EchoLVFM: One-Step Video Generation via Latent Flow Matching for Echocardiogram Synthesis

Emmanuel Oladokun[1], Sarina Thomas[2], Jurica Šprem[2], and Vicente Grau[1]

[1] University of Oxford, Oxford, England
[2] GE HealthCare, Cardiovascular Ultrasound R&D, Oslo, Norway
emmanuel.oladokun@eng.ox.ac.uk

Abstract. Echocardiography is widely used for assessing cardiac function, where clinically meaningful parameters such as left-ventricular ejection fraction (EF) play a central role in diagnosis and management. Generative models capable of synthesising realistic echocardiogram videos with explicit control over such parameters are valuable for data augmentation, counterfactual analysis, and specialist training. However, existing approaches typically rely on computationally expensive multi-step sampling and aggressive temporal normalisation, limiting efficiency and applicability to heterogeneous real-world data.

We introduce EchoLVFM, a one-step latent video flow-matching framework for controllable echocardiogram generation. Operating in the latent space, EchoLVFM synthesises temporally coherent videos in a single inference step, achieving a $\sim$**50**$\times$ improvement in sampling efficiency compared to multi-step flow baselines while maintaining visual fidelity. The model supports global conditioning on clinical variables, demonstrated through precise control of EF, and enables reconstruction and counterfactual generation from partially observed sequences. A masked conditioning strategy further removes fixed-length constraints, allowing shorter sequences to be retained rather than discarded.

We evaluate EchoLVFM on the CAMUS dataset under challenging single-frame conditioning. Quantitative and qualitative results demonstrate competitive video quality, strong EF adherence, and 57.9% discrimination accuracy by expert clinicians which is close to chance. These findings indicate that efficient, one-step flow matching can enable practical, controllable echocardiogram video synthesis without sacrificing fidelity. Code available at: `EchoLVFM`

Keywords: Video Generation · Ultrasound · Echocardiography.

1 Introduction

Echocardiography is a widely used cardiac imaging modality that supports diagnosis, monitoring, and treatment planning across a broad range of cardiovascular diseases [2,22]. From echocardiogram videos, clinically meaningful physiological

parameters such as left-ventricular ejection fraction (EF) can be derived, which play a central role in assessing cardiac function and diagnosing conditions including heart failure [17]. As a result, echocardiography remains a cornerstone of routine clinical practice due to its non-invasive nature, low cost, and portability.

The ability to generate realistic echocardiogram videos while explicitly controlling such clinical parameters is highly desirable. Controllable synthesis enables the creation of counterfactual examples, allowing clinicians to visualise how changes in physiological variables may manifest in imaging. It also supports specialist training by exposing trainees to rarely observed pathological cases, and facilitates dataset rebalancing in settings where real-world data is skewed.

Generative modelling for echocardiography is, however, challenging. Compared to natural video data, public medical datasets are relatively small and heterogeneous, with substantial variation in sequence length, frame rate, and image quality. Early generative approaches based on variational autoencoders (VAEs) [12] and generative adversarial networks (GANs) [7] demonstrated the feasibility of medical image synthesis [3,19], but often suffered from over-smoothing, training instability, or limited diversity. More recently, diffusion models [9] have become the dominant paradigm for high-fidelity image [20] and video generation [10], including latent video synthesis [31], but their reliance on many iterative denoising steps leads to slow and computationally costly inference.

Flow matching [16] has recently emerged as a compelling alternative. By learning a deterministic transport between noise and data distributions, flow matching combines the stability and generative quality of diffusion models with the efficiency of normalising flows [13]. Recent work has shown that this framework can be further simplified to enable generation in as little as a single inference step [4], making it appealing for video generation where efficiency is critical.

Despite these advantages, applications of flow matching to echocardiography remain limited. Yazdani et al. [28] apply linear flow matching for echocardiogram synthesis but restrict generation to key frames, leaving temporal dynamics unmodelled. Reynaud et al. [24] extend flow matching to latent video generation conditioned on EF; however, their approach is limited to frame animation, relies on temporally normalised inputs obtained via cropping and resampling to a fixed number of frames, and requires ~ 100 inference steps.

In practice, echocardiographic data is highly heterogeneous. Sequences vary substantially in length, frame rate, and quality depending on acquisition conditions. For example, the CAMUS dataset contains videos depicting only end-diastole to end-systole in as few as ten frames. Existing approaches [25,24] simplify modelling by enforcing fixed sequence lengths, which either necessitates discarding shorter sequences or upsampling them via temporal interpolation. Both strategies risk losing information about the original temporal dynamics, limiting applicability to real-world clinical datasets. Our contributions are:

1. We introduce `EchoLVFM`, the first one-step latent video flow-matching framework for echocardiogram generation, enabling temporally coherent video synthesis in a single inference step while preserving the sample quality.

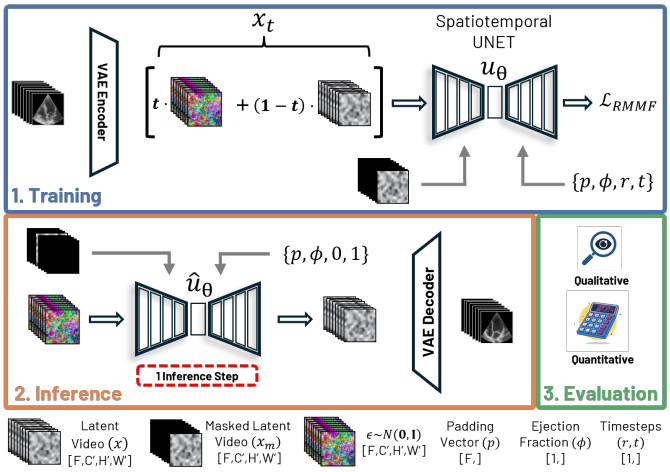

Fig. 1. EchoLVFM. *Training:* Videos are noised and processed in latent space. All but one observed frame are zeroed to form x_m, which serves as conditioning. To support variable-length sequences, a padding vector p indicates which frames are valid observations in the temporally augmented input. The target EF ϕ is provided as global conditioning. r and t denote timesteps with $r < t$, and the model u_θ learns to predict the conditional average velocity over the interval $[r, t]$. *Inference:* A partial video containing as little as a single observed frame, together with a target EF and random noise, is passed to the trained model. One-step integration produces the generated video.

2. `EchoLVFM` enables controllable echocardiogram generation via global conditioning on clinically meaningful variables, demonstrated through precise control of left-ventricular EF, and supports reconstruction and counterfactual synthesis from partially observed sequences.
3. We propose a masked video conditioning strategy that supports variable-length videos, eliminating the need for aggressive temporal normalisation and extending generation beyond single-frame animation to real-world clinical settings.

2 Methods

Flow Matching. Flow matching learns a velocity field $v(x_t, t) = \frac{dx_t}{dt}$ that transports a data sample $x \sim p_{\text{data}}$ to a noise sample $\epsilon \sim \mathcal{N}(0, I)$ as t evolves from 0 to 1. Multiple interpolation paths can be defined between ϵ and x. In linear flow matching, the interpolation $x_t = (1-t)x + t\,\epsilon$ yields a constant target velocity $v(x_t, t) = \epsilon - x$, and a model v_θ is trained by minimising $\mathbb{E}_{\epsilon,x,t}\|v_\theta(x_t, t) - (\epsilon - x)\|_2^2$. At inference, the trajectory is obtained by solving the ODE, e.g. with Euler updates $x_{t+\Delta t} = x_t + \Delta t\, v_\theta(x_t, t)$. An alternative formulation, MeanFlow [4], models the *average* velocity over an interval $[r, t]$, $u(x_t, r, t) = \frac{1}{t-r}\int_r^t v(x_\tau, \tau)\, d\tau$,

which yields the MeanFlow identity

$$(t-r)u(x_t, r, t) = \int_r^t v(x_\tau, \tau)\, d\tau. \tag{1}$$

Differentiating w.r.t. t gives $u(x_t, r, t) = v(x_t, t) - (t-r)\frac{d}{dt}u(x_t, r, t)$. Substituting $\frac{dx_t}{dt} = v(x_t, t) = \epsilon - x$ produces the effective regression target

$$u_{\text{tgt}} \;=\; v(x_t, t) - (t-r)\big(v(x_t, t)\partial_x u_\theta + \partial_t u_\theta\big) \;\;=\;\; (\epsilon - x) - \mathcal{I}(x_t, r, t). \tag{2}$$

A network $u_\theta(x_t, r, t)$ can then be trained to regress u_{tgt} using the objective $\mathcal{L}_{\text{MF}} = \mathbb{E}_{\epsilon,x,r,t}\|u_\theta(x_t, r, t) - \text{sg}(u_{\text{tgt}})\|_2^2$, where $\text{sg}(\cdot)$ is the stop-gradient operator which prevents double backpropagation through the Jacobian–vector product. Once trained, sampling uses the mean velocity $x_r = x_t - (t-r)u(x_t, r, t)$, which can be used to perform one-step sampling like so: $x = \epsilon - u(\epsilon, 0, 1)$.

EchoLVFM. We build upon MeanFlow by introducing `EchoLVFM` (Fig. 1), extending the framework to conditional latent video generation for controllable echocardiogram synthesis. EchoLVFM takes as input a partial video, potentially containing as little as a single observed frame, a target EF (ϕ), and a padding indicator p, and generates a sequence of temporal length $f \leq F$, where F denotes the maximum supported length.

Let $s \in \mathbb{R}^{f\times C\times H\times W}$ denote a sequence with f frames. The sequence is encoded via a VAE into latent space and temporally normalised to length F, yielding $x \in \mathbb{R}^{F\times C'\times H'\times W'}$. If $f > F$, frames are uniformly downsampled; if $f < F$, zero-padding is applied along the temporal dimension, thus preserving shorter sequences. A masked latent counterpart x_m is constructed by masking a subset of observed (non-padding) frames to be used for conditioning. A binary padding vector $p \in \{0,1\}^F$ indicates valid frames ($p_f = 0$) and padded frames ($p_f = 1$). The model predicts a conditional mean velocity $u_\theta(x_t, r, t; x_m, \phi, p)$ in the latent space. Henceforth, we denote the conditioning variables $\{x_m, \phi, p\}$ as $\mathbf{c}$.

Applying vanilla MeanFlow to videos of varying lengths introduces two issues. First, sequences with fewer valid frames contribute fewer terms to the loss, biasing optimisation towards longer videos. Second, padded frames contribute zero targets, encouraging the model to generate blank frames. Both effects intensify as F increases, limiting the range of video lengths that can be generated.

To address this, we introduce a temporal loss mask $LM = 1 - p$ that excludes padded frames from optimisation. We incorporate LM into the MeanFlow loss:

$$\mathcal{L}_{\text{MMF}} = \mathbb{E}_{\epsilon,x,t,r}\big[\alpha \cdot \|M \odot e\|_2^2\big], \quad \alpha = [(\textstyle\sum_{f=1}^{F} LM_f)C'H'W']^{-1}, \tag{3}$$

where $\hat{u}_\theta$ is the prediction, $e = \hat{u}_\theta - \text{sg}(u_{\text{tgt}})$ is the error, M denotes the mask LM broadcast to the shape of e, and α ensures that each video contributes equally to the loss function. Following [6], we adaptively weight the loss function by a factor $w = \big(\|M \odot e\|_2^2 + \varepsilon\big)^{-h}$, with gradients stopped through w, yielding

$$\mathcal{L}_{\text{MMF}}^{\text{adapt}} = \mathbb{E}_{\epsilon,x,t,r}[\text{sg}(w) \cdot \alpha \cdot \|M \odot e\|_2^2]. \tag{4}$$

This weighting upscales gradient contributions when the residual is small, avoiding vanishing gradients as the model approaches the target and stabilising optimisation across noise levels. Recent analyses [29,5] of the MeanFlow framework have highlighted training instability and optimisation challenges. To tackle this, we introduce a masked reconstruction regulariser. Using the predicted mean velocity $\hat{u}_\theta$, we calculate $\hat{x} = x_t - t\big(\hat{u}_\theta(x_t, r, t; \mathbf{c}) + \mathcal{I}(x_t, r, t)\big)$ and penalise

$$\mathcal{L}_{\text{rec}} = \mathbb{E}_{\epsilon,x,t,r}\left[\alpha \cdot \|M \odot (\hat{x} - x)\|_2^2\right]. \tag{5}$$

The final `EchoLVFM` objective is the regularised masked mean flow loss

$$\boxed{\mathcal{L}_{RMMF} = \mathcal{L}_{\text{MMF}}^{\text{adapt}} + \lambda_{\text{rec}}\,\mathcal{L}_{\text{rec}}} \tag{6}$$

combining adaptive masked MeanFlow with reconstruction regularisation.

Geng et al. [4] reported that they achieved the best performance when alternating objectives during training, using linear flow matching 75% of the time and Mean Flow 25% of the time. To adapt this for our setting, we used a masked version of linear flow matching $\mathcal{L}_{MLF}$. Concretely,

$$\mathcal{L}_{MLF} = \mathbb{E}_{\epsilon,x,t}[\alpha \cdot \|M \odot (\hat{v}_\theta(x_t, t; \mathbf{c}) - (\epsilon - x))\|_2^2], \tag{7}$$

Although the experiments of this work focus on the most challenging setting where only a single observed frame is present in x_m at inference, EchoLVFM generalises beyond this regime. By interleaving an existing sequence with padded frames and reconstructing it, our method naturally performs temporal upsampling. Moreover, because the padding indicator p is incorporated during training, the generated sequence length can be controlled up to F, enabling variable-length synthesis unlike prior approaches that always generate a video of fixed length.

Ejection Fraction. For all videos, we assign a proxy EF using the area–length method [23], computed directly from LV segmentation masks. Using the LV cavity area A and long-axis length L, volume is approximated as $V = \frac{8}{3\pi}\frac{A^2}{L}$, yielding $EF \approx 1 - \frac{V_{\text{ES}}}{V_{\text{ED}}} = 1 - \left(\frac{L_{\text{ED}}}{L_{\text{ES}}}\right)\left(\frac{A_{\text{ES}}}{A_{\text{ED}}}\right)^2$. Each video is thus assigned a single proxy EF, used as the conditioning variable ϕ during training and inference.

3 Experiments

Data. We use the CAMUS dataset [15], comprising 1,000 echocardiogram videos from 500 patients, each providing apical two- and four-chamber views. We adopt the original patient-level split: 800 videos (400 patients) for training, 100 for validation, and 100 for testing. The sequences span ED to ES and are heterogeneous in both quality and temporal length (10–42 frames), reflecting realistic clinical acquisition variability. To enable latent video generation, publicly available echocardiogram VAEs were evaluated on a reconstruction task. The '$4f4_{[28,28,4]}$'

model from [24] achieved the best overall performance (FID = 5.16, FVD = 36.1, SSIM = 0.958, LPIPS = 0.0272, PSNR = 35.3, MAE = 1.00%, RMSE = 1.73%) and was therefore used in all experiments. All videos were resized to 112×112 and encoded into its latent space.

Training. We employ a conditional spatio-temporal UNet with four resolution levels and feature dimensions $\{128, 128, 256, 256\}$, integrating self- and cross-attention to jointly model spatial structure and temporal dynamics under conditioning. The model contains 76.8M parameters. A maximum length $F = 32$ was chosen. Following extensive hyperparameter exploration, the final configurations trained for 1000 epochs with $\lambda_{rec} = 1$, cosine annealing with an initial learning rate of 5×10^{-5}, and batch size 2. All experiments, including inference speed measurements, were conducted on a single NVIDIA L40s 48GB GPU. Flash attention with support for `torch.func.jvp` was not implemented at the time of our experiments. Consequently, we implemented a custom attention processor to retain memory-efficient attention, using the `jvp-flash-attention` [18] and `diffusers` [21] libraries.

Evaluation. We compared EchoLVFM against baselines trained solely with $\mathcal{L}_{MLF}$. Preliminary experiments showed Linear performance plateaued after 25 inference steps. Quantitative evaluation was conducted along three axes: *efficiency*, *video quality*, and *EF adherence*, for both reconstruction (Rec) and generation (Gen). In Rec, the true EF was used for conditioning. In Gen, an EF differing by at least 5% from the true value was sampled from the challenging range $[0, 100]$, exceeding the training distribution and typical clinical values.

Sampling efficiency was measured and averaged over 100 iterations. Video quality was assessed using FID [8], FVD [26], SSIM [27], and LPIPS [30] at resolution 112×112. For robustness, three independent noise samples were generated per test video (300 samples total). To evaluate EF adherence, the LV cavity in generated videos was segmented using a pretrained nnUNet[11], and the EF was computed. We report R^2, MAE, and RMSE between requested and observed EF. In line with prior work [25], we also report performance after rejection sampling, leveraging the independent EF estimator. Qualitative evaluation was performed via a blinded real-vs-fake assessment by two expert cardiologists (>15 years of experience each). After calibration with real examples, they classified 120 videos (60 real, 60 generated from the best Linear and EchoLVFM models).

4 Results & Discussion

Quantitative Evaluation. Unless stated otherwise, results correspond to the most challenging setting where only a single observed frame is present in x_m at inference. The strongest diffusion-based baselines are included in Table 1; notably, these methods do not evaluate per-video EF adherence.

As shown in Table 1, EchoLVFM generates 18.5 videos per second using a single sampling step, compared to 0.37 videos per second for linear flow, which

Table 1. Quantitative Results. Comparison of baseline and proposed methods. `Cond.` lists the inputs: Image (I) or single frame in x_m, Text (T), Motion Mask (MM), Video (V), and partial video (pV). Classifier-free guidance (CFG). `Steps (Vid/s)` reports inference steps and video throughput. `Task` denotes reconstruction (Rec) and generation (Gen) . pmf is the proportion of masked frames and h is the exponent in adaptive weight w. μ denotes the mean score; (RS) denotes rejection sampling (three samples per conditioning, best retained). **Bold** and **Blue** indicate best Rec and Gen performance. † Inference steps not reported (DDPM default 1000). ‡ Clip length unspecified (likely FVD_{16}).

Method	Cond.	Steps (Vid/s)	Task	Video Quality				EF Adherence		
				FID ↓	FVD_{10} ↓	SSIM ↑	LPIPS ↓	R^2, % ↑ μ (RS)	MAE, % ↓ μ (RS)	RMSE, % ↓ μ (RS)
$HeartBeat_{2C}$[31]	I, T	1000†	-	36.3	9.6 ‡	0.61	-			
$HeartBeat_{4C}$[31]	I, T	1000†	-	36.0	12.6 ‡	0.61	-			
CES[14]	MM,V	1000	-	26.6	69.6 ‡	0.57	-			
Linear	I, EF	25 (0.37)	Rec	40.4	153.9	**0.73**	**0.128**	**73 (84)**	**5.7 (4.0)**	**7.3 (5.6)**
			Gen	42.2	154.2	-	-	59 (67)	**14.4** (12.2)	18.6 (16.8)
$Linear_{CFG}$	I, EF	25 (0.37)	Rec	46.8	164.7	**0.73**	0.129	49 (64)	7.8 (6.1)	10.1 (8.5)
			Gen	46.8	178.6	-	-	6 (16)	23.0 (21.3)	28.2 (26.7)
$Ours_{h=2}^{\lambda_{rec}=0}$	I, EF	**1** (18.6)	Rec	41.9	169.9	0.72	0.133	42 (77)	8.1 (4.6)	10.7 (6.8)
			Gen	41.8	166.6	-	-	53 (70)	16.0 (11.7)	20.0 (15.9)
Ours ($h = 1$)	I, EF	1 (18.4)	Rec	49.2	162.1	**0.73**	0.133	57 (80)	7.2 (4.2)	9.3 (6.3)
			Gen	49.3	161.5	-	-	**64** (78)	**14.4** (10.8)	**17.5** (13.7)
Ours ($h = 2$)	I, EF	1 (18.5)	Rec	**38.5**	**138.8**	0.70	0.136	32 (75)	9.1 (4.9)	11.6 (7.0)
			Gen	**38.5**	**144.1**	-	-	52 (74)	16.3 (11.4)	20.1 (14.9)
Ours ($h = 2$) $pmf = 50\%$	pV,EF	1 (18.5)	Rec	38.6	155.2	0.82	0.098	93 (96)	3.1 (2.2)	3.8 (2.9)
			Gen	38.9	150.5	-	-	-1 (7)	25.5 (24.3)	29.2 (28.1)

requires 25 steps, representing an approximate 50× improvement in efficiency. Despite this substantial reduction in inference cost, EchoLVFM achieves competitive and, in several cases, superior video quality metrics. The best-performing configuration, $EchoLVFM_{h=2}$, attains the lowest FID (38.5) for both reconstruction and generation, alongside the strongest FVD scores (138.8 Rec, 144.1 Gen). While the Linear model achieves slightly stronger perceptual similarity (LPIPS = 0.128), its distributional metrics remain higher (FID 40.4/42.2; FVD 153.9/154.2). These results demonstrate that a single-step EchoLVFM model can match or exceed the distributional quality of multi-step linear flow models while operating at a fraction of the computational cost, highlighting a favourable efficiency–quality trade-off. Ablating the reconstruction loss ($\lambda_{\text{rec}} = 0$) degrades performance, with FVD worsening by ≈ 30 points, highlighting its importance.

The nnUNet model used for LV segmentation achieved a Dice score of 93% on the test set and was subsequently applied to generated videos for EF estimation. Table 1 shows that Linear performs best in the reconstruction task, whereas EchoLVFM achieves the strongest performance in the generation task. Notably, generation is the more challenging setting: models are conditioned on a conflicting x_m and evaluated on EF values extending beyond the training distribution.

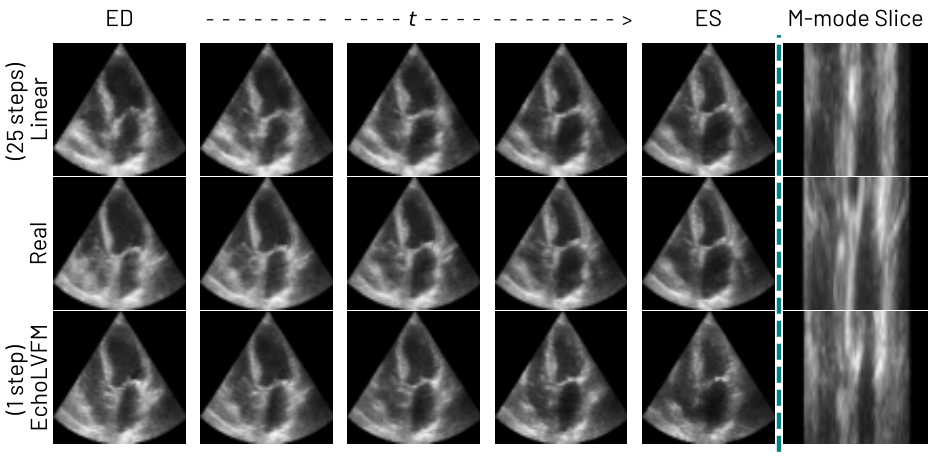

Fig. 2. Qualitative Results. Columns 1-5 show frames sampled between and including ED and ES, while column 6 presents the M-mode slice of the middle row over time.

This suggests that while Linear excels at reconstructing inputs with their original EF, EchoLVFM$_{h=1}$ generalises more effectively under distributional shift.

EchoLVFM$_{h=2}^{pmf=50\%}$ shows when only 50% of valid frames in x_m are masked. In this setting, frame-level fidelity improves, as reflected by higher SSIM and lower LPIPS, while video-level dynamics degrade slightly, as indicated by increased FVD, with FID remaining stable. The strongest effects are observed in EF adherence, with $R^2 = 93\%$ in Rec but an expected drop to -1 in Gen, as x_m acts as a confounder when conditioning on a new EF. This indicates that, while the previous results were obtained under the hardest setting, EF adherence in Rec improves substantially when additional frames are included in x_m.

Collectively, these findings demonstrate that EchoLVFM enables efficient, one-step, controllable video generation while maintaining competitive visual fidelity and strong clinical adherence under challenging conditioning regimes.

Qualitative Results. Across both clinicians, the overall confusion matrix was as follows: {R as R: 59, R as S: 61, S as R: 40, S as S: 80} where R and S correspond to Real and Synthetic, respectively. This corresponds to an overall accuracy of 58%, compared with 50% expected from random guessing in this binary task. Real videos were correctly identified 49% of the time, indicating that clinicians struggled to reliably distinguish real echocardiograms from synthetic ones and suggesting strong perceptual realism in the generated videos. Comparing methods, synthetic videos produced using EchoLVFM$_{h=2}$ were identified as fake in 73% of cases, and those produced using Linear were identified as fake in 60% of cases. This suggests that Linear yields slightly more perceptually convincing outputs. However, EchoLVFM achieves generation in a single inference step, while Linear requires 25 steps, again illustrating the trade-off between efficiency and perceptual fidelity.

5 Conclusion

We introduced EchoLVFM, a one-step latent video flow-matching framework for controllable echocardiogram generation. By developing a novel loss, incorporating masked conditioning, and a padding indicator, our method removes the lower bound on usable sequence length, enabling shorter sequences to be retained rather than discarded. EchoLVFM supports conditioning on an arbitrary number of observed frames and naturally extends to tasks such as temporal upsampling.

Results show that EchoLVFM achieves competitive video quality and EF adherence in a single inference step, yielding substantial efficiency gains. While one-pass sampling has traditionally been a key advantage of GAN-based models, our results demonstrate that continuous-time generative models can now approach this level of efficiency without sacrificing stability or controllability. Our findings suggest that one-step flow matching provides a practical foundation for efficient, controllable modelling of realistic clinical video data.

Acknowledgements. The authors thanks Daria Kulikova and Anna Novikova for participating in the quiz. We thank Phil Wang (lucidrains) [1] for their implementation of the base Mean Flow method. The authors would like to acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility (http://dx.doi.org/10.5281/zenodo.22558)

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

References

1. Phil Wang / rectified-flow-pytorch · GitLab (2026), https://gitlab.com/lucidrains/rectified-flow-pytorch
2. Aly, I., Rizvi, A., Tubbs, R.S., Loukas, M.: Cardiac ultrasound: An Anatomical and Clinical Review. Translational Research in Anatomy **22**, 100083 (1 2021). https://doi.org/10.1016/J.TRIA.2020.100083
3. Friedrich, P., Frisch, Y., Cattin, P.C.: Deep Generative Models for 3D Medical Image Synthesis
4. Geng, Z., Deng, M., Bai, X., Kolter, J.Z., He, K.: Mean Flows for One-step Generative Modeling (5 2025), https://arxiv.org/pdf/2505.13447
5. Geng, Z., Lu, Y., Wu, Z., Shechtman, E., Zico Kolter, J., He, K.: Improved Mean Flows: On the Challenges of Fastforward Generative Models
6. Geng, Z., Pokle, A., Luo, W., Lin, J., Kolter, J.Z.: Consistency Models Made Easy. 13th International Conference on Learning Representations, ICLR 2025 pp. 96638–96665 (10 2024), http://arxiv.org/abs/2406.14548
7. Goodfellow, I.J., Sutskever, I., Duvenaud, D., Doersch, C.: Generative Adversarial Networks. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **11046 LNCS**(NeurIPS), 1–9 (6 2014), https://arxiv.org/abs/1406.2661v1

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. Advances in Neural Information Processing Systems **2017-December**, 6627–6638 (2017). https://doi.org/10.18034/ajase.v8i1.9
9. Ho, J., Jain, A., Abbeel, P.: Denoising Diffusion Probabilistic Models. Advances in Neural Information Processing Systems **2020-December** (6 2020), https://arxiv.org/pdf/2006.11239
10. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video Diffusion Models. Advances in Neural Information Processing Systems **35** (4 2022), https://arxiv.org/pdf/2204.03458
11. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**, 203–211 (2021)
12. Kingma, D.P., Welling, M.: Auto-Encoding Variational Bayes. 2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings (12 2013), https://arxiv.org/abs/1312.6114v11
13. Kobyzev, I., Prince, S.J., Brubaker, M.A.: Normalizing Flows: An Introduction and Review of Current Methods. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(11), 3964–3979 (11 2021), https://ieeexplore.ieee.org/abstract/document/9089305
14. Kondori, N., Liang, H., Vaseli, H., Xie, B., Luong, C., Abolmaesumi, P., Tsang, T., Liao, R.: ControlEchoSynth: Boosting Ejection Fraction Estimation Models via Controlled Video Diffusion (8 2025), https://arxiv.org/pdf/2508.17631
15. Leclerc, S., Smistad, E., Bernard, O.: Deep Learning for Segmentation Using an Open Large-Scale Dataset in 2D Echocardiography. IEEE transactions on medical imaging **38**(9), 2198–2210 (9 2019), https://pubmed.ncbi.nlm.nih.gov/30802851/
16. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow Matching for Generative Modeling. 11th International Conference on Learning Representations, ICLR 2023 (10 2022), https://arxiv.org/pdf/2210.02747
17. McDonagh, T.A., Metra, M., Zeppenfeld, K.: 2023 Focused Update of the 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. European heart journal **44**(37), 3627–3639 (10 2023). https://doi.org/10.1093/EURHEARTJ/EHAD195
18. Morehead, A.: JVP Flash Attention (9 2025), https://github.com/amorehead/jvp_flash_attention
19. Oladokun, E., Abdulkareem, M., Šprem, J., Grau, V.: Transesophageal Echocardiography Generation using Anatomical Models. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **14379 LNCS**, 43–52 (10 2024). https://doi.org/10.1007/978-3-031-58171-7{_}5
20. Oladokun, E., Ou, Y., Novikova, A., Kulikova, D., Thomas, S., Šprem, J., Grau, V.: From Transthoracic to Transesophageal: Cross-Modality Generation using LoRA Diffusion (8 2025), http://arxiv.org/abs/2508.13077
21. von Platen, P., Patil, S., Sayak, P., Thomas Wolf: Diffusers: State-of-the-art diffusion models (2022), https://github.com/huggingface/diffusers
22. Potter, A., Pearce, K., Hilmy, N.: The benefits of echocardiography in primary care. British Journal of General Practice **69**(684), 358–359 (7 2019). https://doi.org/10.3399/BJGP19X704513
23. Pujadas, S., Reddy, G.P., Weber, O., Lee, J.J., Higgins, C.B.: MR imaging assessment of cardiac function. Journal of magnetic resonance imaging : JMRI **19**(6), 789–799 (6 2004). https://doi.org/10.1002/jmri.20079

24. Reynaud, H., Gomez, A., Leeson, P., Meng, Q., Kainz, B.: EchoFlow: A Foundation Model for Cardiac Ultrasound Image and Video Generation. IEEE TRANSACTIONS ON MEDICAL IMAGING **XX** (3 2025), https://arxiv.org/pdf/2503.22357
25. 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. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **14229 LNCS**, 142–152 (2 2024), http://arxiv.org/abs/2303.12644
26. Unterthiner, T., Van Steenkiste, S., Kurach, K., Brain, G., Marinier, R., Michalski, M., Gelly, S.: Towards Accurate Generative Models of Video: A New Metric & Challenges (12 2018), https://arxiv.org/pdf/1812.01717
27. 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). https://doi.org/10.1109/TIP.2003.819861
28. Yazdani, M., Medghalchi, Y., Ashrafian, P., Hacihaliloglu, I., Shahriari, D.: Flow Matching for Medical Image Synthesis: Bridging the Gap Between Speed and Quality (3 2025), https://arxiv.org/pdf/2503.00266
29. Zhang, H., Siarohin, A., Menapace, W., Vasilkovsky, M., Tulyakov, S., Qu, Q., Skorokhodov, I.: ALPHAFLOW: UNDERSTANDING AND IMPROVING MEANFLOW MODELS https://github.com/snap-research/alphaflow.
30. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition pp. 586–595 (2018). https://doi.org/10.1109/CVPR.2018.00068
31. Zhou, X., Huang, Y., Xue, W., Dou, H., Cheng, J., Zhou, H., Ni, D.: HeartBeat: Towards Controllable Echocardiography Video Synthesis with Multimodal Conditions-Guided Diffusion Models. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **15007 LNCS**, 361–371 (6 2024), https://arxiv.org/pdf/2406.14098