npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-024-01332-0

Probing the limits and capabilities of diffusion models for the anatomic editing of digital twins

Karim Kadry[1] ✉, Shreya Gupta[1], Farhad R. Nezami[2] & Elazer R. Edelman[1]

Numerical simulations of cardiovascular device deployment within digital twins of patient-specific anatomy can expedite and de-risk the device design process. Nonetheless, the exclusive use of patient-specific data constrains the anatomic variability that can be explored. We study how Latent Diffusion Models (LDMs) can edit digital twins to create digital siblings. Siblings can serve as the basis for comparative simulations, which can reveal how subtle anatomic variations impact device deployment, and augment virtual cohorts for improved device assessment. Using a case example centered on cardiac anatomy, we study various methods to generate digital siblings. We specifically introduce anatomic variation at different spatial scales or within localized regions, demonstrating the existence of bias toward common anatomic features. We furthermore leverage this bias for virtual cohort augmentation through selective editing, addressing issues related to dataset imbalance and diversity. Our framework delineates the capabilities of diffusion models in synthesizing anatomic variation for numerical simulation studies.

Physics-based simulations of cardiovascular interventions such as endovascular stent expansion or heart valve implantation can help optimize device design and deployment, especially in challenging anatomies[1]. These "virtual interventions" can be modelled on a patient-specific digital twin, which is a computational replication of a real anatomy derived from medical imaging[2–4]. Virtual interventions have been shown to model the mechanical and hemodynamic consequences of implanting heart valves[5,6], atrial appendage occluders[7], and coronary stents[8,9], as well as the electrophysiological consequences of cardiac ablation[10]. Applied to a cohort of digital twins, virtual interventions enable in silico trials of medical devices[11], in which their safety and efficacy can be assessed within a digital environment. Such trials can act as digital evidence for regulatory agencies, reducing the exorbitant cost and failure rates involved with bringing a device to market[12,13]. Virtual interventions also enable the simulation of hypothetical scenarios, such as implanting alternative devices or modeling different physiological conditions within the same patient[1]. This experimental framework provides mechanistic insight regarding what factors concerning device design and physiology critically influence deployment. Such insights can influence both regulatory and development processes, enhancing future designs and guiding recruitment for clinical trials[1,14].

In contrast, our ability to extract mechanistic insight involving alternative anatomic variants is highly limited. Specifically, we delineate three phenomena critical to device development and regulatory evaluation that digital twin frameworks are unable to properly address. First, the uniqueness of each digital twin complicates the assessment of uncertainty in device performance attributable to scale-specific anatomic variation. Small scale anatomic features can be highly influential on both hemodynamics and biomechanics. Examples include coronary plaque rupture being influenced by thin fibrous caps[15], ventricular trabeculae influencing cardiac hemodynamics[16], and coronary branches affecting blood-flow through the aortic root[17]. Second, due to the complex correlations between local anatomic features within digital twin cohorts, it remains difficult to disentangle the causal relationships and interaction effects exerted by localized anatomic regions on device failure. Localized anatomic features have been widely known to interact in influencing cardiovascular physics, examples include the interactions between lipid and calcium in determining plaque rupture risk[2,4], mitral valve pathology on aortic valve replacements[18], and aortic valve replacements on coronary flow[19]. Lastly, the reliance on digital twin cohorts for in silico trials can compromise device evaluation on less common or pathological anatomic shapes[11,14]. Accordingly, current digital twin paradigms are unable to fully or precisely explore anatomic space, limiting the broader applicability of virtual interventions for device development and regulatory review.

To address such issues, generative models of virtual anatomies have been proposed but typically struggle to balance between producing outputs that are both realistic and controllable. The gold standard method is

[1]Massachusetts Institute of Technology (MIT), Cambridge, MA, 02139, USA. [2]Brigham and Women's Hospital, Boston, MA, 02115, USA.
✉e-mail: kkadry@mit.edu

principal component analysis (PCA), which has traditionally been used to generate virtual cohorts for biomechanical and hemodynamic simulations[20]. Despite its utility, PCA is unable to accurately model the highly nonlinear anatomic variation inherent to human anatomy. As such, there has been a rising interest in deep learning approaches for producing virtual anatomies. State-of-the-art deep learning architectures for this purpose have been variational autoencoders (VAE) and generative adversarial networks (GANs), which exhibit improved performance compared to PCA[21–23]. While such architectures have demonstrated the ability to produce variations of anatomy by exploring their latent space[21], as of yet current approaches are limited in their ability to precisely edit patient-specific models. This is because such methods represent anatomic shape in terms of global shape vectors, which are not expressive enough to control anatomic variation at different spatial scales or within localized regions while keeping others constant. To overcome this limitation, a previous study by Kong et al. found that representing anatomy in terms of a higher-dimensional and spatially extended latent grid enabled higher expressiveness but decreased generation quality under an auto-decoder paradigm[24].

In contrast, diffusion models are a novel class of generative models that can synthesize 2D and 3D medical images with high quality and diversity[25–27]. However, their use in generating virtual anatomy in the form of anatomic label maps is still in its infancy. Preliminary studies used unconditional diffusion models to produce multi-label 2D segmentations of the brain and retinal fundus vasculature respectively in order to train downstream computer vision algorithms[28,29]. The ability of diffusion models to flexibly edit natural images is also well-characterized. For example, diffusion models can create variations of natural images through a perturb-denoise process, partially corrupting a seed image and restoring it through iterative denoising[30]. The level of added noise can control whether the model synthesizes global or local features[31]. Furthermore, diffusion models can be used to locally in-paint regions within an image by specifying a spatially extended mask[32–36], either by directly replacing the masked portion during each denoising step or using the gradient of a masked similarity loss. While these technique has been used in the context of medical images for anomaly detection[37,38] and data augmentation[39], their use in modifying virtual anatomy has not been studied.

Lastly, research into generative models for virtual anatomy is hampered by the lack of appropriate evaluation frameworks to assess the quality of synthetic cohorts for in silico trials. For example, the Fréchet inception distance (FID)[40] is difficult to use for evaluating generative models of virtual anatomies, as no standard pre-trained network for 3D anatomic segmentations is available. Moreover, point cloud-based metrics delineate 3D shape quality and diversity but do not measure the interpretable morphological metrics necessary to understand device performance, nor do they measure topological correctness, a critical factor to ensure compatibility with numerical simulation. Recent studies attempt to address this by visualizing the 1D distributions of clinically relevant morphological variables such as tissue volumes[23,41], but fail to study the multi-dimensional relationship between morphological metrics, nor do they investigate morphological bias due to imbalanced data distributions.

In this study, we develop an experimental framework to study how latent diffusion models (LDMs) can act as a controllable source of anatomic variants for in silico trials to fulfill two main functionalities. The first functionality centers on the controlled synthesis of informative anatomies through editing digital twins, which we term "digital siblings". As opposed to a digital twin, which is a computational replication of a patient-specific anatomy, a digital sibling would resemble the corresponding twin, but exhibit subtle differences in anatomic form. Comparative simulation studies using twins and their siblings would yield insight regarding how scale-specific and region-specific anatomic variation can influence simulated deployment. The second functionality revolves around virtual cohort augmentation by creating digital siblings from a curated subpopulation of digital twins. This would enrich virtual cohorts with specified anatomic attributes, addressing issues related to cohort imbalance and diversity. We accordingly develop a latent diffusion model to generate 3D cardiac label maps and introduce a novel experimental framework to study the synthesis of anatomic variation (Fig. 1). We first characterize the baseline performance of the model through generating de-novo cardiac label maps (Fig. 2). We then investigate two methods to generate digital siblings with diffusion models: (1) perturbational editing of cardiac digital twins to enable scale-specific variation; and (2) localized editing of cardiac digital twins to enable region-specific variation. In our experimental framework, we select various digital twins to act as "seed" volumes and produce several digital siblings through editing. We then apply this procedure over different hyperparameters and seed characteristics to study how generative editing can alter the morphological and topological attributes of digital twins. Lastly, we study how such editing methods can be used to augment virtual cohorts with less common anatomic features. Our main contributions and insights are as follows:

1. We develop and train a latent diffusion model to generate 3D cardiac label maps and introduce a novel experimental framework to study how generative editing techniques can produce scale-and-region specific variants of digital twins.
2. We demonstrate that latent diffusion models can introduce topological violations during generation and editing, where the number of violations is influenced by editing methodology and seed characteristics.
3. We find that dataset imbalance induces a bias within the generation process towards common anatomic features. This anatomic bias extends to scale-and-region specific editing. The degree and spatial distribution of this bias is influenced by editing hyperparameters and seed characteristics.
4. We demonstrate that this anatomic bias can be leveraged to enhance virtual cohort diversity in two manners. Virtual cohort augmentation with scale-specific variation can help explore less populated spaces within the anatomic distribution bounded by the training set. Similarly, augmentation with region-specific variation can augment the cohort with anatomic forms outside the anatomic distribution.

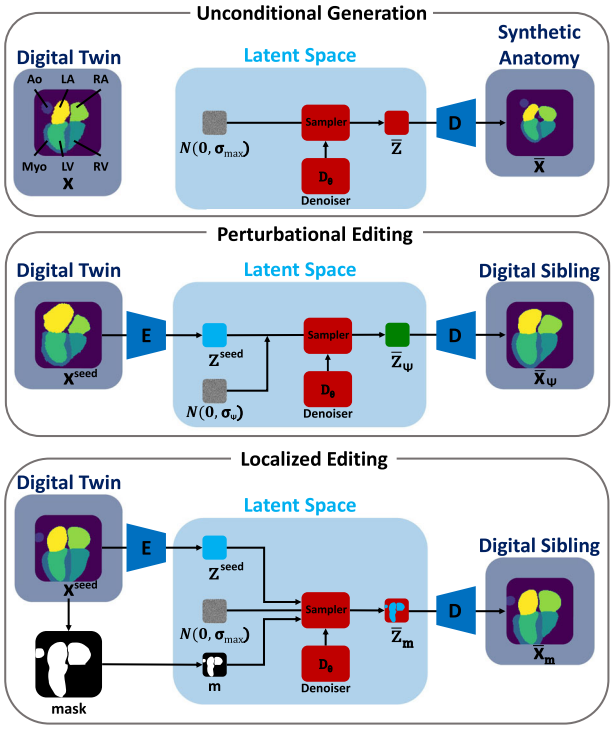

Fig. 1 | We study the ability of diffusion models to generate digital siblings for virtual interventions and augment in silico trials. Top row: we unconditionally generate latent codes ($\bar{\mathbf{z}}$) which are decoded ($\mathbf{D}$) into cardiac label maps ($\bar{\mathbf{x}}$). Middle row: We encode ($\mathbf{E}$) patient-specific digital twins ($\mathbf{x}$) into a latent space ($\mathbf{z}$) and apply a partial perturb-denoise process to achieve scale-specific variations ($\bar{\mathbf{x}}_{\psi}$). Bottom row: We locally edit pre-specified tissues to achieve region-specific variations ($\bar{\mathbf{x}}_{\mathbf{m}}$).

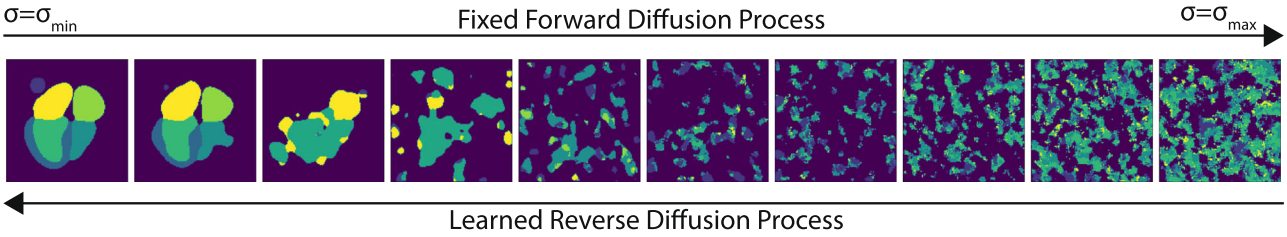

Fig. 2 | Schematic for the forward and reverse diffusion process. The decoded cardiac label maps for several intermediately noised latent representations $\mathbf{z}_\sigma$. During training, a neural denoiser learns to approximate the incremental reverse process at each noise level σ. During sampling, the network is recursively applied to produce de-novo cardiac label maps.

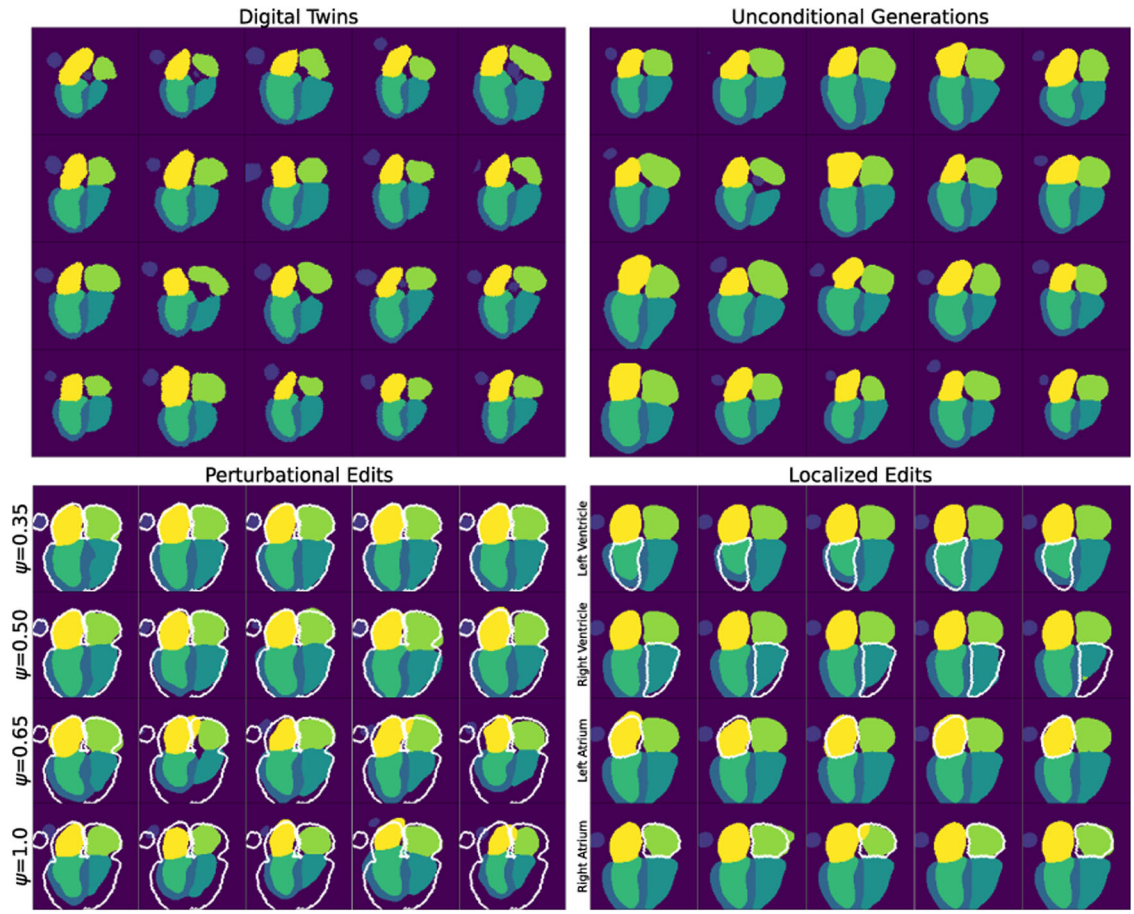

Fig. 3 | Example 2D slices from 3D cardiac label maps. Top left: digital twin label maps from the training set. Top right: unconditionally generated label maps generated by the diffusion model. Bottom left: perturbational edits of a single cardiac digital twin over various sampling ratios. Bottom right: localized edits of cardiac digital twins over various tissue masks. Bottom row has a white outline of the edited twin for perturbational edits (left) and an outline of the edited tissue region for localized edits (right).

Results

Unconditional sampling of virtual anatomies

We conducted a sensitivity analysis of cohort quality with respect to the sampling steps and cohort size and and use a minimum of 50 sampling steps and a cohort size of 120 (Supplementary Figs. 7 and 8). We then sample 360 label maps with a diffusion model for 50 steps for analysis and visualization. We also train a baseline generative VAE that samples a global latent vectors which can be decoded into cardiac label maps. Example label maps can be seen in Fig. 3. We further visualize the reconstructed and synthetic cardiac label maps in 3D within Supplementary Figs. 5 and 6. The scatterplot (Fig. 4) and the difference heatmap (Fig. 5) show the morphological distribution of the synthetic anatomies generated by the diffusion model on a global and local scale respectively. Both figures demonstrate that unconditional sampling tends to generate mean-sized cardiac label maps, but fails to sample rarer anatomic configurations on the periphery of the distribution. This is especially the case for the baseline VAE, which learns a much more constrained distribution concentrated around the anatomic mean. This bias also exists on a local level as seen in the difference heatmap $\mathbf{P}_{\text{diff}}$ in Fig. 5. The heatmaps for each individual chamber are shown in Supplementary Fig. 4.

Table 1 shows the morphological and 3D shape based metrics for the generative VAE and diffusion model. Our diffusion model is able to sample from a wider distribution of anatomy due to it's expressive latent grid, with higher recall and coverage values. However, the generative VAE exhibits higher morphological precision, MMD, and 1-NNA values, likely due to sampling common anatomies near the center of the distribution. Table 2 indicates the primary source of topological violation stems from the initial

segmentations used to train the generative models. Violations in the real dataset stem from the segmentation network used to create the original dataset, in which small clusters of misclassified tissues contribute to the amount of topological violations (Fig. 3 and Supplementary Fig. 2).

Scale specific variation through perturbational editing

We select four seed label maps that represent different types of cardiac anatomy: a seed with a large LV and RV ($L^{\uparrow}R^{\uparrow}$), a seed with a small LV and RV ($L^{\downarrow}R^{\downarrow}$), a seed with a large LV but mean sized RV ($L^{\uparrow}R^{\leftrightarrow}$), and a seed with a mean sized LV and RV ($L^{\leftrightarrow}R^{\leftrightarrow}$). For each seed, we generate synthetic anatomies with varying sampling ratios, corresponding to $\psi = [0.35, 0.50, 0.65, 0.8, 1]$, leading to a total of 20 virtual cohorts of 120 anatomies each. Example label maps can be seen in Fig. 3.

Figure 6 shows that the cohorts generated by perturbational editing are increasingly biased towards the most common anatomies with increasing noise. Figure 7 further shows that the amount of injected noise corresponds to spatial scale, as the bias exhibited by the spatial heatmap $\mathbf{P}_{\text{diff}}$ expands with increasing noise. Table 3 demonstrates that the topological quality of the sampled label map can degrade when editing outlier twins, as can be seen when perturbationally editing seed $L^{\uparrow}R^{\leftrightarrow}$ with a sampling ratio ψ of 0.35. This is because the seed occupies a sparsely populated region of the anatomic distribution. A visualization of the topological violations exhibited after perturbational editing can be found in Supplementary Fig. 3

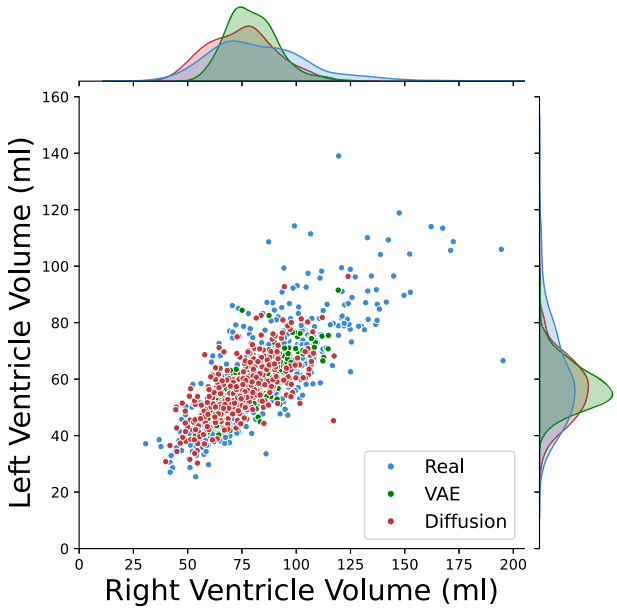

Fig. 4 | Unconditional generation captures common anatomic variations but fail to capture outliers. Diagonal plot shows the 2D morphological distribution (shown as a scatterplot) exhibited by real cohorts and synthetic cohorts generated by unconditional sampling from a generative VAE and diffusion model. Marginal plots show the equivalent 1D morphological distributions for the virtual anatomy cohorts, visualized as a kernel density estimate plot.

Region specific variation through localized editing

For each of the previously mentioned seeds, we specify two masks designed to edit the RV and LV respectively. The myocardium was not included for each tissue mask, allowing it to vary with each ventricular chamber. This process resulted in eight synthetic cohorts of 120 anatomies each. Example label maps can be seen in Fig. 3.

Figure 8 shows that the 1D distributions of edited ventricular volumes are biased towards most common values of the real cohort. This can be seen most prominently with seed $L^{\uparrow}R^{\leftrightarrow}$ where the edited LVs have a substantially lower volume as compared to the seed label map. From the spatial difference heatmaps $\mathbf{P}_{\text{diff}}$ visualized in Fig. 9, we further observe that localized editing can change individual chambers while maintaining others as constant, where the edited chambers are biased towards a mean anatomic shape. With the exception of editing the RV of seed $L^{\uparrow}R^{\leftrightarrow}$, locally editing the seed label maps increased the percentage of topological violations as compared to the seeds, as can be seen in Table 4. A visualization of the topological violations exhibited after localized editing can be found in Supplementary Fig. 3, and a comparison of our replacement-based inpainting method to guidance-based inpainting can be found in Supplementary Fig. 1.

Virtual cohort augmentation through selective editing

In this experiment we contrast and compare three strategies that can augment virtual cohorts with rare anatomies to improve dataset imbalance and diversity. In this case, we aim to enrich a target cohort with rare patient-

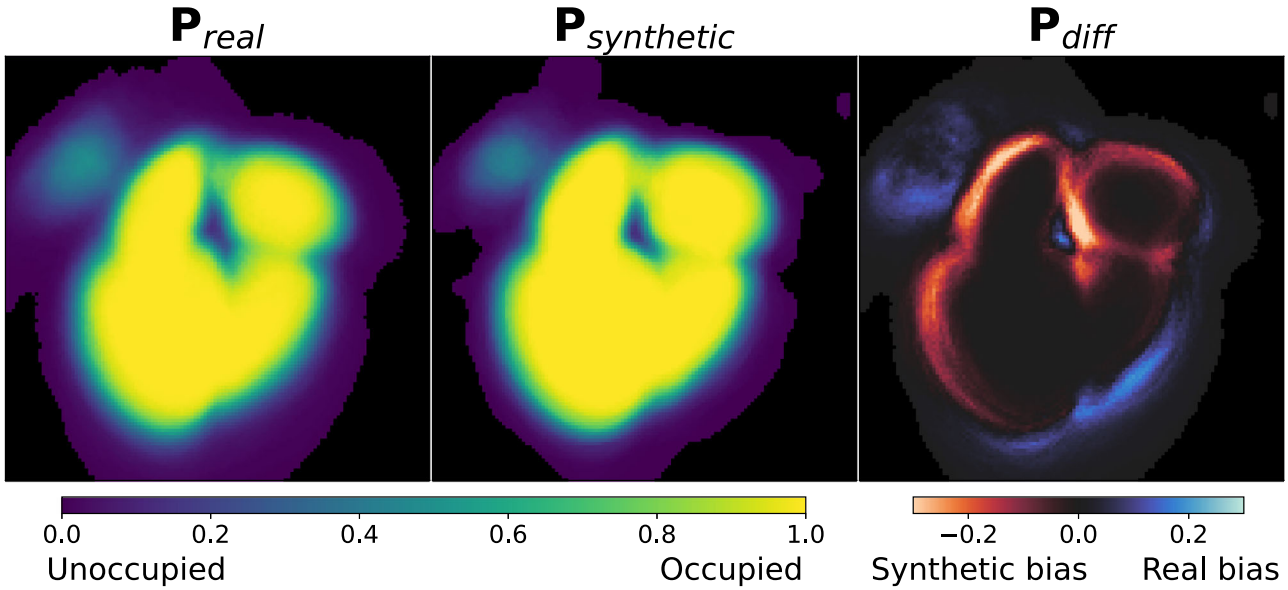

Fig. 5 | The distribution of label maps synthesized by the diffusion model exhibits spatially varying differences against that of real label maps. Spatial occupancy heatmaps show the distribution of real ($\mathbf{P}_{\text{real}}$) and synthetic ($\mathbf{P}_{\text{synthetic}}$) label maps, as well as the difference in occupation ($\mathbf{P}_{\text{diff}}$). Heatmaps are masked out where $\mathbf{P}_{\text{real}}$ or $\mathbf{P}_{\text{synthetic}}$ are zero. Real or synthetic bias correspond to increased relative occupancy by real or synthetic anatomies respectively.

Table 1 | Morphological and shape based metrics comparing a baseline generative VAE and diffusion model for unconditional cardiac generation

Model	Morphological metrics			MMD (↓)		COV (%, ↑)		1-NNA (%, ↓)	
	Prec. (↑)	Rec. (↑)	FMD (↓)	CD	EMD	CD	EMD	CD	EMD
VAE	**0.99**	0.46	3.79	**1.25**	**4.06**	22.16	23.40	**65.37**	**66.12**
Diffusion (ours)	0.85	**0.74**	**2.28**	1.32	4.20	**28.72**	**31.91**	77.05	76.62

Generative VAE's suffer from reduced generation diversity in terms of both morphology and 3D shape, while diffusion models are able to sample a wide variety of cardiac segmentations at the cost of reduced fidelity. MMD-CD and MMD-EMD values were multiplied by 1000 and 100 respectively.

Table 2 | Topological violations exhibited by real, reconstructed real, and synthetic cohorts respectively

	Real	Recon	Synthetic
TV (%)	15.6	12.0	13.3

specific cardiac label maps distinguished by an RV volume larger than a threshold value of 115 ml. Our first strategy is to unconditionally sample 7200 label maps and filter all outputs with RV volumes less the threshold. In our second strategy, we utilize the bias inherent to perturbational editing and modify digital twins from the target cohort to create digital sibling cohorts. Half of the digital twins received a large perturbation ($\psi = 0.5$) and the other half received a small perturbation ($\psi = 0.35$). Following the editing process, digital siblings with an RV volume below the threshold were excluded. Our third strategy leverages the bias inherent to localized editing, in which half of the target cohort was locally edited to have different LV shapes, while the other half were edited to have different RV shapes. Similarly, outputs that do not meet the RV volume threshold were excluded. All three strategies resulted in filtered cohorts of size 140 each. The evaluation metrics, namely Frechet morphological distance, morphological precision, and morphological recall, were computed against the target cohort consisting of 50 cardiac label maps from the train set as the reference standard.

Figure 10 demonstrates that unconditional generation does not fully explore the peripheries of the target cohort distribution, where it can be seen that the largest RV volumes within the target cohort (black stars) are not represented. In contrast, perturbational editing excels in filling sparsely populated peripheries of the distribution (producing anatomies with both ventricles enlarged). Table 5 reinforces these insights, demonstrating that augmentation through perturbational editing enhances diversity through exhibiting higher recall and COV values as compared to unconditional generation. Augmenting cohorts with localized editing yields cardiac label maps with morphological features that conform to the distribution of individual morphological metrics but deviate from the multidimensional distribution, producing anatomies with only a single large ventricle. Table 5 shows that localized editing results in increased diversity metrics. Furthermore, both editing-based strategies yield similar or better fidelity metrics such as precision and MMD when compared to unconditional sampling. Table 5 also demonstrates that virtual cohorts produced by the all augmentation strategies exhibit similar topological quality.

Discussion

In this study we developed an experimental framework to investigate how generative diffusion models of human anatomy can be integrated into virtual intervention workflows through the precision editing of digital twins. This novel paradigm is designed to facilitate the generation of mechanistic insights for device development as well as digital evidence for regulatory assessment. Specifically, we trained a diffusion model on a dataset of 3D cardiac label maps and leveraged the model to edit digital twins under various hyperparameters. By examining the 3D shape, morphological attributes and topological quality of the label maps post-editing, we find that diffusion model-based editing techniques can generate insightful morphological variants of digital twins for virtual interventions. Perturbational editing can produce scale-specific variations of digital twins, which can isolate the sensitivity of device deployment to both small and large-scale variations. In contrast, localized editing can produce region-specific variations of digital twins, which can elucidate the localized effect of anatomic features on device deployment. Such insights can streamline the development of novel medical devices and provide a more comprehensive assessment of device performance for regulatory agencies.

While the integration of generative editing with virtual interventions has the potential to produce mechanistic insight and augment in silico trials, they should be employed with caution. For example, we find that generative editing can produce anatomies with topologically incorrect features, such as connected atria or several left ventricle components, which induce non-physiological phenomena within numerical simulations of cardiovascular physics. Moreover, we demonstrate that diffusion models exhibit a bias towards generating the more common anatomic features within the dataset, a bias that extends to diffusion model-based editing techniques. Anatomic variants with low morphological plausibility can induce inaccuracies in the regulatory assessment of device safety and fail to capture possible failure modes. As such, methods that evaluate and control anatomic bias will be critical to the integration of generative artificial intelligence within workflows regarding device development and regulatory review. We nevertheless demonstrate that such anatomic bias can be leveraged to enhance the digital evidence produced by in silico trials. This is achieved by augmenting cohorts with digital siblings, thereby improving factors critical to regulatory approval, such as cohort balance and diversity. Specifically, we found that perturbational editing can fill the sparsely populated regions within the anatomic distribution, potentially improving device assessment for realistic anatomies. Similarly, localized editing can expand the space of plausible anatomies that can be probed with virtual interventions, enabling the assessment of possible failure modes, at the expense of decreased anatomic realism.

However, while our experimental framework can derive novel insights regarding the morphological and topological behaviour of generative editing for virtual interventions, it exhibits a number of limitations. First, it does not quantitatively analyze morphology on multiple scales, instead measuring global level metrics such as 3D shape, volumes and axis lengths. Second, the influence of the diffusion model architecture or sampling methodology on generative editing was not explored. Lastly, the validity of visualizing spatial heatmaps depends on spatial correspondence between anatomic features, and would not apply to anatomies that have a variable topology such as organs with multi-component inclusions. All of these limitations present exciting directions for future work on evaluation metrics and experimental frameworks regarding the generative editing of digital twins for device development and regulation.

Methods

Dataset

We used the TotalSegmentator dataset[42], consisting of 1204 Computed Tomography (CT) images, each segmented into 104 bodily tissues. We filtered out all patient label maps that do not have complete and adequate-quality segmentations for all four cardiac chambers. This resulted in a dataset of 512 3D cardiac label maps, where each label map consisted of 6 tissues: aorta (Ao), myocardium (Myo), right ventricle (RV), left ventricle (LV), right atrium (RA), and left atrium (LA). All cardiac label maps were

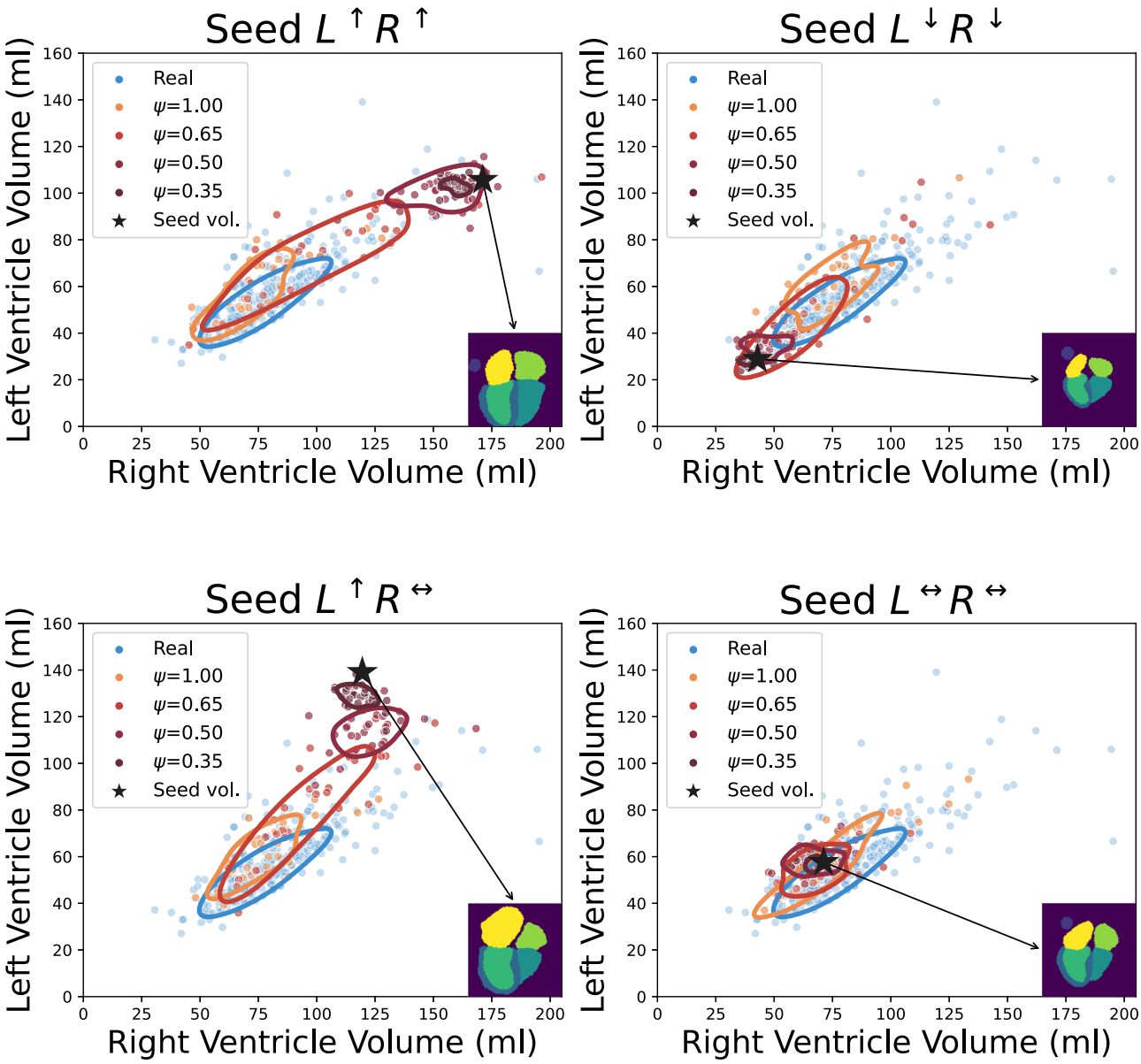

Fig. 6 | Perturbationally editing seed cardiac label maps (star marker) with increasing levels of injected noise ψ produces cohorts that are biased towards the most common anatomies (blue contour). Each scatterplot corresponds to a different seed label map, showing multiple cohorts synthesized by editing the same seed with different sampling ratios (ψ). For improved visual clarity, scatterplots are supplemented with kernel density estimate plots, and the number of data points displayed per cohort is reduced by half.

cropped and resampled to a size of $7 \times 128 \times 128 \times 128$, with an isotropic voxel size of 1.4 mm^3. We then reoriented each cardiac segmentation so that the axis between the LV and LA centroids is aligned with the positive z-axis. Lastly, we rigidly registered all segmentations to a reference label map using the methodology described by Avants et al.[43].

Latent diffusion model training

We employed a latent diffusion model (LDM), consisting of a variational autoencoder (VAE) and a denoising diffusion model. The VAE encodes cardiac label maps $\mathbf{x}$ into latent representations $\mathbf{z}$, which can be decoded into label maps $\bar{\mathbf{x}}$. The training process for our diffusion model is done in the latent space of the trained autoencoder, we represent the probability distribution of cardiac anatomy by $p_{\text{data}}(\mathbf{z})$ and consider the joint distribution $p(\mathbf{z}_\sigma; \sigma)$ obtained through a forward diffusion process, in which i.i.d Gaussian noise of standard deviation σ is added to the data, where at $\sigma = \sigma_{\max}$ the data is indistinguishable from Gaussian noise. The driving principle of diffusion models is to sample pure Gaussian noise and approximate the reverse diffusion process through using a neural network to sequentially denoise the latent representations $\mathbf{z}_\sigma$ with noise levels $\sigma_0 = \sigma_{\max} > \sigma_1 > \cdots > \sigma_N = \sigma_{\min}$ such that the final denoised latents correspond to the clean data distribution. Following Karras et al.[44], we represent the reverse diffusion process as the solution to the following stochastic differential equation

$$d\mathbf{z}_\sigma = -2\sigma \nabla_{\mathbf{z}} \log p(\mathbf{z}_\sigma; \sigma)\, dt + \sqrt{2\sigma} d\mathbf{w} \tag{1}$$

Where the score function $\nabla_{\mathbf{z}_\sigma} \log p(\mathbf{z};\ \sigma)$ denotes the direction in which the rate of change for the log probability density function is greatest and $d\mathbf{w}$ is the standard Wiener process. Since the data distribution is not analytically

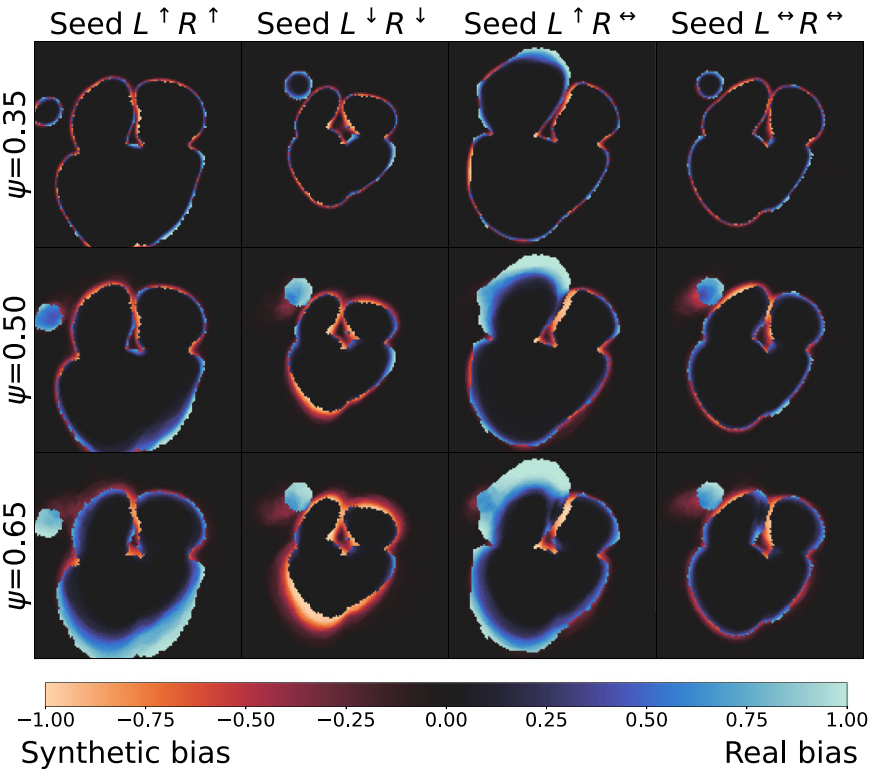

Fig. 7 | Perturbationally editing seed cardiac label maps (columns) with increasing levels of injected noise ψ (rows) enables scale-specific variation. Difference heatmaps $\mathbf{P}_{\text{diff}}$ show spatially varying discrepancies between the seed and synthetic cohorts generated by perturbationally editing various seed label maps.

Table 3 | Topological violations exhibited by each cohort produced by perturbationally editing various seed label maps for different sampling ratios ψ

	Seed Label Map			
TV (%)	$L^{\uparrow}R^{\uparrow}$	$L^{\downarrow}R^{\downarrow}$	$L^{\uparrow}R^{\leftrightarrow}$	$L^{\leftrightarrow}R^{\leftrightarrow}$
Recon	8.3	8.3	8.3	8.3
$\psi = 0.35$	9.7	10.4	23.3	11.1
$\psi = 0.50$	11.4	10.6	11.9	10.7
$\psi = 0.65$	10.6	11.4	10.4	11.4
$\psi = 0.80$	11.3	12.5	10.8	12.1
$\psi = 1.00$	12.0	10.5	11.3	11.7

tractable we train a neural network to approximate the score function. We start with clean latent representations $\mathbf{z}$ and model a forward diffusion process that produces intermediately noised latents $\mathbf{z}_\sigma = \mathbf{z} + \mathbf{n}$ where $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, parameterized by a noise level σ. The diffusion model is parameterized as a function F_θ, encapsulated within a denoiser D_θ, that takes as input an intermediately noised output $\mathbf{z}_\sigma$ and a noise level σ to predict the clean data $\mathbf{z}$.

$$D_\theta(\mathbf{z}_\sigma; \sigma) = c_{\text{skip}}(\sigma)\,\mathbf{z}_\sigma + c_{\text{out}}(\sigma)\,F_\theta(c_{\text{in}}(\sigma)\,\mathbf{z}_\sigma;\ c_{\text{noise}}(\sigma)), \tag{2}$$

where c_{skip} controls the skip connections that allow the F_θ to predict the noise $\mathbf{n}$ at low σ and the training data $\mathbf{z}$ at high σ. This parametrization has been shown to improve convergence speed and performance[44]. The variables c_{out} and c_{in} scale the input and output magnitudes to be within unit variance, and the constant c_{noise} maps the noise level σ to a conditioning input to the network[44]. The denoiser output is related to the score function through the relation $\nabla_{\mathbf{z}_\sigma} \log p(\mathbf{z}_\sigma; \sigma) = \left(D_\theta(\mathbf{z}_\sigma; \sigma) - \mathbf{z}\right)/\sigma^2$ and F_θ is chosen to be a 3D U-net with both convolutional and self-attention layers, similar to previous approaches[28,30,44,45]. The loss L is then specified based on the agreement between the denoiser output and the original training data:

$$L = \mathbb{E}_{\sigma,\mathbf{z},\mathbf{n}}[\lambda(\sigma)||D_\theta(\mathbf{z}_\sigma; \sigma) - \mathbf{z}||_2^2], \tag{3}$$

such that the loss weighting $\lambda(\sigma) = 1/c_{\text{out}}(\sigma)^2$ ensures an effective loss weight that is uniform across all noise levels, and σ is sampled from a log-normal distribution with a mean of 1 and standard deviation of 1.2.

Once the denoiser has been sufficiently trained, we define a specific noise level schedule governing the reverse process, in which the initial noise level, σ, starts at $\sigma_{\max}$ and decreases to $\sigma_{\min}$:

$$\sigma_i = \left(\sigma_{\max}^{\frac{1}{\rho}} + \frac{i}{N-1}\left(\sigma_{\min}^{\frac{1}{\rho}} - \sigma_{\max}^{\frac{1}{\rho}}\right)\right)^{\rho} \tag{4}$$

where ρ, $\sigma_{\min}$ and $\sigma_{\max}$ are hyperparameters that were set to 3, 2e−3, and 80 respectively. We specifically leverage a stochastic variant of the solver detailed in Karras et al.[44] to sequentially denoise the latent representations $\mathbf{z}_\sigma$ and solve the reverse diffusion process detailed in Eq. (1) (Fig. 1).

Latent diffusion model implementation

We trained the variational autoencoder with an MSE reconstruction loss and a KL divergence loss with a relative weight of 1e−6. We modified the architecture from Rombach et. al.[45] to ensure compatibility with 3D voxel grids and adjusted the number of channels to [64,128,192]. We augmented our data with random scaling (0.5–1.5), rotations (0–180°), and translations (0–20 voxels) in each direction. For the denoising diffusion model, we modified the original architecture of the specified by Rombach et al.[45] to ensure compatibility with 3D voxel grids and adjusted the model channels to [64,128,192]. We used the Adam optimizer[46] for the VAE and diffusion model, using learning rates of 1e−4 and 2.5e−5 respectively.

Perturbational editing

To create digital siblings by perturbational editing, we first encode a seed cardiac label map $\mathbf{x}^{\text{seed}}$ into the latent representation $\mathbf{z}^{\text{seed}}$. Instead

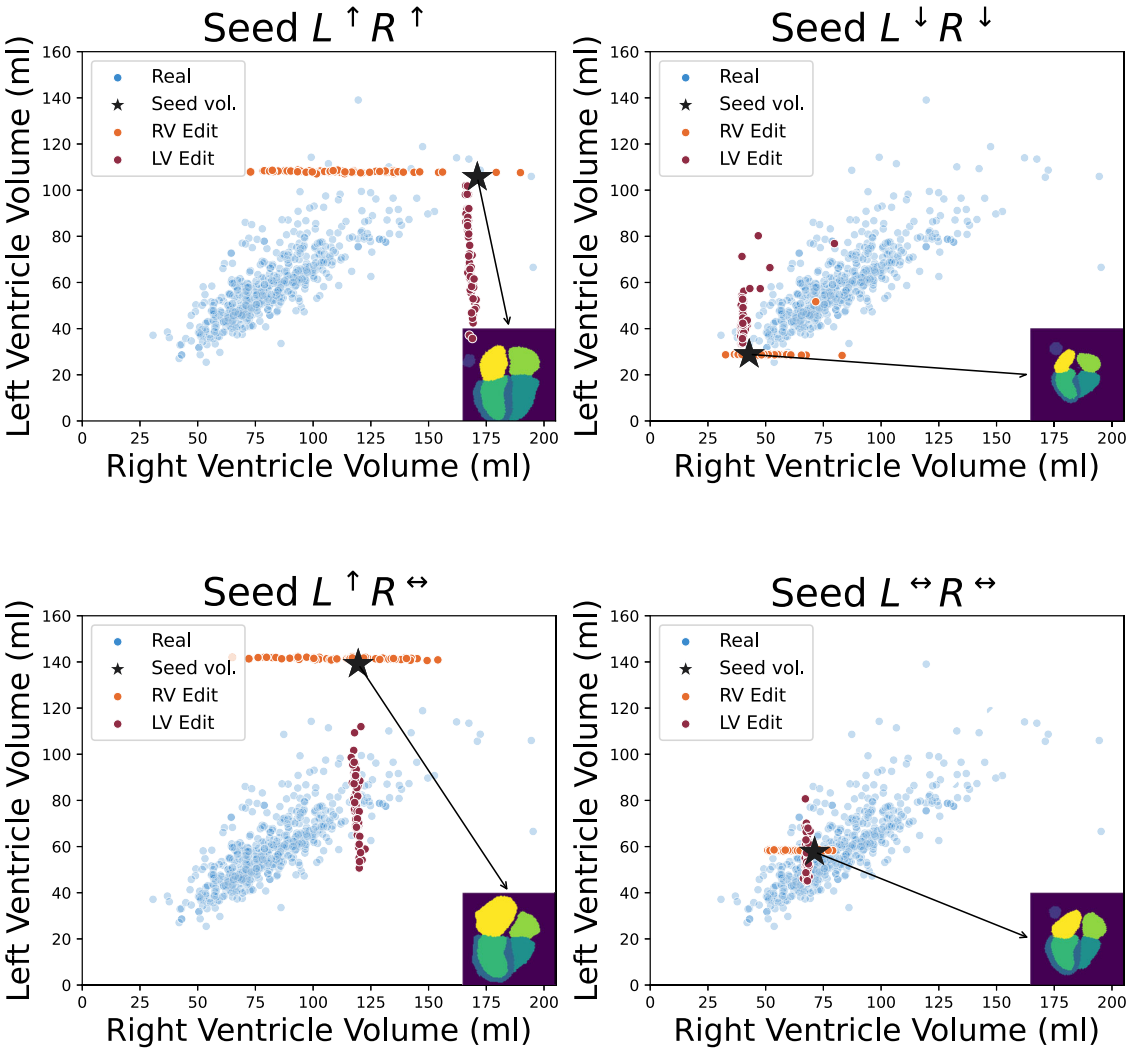

Fig. 8 | Localized editing of seed cardiac label maps (star marker) produces cohorts with region-specific variation that is biased towards those of the most common anatomies. Each scatterplot corresponds to a different seed, showing multiple cohorts synthesized by locally editing the same seed label map with different tissue masks **m**.

of sampling from pure Gaussian noise, we recursively apply the denoiser using the intermediately noised latent $\mathbf{z}_\sigma$ as the starting point (Fig. 1) to produce $\bar{\mathbf{z}}_\psi$. The latent $\bar{\mathbf{z}}_\psi$ is then decoded into the cardiac label map $\bar{\mathbf{x}}_\psi$ using the autoencoder. The intermediate step $i < N$ is a hyperparameter that determines how much of the sampling process is recomputed. We express this hyperparameter in terms of the the sampling ratio $\psi = (N - i)/N$ in our experiments, such that $\psi = 0$ is equivalent to the reconstruction of the original label map, and $\psi = 1$ is equivalent to the unconditional generation of cardiac label maps.

Localized editing

To create digital siblings by localized editing, we first encode a seed cardiac label map $\mathbf{x}^{\text{seed}}$ into the latent representation $\mathbf{z}^{\text{seed}}$. To better preserve the masked region, we set $\mathbf{z}^{\text{seed}}$ as the mean prediction of the encoder without sampling from the Gaussian prior. A tissue-based mask, **m**, denoting which cardiac tissues are to be preserved, was created and downsampled to the same size as the latent representation. The mask was then dilated twice to ensure that tissue interfaces remain stable during editing. The sampling process is similar to that of unconditional sampling, with the addition of an update step that replaces the unmasked portion of the intermediately denoised image with an equivalently corrupted latent representation belonging to the seed label map:

$$\mathbf{z}_\sigma = \mathbf{m} \odot (\mathbf{z}^{\text{seed}} + \mathbf{n}(\sigma)) + (1 - \mathbf{m}) \odot \mathbf{z}_\sigma \tag{5}$$

At the end of sampling, the denoised latent $\bar{\mathbf{z}}_\mathbf{m}$ is then decoded into the cardiac label map $\bar{\mathbf{x}}_\mathbf{m}$ through the decoder (Fig. 1).

Shape evaluation

To evaluate virtual cohorts in terms of 3D shape, we use point cloud-based metrics as proposed by Yang et al.[47]. These metrics include (1) minimum matching distance (MMD), which measures shape fidelity, (2) coverage (COV), which measures shape diversity, and (3) 1-nearest-neighbor accuracy (1-NNA), which measures distributional similarity. To convert label maps into point clouds, we group the main cardiac chambers and myocardium into a single shape and use marching cubes[48] to obtain a 3D surface mesh from which we randomly sample a point cloud of 1024 points. To calculate the shape metrics, We first define the similarity between point clouds in terms of Chamfer distance (CD) and earth mover's distance (EMD) as follows:

$$\text{CD}(\text{X}, \text{Y}) = \sum_{\text{x}\in\text{X}} \min_{\text{y}\in\text{Y}} \| \text{x} - \text{y} \|_2^2 + \sum_{\text{y}\in\text{Y}} \min_{\text{x}\in\text{X}} \| \text{x} - \text{y} \|_2^2, \tag{6}$$

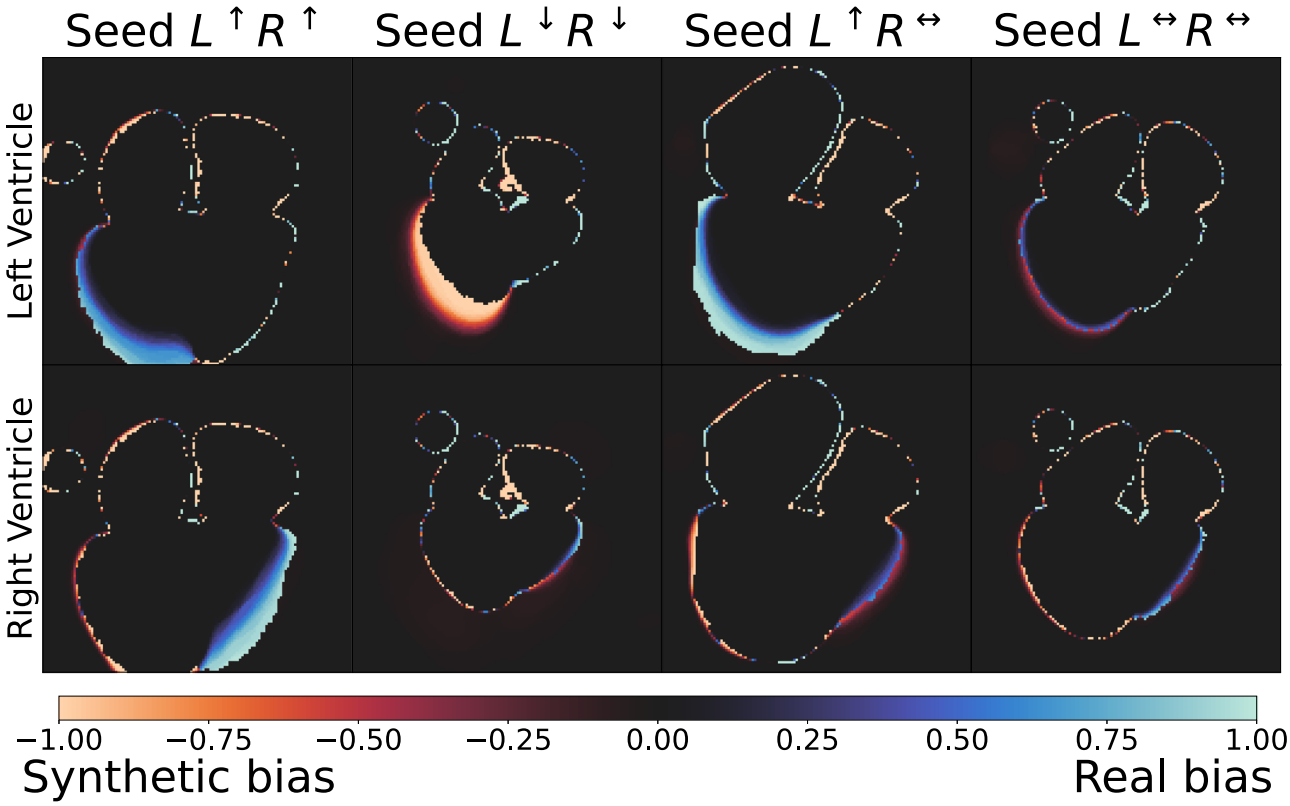

Fig. 9 | Locally editing seed cardiac label maps (columns) with different tissue masks m (rows) enables region-specific variation. Difference heatmaps $\mathbf{P}_{\text{diff}}$ show spatially varying discrepancies between the seed and synthetic cohorts generated by locally editing 4 seed label maps.

Table 4 | Topological violations exhibited by each cohort produced by localized editing of various seed label maps with different tissue masks

	Seed Label Map			
TV (%)	$L^{\uparrow}R^{\uparrow}$	$L^{\downarrow}R^{\downarrow}$	$L^{\uparrow}R^{\leftrightarrow}$	$L^{\leftrightarrow}R^{\leftrightarrow}$
Recon	**8.3**	**8.3**	**8.3**	**8.3**
RV Edit	10.9	13.2	14.9	9.7
LV Edit	10.9	10.7	9.6	9.0

$$\mathrm{EMD}(\mathrm{X},\mathrm{Y}) = \min_{\psi:\mathrm{X}\to\mathrm{Y}} \sum_{\mathrm{x}\in\mathrm{X}} \| \mathrm{x} - \psi(\mathrm{x}) \|_2, \tag{7}$$

where X and Y are two point clouds with the same number of points and ψ is a bijection between them. Given a set of generated (S_g) and real (S_r) point clouds, we measure shape fidelity through MMD as follows:

$$\mathrm{MMD}(S_g, S_r) = \frac{1}{|S_r|} \sum_{Y\in S_r} \min_{X\in S_g} D(X, Y), \tag{8}$$

where lower values indicate the generated shapes are of higher fidelity. To measure shape diversity, we compute COV as the fraction of point clouds in the real set that are matched to at least one point cloud in the generated set. Mathematically we compute COV as follows:

$$\mathrm{COV}(S_g, S_r) = \frac{\left|\left\{\arg\min_{Y\in S_r} D(X, Y) | X \in S_g\right\}\right|}{|S_r|}, \tag{9}$$

where higher values indicate better diversity or coverage in terms of 3D shape. Finally, we use 1-NNA to compare the distribution of real and generated shapes, to do this we let $S_{-X} = S_r \cup S_g - \{X\}$ and N_X be the nearest neighbor of X in S_{-X}. 1-NNA is the leave-one-out accuracy of the 1-NN classifier:

$$1 - \mathrm{NNA}(\mathrm{S_g}, \mathrm{S_r}) = \frac{\sum_{X\in S_g} \mathbb{I}[N_X \in S_g] + \sum_{Y\in S_r} \mathbb{I}[N_Y \in S_r]}{|S_g| + |S_r|}, \tag{10}$$

where $\mathbb{I}$ is the indicator function. A value close to 50% implies that S_g and S_r are sampled from the same distribution.

Lastly, to analyze anatomic bias on a local scale, a voxel-wise mean was computed over all virtual anatomies within a cohort. This results in a spatial heat map $\mathbf{P}$ of size $7 \times 128 \times 128 \times 128$ for the real and synthetic cohorts. The inverse of the background channel was chosen for further visualization.

Morphological evaluation

To assess the morphological quality of a virtual cohort, we represent each virtual anatomy in terms of a 12-dimensional morphological feature vector. For each cardiac label map, we calculate the volume, major axis length, and minor axis length for the LV, RV, LA, and RA. Two of these metrics (LV and RV volumes) were further chosen to plot the global morphological distribution of each cohort. To calculate measures of morphological fidelity and diversity, we adapt the improved precision and recall metrics defined for generative image models[49]. Our key idea is to form explicit non-parametric representations of the manifolds of real and generated data within *morphological* space, rather than the feature space of a neural classifier.

Following Kynkaanniemi et al.[49], we embed our real and synthetic anatomy (in the form of multi-label segmentations) into morphological feature space. We denote the morphological feature vectors of the real and generated anatomies by φ_r and φ_g, respectively, and the corresponding sets of morphological feature vectors by Φ_r and Φ_g.

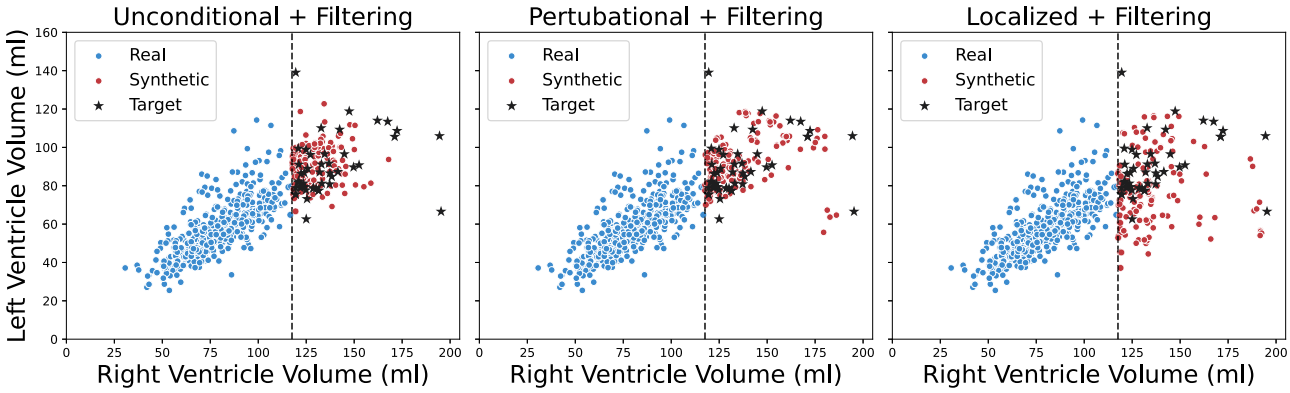

Fig. 10 | Scatterplots demonstrating three augmentation strategies for a target cohort of real cardiac label maps distinguished by right ventricle volumes larger than a minimum threshold (dashed lines). The first strategy uses unconditional generation while second and third strategies utilized generative editing applied to a cohort of seed label maps. All generated cohorts underwent filtering to ensure a minimum right ventricular volume.

Table 5 | Comparison of various metrics across different virtual cohort augmentation strategies

Augmentation strategy	Topological violations (%)	Morphological metrics			MMD (↓)		COV (%, ↑)		1-NNA (%, ↓)	
		Prec. (↑)	Rec. (↑)	FMD (↓)	CD	EMD	CD	EMD	CD	EMD
Uncond + Filter	**11.2**	**0.98**	0.38	5.40	2.6	6.0	20.00	20.00	95.26	94.74
Pert. + Filter	12.2	**0.98**	**0.92**	**2.11**	1.19	**3.95**	82.00	84.00	64.11	54.21
Local. + Filter	12.0	0.96	0.90	2.94	**1.16**	3.96	**94.00**	**94.00**	**49.47**	**53.69**

All comparison metrics were calculated using the target cohort as a reference. MMD-CD and MMD-EMD values were multiplied by 1000 and 100 respectively.

For each set of feature vectors $\Phi \in \{\Phi_r, \Phi_g\}$, we estimate the corresponding manifold in the feature space. We obtain the estimate by forming a hypersphere around each feature vector with a radius equal to the distance to its kth nearest neighbor. Together, these hyperspheres define a volume in morphological feature space that serves as an estimate of the true manifold. To determine whether a given sample φ is located within this volume, we define a binary function

$$f(\varphi, \Phi) = \begin{cases} 1, & \text{if } \| \varphi - \varphi' \|_2 \leq \| \varphi' - \mathrm{NN}_k(\varphi', \Phi) \|_2 \\ & \text{for at least one } \varphi' \in \Phi \\ 0, & \text{otherwise} \end{cases} \quad (11)$$

where $\mathrm{NN}_k(\varphi', \Phi)$ returns the kth nearest feature vector of φ' from set Φ. As such, $f(\varphi, \Phi_r)$ provides information on whether an individual anatomy is morphologically realistic, whereas $f(\varphi, \Phi_g)$ determines if an anatomy could be reproduced by the diffusion model. We can now define our metrics as

$$\mathrm{precision}(\Phi_r, \Phi_g) = \frac{1}{|\Phi_g|} \sum_{\varphi_g \in \Phi_g} f(\varphi_g, \Phi_r) \quad (12)$$

$$\mathrm{recall}(\Phi_r, \Phi_g) = \frac{1}{|\Phi_r|} \sum_{\varphi_r \in \Phi_r} f(\varphi_r, \Phi_g), \quad (13)$$

where precision denotes the fraction of morphologically 'realistic' anatomies in the generated dataset, while recall denotes the fraction of real anatomies that could have been generated by the diffusion model.

Lastly, we implement a variant of the Frechet Inception Distance[50] which we call "Frechet Morphological Distance" (FMD). The key difference is that instead of using the features of a pretrained Inception v3 model, we utilize morphological features. Given the set of real and generated morphological feature vectors Φ_g and Φ_r, we calculate the means (μ_g, μ_r) and standard deviations (Σ_g, Σ_r) and compute FMD as follows:

$$\begin{aligned} \mathrm{FMD}(\mu, \mu', \Sigma, \Sigma') &= \| \mu - \mu' \|_2^2 \\ &+ \mathrm{tr}(\Sigma + \Sigma') \\ &- 2\mathrm{tr}\left((\Sigma\Sigma')^{\frac{1}{2}} \right), \end{aligned} \quad (14)$$

where a lower FMD indicates the morphological distribution of real and generated anatomies are similar.

Topological evaluation

In order to study how well anatomic constraints and compatibility with numerical simulation are respected, we assess the topological quality of each label map. Clinically, topological defects such as a septal defect between the right and left hearts can have a significant effect on electrophysiology[51] and hemodynamics[52]. Specifically, for each generated anatomy we evaluate 12 different topological violations and calculate the percentage of topological violations exhibited by the cohort. We assess three types of topological violations. The first five metrics checks for the correct number of connected components for the Myo, LV, RV, LA, and RA channels. The next five metrics assesses the required adjacency relations between the following tissues: LV & Ao, LV & Myo, LV & LA, RV & Myo, RV & RA. The final two metrics examine the absence of adjacency relations between the LV & RV as well as the LA & RA. Multi-component topological violations were found by determining the presence of critical voxels as described in Gupta et al.[53], while the number of connected components was assessed by the method described by Silversmith et al.[54]. For computational efficiency, the label maps were subsampled by a factor of two before calculating all topological violations.

Generative autoencoder baseline

To establish a baseline for comparison with our diffusion model approach, we trained a generative VAE that encodes the cardiac label map into a global latent vector and decodes it back into voxel space. We generate new label maps with the generative VAE by sampling the global latent from a gaussian distribution and using it as input to the decoder. The generative VAE

architecture is similar to our reconstructive VAE, except we use 6 downsampling blocks to reduce the $128 \times 128 \times 128$ voxel resolution to $4 \times 4 \times 4$ before flattening the latent grid and feeding it as input to a fully connnected layer to produce a global latent vector of size 128. We also change the number of channels in the encoder and decoder to [64,128,196,256,256,512]. We trained the VAE with the same reconstruction and KL divergence loss, but increased the KL loss term to 1e−3 to more strongly enforce a Gaussian distribution on the global latent vectors for improved sampling.

Data availability

The label map data used for this study was derived from the TotalSegmentatorv1 dataset, which is available at the following URL https://zenodo.org/records/6802614.

Code availability

The code for training and evaluation will be available at https://github.com/kkadry/AnatomicEditing/.

Received: 22 January 2024; Accepted: 9 November 2024;
Published online: 05 December 2024

References

1. Sarrami-Foroushani, A. et al. In-silico trial of intracranial flow diverters replicates and expands insights from conventional clinical trials. *Nat. Commun.* **12**, 3861 (2021).
2. Kadry, K., Olender, M. L., Marlevi, D., Edelman, E. R. & Nezami, F. R. A platform for high-fidelity patient-specific structural modelling of atherosclerotic arteries: from intravascular imaging to three-dimensional stress distributions. *J. R. Soc. Interface* **18**, 20210436 (2021).
3. Rouhollahi, A. et al. Cardiovision: a fully automated deep learning package for medical image segmentation and reconstruction generating digital twins for patients with aortic stenosis. *Comput. Med. Imaging Graph.*https://doi.org/10.1016/j.compmedimag.2023.102289 (2023).
4. Straughan, R., Kadry, K., Parikh, S. A., Edelman, E. R. & Nezami, F. R. Fully automated construction of three-dimensional finite element simulations from optical coherence tomography. *Comput. Biol. Med.* **165**, 107341 (2023).
5. Bianchi, M. et al. Patient-specific simulation of transcatheter aortic valve replacement: impact of deployment options on paravalvular leakage. *Biomech. Model. Mechanobiol.* **18**, 435–451 (2019).
6. Kusner, J. et al. Understanding tavr device expansion as it relates to morphology of the bicuspid aortic valve: a simulation study. *PloS ONE* **16**, e0251579 (2021).
7. Ranard, L. S. et al. Feops heartguide patient-specific computational simulations for watchman flx left atrial appendage closure: a retrospective study. *JACC* **1**, 100139 (2022).
8. Karanasiou, G. S. et al. Design and implementation of in silico clinical trial for bioresorbable vascular scaffolds. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2675–2678 (IEEE, 2020).
9. Conway, C. et al. Acute stent-induced endothelial denudation: biomechanical predictors of vascular injury. *Front. Cardiovasc. Med.* **8**, 733605 (2021).
10. Roney, C. H. et al. In silico comparison of left atrial ablation techniques that target the anatomical, structural, and electrical substrates of atrial fibrillation. *Front. Physiol.* **11**, 1145 (2020).
11. Viceconti, M. et al. Possible contexts of use for in silico trials methodologies: a consensus-based review. *IEEE J. Biomed. Health Inform.* **25**, 3977–3982 (2021).
12. Sertkaya, A., DeVries, R., Jessup, A. & Beleche, T. Estimated cost of developing a therapeutic complex medical device in the us. *JAMA Netw. Open* **5**, e2231609–e2231609 (2022).
13. Niederer, S. et al. Creation and application of virtual patient cohorts of heart models. *Philos. Trans. R. Soc. A* **378**, 20190558 (2020).
14. Fogel, D. B. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: a review. *Contemp. Clin. Trials Commun.* **11**, 156–164 (2018).
15. Fabris, E. et al. Thin-cap fibroatheroma rather than any lipid plaques increases the risk of cardiovascular events in diabetic patients: Insights from the combine oct–ffr trial. *Circulation: Cardiovasc. Interv.* **15**, e011728 (2022).
16. Sacco, F. et al. Left ventricular trabeculations decrease the wall shear stress and increase the intra-ventricular pressure drop in cfd simulations. *Front. Physiol.* **9**, 458 (2018).
17. Moore, B. L. & Dasi, L. P. Coronary flow impacts aortic leaflet mechanics and aortic sinus hemodynamics. *Ann. Biomed. Eng.* **43**, 2231–2241 (2015).
18. Keshavarz-Motamed, Z. et al. Mixed valvular disease following transcatheter aortic valve replacement: quantification and systematic differentiation using clinical measurements and image-based patient-specific in silico modeling. *J. Am. Heart Assoc.* **9**, e015063 (2020).
19. Garber, L., Khodaei, S., Maftoon, N. & Keshavarz-Motamed, Z. Impact of tavr on coronary artery hemodynamics using clinical measurements and image-based patient-specific in silico modeling. *Sci. Rep.* **13**, 8948 (2023).
20. Williams, J. G. et al. Aortic dissection is determined by specific shape and hemodynamic interactions. *Ann. Biomed. Eng.* **50**, 1771–1786 (2022).
21. Beetz, M. et al. Interpretable cardiac anatomy modeling using variational mesh autoencoders. *Front. Cardiovasc. Med.* **9**, 983868 (2022).
22. Beetz, M., Banerjee, A. & Grau, V. Generating subpopulation-specific biventricular anatomy models using conditional point cloud variational autoencoders. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, 75–83 (Springer, 2021).
23. Qiao, M. et al. Cheart: a conditional spatio-temporal generative model for cardiac anatomy. (IEEE transactions on medical imaging, 2023).
24. Kong, F. et al. Sdf4chd: generative modeling of cardiac anatomies with congenital heart defects. *Med. Image Anal.* **97**, 103293 (2024).
25. Pinaya, W. H. et al. Brain imaging generation with latent diffusion models. In *Deep Generative Models: Second MICCAI Workshop, DGM4MICCAI 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings*, 117–126 (Springer, 2022).
26. Müller-Franzes, G. et al. Diffusion probabilistic models beat gans on medical images. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2212.07501 (2022).
27. Khader, F. et al. Denoising diffusion probabilistic models for 3d medical image generation. *Sci. Rep.* **13**, 7303 (2023).
28. Fernandez, V. et al. Can segmentation models be trained with fully synthetically generated data? In *Simulation and Synthesis in Medical Imaging: 7th International Workshop, SASHIMI 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18, 2022, Proceedings*, 79–90 (Springer, 2022).
29. Go, S., Ji, Y., Park, S. J. & Lee, S. Generation of structurally realistic retinal fundus images with diffusion models. In *Proceedings of the IEEE/CVF Conferenceon Computer Vision and Pattern Recognition*, p 2335–2344 (2024).
30. Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **33**, 6840–6851 (2020).
31. Meng, C. et al. Sdedit: guided image synthesis and editing with stochastic differential equations. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2108.01073 (2021).
32. Nichol, A. et al. Glide: towards photorealistic image generation and editing with text-guided diffusion models. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2112.10741 (2021).

33. Song, Y. et al. Score-based generative modeling through stochastic differential equations. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2011.13456 (2020).
34. Song, Y., Shen, L., Xing, L. & Ermon, S. Solving inverse problems in medical imaging with score-based generative models. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2111.08005 (2021).
35. Song, J., Vahdat, A., Mardani, M. & Kautz, J. Pseudoinverse-guided diffusion models for inverse problems. In *International Conference on Learning Representations* (2023).
36. Chung, H., Kim, J., Mccann, M. T., Klasky, M. L. & Ye, J. C. Diffusion posterior sampling for general noisy inverse problems. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2209.14687 (2022).
37. Bercea, C. I., Neumayr, M., Rueckert, D. & Schnabel, J. A. Mask, stitch, and re-sample: enhancing robustness and generalizability in anomaly detection through automatic diffusion models. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2305.19643 (2023).
38. Fontanella, A., Mair, G., Wardlaw, J., Trucco, E. & Storkey, A. Diffusion models for counterfactual generation and anomaly detection in brain images. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2308.02062 (2023).
39. Rouzrokh, P. et al. Multitask brain tumor inpainting with diffusion models: a methodological report. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2210.12113 (2022).
40. Ho, J. & Salimans, T. Classifier-free diffusion guidance. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2207.12598 (2022).
41. Romero, P. et al. Clinically-driven virtual patient cohorts generation: an application to aorta. *Front. Physiol.* 1375 (2021).
42. Wasserthal, J. et al. Totalsegmentator: robust segmentation of 104 anatomical structures in ct images. *Radiol Artif Intell*. 5 (2023).
43. Avants, B. B., Epstein, C. L., Grossman, M. & Gee, J. C. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Med. Image Anal.* **12**, 26–41 (2008).
44. Karras, T., Aittala, M., Aila, T. & Laine, S. Elucidating the design space of diffusion-based generative models. *Adv. Neural Inf. Process. Syst.* **35**, 26565–26577 (2022).
45. Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695 (2022).
46. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*. (San Diega, CA, USA, 2015).
47. Yang, G. et al. Pointflow: 3d point cloud generation with continuous normalizing flows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4541–4550 (2019).
48. Lorensen, W. E. & Cline, H. E. Marching cubes: A high resolution 3d surface construction algorithm. In *Seminal graphics: pioneering efforts that shaped the field*, 347–353 (1998).
49. Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J. & Aila, T. Improved precision and recall metric for assessing generative models. *Adv. Neural Inf. Process. Syst.* **32**, 3927–3936 (2019).
50. 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. *Adv. Neural Inf. Process. Syst.* **30**, 6626–6637 (2017).
51. Williams, M. R. & Perry, J. C. Arrhythmias and conduction disorders associated with atrial septal defects. *J. Thorac. Dis.* **10**, S2940 (2018).
52. Shah, S. R. et al. The impact of an atrial septal defect on hemodynamics in patients with heart failure. *US Cardiol. Rev.* **11**, 72 (2017).
53. Gupta, S. et al. Learning topological interactions for multi-class medical image segmentation. In *European Conference on Computer Vision*, 701–718 (Springer, 2022).
54. Silversmith, W. cc3d: Connected components on multilabel 3D & 2D images. (2021).

Acknowledgements

The authors would like to thank Vivek Gopalakrishnan and Payal Chandak for their helpful feedback on the manuscript and figure design. This work was supported in part by grants to E.R.E. and F.R.N. from the National Institutes of Health (R01 HL161069) and a Henri Termeer Fellowship to K.K.

Author contributions

K.K. curated the dataset, wrote the paper, and performed all experiments. K.K. and S.G. performed experiments. K.K., S.G., F.R.N., and E.R.E. reviewed the results.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-024-01332-0.

Correspondence and requests for materials should be addressed to Karim Kadry.