EchoNet-Synthetic: Privacy-preserving Video Generation for Safe Medical Data Sharing

Hadrien Reynaud[1,2], Qingjie Meng[3], Mischa Dombrowski[4], Arijit Ghosh[4], Thomas Day[5,6], Alberto Gomez[5,7], Paul Leeson[7,8], and Bernhard Kainz[2,4]

[1] UKRI CDT in AI for Healthcare, Imperial College London, London, UK
[2] Department of Computing, Imperial College London, London, UK
hadrien.reynaud19@imperial.ac.uk
[3] School of Computer Science, Univeristy of Birmingham, Birmingham, UK
[4] Friedrich–Alexander University Erlangen–Nürnberg, DE
[5] School of BMEIS, King's College London, London, UK
[6] Guy's and St Thomas' NHS Foundation Trust, London, UK
[7] Ultromics Ltd, Oxford, UK
[8] John Radcliffe Hospital, Cardiovascular Clinical Research Facility, Oxford, UK

Abstract. To make medical datasets accessible without sharing sensitive patient information, we introduce a novel end-to-end approach for generative de-identification of dynamic medical imaging data. Until now, generative methods have faced constraints in terms of fidelity, spatio-temporal coherence, and the length of generation, failing to capture the complete details of dataset distributions. We present a model designed to produce high-fidelity, long and complete data samples with near-real-time efficiency and explore our approach on a challenging task: generating echocardiogram videos. We develop our generation method based on diffusion models and introduce a protocol for medical video dataset anonymization. As an exemplar, we present EchoNet-Synthetic, a fully synthetic, privacy-compliant echocardiogram dataset with paired ejection fraction labels. As part of our de-identification protocol, we evaluate the quality of the generated dataset and propose to use clinical downstream tasks as a measurement on top of widely used but potentially biased image quality metrics. Experimental outcomes demonstrate that EchoNet-Synthetic achieves comparable dataset fidelity to the actual dataset, effectively supporting the ejection fraction regression task. Code, weights and dataset are available at https://github.com/HReynaud/EchoNet-Synthetic.

Keywords: Dataset Generation · Video Diffusion · Echocardiography

1 Introduction

Medical datasets play a crucial role for learning-based medical image analysis [21, 27] and form the basis for a potential future medical foundation model [4]. However, the confidential nature of medical data, coupled with concerns over privacy and extensive individual patient-consent requirements, often restricts dataset

releases. To address these challenges, we propose a new protocol for generative medical dataset de-identification. In our protocol, we first train a generative model, *e.g.*, diffusion models, to learn the data distribution from a real training set. Then, we sample our generative model to generate a synthetic dataset, with the same properties as the real training set, *e.g.*, diversity, quantity, population statistics, etc. This synthetic dataset is then filtered to preserve patient privacy and to safeguard against model memorization. To verify the quality of this privacy-preserving synthetic data, we train a downstream model (regression, classification, segmentation, ...) on the generated data, and evaluate it on a held out set of real data. We compare the model's performance against the same downstream model trained on the real training data. If their performance difference is within an acceptable range, it proves that the synthetic dataset is a valid substitute to the real training dataset, and can be shared as such without data privacy constraints.

Contributions: (1) We introduce the first Latent Video Diffusion Model (LVDM) capable of generating high-fidelity, long echocardiogram videos at near real-time speeds. (2) We propose a comprehensive protocol for generating useful medical datasets. This protocol emphasizes the importance of dataset quality, validated through the training of downstream models (*e.g.*, ejection fraction regression) and ensures that synthetic datasets can effectively support medical imaging research and clinical translation. (3) We release EchoNet-Synthetic, a pioneering fully synthetic echocardiography dataset that maintains the quality and diversity required for effective downstream model training, while protecting patient privacy. Models trained on EchoNet-Synthetic exhibit performance metrics comparable to those trained on real datasets, thereby validating the efficacy of our dataset generation protocol.

Related Works: Video generation is an important research area within computer vision. Recently, several works started studying diffusion models for video generation [1, 3, 10, 12, 14, 15, 17, 19, 20, 24, 30, 34–36, 39]. These diffusion-based video generation methods can achieve outstanding modelling capabilities, but they suffer from excessive computational requirements. To make them more tractable, concurrent works explore Latent Diffusion Models (LDMs) for image generation [26] and for video generation [3, 11, 37, 39]. VideoLDM [3] extends LDMs to high-resolution video generation by turning pre-trained image LDMs into video generators by inserting temporal layers. Latent Video Diffusion Models (LVDM) [11] proposes a hierarchical diffusion model operating in the video latent space, allowing long video generation through a second model.

In the field of ultrasound generation, some works focus on physics-based simulators [16, 29] while others use Generative Adversarial Network (GAN)-based methods for individual images. These works condition their models on Magnetic Resonance Imaging (MRI), Computed Tomography (CT) [31, 33] or simulated [8, 32] images. Other works focus on ultrasound video generation. [18] presents a motion-transfer-based method for pelvic ultrasound videos, while [25] introduces a causal model for echocardiogram video generation. [24] introduces a diffusion model-based video generation method for echocardiogram synthesis, which how-

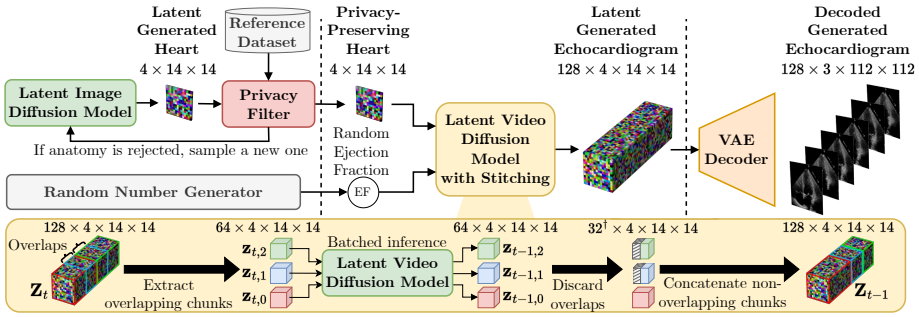

Fig. 1. Our inference video generation pipeline, consisting of a latent image diffusion model with a privacy filter, a latent video diffusion model with video stitching and a VAE decoder. All dimensions are given in (time ×) channels × height × width. † In the video stitching method (yellow block), the first half of $\mathbf{z}_{t-1,0}$ is not discarded.

ever requires extensive sampling times. In this work, we build upon LDMs for fast and temporally-consistent echocardiogram generation.

Privacy issues have recently gained a lot of attention due to diffusion models' ability of to memorize training samples [5]. Current approaches focus on two aspects, privatizing the generative model [7] and filtering out the generated samples that raise privacy concerns [6, 22]. In this work, we focus on the second approach to remove samples that can be linked back to the training set.

2 Method

The goal of our method is to generate videos suitable to train downstream models while preserving patient privacy and allowing unlimited video duration. In this section, we present our de-identification protocol, consisting of (1) video generation, (2) privacy filtering, (3) video stitching and (4) evaluation on a downstream task. Figure 1 summarizes our video generation pipeline.

Video Generation: In this work, we implement an LDM pipeline for video generation. To do so, we train three models, a Variational Auto-Encoder (VAE), a Latent Image Diffusion Model (LIDM) and a Latent Video Diffusion Model (LVDM), independently.

We start by training a VAE on an image reconstruction task. The VAE is made of four downsampling blocks, based on attention and convolutional layers. Given an input image $I \in \mathbb{R}^{C \times H \times W}$, the variational autoencoder (VAE) encodes it to produce two 3D latent tensors: a mean μ and a standard deviation σ, such that $\{\mu, \sigma\} \in \mathbb{R}^{4 \times H/8 \times W/8}$. These latent tensors are used to define an n-dimensional Gaussian distribution, which is sampled to produce latent representations of the image. From this latent representation, the VAE decoder reconstructs the input image. The VAE is trained by using a combination of MSE, LPIPS [38] and GAN [9] loss, and a KL-divergence loss which constrains the latent space to be Gaussian distributed. After the VAE is trained, we encode

the datasets into the VAE latent space, and train the diffusion models on these latent representations.

The LIDM and LVDM are trained as discrete-time Denoising Diffusion Probabilistic Models (DDPMs) [13] with a v-prediction [28] objective. The noising process of the DDPM is defined as a Markov chain, where

$$p(\mathbf{z}_{1:T}|\mathbf{z}_0) = \prod_{t=1}^{T} p(\mathbf{z}_t|\mathbf{z}_{t-1}). \tag{1}$$

The noise at each step is Gaussian, such that $p(\mathbf{z}_t|\mathbf{z}_{t-1}) = \mathcal{N}(\mathbf{z}_t; \sqrt{1-\beta_t}\mathbf{z}_{t-1}, \beta_t\mathbf{I})$, and β_t specifies the noise variance. Here, $\mathbf{z}_t$ is the noisy latent and T is the total number of diffusion steps. For the denoising process, a neural network is trained to estimate the velocity $\mathbf{v}$, such that $\mathbf{v} = \alpha_t\epsilon - \sigma_t\mathbf{z}$, and updates $\mathbf{z}$ through

$$\mathbf{z}_{t-1} = \alpha_t\mathbf{z}_t - \sigma_t\mathbf{v}_\theta(\mathbf{z}_t), \tag{2}$$

with α_t and σ_t as time-dependent coefficients, and $\mathbf{v}_\theta$ as the v-prediction model.

The difference between the LIDM and LVDM resides in their backbone neural networks and their training strategy. The LIDM is implemented as an unconditional UNet with residual downsampling blocks containing convolutional layers and self-attention layers. We set the number of residual blocks to four, with channel sizes 128, 256, 256, 512. The LIDM is trained to generate latent images by minimizing an MSE loss over its outputs and the VAE pre-encoded images. The LVDM consists of a Spatio-Temporal UNet [2], with four residual blocks and channel sizes 128, 256, 256, 512. The residual blocks contain convolutional layers and cross-attention layers, both with space-time separation. To train the LVDM we use an MSE loss, and condition the model on an encoded real heart image and a given Left Ventricular Ejection Fraction (LVEF) score.

During inference, we combine the three models into a video generation pipeline, as shown in Figure 1. We start by sampling Gaussian noise, and use the LIDM to progressively generate a random latent heart. Then, this latent heart, a randomly sampled LVEF score and some Gaussian noise are passed to the LVDM. The LVDM produces a latent video of the latent heart, where the motion of the heart is conditioned by the given LVEF. Finally, we use the VAE-decoder to decode the latent video frames into a pixel-space echocardiogram.

Video Stitching: Generating long videos with diffusion models remains a challenging task, even in a latent space, because of the training computational cost. Naively extending the number of video frames during inference does not yield realistic temporal consistency. This is because the attention layers in the backbone model cannot handle more than the time window it was trained on. To circumvent this limitation and allow the generation of longer videos during inference, we split and denoise a long noisy latent video into overlapping pieces, and then stitch them back together (see Figure 1). As these overlapping pieces are aware of the content of their predecessor, our method ensures seamless temporal continuity in long videos. Formally, we start from a noisy latent video $\mathbf{Z}_t \in \mathcal{R}^{l_v \times c \times h \times w}$, where l_v is the video length (*e.g.*, 128). The LVDM $\mathbf{v}_\theta$ is

trained on videos with length l_m (*e.g.*, 64), and $l_v > l_m$. We define an overlapping factor $o = \frac{l_m}{2}$ and split the $\mathbf{Z}_t$ tensor into k overlapping chunks, $k = \frac{l_v - l_m}{o} + 1$. This produces $\mathbf{z}_{t,i} = \mathbf{Z}_t[i \cdot (l_m - o) : i \cdot (l_m - o) + l_m], \forall i \in \{0, 1, \ldots, k-1\}$. Then, we denoise these overlapping noisy latent videos with Equation (2), *i.e.*, $\mathbf{z}_{t-1,i} = \alpha_t \cdot \mathbf{z}_{t,i} - \sigma_t \cdot \mathbf{v}_\theta(\mathbf{z}_{t,i})$. Finally, to reconstruct $\mathbf{Z}_{t-1}$ from all $\mathbf{z}_{t-1,i}$, we discard the overlaps over the time axis, *i.e.*, for $\forall i > 0$, $\mathbf{z}_{t-1,i} = \mathbf{z}_{t-1,i}[o : l_m]$, and then, we concatenate all updated $\mathbf{z}_{t-1,i}$ over the time axis.

Privacy filtering: Since the LVDM only animates the heart images it is given, ensuring that those images are privacy compliant is enough to guarantee that the corresponding generated videos are also privacy compliant. Thus, we apply the privacy filter on the latents generated by the LIDM. Specifically, we train a re-identification model [22] on the encoded real training set. The re-identification model is trained with a contrastive loss. We set the positive pairs as different frames from the same video and the negative pairs as two frames from different videos. The penultimate layer of the re-identification model is used to compute the distance between two samples. Following [6], we use the Pearson correlation score to measure that distance. During inference, we use the first frame of each video to compute distances. We calculate the distance between all real training samples and all real validation samples. From this distribution of distances, we find the threshold τ, where 95% of the distances are above τ.Then, we compute the distances between all synthetic and all real training samples. For each synthetic sample, if the shortest distance to a real samples is under τ, we consider the synthetic sample to be 'memorized' and exclude it.

Evaluation on downstream tasks: Current generated medical datasets are usually evaluated with standard generative metrics such as Fréchet Inception Distance (FID), Fréchet Video Distance (FVD) and Inception Score (IS). However, these metrics do not directly correlate with performance on downstream tasks such as classification, regression, or segmentation. To address this, following our de-identification protocol, we propose to generate synthetic datasets and assess them on established downstream benchmarks to confirm their efficacy. Specifically, we first use our privacy-preserving video generation pipeline to generate a synthetic echocardiogram dataset. Then, we compare the performance of an LVEF regression model trained on real data, against the same model trained on our generated dataset. After training, both models are evaluated on the same real test set to ensure that the surrogate dataset's distribution aligns with the real data, thereby validating its quality and diversity.

3 Experiments

Data: To facilitate reproducibility, we use the publicly available echocardiogram datasets, EchoNet-Dynamic (Dyn) [21] and the EchoNet-Pediatric (Ped) [23], as real datasets. Every echocardiogram in the real datasets has a manually labelled LVEF. Dyn contains exclusively Apical 4 Chamber (A4C) views, split into training, validation and test sets. Ped contains A4C and Parasternal Short Axis (PSAX) views, with each view split into 10 folds. To match the datasets

Table 1. Performance of the VAE. The Ped datasets do not have enough videos with more than 128 frames to enable FVD_{128} computation.

	Reconstruction Metrics					**Generative Metrics**			
	MSE↓	MAE↓	SSIM↑	PSNR(dB)↑	LPIPS↓	FID↓	FVD_{16}↓	FVD_{128}↓	IS↑
Dyn	$3e\text{-}3_{\pm 8e\text{-}4}$	$0.03_{\pm 5e\text{-}3}$	$0.78_{\pm 0.05}$	$24.9_{\pm 1.08}$	$0.08_{\pm 0.01}$	16.9	87.4	69.2	$2.33_{\pm 0.13}$
Ped(A4C)	$1e\text{-}3_{\pm 7e\text{-}4}$	$0.02_{\pm 6e\text{-}3}$	$0.86_{\pm 0.05}$	$28.9_{\pm 1.95}$	$0.07_{\pm 0.02}$	6.9	21.8	-	$2.90_{\pm 0.09}$
Ped(PSAX)	$2e\text{-}3_{\pm 8e\text{-}4}$	$0.02_{\pm 6e\text{-}3}$	$0.85_{\pm 0.05}$	$28.5_{\pm 2.14}$	$0.08_{\pm 0.02}$	8.1	22.1	-	$3.16_{\pm 0.13}$

structure, we split Ped into Ped(A4C) and Ped(PSAX), and arbitrarily use the first 8 folds for training, the 9^{th} for validation and the 10^{th} for test. All videos have 3 channels, and we preprocess all videos to greyscale.

Model training: Our VAE is trained on all the frames from the training sets of Dyn and Ped. We train the model from scratch for 5 days on 8×A100 (80GB), with a total batch size of 256, and a learning rate of $5e-4$. The VAE compresses the frames from $3 \times 112 \times 112$ to $4 \times 14 \times 14$.

We train three independent LIDMs for Dyn, Ped(A4C) and Ped(PSAX), by using the VAE-encoded train sets of each dataset. The backbone UNet operates in a $4 \times 16 \times 16$ configuration to allow three downsampling steps, so we use replication padding on the latents produced by the VAE to match those dimensions. The three LIDMs are trained for 24h on a single A100 (80Gb), with a batch size of 256, and a learning rate of $3e-4$.

We train a re-identification model for each real dataset on a single A100 with batch size 128 for 1000 epochs. All other parameters are taken from [6].

The LVDM is trained on all the training sets of the three datasets. During training, we sample a pair of latent video and LVEF from the training set. We use a random frame from the latent video as the input heart. Conditioned on this input heart and the paired LVEF, our LVDM is trained to reconstruct the sampled video with a fixed length of 64 frames. We train the model for 2 days with a learning rate of $1e-4$ on 8×A100 (80GB), with a total batch size of 128.

Model evaluation: We evaluate the VAE by computing the usual reconstruction metrics and generative metrics. The results in Table 1 show that our VAE is capable of reconstructing echocardiogram frames with high fidelity, and thus learns a meaningful latent space.

The LIDMs are evaluated using generative metrics FID and IS. We generate 100,000 samples with each LIDM, and apply our privacy filter to remove any sample that has a close match in the corresponding real training set. Then, we compute FID and IS on the unfiltered and filtered VAE-decoded latents. Table 2 shows that our privacy filter maintains the overall quality of the generated images. The 'Rejected Samples' shows that our LIDMs have limited memorization.

The privacy filtering models all reach Recall scores above 99%, ensuring a low false-negative rate. Combined with our distance thresholding method, they drastically reduce the risk of private data being present in the remaining samples.

To evaluate the LVDM, we use generative metrics FID, FVD_{16}, FVD_{128} and IS. We generate 2048 videos with 192 frames for Dyn and 128 frames for Ped. We use two types of latent images to condition the video generation: (1) en-

Table 2. Performance of the LIDM before and after applying privacy filtering.

	Before Filtering		After Filtering		Rejected Samples
	FID↓	IS$_{\pm \mathbf{std}}$↑	FID↓	IS$_{\pm \mathbf{std}}$↑	
Dyn	17.3	$2.42_{\pm 0.02}$	16.4	$2.37_{\pm 0.02}$	11.25%
Ped(A4C)	13.7	$2.86_{\pm 0.03}$	11.0	$2.95_{\pm 0.03}$	37.06%
Ped(PSAX)	16.8	$3.05_{\pm 0.03}$	14.5	$3.03_{\pm 0.02}$	27.45%

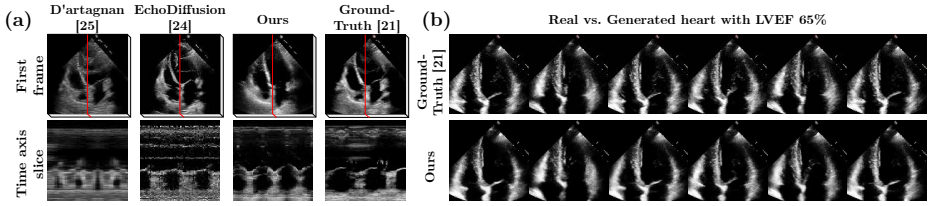

Fig. 2. (a) Temporal consistency of our generated videos against previous works [24, 25] and the real data [21]. (b) Qualitative results of our generated videos against the real data [21], animating the same heart with the same LVEF.

coded real images, and (2) LIDM-generated latent images. For the Dyn dataset, Table 3 shows that our LVDM performs better when using encoded real images than when using LIDM-generated latent images. Nevertheless, our LVDM using LIDM-generated latent images achieves competitive results compared to the state-of-the-art echocardiogram generation method [24], in terms of image quality (FID) and video quality (FVD). Qualitatively, Figure 2 (b) shows that our LVDM can generate videos which are indistinguishable from real ones. For the Ped dataset, Table 3 shows that our LVDM achieves good overall performance. Regarding video generation sampling time, our method (2.4s for 64 frames) is two orders of magnitude faster than the state-of-the-art method [24] (279s for 64 frames). In addition, with the proposed video stitching strategy, we can generate infinitely long temporally-consistent videos. To demonstrate the scalability of our method, we generate a 10-minute-video (19,200 frames at 32 fps) in 14 minutes on a A100 GPU.

EchoNet-Synthetic: We generate EchoNet-Synthetic (Syn) with our video generation pipeline. Syn contains three sub-datasets, corresponding to the three real datasets. Each sub-dataset has the same number of videos as its corresponding real dataset, and the number of frames in each video is the average number of frame in the corresponding real dataset. The visual quality of Syn is competitive with previous state-of-the-art, and is indistinguishable from real data (Figure 2 (a) top row). In terms of temporal consistency, Syn outperforms all previous works, and matches the real data (Figure 2 (a) bottom row).

Downstream Evaluation: We train LVEF regression models on the Syn datasets. We use LVEF prediction metrics (R2, MAE, RMSE) from [21, 23] as baselines, where the models have been trained on the real data. We report the

Table 3. Performance of the LVDM when generating videos from encoded real hearts (left) and LIDM-generated hearts (right). Sampling time for 64 frames, or 2s of video.

	Real Hearts				**LIDM Generated Hearts**				**Sampling**
	FID↓	FVD_{16} ↓	FVD_{128} ↓	$IS_{\pm \mathbf{std}}$↑	FID↓	FVD_{16} ↓	FVD_{128} ↓	$IS_{\pm \mathbf{std}}$↑	**Time**
EchoDiff.[24]	24.0	228	-	$2.59_{\pm 0.06}$	28.0	294.4	-	$2.47_{\pm 0.04}$	279s
Dyn	17.4	71.4	168.3	$2.31_{\pm 0.08}$	28.8	103.5	280.2	$2.26_{\pm 0.08}$	2.4s
Ped(A4C)	24.8	112.2	-	$2.69_{\pm 0.18}$	22.6	86.6	-	$2.77_{\pm 0.23}$	2.4s
Ped(PSAX)	33.0	126.9	-	$2.49_{\pm 0.09}$	27.9	85.1	-	$2.86_{\pm 0.18}$	2.4s

Table 4. Comparison for LVEF regression models performance. 'Claimed' and 'Reproduced' are the claimed results in [21, 23] and our reproduced results, respectively. 'VAE rec.', 'EchoDiff. [24]' and 'Ours' are the results where the regression model is trained on VAE-reconstructed data, generated data from [24] and our generated data, respectively, and evaluated on the real data. The last row shows the regression model trained and tested on our Syn datasets.

	Dynamic [21]			**Ped (A4C)** [23]			**Ped (PSAX)** [23]		
	R2↑	MAE↓	RMSE↓	R2↑	MAE↓	RMSE↓	R2↑	MAE↓	RMSE↓
Claimed	0.81	4.05	5.32	0.70	4.15	5.70	0.74	3.80	5.14
Reproduced	0.81	3.98	5.29	0.68	4.19	5.71	0.71	3.79	5.14
VAE rec.	0.80	4.12	5.48	0.68	4.21	5.71	0.72	3.82	5.14
EchoDiff. [24]	0.55	6.02	8.21	-	-	-	-	-	-
Ours	0.75	4.55	6.10	0.70	5.06	7.07	0.68	4.82	6.95
	Syn Dynamic			**Syn Ped (A4C)**			**Syn Ped (PSAX)**		
Ours (Syn)	0.93	1.67	2.24	0.94	1.24	1.90	0.96	0.94	1.51

results from these works ('Claimed' in Table 4) and reproduce them exactly with the same models and training setup ('Reproduced' in Table 4). To test the information loss caused by VAE encoding/decoding, we train the regression models on the VAE-reconstructed real data ('VAE rec.' in Table 4). We also train the regression model on synthetic data generated with EchoDiffusion [24], the most recent ultrasound synthesis diffusion model ('EchoDiff. [24]' in Table 4). The results on our Syn datasets are shown in row 'Ours' in Table 4. From Table 4, we observe that the regression models trained on Syn perform notably better than the state-of-the-art (*i.e.*, 'EchoDiff. [24]' vs. 'Ours'). The models trained on Syn overall perform well on the real data, although they are not equivalent to the models trained directly on the real data. We attribute this performance difference to residual domain shift, given the high performance of the regression model trained and tested on the Syn dataset ('Ours (Syn)' in Table 4).

4 Conclusion

In this paper, we introduced a new end-to-end protocol for surrogate medical dataset generation, showcasing its adaptability through the provision of open-source code and a synthetic dataset that facilitates replication and further investigation by the research community. Our work not only highlights a novel

approach to generating extended video sequences beyond the initial training scope of the video diffusion model but also integrates a cutting-edge privacy preservation technique to ensure the synthetic datasets can be shared safely. We believe this to be a step forward in leveraging synthetic datasets, offering a method that maintains dataset attributes in terms of size, quality and diversity. While we acknowledge the complexity surrounding intellectual property, it is primarily a political challenge, and our focus here remains on the technical advancements that should mitigate the legal constraints around patient data safety. We are committed to employing the here presented protocol to release synthetic surrogates of private datasets in the future.

Acknowledgments. This work was supported by Ultromics Ltd., the UKRI Centre for Doctoral Training in Artificial Intelligence for Healthcare (EP / S023283/1) and HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) under the NHR project b180dc. NHR and high-tech agenda Bavaria (HTA) funding is partly provided by federal and Bavarian state authorities. NHR@FAU hardware is partially funded by the German Research Foundation (DFG) – 440719683. Support was also received from the ERC - project MIA-NORMAL 101083647 and DFG KA 5801/2-1, INST 90/1351-1.

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

References

1. An, J., Zhang, S., Yang, H., Gupta, S., Huang, J.B., et al.: Latent-shift: Latent diffusion with temporal shift for efficient text-to-video generation. arXiv preprint arXiv:2304.08477 (2023)
2. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
3. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., et al.: Align your latents: High-resolution video synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22563–22575 (2023)
4. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al.: On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021)
5. Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramer, F., Balle, B., Ippolito, D., Wallace, E.: Extracting training data from diffusion models. In: 32nd USENIX Security Symposium (USENIX Security 23). pp. 5253–5270 (2023)
6. Dar, S.U.H., Seyfarth, M., Kahmann, J., Ayx, I., Papavassiliu, T., Schoenberg, S.O., Engelhardt, S.: Unconditional latent diffusion models memorize patient imaging data. arXiv preprint arXiv:2402.01054 (2024)
7. Dombrowski, M., Kainz, B.: Quantifying sample anonymity in score-based generative models with adversarial fingerprinting (2023)

8. Gilbert, A., Marciniak, M., Rodero, C., Lamata, P., Samset, E., Mcleod, K.: Generating synthetic labeled data from existing anatomical models: an example with echocardiography segmentation. IEEE Transactions on Medical Imaging **40**(10), 2783–2794 (2021)
9. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
10. Harvey, W., Naderiparizi, S., Masrani, V., Weilbach, C., Wood, F.: Flexible Diffusion Modeling of Long Videos (December 2022), arXiv:2205.11495
11. He, Y., Yang, T., Zhang, Y., Shan, Y., Chen, Q.: Latent video diffusion models for high-fidelity video generation with arbitrary lengths. arXiv preprint arXiv:2211.13221 (2022)
12. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., et al.: Imagen Video: High Definition Video Generation with Diffusion Models (October 2022), arXiv:2210.02303
13. Ho, J., Jain, A., Abbeel, P.: Denoising Diffusion Probabilistic Models. In: Advances in Neural Information Processing Systems. vol. 33, pp. 6840–6851 (2020)
14. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., et al.: Video Diffusion Models (June 2022), arXiv:2204.03458
15. Hoeppe, T., Mehrjou, A., Bauer, S., Nielsen, D., Dittadi, A.: Diffusion models for video prediction and infilling. arXiv preprint arXiv:2206.07696 (2022)
16. Jensen, J.: Simulation of advanced ultrasound systems using Field II. In: 2004 2nd IEEE International Symposium on Biomedical Imaging: Nano to Macro (IEEE Cat No. 04EX821). pp. 636–639 Vol. 1 (April 2004)
17. Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., et al.: Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2023)
18. Liang, J., Yang, X., Huang, Y., Li, H., He, S., et al.: Sketch guided and progressive growing GAN for realistic and editable ultrasound image synthesis. Medical Image Analysis **79**, 102461 (July 2022)
19. Luo, Z., Chen, D., Zhang, Y., Huang, Y., Wang, L., et al.: Videofusion: Decomposed diffusion models for high-quality video generation. In: CVPR (2023)
20. Nikankin, Y., Haim, N., Irani, M.: Sinfusion: Training diffusion models on a single image or video. arXiv preprint arXiv:2211.11743 (2022)
21. Ouyang, D., He, B., Ghorbani, A., Yuan, N., Ebinger, J., et al.: Video-based AI for beat-to-beat assessment of cardiac function. Nature **580**, 252–256 (April 2020)
22. Packhäuser, K., Gündel, S., Münster, N., Syben, C., Christlein, V., Maier, A.: Deep learning-based patient re-identification is able to exploit the biometric nature of medical chest x-ray data. Scientific Reports **12**(1), 14851 (2022)
23. Reddy, C.D., Lopez, L., Ouyang, D., Zou, J.Y., He, B.: Video-based deep learning for automated assessment of left ventricular ejection fraction in pediatric patients. Journal of the American Society of Echocardiography **36**(5), 482–489 (2023)
24. 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. In: MICCAI. pp. 142–152 (2023)
25. Reynaud, H., Vlontzos, A., Dombrowski, M., Gilligan Lee, C., Beqiri, A., et al.: D'ARTAGNAN: Counterfactual Video Generation. In: MICCAI. pp. 599–609 (2022)
26. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-Resolution Image Synthesis with Latent Diffusion Models (April 2022), arXiv:2112.10752

27. Rueckert, D., Glocker, B., Kainz, B.: Learning clinically useful information from images: Past, present and future (2016)
28. Salimans, T., Ho, J.: Progressive Distillation for Fast Sampling of Diffusion Models (June 2022), arXiv:2202.00512
29. Shams, R., Hartley, R., Navab, N.: Real-Time Simulation of Medical Ultrasound from CT Images. In: MICCAI. pp. 734–741. Springer (2008)
30. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., et al.: Make-A-Video: Text-to-Video Generation without Text-Video Data (September 2022), arXiv:2209.14792
31. Teng, L., Fu, Z., Yao, Y.: Interactive Translation in Echocardiography Training System With Enhanced Cycle-GAN. IEEE Access **8**, 106147–106156 (2020)
32. Tiago, C., Gilbert, A., Beela, A.S., Aase, S.A., Snare, S.R., Šprem, J., McLeod, K.: A data augmentation pipeline to generate synthetic labeled datasets of 3d echocardiography images using a gan. IEEE Access **10**, 98803–98815 (2022)
33. Tomar, D., Zhang, L., Portenier, T., Goksel, O.: Content-Preserving Unpaired Translation from Simulated to Realistic Ultrasound Images. In: MICCAI. pp. 659–669 (2021)
34. Voleti, V., Jolicoeur-Martineau, A., Pal, C.: Masked conditional video diffusion for prediction, generation, and interpolation. arXiv preprint arXiv:2205.09853 (2022)
35. Wang, W., Yang, H., Tuo, Z., He, H., Zhu, J., et al.: Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation. arXiv preprint arXiv:2305.10874 (2023)
36. Yang, R., Srivastava, P., Mandt, S.: Diffusion Probabilistic Modeling for Video Generation (December 2022), arXiv:2203.09481
37. Yu, S., Sohn, K., Kim, S., Shin, J.: Video probabilistic diffusion models in projected latent space. In: CVPR (June 2023)
38. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
39. Zhou, D., Wang, W., Yan, H., Lv, W., Zhu, Y., et al.: Magicvideo: Efficient video generation with latent diffusion models. arXiv preprint arXiv:2211.11018 (2022)