This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

EchoFID: A Multi-Teacher Distilled Evaluation Network for Cardiac Ultrasound Image Generation

Junde Wu[1], Renee Miller[2], Jurica Šprem[2], and Vicente Grau[1]

[1] University of Oxford, Oxford OX1 2JD, UK
[2] GE HealthCare, 500 W Monroe St, Chicago, IL 60661, USA
vicente.grau@eng.ox.ac.uk

Abstract. Reliable evaluation remains a fundamental bottleneck in cardiac ultrasound image generation. While diffusion and GAN-based models have demonstrated increasingly realistic synthesis of speckle patterns and anatomical structures, existing metrics such as FID, KID, and LPIPS rely on generic natural-image encoders or single-task medical backbones, which fail to capture the multi-faceted clinical information embedded in echocardiography. In this work, we propose EchoFID, a clinically grounded evaluation framework built upon a multi-teacher distilled feature backbone for cardiac ultrasound generation assessment. Our approach distills knowledge from a diverse ensemble of pretrained expert networks spanning segmentation, prediction, and functional measurement tasks. Instead of enforcing direct feature matching, we formulate evaluation learning as a dual decodability problem with asymmetric alignment flows: a student encoder is trained such that its latent representation can be reconstructed into each teacher's feature space through lightweight converters, with asymmetric gradient routing to resolve conflicting supervision. Extensive controlled experiments across variations in generator training epochs, training data quota, and model capacity demonstrate that EchoFID exhibits smoother monotonic behavior, stronger discriminative power, and improved real-to-real compactness compared to FID, RadFID, IS, KID, LPIPS, and other medical-specific metrics. These results establish EchoFID as a principled and practical benchmark for cardiac ultrasound image generation, enabling more reliable comparison and accelerating progress in medical generative modeling.

1 Introduction

Cardiac ultrasound (echocardiography) is among the most widely used imaging modalities in routine care due to its non-invasiveness, real-time acquisition, and ability to capture both anatomy and function. It is central to diagnosing and monitoring cardiovascular disease, a leading cause of mortality worldwide. While recent deep learning methods have advanced ultrasound segmentation, diagnosis, and functional measurement, their performance and generalization remain

constrained by the limited availability of large-scale, diverse, well-annotated datasets. Compared with natural images, echocardiography data are expensive to collect, and highly variable across devices, views, and patients, motivating generative modeling for augmentation and representation learning [11].

A key unresolved bottleneck is evaluation: Expert visual inspection is subjective and non-scalable, while downstream-task evaluation depends on specific predictive models and cannot serve as a general-purpose metric. Standard distributional measures such as FID [8] and KID [2] rely on a single pretrained backbone, implicitly assuming a unified feature space. However, echocardiography encodes clinically meaningful information across heterogeneous dimensions—including anatomical structure, functional morphology, and diagnostic texture—each emphasized differently by segmentation, diagnosis, and measurement models. A single-task encoder is therefore inherently incomplete, motivating a multi-perspective evaluation backbone tailored to cardiac ultrasound.

A naïve solution would be to aggregate features from multiple expert networks. However, simple feature averaging collapses heterogeneous representations into a single mean space. Segmentation models prioritize spatial boundary precision, and measurement networks capture scale-sensitive morphology. Averaging these features blurs their distinct inductive biases. Conversely, strict multi-task feature matching forces the student to replicate each teacher representation simultaneously, introducing mutually incompatible constraints that could not converge in training.

To overcome this structural conflict, we introduce a novel multi-teacher decodability distillation framework. Instead of enforcing feature equality, we require that each teacher representation be recoverable from the student embedding through lightweight neural converters. This shifts supervision from strict alignment to information sufficiency: the student must encode enough structure so that different teacher features can be reconstructed back through very simple, low-parameter networks, ensuring that diverse and essential information is preserved in the student embedding itself. To jointly train the student and the converters, we design two complementary flows. A forward decodability flow preserves teacher-relevant information by updating the student, while an asymmetric alignment flow stabilizes the shared space without allowing competing gradients to interfere with the student encoder. The resulting embedding forms a coherent evaluation-specific feature space designed explicitly for generative assessment.

Finally, we validate the proposed evaluation backbone using a controlled protocol that systematically varies generator quality through early stopping, training data constraints, and model capacity. By probing the metric under progressively degraded conditions, we demonstrate that the distilled evaluator produces smoother, more discriminative, and more consistent scores than traditional metrics such as FID, KID, and RadFID.

2 Related Work

Generative Modeling in Medical Imaging. Medical image synthesis has rapidly evolved from GAN-based frameworks [5] to diffusion models [9, 4]. In medical domains, GANs have been applied to MRI and ultrasound synthesis [6, 22], while diffusion models increasingly dominate due to improved fidelity and stability [14, 12]. In echocardiography specifically, recent works explore diffusion-based image and video generation [19, 3, 26], demonstrating realistic anatomical and motion synthesis. However, evaluation protocols largely inherit natural-image metrics.

Evaluation Network for Generative Models Standard metrics such as FID [8], KID [2], and Inception Score [21] rely on pretrained natural-image encoders. Although widely adopted, these metrics exhibit bias and instability on medical images. In medical imaging, some works replace natural encoders with radiology-trained networks or adopt task-based evaluation strategies [11, 1, 16]. But they remain confined to single-task feature spaces, limiting their applicability when distilling knowledge from multiple teachers. Multi-task learning and conflict-aware optimization have been studied in representation learning [13, 15, 17], but conventional fusion strategies enforce direct feature alignment, which can introduce supervision conflicts when tasks encode incompatible invariances.

3 Method

3.1 Problem Formulation and Motivation

Let $x \in \mathcal{X}$ denote a real cardiac ultrasound image and $x^g \sim p_G$ a generated sample. Generative quality is typically evaluated by comparing real and generated distributions in a learned feature space. We assume access to pretrained teacher networks $\{T_k\}_{k=1}^{K}$, spanning segmentation, prediction, and functional measurement tasks. Each teacher produces frozen intermediate features $h_k(x) \in \mathbb{R}^{d_k}$.

Our objective is to distill information from all teachers into a single student encoder $S_\theta : \mathcal{X} \to \mathbb{R}^d$, which serves solely as an evaluation backbone. The student is not optimized for any downstream task; instead, it defines a feature space in which distributional distances between real and generated samples can be meaningfully computed. The key challenge is therefore to construct a feature representation that captures clinically meaningful variations in cardiac ultrasound while remaining stable and general-purpose.

A straightforward idea is to aggregate features from multiple pretrained expert networks and measures similarity via averaging. However, this suppresses rare yet clinically important activations, such as subtle anatomical or pathological patterns, leading to metrics that overlook critical but infrequent signals. Conversely, directly forcing a single representation to match multiple task-specific teacher features introduces conflicting constraints: segmentation models emphasize boundary precision, diagnostic models prioritize discriminative texture, and

measurement models encode morphology-sensitive cues. Joint feature matching across such heterogeneous objectives often results in unstable optimization.

To address these issues, we replace strict feature alignment with a weaker but principled constraint: *decodability*. Rather than requiring exact feature equality, we require that each expert representation can be recovered from a unified embedding through a simple transformation. This preserves heterogeneous clinical signals without imposing incompatible invariances.

3.2 Decodability-based feature fusion

The student encoder produces a latent representation $s(x) = S_\theta(x)$. Rather than enforcing $s(x)$ to directly approximate each $h_k(x)$, we introduce a collection of lightweight converter networks that mediate between the student and teacher feature spaces.

For each teacher k, we define converters

$$C_{\psi_k}^{s\to k} : \mathbb{R}^d \to \mathbb{R}^{d_k}, \quad C_{\varphi_k}^{k\to s} : \mathbb{R}^{d_k} \to \mathbb{R}^d.$$

These converters are intentionally small, implemented as shallow multilayer perceptrons with a narrow bottleneck, so that successful reconstruction requires the relevant information to be present in the input representation rather than encoded in the converter parameters.

The primary constraint imposed on the student representation is that each teacher feature must be recoverable from it through a simple converter. This is enforced by minimizing

$$\mathcal{L}_{s\to t} = \sum_{k=1}^{K} \mathbb{E}_{x\sim\mathcal{D}} \left[\left\| C_{\psi_k}^{s\to k}(S_\theta(x)) - h_k(x) \right\|_2^2 \right],$$

where $\mathcal{D}$ denotes the training distribution. Gradients from this loss are propagated into both the student encoder S_θ and the student-to-teacher converters $C_{\psi_k}^{s\to k}$. Intuitively, this encourages the student to retain clinically relevant signals identified by each teacher while avoiding hard alignment to any single task-specific inductive bias. Because the student is trained to preserve decodability across multiple complementary teachers, anatomically inconsistent or clinically implausible patterns cannot be supported unless they are jointly coherent across tasks, reducing the risk of encoding semantically correct yet structurally impossible features.

To stabilize the shared latent space and avoid degenerate solutions in which converters simply overfit to the student features, we impose a complementary constraint in the reverse direction. Specifically, we require that teacher features can be mapped into the student space by a simple converter, but without altering the student representation itself. This is achieved by minimizing

$$\mathcal{L}_{t\to s} = \sum_{k=1}^{K} \mathbb{E}_{x\sim\mathcal{D}} \left[\left\| C_{\varphi_k}^{k\to s}(h_k(x)) - \mathrm{sg}(S_\theta(x)) \right\|_2^2 \right],$$

where $\mathrm{sg}(\cdot)$ denotes the stop-gradient operator. Gradients from this loss update only the teacher-to-student converters φ_k, while the student encoder remains fixed. This asymmetric gradient routing is critical: it allows each converter to align teacher features to the student space, while preventing conflicting teacher objectives from directly perturbing the student representation.

Together, these two losses enforce a complementary decodability constraint through asymmetric optimization flows. The student representation must be sufficiently expressive to reconstruct all teacher features through simple decoders, and teacher features must admit a low-complexity projection into the student space. Importantly, the student is never required to satisfy mutually contradictory feature equalities, resolving the optimization conflicts inherent in direct multi-teacher supervision.

3.3 Training and EchoFID

The overall objective is

$$\mathcal{L} = \mathcal{L}_{s\to t} + \lambda \mathcal{L}_{t\to s},$$

where the first term enforces forward decodability and the second stabilizes the shared latent space via asymmetric alignment.

After training, the student encoder defines a fixed embedding $z(x)$. We compute a Fréchet-style distributional distance between real and generated sets, denoted $\mathrm{DIS}(\mathcal{T}, \mathcal{G})$, and estimate intrinsic real-data variability as

$$\Delta_{\mathrm{real}} = \frac{1}{J} \sum_{j} \mathrm{DIS}(\mathcal{T}, \mathcal{R}_j).$$

The final evaluation metric, **EchoFID**, is then defined as

$$\mathrm{EchoFID}(G) = \mathrm{DIS}(\mathcal{T}, \mathcal{G}) - \Delta_{\mathrm{real}}.$$

This margin measures how far generated samples deviate from the real manifold beyond intrinsic variability, yielding a geometrically calibrated evaluation score.

4 Experimental Setting

4.1 Datasets and Splits

All experiments are conducted on a mixture of private echocardiography studies and the public EchoNet-Dynamic dataset [18]. We randomly split the dataset at the video level, reserving 20% as a held-out test set for all distributional evaluations. All metrics are computed against this fixed real test set.

4.2 Controlling Generation Quality

To systematically probe metric sensitivity, we control generation quality along three axes: (i) generator training epochs, (ii) generator model capacity, and (iii) generator training data scale. Generators are trained for increasing numbers of epochs to produce progressively refined samples, while model capacity is varied through parameter count to obtain generators of different representational strength. In addition, generators are trained under data quotas of 20%, 40%, 60%, 80%, and 100%. Empirically, smaller models, limited training data, and shorter training schedules consistently yield lower anatomical fidelity and reduced diversity, thereby forming a controlled spectrum of generation quality. For each configuration, a fixed number of generated samples is compared against the same held-out real test set.

4.3 Distillation Teachers and Student Backbone

For multi-teacher distillation, we use 15 pretrained cardiac ultrasound models spanning segmentation and functional measurement tasks. These include 6 state-of-the-art segmentation networks trained for chamber delineation [7, 10, 25, 23, 24, 27], and 9 networks for regression models for ejection fraction and structural measurements [20]. All teacher networks remain frozen during training.

The student evaluation backbone is using a standard Inception-V3 network with output embedding dimension $d = 2048$. Converter modules are lightweight two-layer MLPs which transfer the teacher embedding size to that of the student. All student variants share the same architecture and training recipe to ensure fair comparison across ablations.

5 Experimental Results and Analysis

Table 1 and Fig. 1 compares **EchoFID** with conventional metrics across three controlled factors: training epochs, training data quota, and generator type. As generator training increases from 20 to 320 epochs, image quality improves progressively. While FID and RadFID generally decrease, their trajectories are irregular; for example, RadFID degrades at 160 epochs and KID fluctuates despite visible quality gains. EchoFID instead decreases consistently from 165.22 to 11.86, forming a smooth curve aligned with progressive anatomical refinement, indicating higher sensitivity to incremental structural improvements.

When reducing the training data quota, generation quality deteriorates due to blurred anatomy and reduced diversity. EchoFID responds monotonically from 43.22 at 20% data to 11.86 at 100%, clearly separating intermediate regimes. In contrast, FID and RadFID show weaker separation, and LPIPS and ASW exhibit erratic trends, suggesting lower robustness under data scarcity. Across generator architectures, from GAN to increasingly large diffusion models, EchoFID decreases smoothly and preserves correct ranking, reaching 11.86 for the strongest model. Several baseline metrics fail to maintain consistent ordering or saturate at

Table 1. Quantitative comparison of EchoFID with existing evaluation metrics under controlled variations of training epochs, training data quota, and generator model size. Red values indicate violations of expected monotonic behavior.

Model	FID	RadFID	IS	KID %	LPIPS	ASW	FPD	MAID	RS	CMMD	EchoFID
On training epochs											
20 epochs	156.73	113.32	1.1	93.0	228.4	-0.52	77.8	52.9	145.8	3.7	**165.22**
40 epochs	144.63	108.21	1.7	86.7	176.5	-0.13	84.6	41.6	116.5	2.1	**125.42**
80 epochs	118.90	72.22	1.6	37.8	128.0	0.48	35.5	44.3	81.7	1.1	**84.53**
160 epochs	45.6	80.71	2.6	41.2	68.2	0.40	46.2	21.3	21.0	0.6	**42.13**
320 epochs	27.9	23.11	2.4	10.8	16.8	0.52	37.5	14.7	15.7	0.8	**11.86**
Training Data Quota											
20 %	112.45	41.25	1.7	71.8	140.2	-0.58	64.5	48.9	88.7	3.5	**43.22**
40 %	109.67	29.41	1.5	65.7	110.9	-0.26	61.0	44.1	45.8	1.4	**28.19**
60 %	40.43	25.21	1.8	41.2	62.8	0.22	43.3	34.8	55.0	0.6	**21.75**
80 %	62.88	22.68	2.2	46.9	14.1	0.63	32.8	20.1	32.9	0.9	**19.44**
100 %	27.9	23.11	2.4	10.8	16.8	0.52	37.5	14.7	15.7	0.8	**11.86**
Generation Models											
GAN	133.78	67.13	1.5	96.6	134.9	-0.26	56.3	49.8	87.8	2.5	**48.46**
Diffusion (36M)	126.65	26.78	1.7	89.3	87.5	0.35	52.2	51.2	56.9	2.1	**25.77**
Diffusion (95M)	68.74	29.41	2.4	56.4	54.3	0.47	46.9	44.2	31.8	1.3	**23.20**
Diffusion (144M)	72.31	25.21	2.8	56.1	65.5	0.64	38.5	23.8	21.9	0.7	**22.11**
Diffusion (328M)	42.28	26.25	2.6	28.2	15.6	0.49	32.1	26.7	24.2	1.2	**21.15**
Diffusion (676M)	27.9	21.11	2.4	10.8	16.8	0.58	37.5	14.7	15.7	0.8	**11.86**
Real	31.26	28.23	2.1	8.1	9.7	0.62	35.0	16.0	8.4	0.5	**7.44**

higher quality levels. Finally, real-to-real comparisons yield a compact EchoFID value of 7.44, establishing a stable lower bound. Some baseline metrics produce inflated real-to-real distances, reflecting domain mismatch between their pretrained encoders and ultrasound data. Together, these results indicate that EchoFID constructs a calibrated and clinically aligned feature space rather than merely minimizing absolute distances.

5.1 Ablation study

Table 2 analyzes the contribution of each component under a fixed high-quality generator. A reliable evaluation metric should produce a compact real-data manifold and maintain a clear separation between real–real and real–generated distances.

The full EchoFID model satisfies this criterion, with a real–generated distance of 11.86 and a real–real distance of 7.44, indicating compactness and stable discriminative power. Removing the teacher→student alignment term or replacing asymmetric routing with symmetric updates breaks this calibration: the real manifold becomes overly dispersed (e.g., 43.33 vs. 24.91), and separation reliability deteriorates due to conflicting supervision. Direct multi-teacher feature matching and naïve feature averaging further collapse the metric's calibration, as real–real and real–generated distances become nearly identical (31.76 vs. 31.84 and 33.15 vs. 32.62), eliminating meaningful separation. Restricting supervision to a single task preserves inequality but weakens the separation gap, indicating

Table 2. Ablation study of EchoFID. Lower values indicate better generation quality for that evaluation.

Model Variant	generated↓	real↓
Full Model (EchoFID)	**11.86**	**7.44**
w/o Teacher→Student Alignment ($\mathcal{L}_{t\to s}$)	21.58	23.47
w/ Symmetric Gradient Routing (no stop-grad)	24.91	43.33
Direct Multi-Teacher Feature Matching	31.84	31.76
Naïve Feature Averaging (No Distillation)	32.62	33.15
Segmentation-Only Teachers	19.47	18.28
Measurement-Only Teachers	22.63	22.94

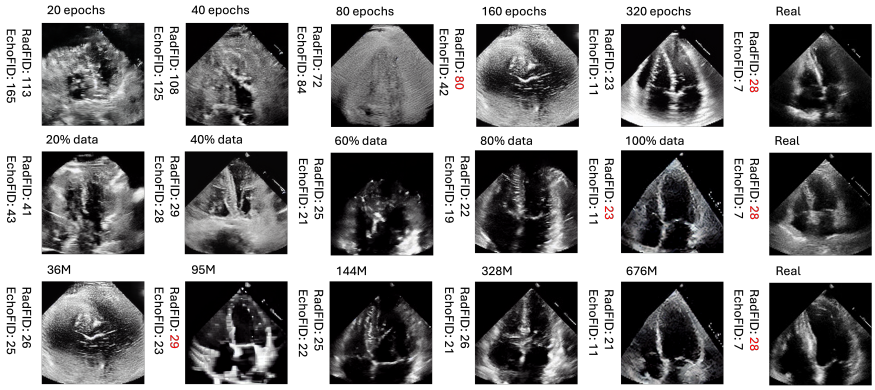

Fig. 1. Visual comparison of generated cardiac ultrasound images under increasing training epochs, data quota, and model capacity, with corresponding RadFID and EchoFID scores. EchoFID decreases consistently with perceptual quality improvements, while RadFID shows erratic behavior (red).

incomplete structural coverage. Overall, these results show that EchoFID's components collectively enforce compact real embeddings and stable real–generated separation, rather than merely reducing absolute distances.

6 Conclusion

We proposed EchoFID, a clinically grounded evaluation backbone for cardiac ultrasound generation. The method distills knowledge from multiple pretrained expert networks, covering segmentation and measurement into a unified student encoder via a decodability-based multi-teacher distillation framework with lightweight converters and asymmetric gradient routing. Extensive controlled experiments and ablation studies demonstrate that EchoFID yields smoother, more discriminative, and better-calibrated behavior than conventional metrics such as FID and RadFID. EchoFID provides a more reliable and principled foundation for evaluating generative models in cardiac ultrasound.

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

References

[1] Abdusalomov, A.B., Nasimov, R.R., Nasimova, N.R., Muminov, B.M., Whangbo, T.K.: Evaluating synthetic medical images using artificial intelligence with the gan algorithm. Sensors **23**(7), 3440 (2023). https://doi.org/10.3390/s23073440

[2] Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying MMD GANs (2021), https://arxiv.org/abs/1801.01401

[3] Chen, T., et al.: Ultrasound image-to-video synthesis via latent dynamic diffusion models. In: MICCAI. pp. 764–774. Springer (2024)

[4] Dhariwal, P., et al.: Diffusion models beat gans on image synthesis. NeurIPS **34** (2021)

[5] Goodfellow, I., et al.: Generative adversarial networks. NeurIPS **27** (2014)

[6] Han, C., Hayashi, H., Rundo, L., Araki, R., Shimoda, W., Muramatsu, S., Furukawa, Y., Mauri, G., Nakayama, H.: GAN-based synthetic brain MR image generation. In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018). pp. 734–738. IEEE (2018)

[7] Hansegård, J., Orderud, F., Rabben, S.I.: Real-time active shape models for segmentation of 3d cardiac ultrasound. In: International Conference on Computer Analysis of Images and Patterns. pp. 157–164. Springer (2007)

[8] 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. Adv Neural Inf Process Syst **30** (2017)

[9] Ho, J., et al.: Denoising Diffusion Probabilistic Models. In: NeurIPS. vol. 33, pp. 6840–6851 (2020)

[10] Hsu, W.Y.: Automatic left ventricle recognition, segmentation and tracking in cardiac ultrasound image sequences. IEEE Access **7**, 140524–140533 (2019)

[11] Jha, A.K., Myers, K.J., Obuchowski, N.A., Liu, Z., Rahman, M.A., Saboury, B., Rahmim, A., Siegel, B.A., Frey, E.C., Samei, E.: Objective task-based evaluation of artificial intelligence-based medical imaging methods: Framework, strategies, and role of the physician. Journal of Medical Imaging **9**(1), 011004 (2022). https://doi.org/10.1117/1.JMI.9.1.011004

[12] 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**, 102846 (2023)

[13] Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: Proc. IEEE Conf. Comput. Vision Pattern Recog. pp. 7482–7491 (2018)

[14] Khader, F., Mueller-Franzes, G., Arasteh, S., Han, T., Haarburger, C., Schulze-Hagen, M., Schad, P., Engelhardt, S., Baessler, B., Foersch, S., Stegmaier, J.: Medical diffusion–denoising diffusion probabilistic models for

3D medical image generation. arXiv.org **abs/2211.03364** (2022). https://doi.org/10.48550/arXiv.2211.03364
[15] Liu, B., Liu, X., Jin, X., Stone, P., Liu, Q.: Conflict-averse gradient descent for multi-task learning. Adv. Neural Inf. Process. Syst. **34**, 18878–18890 (2021)
[16] Mei, X., Liu, Z., Robson, P.M., Marinelli, B., Huang, M., Doshi, A., Jacobi, A., Cao, C., Link, K.E., Yang, T., et al.: Radimagenet: an open radiologic deep learning research dataset for effective transfer learning. Radiology: Artificial Intelligence **4**(5), e210315 (2022)
[17] Navon, A., Shamsian, A., Achituve, I., Maron, H., Kawaguchi, K., Chechik, G., Fetaya, E.: Multi-task learning as a bargaining game. arXiv.org **abs/2202.01017** (2022). https://doi.org/10.48550/arXiv.2202.01017
[18] Ouyang, D., He, B., Ghorbani, A., Lungren, M.P., Ashley, E.A., Liang, D.H., Zou, J.Y.: Echonet-dynamic: a large new cardiac motion video data resource for medical machine learning. In: NeurIPS ML4H Workshop: Vancouver, BC, Canada. vol. 5 (2019)
[19] Reynaud, H., et al.: Feature-Conditioned Cascaded Video Diffusion Models for Precise Echocardiogram Synthesis. In: MICCAI. pp. 142–152. Springer (2023)
[20] Sahashi, Y., Ieki, H., Yuan, V., Christensen, M., Vukadinovic, M., Binder-Rodriguez, C., Rhee, J., Zou, J., He, B., Cheng, P., Ouyang, D.: Artificial intelligence automation of echocardiographic measurements. Journal of the American College of Cardiology **86**(13), 964–978 (2025). https://doi.org/10.1016/j.jacc.2025.07.053
[21] Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training GANs. Adv Neural Inf Process Syst **29** (2016)
[22] Tiago, C., et al.: A data augmentation pipeline to generate synthetic labeled datasets of 3d echocardiography images using a gan. IEEE Access **10**, 98803–98815 (2022)
[23] Wu, J., Fu, R., Fang, H., Zhang, Y., Yang, Y., Xiong, H., Liu, H., Xu, Y.: Medsegdiff: Medical image segmentation with diffusion probabilistic model. In: Medical imaging with deep learning. pp. 1623–1639. PMLR (2024)
[24] Wu, J., Wang, Z., Hong, M., Ji, W., Fu, H., Xu, Y., Xu, M., Jin, Y.: Medical sam adapter: Adapting segment anything model for medical image segmentation. Medical image analysis **102**, 103547 (2025)
[25] Zhang, H., Wu, Y., Zhao, M., Chen, Z., Li, R., Zhu, F., Zhao, H., Yuan, X., Yang, M., Qiu, C., et al.: A fully open and generalizable foundation model for ultrasound clinical applications. arXiv preprint arXiv:2509.11752 (2025)
[26] Zhou, X., Huang, Y., Xue, W., Dou, H., Cheng, J., Zhou, H., Ni, D.: Heartbeat: Towards controllable echocardiography video synthesis with multi-modal conditions-guided diffusion models. arXiv preprint arXiv:2406.14098 (2024), https://arxiv.org/abs/2406.14098
[27] Zhu, J., Qi, Y., Wu, J.: Medical sam 2: Segment medical images as video via segment anything model 2. arXiv preprint arXiv:2408.00874 (2024)