Multi-Factor Conditioning
(1) Class Label ϵ {1, 0}
(2) Temporal Index ϵ {0,...,32}
(3) Doppler-Mode FOV
(4) B-mode (if available)
Denoising U-Net
Down Blocks
ResNet -> Down Sampling
Spatial Transformer (Cross-Attention) -> ResNet -> Down Sampling
Spatial Transformer (Cross-Attention) -> ResNet
Middle Block
ResNet1 -> Spatial Transformer (Cross-Attention) -> ResNet2
Skip connections
Up Blocks
2× Spatial Transformer (Cross-Attention) -> 2× ResNet (with skip connections)-> Up Sampling
2× ResNet (with skip connections)

Fig. 2. Overview of dual-channel color Doppler echocardiogram synthesis using a 3.5D-DM, conditioned on class label, temporal index, Doppler-mode FOV, and optional B-mode input.

3.1 3.5D-Conditional Diffusion Model

Our approach employed a conditional denoising diffusion probabilistic model (DDPM) [16] to generate paired B-mode and color Doppler echocardiograms. When available, the model conditioned on B-mode ultrasound images to synthesize the corresponding color Doppler echocardiograms from pure noise. If B-mode was unavailable, the model synthesized both B-mode and color Doppler echocardiograms from noise alone. Additional conditioning factors, including class label, temporal index, and Doppler-mode FOV, further guided the synthesis process. During training, a U-Net-based denoiser progressively refined noise-corrupted echocardiographic images by estimating and removing Gaussian noise. The U-Net architecture had three layers with channels (128, 256, 256), each containing residual blocks, and attention in the last two layers (Fig. 2). The model processed 128×128-pixel inputs over 32-frame videos, with image and frame sizes selected to accommodate memory constraints. B-mode conditioning was optional and could originate from the same or different views to depict anatomical structure. The denoising U-Net function is defined as $f_\theta\ (x_t, t, y_{cls}, y_{temp}, M_{Bmode}, M_{FOV})$, where x_t is the noisy image at time step t, M_{Bmode} denotes the B-mode ultrasound image used as a conditional input, and M_{FOV} corresponds to the Doppler-mode FOV. Class label y_{cls} (e.g., 0: Normal, 1: RHD, it can be adapted to other pathologies) was embedded into the model using a learned embedding to condition the generation process. The temporal index y_{temp} was normalized by the total number of diffusion timesteps and used to condition multiple U-Net stages via the time embedding network. B-mode and Doppler-mode FOV images were concatenated with the input echocardiograms to provide image conditioning for the DM. The forward diffusion process followed a variance-preserving stochastic model, gradually adding Gaussian noise at each timestep. The model learned to reverse this process by predicting and removing the noise component to reconstruct the original image x_0 from x_t. Training minimized the mean squared error between the predicted and actual noise, using the Adam optimizer (learning rate: 2.5e^{-5}, batch size: 1) over 200 epochs.

To improve sampling quality, inference-time computation was necessary, as highlighted in [17]. In our RHD detection application, overlap between normal and diseased distributions (Fig. 1) required iterative sampling and validation to ensure sample quality prior to classification. Multiple samples were generated and evaluated using a pre-trained RHD classifier, which was used solely for filtering and not reused in downstream evaluation, ensuring no performance inflation. Samples were retained if the predicted label matched the conditioned class and included at least six frames during ventricular systole—the dataset average. Otherwise, the process was repeated for up to 10 iterations. If inconsistencies persisted, an expert manually reviewed the data to ensure clinical relevance and adherence to classification criteria. The misclassified samples were then added to the training data to improve model robustness.

3.2 RHD Detection

We demonstrated the utility of our general synthesis model for detecting a neglected disease with limited data, specifically RHD. A separate classifier was trained from scratch for RHD classification. We selected the PLAXC view due to its ease of acquisition and minimal expertise requirement, making it practical for healthcare workers with limited echocardiography training. Our end-to-end network first analyzed ventricular systolic frames—when the heart contracts and MR, a key RHD indicator, is most visible—using a ResNet-50 model. The selected systolic frames were then processed by an 18-layer 3D residual network (R3D-18) [18] to classify RHD presence. The R3D-18 network consisted of 3D convolutional layers, four dense blocks, adaptive average pooling, and a fully connected layer for final classification. Both models were trained jointly for 150 epochs using the Adam optimizer (learning rate of 10^{-4}) and a binary cross-entropy loss function, with a batch size of 3.

3.3 Validation and Performance Metrics

Our evaluation included two key components: (1) synthesized data quality assessment using Fréchet inception distance (FID) [19], Fréchet video distance (FVD) [20], and structural similarity index measure (SSIM) [21], and (2) RHD detection performance based on accuracy, sensitivity, and specificity. FID was measured using a pre-trained ResNet50 on RadImageNet [22], while FVD used a pre-trained R3D-18 on dataset D1. Since this is the first image synthesis framework for color Doppler overlaid with B-mode ultrasound, no existing benchmarks are available for FID, FVD, or SSIM. We conducted ablation studies to evaluate the impact of each conditioning factor on the synthesis process. Our model was compared against 3D LDMs (adapted from [23]) and the latent video DM (LVDM) [6]. These methods were originally developed for 3D brain MRI and grayscale B-mode datasets; we adapted them for color Doppler. The RHD model was trained on 462 real cases from D1, tested on 188 real PLAXC videos from D2, and compared to a model trained on 1,414 synthesized PLAXC samples (480 from D1, 466 from D2, and 468 from CA). To evaluate the impact of synthetic data, we started with 462 samples—matching the real dataset—and progressively increased the number (e.g., to 1,414) to enable significant gains in model performance.

4 Experimental Results

We implemented the network using PyTorch v2.5.1 and MONAI v1.4.0 [24] on an NVIDIA H100 80GB GPU. Training took 516 minutes for conditional DM and 850 minutes for RHD detection. Inference times were 6.5 minutes per video (synthesis) and <1 minute for RHD detection on a CPU laptop. RHD model parameters were optimized over 80–200 epochs using input sizes of 128×128×3×32. This compact resolution preserved key diagnostic features, as confirmed by the downstream classification task. Other parameters were derived from prior work or adjusted for memory constraints.

Quantitative Assessment of Synthesis: Table 1 quantifies the impact of key modeling components on synthesized data quality for A4CC and PLAXC views. Incorporating Doppler-mode FOV and temporal indexing improved FID, FVD, and SSIM, emphasizing their role in spatiotemporal modeling. The best performance was achieved by integrating B-mode anatomical structure. Our approach significantly outperformed 3D LDM and LVDM adapted for color Doppler (p-values < 0.001, Wilcoxon signed-rank test), as shown in Fig. 3. The results demonstrate that although the outputs produced by LVDM and 3D LDM have anatomical correctness, their MR jets do not resemble realistic pathological features of RHD, whereas our model's output exhibits characteristics of RHD, as confirmed by clinicians. Fig. 4 presents A4CC and PLAXC samples labeled as RHD and normal across different datasets, generated using dual synthesis (a, b), B-mode conditioning (c, d, e, f, i), and view translation (g, h).

RHD Detection: Using 462 real data points, the model achieved an accuracy of 0.72 ± 0.4, sensitivity of 0.73 ± 0.4, and specificity of 0.68 ± 0.5. Training with only 1,414 synthetic data significantly improved performance, increasing accuracy to 0.82 ± 0.4, sensitivity to 0.79 ± 0.4, and specificity to 0.89 ± 0.3 (p-value = 0.008; McNemar test [25]). These results highlight the limitations of generalizing a model trained on a small, single-site dataset acquired with one device (D1), especially when tested on data from a different device (D2) and a population with greater age variability. In contrast, synthetic data introduced greater variability, helping bridge these gaps and significantly improving model generalization. Additionally, two expert cardiologists from Children's National Hospital independently and blindly reviewed 20 synthetic cases to assess realism and classify them as RHD or normal. Without prior exposure to synthetic data, they were unable to distinguish real from synthetic cases and identified RHD with 0.95 ± 0.05 accuracy, demonstrating strong clinical fidelity.

5 Discussion

RHD remains a major global health challenge, particularly in low- and middle-income countries with limited access to expert clinicians. Color Doppler interpretation is complex due to image variability, intricate blood flow, and diverse pediatric presentations. A key hurdle is the scarcity of pathological data, with few MR cases and almost no aortic regurgitation cases. Additionally, currently no public pediatric color Doppler datasets exist, limiting AI model development. To address these limitations, we introduced a 3.5D-conditional DM for synthesizing high-quality echocardiographic data.

We demonstrated the impact of our novel Doppler ultrasound data synthesis approach on early RHD detection in resource-limited settings. Our model is scalable and adaptable to other imaging views and cardiac conditions, making it valuable for rare disease detection in data-scarce settings. While we focused on generating color Doppler from B-mode ultrasound, our approach can also reversely synthesize B-mode from color Doppler. Although this is beyond the scope of this paper, it could be useful for training other disease models with anatomical pathology. A key innovation is multi-factor conditioning, where B-mode ultrasound provides anatomical structure to ensure Doppler flow aligns with real cardiac anatomy. By leveraging public B-mode datasets, our approach generated pathological cases and enabled conversion to the PLAXC view, which is easier to acquire than A4CC, especially for non-expert clinical staff.

Compared to state-of-the-art methods such as LVDM and 3D LDM, our 3.5D-conditional DM is more computationally efficient, requiring fewer Giga Multiply-Accumulate operations (GMac) [26] (112 vs. 502–19,890 GMac). LDMs require extensive parameter tuning and multi-stage training, while 3D U-Nets with volumetric convolutions—suitable for static data like CT or MRI—demand significant memory and compute resources. In contrast, our model embeds temporal indices for joint spatial–temporal modeling without added computational overhead. Instead of frame-wise generation with temporal concatenation, we use temporal conditioning to enable dynamic interactions and maintain temporal consistency with lower complexity.

While our approach offers significant advantages, it also has limitations. The quality and consistency of synthesized Doppler flow depend on conditioning inputs. Lower-quality B-mode images may introduce structural inconsistencies, while omitting B-mode conditioning and relying solely on other factors may potentially affect realism across frames. Future work could explore temporal refinement techniques to enhance frame-to-frame consistency, particularly when B-mode conditioning is unavailable. We also aim to estimate the optimal number of synthetic samples for robust classification and to assess performance limits. Additionally, we plan to extend this methodology to handheld ultrasound devices to broaden clinical utility through low-cost technology.

Table 1. Quantitative evaluation of synthesized data compared to real data from D1.

	Multi-Factor Conditioning								
Methods	Class	Doppler FOV	Temp. Index	B-mode	**View**	**Output**	**FID ↓**	**FVD ↓**	**SSIM ↑**
2D-DM	✓				PLAXC	Doppler	30.28±8.5	5.66±0.2	0.55±0.1
2D-DM	✓	✓			PLAXC	Doppler	29.04±8.7	3.73±0.5	0.54±0.1
2.5D-DM	✓	✓	✓		PLAXC	Doppler	23.93^±8.4	3.26±0.2	0.58±0.1
3.5D-DM*	✓	✓	✓		PLAXC	Doppler, B-mode	23.4±6.5	5.41±0.8	0.58±0.1
3.5D-DM	✓	✓	✓	✓	**PLAXC**	Doppler	**8.69±2.5**	**2.77±0.2**	**0.78±0.1**
3.5D-DM	✓	✓	✓	✓	**A4CC**	Doppler	**8.87±3**	**1.05±0.2**	**0.75±0.1**
3D LDM	✓				PLAXC	Doppler	41.6^±10	10.67±0.7	0.49±0.1
LVDM	✓				PLAXC	Doppler	30.11^±8.7	7.05±0.3	0.51±0.1

*Dual-channel synthesis; ^p-values < 0.001, compared to our approach; Temp.: Temporal.

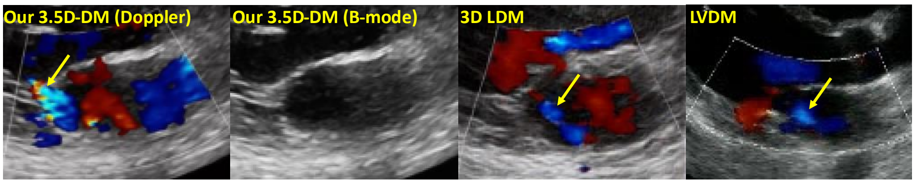
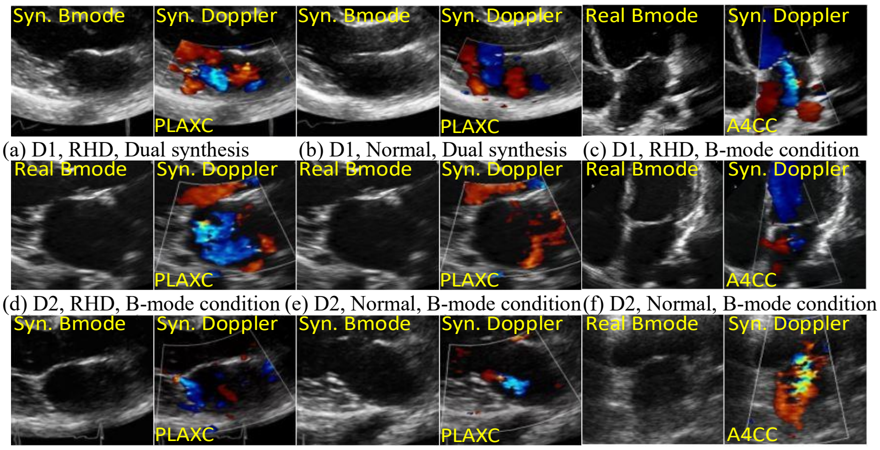


Fig. 3. Synthesized RHD samples from our approach, 3D LDM, and LVDM. The results from LVDM and 3D LDM lack RHD characteristics, as indicated by the arrow.



(a) D1, RHD, Dual synthesis (b) D1, Normal, Dual synthesis (c) D1, RHD, B-mode condition

(d) D2, RHD, B-mode condition (e) D2, Normal, B-mode condition (f) D2, Normal, B-mode condition

(g) CA, RHD, View translation* (h) CA, Normal, View translation* (i) CA, RHD, B-mode condition

Fig. 4. Synthesized PLAXC and A4CC samples labeled as RHD or normal, from D1 (a–c), D2 (d–f), and CA (g–i). *The B-mode image in (i) served as the input for generating PLAXC views in (g) and (h).

6 Conclusion

We introduced a 3.5D-conditional diffusion model to synthesize dual-channel color Doppler echocardiograms coupled with B-mode ultrasound, addressing cardiac data scarcity in resource-limited settings. To our knowledge, this is the first model to generate color Doppler overlaid on B-mode, capturing both structural and functional cardiac features. By incorporating temporal indexing, class label, and Doppler-mode FOV conditioning, the model produced clinically relevant data across imaging conditions and views. It synthesized paired B-mode and Doppler sequences—both healthy and pathological—as well as new acquisition views with clinical utility. Applied to RHD detection, the synthetic data improved model performance on real-world datasets. This approach enables high-fidelity synthesis of rare pathological cases, advancing AI-driven diagnostics and broadening access to cardiac care. Upon institutional approval, the synthetic data and model will be released at: https://github.com/Pediatric-Accelerated-Intelligence-Lab/Dual-Channel_Color_Doppler_Synthesis.

Acknowledgments. This work was supported through funding from the Karp Family Foundation, Gift of Life International, Elias/Ginsberg family, AT&T, American Heart Association, and Children's National Hospital Race for Every Child/Team Jocelyn.

Disclosure of Interests. Children's National Hospital holds the intellectual property rights related to the work disclosed in this paper. Marius George Linguraru is the President of MICCAI.

References

1. Rheumatic Heart Disease. Available: https://www.who.int/news-room/fact-sheets/detail/rheumatic-heart-disease, last accessed 2025/02/27
2. Liu, M., Lu, L., Sun, R., Zheng, Y., Zhang, P.: Rheumatic heart disease: causes, symptoms, and treatments. Cell Biochem. Biophys. 72, 861–863 (2015). https://doi.org/10.1007/s12013-015-0552-5
3. Lu, J.C., Sable, C., Ensing, G.J., Webb, C., Scheel, J., Aliku, T., et al.: Simplified rheumatic heart disease screening criteria for handheld echocardiography. J. Am. Soc. Echocardiogr. 28(4), 463–469 (2015). https://doi.org/10.1016/j.echo.2015.01.001
4. Mitra, J., Qiu, J., MacDonald, M., Venugopal, P., Wallace, K., Abdou, H., et al.: Automatic hemorrhage detection from color Doppler ultrasound using a GAN-based anomaly detection method. IEEE J. Transl. Eng. Health Med. 10, 1–9 (2022). https://doi.org/10.1109/JTEHM.2022.3199987
5. 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: Med. Image Comput. Comput. Assist. Interv. – MICCAI 2023, Lect. Notes Comput. Sci. vol. 14337, pp. 34–43, Springer, Cham, Vancouver, British Columbia, Canada (2023). https://doi.org/10.1007/978-3-031-44521-7_4
6. Reynaud, H., Meng, Q., Dombrowski, M., Ghosh, A., Day, T., Gomez, A., et al.: EchoNet-Synthetic: privacy-preserving video generation for safe medical data sharing. In: Med. Image Comput. Comput. Assist. Interv. – MICCAI 2024, Lect. Notes Comput. Sci. vol. 15007,

pp. 285–295. Springer, Cham, Marrakesh, Morocco (2024). https://doi.org/10.1007/978-3-031-72104-5_28.

7. Mishra, D., Zhao, H., Saha, P., Papageorghiou, A.T., Noble, J.A.: Dual conditioned diffusion models for out-of-distribution detection: application to fetal ultrasound videos. In: Med. Image Comput. Comput. Assist. Interv. – MICCAI 2023, Lect. Notes Comput. Sci. vol. 14220, pp. 216–226. Springer, Cham, Vancouver, British Columbia, Canada (2023). https://doi.org/10.1007/978-3-031-43907-0_21
8. Zhou, X., Huang, Y., Xue, W., Dou, H., Cheng, J., Zhou, H., et al.: HeartBeat: Towards controllable echocardiography video synthesis with multimodal conditions-guided diffusion models. In: Med. Image Comput. Comput. Assist. Interv. – MICCAI 2024, Lect. Notes Comput. Sci. vol. 15007, pp. 361–371. Springer, Cham, Marrakesh, Morocco (2024). https://doi.org/10.1007/978-3-031-72104-5_35
9. Cohen, J.P., Luck, M., Honari, S.: Distribution matching losses can hallucinate features in medical image translation. In: Med. Image Comput. Comput. Assist. Interv. – MICCAI 2018, Lect. Notes Comput. Sci. vol. 11070, pp. 529–536. Springer, Cham, Granada, Spain (2018). https://doi.org/10.1007/978-3-030-00928-1_60
10. Bond-Taylor, S., Leach, A., Long, Y., Willcocks, C.G.: Deep generative modelling: a comparative review of VAEs, GANs, normalizing flows, energy-based and autoregressive models. IEEE Trans. Pattern Anal. Mach. Intell. 44(11), 7327–7347 (2022). https://doi.org/10.1109/TPAMI.2021.3116668
11. Kazerouni, A., Aghdam, E.K., Heidari, M., Azad, R., Fayyaz, M., Hacihaliloglu, I., Merhof, D.: Diffusion models in medical imaging: A comprehensive survey. Med. Image Anal. 88, 1–33 (2023). https://doi.org/10.1016/j.media.2023.102846
12. Croitoru, F.A., Hondru, V., Ionescu, R.T., Shah, M.: Diffusion models in vision: a survey. IEEE Trans. Pattern Anal. Mach. Intell. 45(9), 10850–10869 (2023). https://doi.org/10.1109/TPAMI.2023.3261988
13. Sun, Y., Vixège, F., Faraz, K., Mendez, S., Nicoud, F., Garcia, D., Bernard, O.: A pipeline for the generation of synthetic cardiac color Doppler. IEEE Trans. Ultrason. Ferroelectr. Freq. Control. 69(3), 932–941 (2022). https://doi.org/10.1109/TUFFC.2021.3136620
14. Leclerc, S., Smistad, E., Pedrosa, J., Ostvik, A., Cervenansky, F., Espinosa, F., et al.: Deep learning for segmentation using an open large-scale dataset in 2D echocardiography. IEEE Trans. Med. Imaging. 38(9), 2198-2210 (2019). https://doi.org/10.1109/TMI.2019.2900516
15. Avants, B.B., Tustison, N.J., Song, G., Cook, P.A., Klein, A., Gee, J.C.: A reproducible evaluation of ANTs similarity metric performance in brain image registration. Neuroimage. 54(3), 2033-2044 (2011). https://doi.org/10.1016/j.neuroimage.2010.09.025
16. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Adv. Neural Inf. Process. Syst. (NeurIPS), pp. 6840–6851. Curran Associates, Inc., Vancouver, British Columbia, Canada (2020). https://arxiv.org/abs/2006.11239
17. Xie, E., Chen, J., Zhao, Y., Yu, J., Zhu, L., Lin, Y., et al.: SANA 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. arXiv preprint arXiv:2501.18427, 1-21 (2025). https://doi.org/10.48550/arXiv.2501.18427
18. Tran, D., Wang, H., Torresani, L., Ray, J., Lecun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 6450–6459. IEEE, Salt Lake City, UT, USA (2018). https://doi.org/10.1109/CVPR.2018.00675
19. 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: Adv. Neural Inf. Process. Syst. (NIPS), pp. 6629–6640. Curran Associates, Inc., Long Beach, CA, USA (2017). https://arxiv.org/abs/1706.08500

20. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: FVD: A new metric for video generation. In: Deep Generative Models for Highly Structured Data, ICLR 2019 Workshop, New Orleans, LA, USA, pp. 1-9. (2019). https://doi.org/10.48550/arXiv.1812.01717
21. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. IEEE Trans. Image Process. 13(4), 600-612 (2004). https://doi.org/10.1109/TIP.2003.819861
22. Mei, X., Liu, Z., Robson, P.M., Marinelli, B., Huang, M., Doshi, A., et al.: RadImageNet: an open radiologic deep learning research dataset for effective transfer learning. Radiol. Artif. Intell. 4(5), 1–9 (2022). https://doi.org/10.1148/RYAI.210315
23. Pinaya, W.H.L., Tudosiu, P.D., Dafflon, J., Da Costa, P.F., Fernandez, V., Nachev, P., et al.: Brain imaging generation with latent diffusion models. In: Med. Image Comput. Comput. Assist. Interv. – MICCAI 2022, Lect. Notes Comput. Sci. 13609, 117–126. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-18576-2_12
24. Cardoso, M.J., Li, W., Brown, R., Ma, N., et al.: MONAI: An open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701, 1-25 (2022). https://doi.org/10.48550/arXiv.2211.02701
25. McNemar, Q.: Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika. 12(2), 153–157 (1947). https://doi.org/10.1007/BF02295996
26. Sovrasov, V.: Ptflops: a flops counting tool for neural networks in pytorch framework. https://github.com/sovrasov/flops-counter.pytorch (2018-2024), last accessed 2025/02/27