AnatoDiff: Synthesizing Anatomically Truthful Radiographs With Limited Training Images

Ka-Wai Yung, Jayaram Sivaraj, Lodovico di Giura, Simon Eaton, Paolo De Coppi, Danail Stoyanov, *Fellow, IEEE*, Stavros Loukogeorgakis and Evangelos B. Mazomenos, *Member, IEEE*

***Abstract*— Rapid advancements in diffusion models have enabled synthesis of realistic and anonymized imagery in radiography. However, due to their complexity, these models typically require large training volumes, often exceeding 10,000 images. Pre-training on natural images can partly mitigate this issue, but often fails to generate anatomically accurate shapes due to the significant domain gap. This prohibits applications in specialized medical conditions with limited data. We propose AnatoDiff, a diffusion model synthesizing high-quality X-Ray images with accurate anatomical shapes using only 500 to 1,000 training samples. AnatoDiff incorporates a Shape Prototype Module and Anatomical Fidelity loss, allowing for smaller training volumes through targeted supervision. We extensively validate AnatoDiff across three open-source datasets from distinct anatomical regions: Neonatal Abdomen (1,000 images); Adult Chest (500 images); and Humerus (500 images). Results demonstrate significant benefits, with an average improvement of 14.9% in Fréchet Inception Distance, 9.7% in Improved Precision, and 2.3% in Improved Recall compared to state-of-the-art (SOTA) few-shot and data-limited natural image synthesis methods. Unlike other models, AnatoDiff consistently generates anatomically correct images with accurate shapes. Additionally, a ResNet-50 classifier trained on AnatoDiff-generated images shows a 2.1% to 5.3% increase in F1-score, compared to being trained on SOTA diffusion images, across 500 to 10,000 samples. A survey with 10 medical professionals reveals that images generated by AnatoDiff are challenging to distinguish from real ones, with a Matthews correlation coefficient of 0.277 and Fleiss' Kappa of 0.126, highlighting the effectiveness of AnatoDiff in generating high-quality, anatomically accurate radiographs. Our code is available at https://github.com/KawaiYung/AnatoDiff.**

***Index Terms*— Diffusion Models, X-Ray, Generative Modeling, Topological Data Analysis**

I. Introduction

DENOISING diffusion models have gained increasing attention in recent years as an effective approach for high-quality synthesis. [1]–[3]. This has spurred a variety of applications, including image generation [4]–[6], video synthesis [7]–[9], and 3D object creation [10]–[12], showcasing the versatility of these methodologies. In many studies, diffusion models clearly surpass Generative Adversarial Networks (GANs) in natural image synthesis [2], [4], [13]. Notably, their adoption within the field of medical imaging synthesis holds significant potential, especially in facilitating the generation of de-identified imagery for dataset anonymization [14]–[16]. Diffusion models have been explored in various medical imaging applications, including radiography [17]–[22], histopathology [23]–[25] and retinal imaging [26], [27]. Focusing on radiography, the development of diffusion models is typically carried out with large-size datasets ($\geq$10,000 images), such as CheXpert [28] and MIMIC-CXR [29]. While outstanding results have been reported [19]–[22], the practicality of obtaining such large-volume, high-quality datasets is often limited, particularly in rare diseases with low incidence.

In such cases, datasets are often larger than a few-shot setting (10-100 images), but much smaller than large-scale datasets ($\geq$10,000 images). This poses challenges in applying diffusion models when data availability is restricted, and remains to date an underexplored area. When adapting diffusion models for radiographic image generation, existing approaches often rely on direct applications of the Latent Diffusion Model (LDM) [4] as the generative backbone [30]–[34]. However, LDMs are typically pre-trained on large-scale natural image datasets and often require substantial quantities of diverse training data to adapt effectively. When applied to data-limited medical domains, this can result in poor anatomical consistency and a failure to capture domain-specific structures that are critical for clinical applicability. These limitations highlight the need for a tailored approach that operates effectively under data scarcity while preserving anatomical fidelity.

Fig. 1 illustrates the generation performance of fully fine-tuned LDMs using varying numbers of CheXpert training samples. For larger sample sizes, performance remains stable across all three metrics up to approximately 1,000 samples. As the number of training samples decreases, both Improved Precision (IP) and Improved Recall (IR) decline, and Fréchet Inception Distance (FID) increases—indicating degradation in both diversity and fidelity. For instance, at 100 training images, IP drops significantly. At only 10 images, both IP and

This work was funded in whole, or in part, by the Department of Science, Innovation and Technology (DSIT) and the Royal Academy of Engineering under the Chair in Emerging Technologies [CiET1819/2/36]. For the purpose of open access, the authors have applied a CC BY public copyright licence to any author accepted manuscript version arising from this submission.

(*Corresponding authors: K-W. Yung, E. B. Mazomenos)*

K.-W. Yung, D. Stoyanov and E. Mazomenos are with the UCL Hawkes Institute and the Department of Medical Physics and Biomedical Engineering, University College London, WC1E 6BT, UK

J. Sivaraj, L. di Giura, S. Eaton, P. De Coppi and S. Loukogeorgakis are with the Department of Specialist Neonatal and Paediatric Surgery and the NIHR Great Ormond Street Biomedical Research Centre, WC1N 3JH, UK

(email:{ka.yung.22; e.mazomenos}@ucl.ac.uk)

IR approach zero, and FID rises sharply, reflecting severely impaired generation quality. A knee point emerges around 500 samples, where IP begins to decline while IR and FID remain relatively stable. This suggests that although global realism and diversity are preserved, there is a drop in local anatomical fidelity. In this regime, generated images may appear globally realistic but often exhibit inaccurate anatomical details, such as distorted contours or misplaced structures.

To address this, we propose AnatoDiff, a method capable of generating anatomical realistic radiographs using only 500-1,000 images. Building on the transformer-based DiT model [35], AnatoDiff achieves realistic anatomical fidelity by introducing two pivotal innovations: Shape Prototype Module (SPM) and Anatomical Fidelity (AF) loss.

Specifically, in data-limited scenarios, the model may struggle to infer anatomical priors solely from sparse and undersampled examples. By clustering a given set of training images, we obtain cluster centers, which are initialized as shape prototypes at the start of training to enhance shape fidelity. SPM leverages available shapes from the training set, allowing the model to receives consistent guidance toward anatomically plausible outputs, reducing the risk of unrealistic shape generations caused by insufficient structural coverage.

Meanwhile, the commonly used Mean Squared Error (MSE) loss—though effective in large-scale settings—assigns uniform importance across all pixels. This uniform weighting can dilute the training signal in small-data regimes, where semantically important but spatially small structures may be overlooked. To address this, we introduce the Anatomical Fidelity (AF) loss, which leverages Topological Data Analysis (TDA) to extract and emphasize persistent topological features from both target and generated images. These features typically correspond to key anatomical regions—such as organ boundaries, lung fields, or bony contours—and are assigned higher priority during optimization. The AF loss thus functions as a form of importance-aware supervision, guiding the model to reconstruct structures that are both topologically stable and clinically meaningful, even when sparsely represented in the training data.

Together, the SPM and AF loss address complementary aspects of the limited-data problem: SPM provides a strong inductive bias on global anatomical shape through prototype guidance, while AF loss promotes accurate reconstruction of local and persistent anatomical details.

We benchmark AnatoDiff against state-of-the-art (SOTA) image generation models in limited-size datasets. We evaluate three X-ray datasets representing different anatomical structures: A small subset of the CheXpert dataset (500 images), the MURA bone dataset (500 images) [36], and the GOSH Necrotizing Enterocolitis (NEC) dataset (1,000 images) [37]. Our findings reveal that existing methods fall short in medical image synthesis under data scarcity. GAN-based methods fail to generate realistic images and often collapse due to insufficient training samples relative to the complexity of X-ray images. While diffusion-based methods (LDM [4], FSDM [38], DiffFit [39]) generate semantically higher-quality images, the limited training set hinders their ability to capture general shape and anatomy accurately, resulting in untruthful appearances. AnatoDiff significantly outperforms SOTA methods, showing an average improvement of 14.9% in FID, 9.7% in IP, and 2.3% in IR. Qualitative analysis further confirms that our method generates radiographs with truthful anatomical shapes and higher semantic quality. In summary, our contributions are:

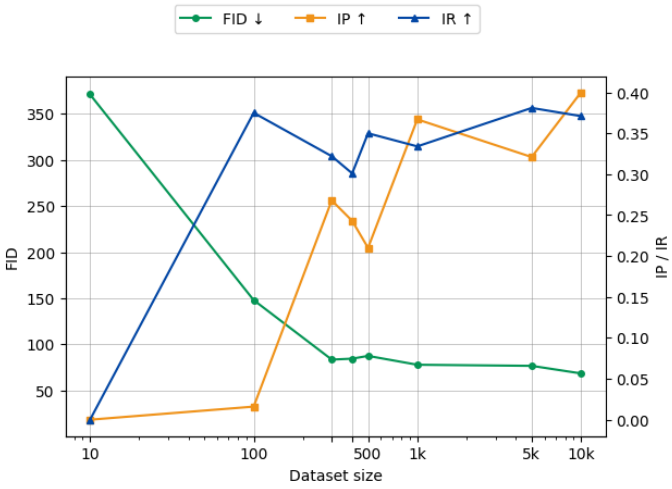

Fig. 1. Performance of full fine-tuning LDM with varying numbers of CheXpert training samples

- We address the underexplored task of radiography synthesis under data-limited conditions (500-1,000 images). Conducting over 150 experiments, we benchmark existing few-shot and data-limited methods.
- We propose a novel framework - AnatoDiff, which enhances anatomical consistency and generates high-fidelity radiographs by incorporating the SPM and AF loss.
- Experiments on three open-source datasets from different anatomical structures demonstrate the superiority of our approach. AnatoDiff achieves SOTA performance across all datasets, with improvements of 14.5%, 15.4% and 14.7% in FID, 13%, 9% and 7% in IP, and 5%, 1% and 1% in IR, while also achieving excellent shape and anatomical truthfulness. A classifier (ResNet-50) trained on images generated by AnatoDiff consistently outperforms models trained on SOTA diffusion images, with 2.1% to 5.3% F1 score across 500 to 10,000 samples.
- A survey involving 10 medical professionals revealed that images generated by AnatoDiff are challenging to distinguish from real images, with a Matthews correlation coefficient of 0.277 and Fleiss' Kappa of 0.126.

II. Related Works

A. Diffusion models

Diffusion models iteratively add Gaussian noise into data x, transforming it from an ordered state x_0 to a disordered state x_T. These models are trained to reverse this trajectory, reconstructing the original data from the noisy state. The transformation and reversal are governed by conditional probabilities modeling transitions between consecutive states. The forward process is described by a Markov chain with Gaussian

transition probabilities:

$$q(x_1, \ldots, x_T|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) \tag{1}$$

x_0 represents the initial, noise-free data point, and x_T represents the data point after T steps of noise addition. The transition at each time step t is defined as:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{2}$$

$\mathcal{N}$ denotes the normal (Gaussian) distribution, β_t represents the variance of the noise added at step t, and I is the identity matrix. This formulation represents the process of incrementally adding noise at each step, gradually transforming the data from a structured state to a less structured one. Conversely, the reverse process is modeled by a neural network with parameters θ, which defines the conditional distribution of earlier data points given later ones and learns to reverse the noise addition process. The reverse process is given as:

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t) \tag{3}$$

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{4}$$

Here, $p_\theta(x_{t-1}|x_t)$ denotes the learned distribution at step $t-1$ given x_t, where μ_θ and Σ_θ are the mean and covariance of the Gaussian distribution learned by the model.

The objective of the model during training is to minimize the difference between the actual noise added during the forward process and the noise estimated by the model during the reverse process, represented by a loss function:

$$L_{\text{simple}}(\theta) = \mathbb{E}_{t,x_0,\epsilon}\left[\left\|\epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t\right)\right\|^2\right] \tag{5}$$

$\mathbb{E}$ is the expectation over the variables t, x_0, the added noise ϵ, and the noise estimated by the model ϵ_θ. The term $\bar{\alpha}_t$ represents the cumulative product of $(1-\beta_t)$ up to time t.

Previous studies on diffusion models for radiographs have predominantly relied on large-scale datasets for training. Müller-Franzes *et. al.* utilized a filtered subset of the CheXpert dataset with 191,027 images, comparing the quality of radiographs generated using diffusion and GAN models [20]. Chambon *et. al.* used two training sets from the MIMIC-CXR dataset, with 38,009 and 175,622 images, to generate synthetic chest X-rays (CXRs) from text prompts [30]. Weber *et. al.* compiled a multi-source dataset totaling 651,471 images for large-scale CXR generation [19]. Packhauser *et. al.* used the ChestX-ray14 dataset with 112,120 images for generating anonymous CXRs [40]. The smallest dataset in this context, used by Ali *et. al.*, comprised 3,165 images to train a stable diffusion model to generate synthetic lung X-rays. In their study, two radiologists determined if an image is real or synthetic; however, no quantitative comparison of the generated images' fidelity was performed [41]. In contrast, motivated by the need to understand model performance in scenarios where large-scale datasets might not be feasible or accessible - an often situation in specialized medical conditions - our study focuses on the potential and challenges of employing diffusion models on limited data settings of 500-1,000 X-ray images.

B. Data-Limited Image Generation

Prior works in image generation under data-limited conditions have primarily utilized GAN models. For instance, CDC enhanced diversity transfer by preserving cross-domain distance consistency [42]. MoCA improved image generation quality with memory prototypes [43]. RICK tackled incompatible knowledge transfer through knowledge truncation, selectively utilizing relevant generative knowledge [44].

Research on diffusion models in data-limited scenarios is scarce. Giannone *et. al.* introduced FSDM for few-shot image generation by integrating a Vision Transformer as a set encoder for reference images, providing extra conditioning to the diffusion model [38]. Zhu *et. al.* developed DDPM-PA to preserve information from source domains during few-shot adaptation [45]. Parameter-efficient fine-tuning from a pre-trained model sourced from large datasets has been also considered, as seen in BitFit, where only the biases of the model are fine-tuned [46], and LoRA, which employs low-rank adaptations for efficient training [47]. Distinct from these methods, our work represents the original study on diffusion models for radiography generation in a data-limited scenario.

III. Methods

A. Background

Fig. 2 illustrates the architecture of our proposed AnatoDiff, which builds upon the DiT. An input image i is initially encoded by a Variational Autoencoder (VAE) into a latent space representation $z \in \mathbb{R}^{C\times W\times H}$. This is then transformed into patches of dimension $\mathbb{R}^{T\times D}$, where T is the number of patches and D is the dimension of each patch's latent representation. The transformed patches serve as input into the DiT along with a class conditioning embedding $c \in \mathbb{R}^D$. Within each DiT block, the conditioning c is processed by a Feed-Forward Network (FFN) to compute six scaling and shifting parameters (Eq. 6), utilized to scale the attention (Eq. 7) and feed-forward outputs (Eq. 8), where Attn represents Multi-head Attention.

$$\alpha_1, \tau_1, \beta_1, \alpha_2, \tau_2, \beta_2 = \text{FFN}(c) \tag{6}$$

$$z_i^{attn} = \alpha_1 \text{Attn}(\tau_1 z_i + \beta_1) + z_i \tag{7}$$

$$z_{i+1} = \alpha_2 \text{FFN}(\tau_2 z_i^{attn} + \beta_2) + z_i^{attn} \tag{8}$$

The default DiT has a large number of trainable parameters, requiring significant computation. Both BitFit and DiffFit have shown that freezing all parameters except for the bias terms enables the model to achieve comparable or even superior performance during fine-tuning, while substantially reducing the number of trainable parameters [39], [46]. We therefore follow the approach of Difffit [39], freezing all weights of a pre-trained DiT except for the bias terms to reduce the risk of overfitting, and introduce learnable scaling factors, γ_1, γ_2:

$$z_i^{attn} = \gamma_1\alpha_1 \text{Attn}(\tau_1 z_i + \beta_1) + z_i \tag{9}$$

$$z_{i+1} = \gamma_2\alpha_2 \text{FFN}(\tau_2 z_i^{attn} + \beta_2) + z_i^{attn} \tag{10}$$

Both scaling factors are initialized to 1.0 and modulate the outputs of the Attention and Feed-Forward blocks by applying element-wise scaling. Scaling factors allows for direct control

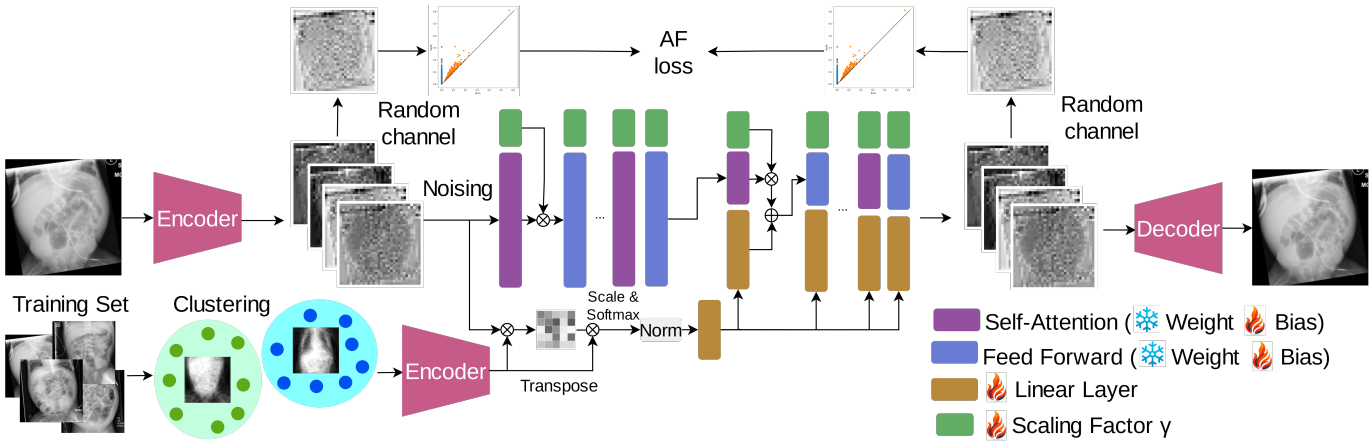

Fig. 2. Illustration of the proposed AnatoDiff. We introduce the SPM that provides a set of shape prototypes to guide the generation of anatomically accurate images. Additionally, we incorporate the AF loss to provide strong regularization on the topological structure of the generated images. This ensures the preservation of key structural elements throughout the training process, enabling the model to produce high-quality images that accurately represent the underlying shape and anatomy.

over layer outputs, effectively allowing the network to suppress or amplify specific layers more efficiently, resulting in faster convergence.

By freezing all parameters except the bias terms and introducing scaling factors γ, computational and memory requirements are significantly reduced. This strategy also mitigates the risk of overfitting, particularly when training on small datasets with models that have a large number of parameters.

B. Shape Prototype Module (SPM)

Unlike natural images, radiographs contain complex anatomical structures. While previous works have shown that diffusion models can capture and generate images with well-defined structures given sufficient training data [19]–[22], our experiments shows that these methods struggle in data-limited settings, where insufficient supervision leads to a poor understanding of the object's anatomy. Considering that objects in radiography, such as the chest and abdomen have well-defined consistent shapes and anatomical structures, we propose an SPM to emphasize this information. This module serves as a prototype during image generation, contributing in maintaining anatomical accuracy despite limited training data.

The dominant anatomical structures within images—such as lung fields, abdominal contours, and joint outlines—are often less distinct and in a soft, continuous representations. To emphasize them, we first apply binarization to the training images, resulting in a sharper and more interpretable cues. Additionally, the binarization process suppresses low-activation noise, resulting in spatially cleaner prototypes.

Specifically, given a set of training images $I = \{i_1, i_2, ..., i_n\}$, where n is the total number of images, we first binarize I by thresholding, followed by applying K-Means clustering to compute K cluster centers:

$$I_c = \{i_{c1}, i_{c2}, ..., i_{ck}\} = KMeans(I, K) \tag{11}$$

where i_c is a cluster center for I. To mitigate potential initialization instability, we adopt k-means++ [48] as our initialization strategy.

Vectorization is then performed using the same VAE to convert I_c into latent $\theta_c \in \mathbb{R}^{KT\times D}$. During training and inference, θ_c is used as input to the SPM to guide the image generation process. Initially, the noised latent z is treated as query and interacts with θ_c through cross-attention to compute a weighted prototype representation z_c, where z_c represents a soft combination of multiple prototypes:

$$z_c = \text{softmax}\left(\frac{(zW^Q)(\theta_c W^K)^T}{\sqrt{d_k}}\right)(\theta_c W^V) \tag{12}$$

Following this, we first pass z_c through lightweight linear layers. To integrate z_c with the generation input z, approaches such as concatenation followed by a feed-forward layer or applying a cross-attention mechanism could offer greater expressiveness. However, these approaches introduce additional trainable parameters, conflicting with our goal of maintaining parameter efficiency. In contrast, we found that simple element-wise addition was sufficient to achieve performance gains without introducing more parameters, thereby preserving the lightweight design of our model.

Furthermore, to retain the parameter-efficient nature of AnatoDiff, we do not apply z_c across all DiT blocks. Instead, we limit its application to the last N blocks. Specifically, in each of the final N DiT blocks, z_c is passed through a linear layer and then aggregated with z via element-wise addition:

$$z_i^{attn} = \begin{cases} \gamma_1\alpha_1\text{Attn}(\tau_1 z_i + \beta_1) + z_i & i < N, \\ \gamma_1\alpha_1\text{Attn}(\tau_1 z_i + \beta_1) + \text{FFN}(z_c) + z_i & else. \end{cases} \tag{13}$$

$$z_{i+1} = \begin{cases} \gamma_2\alpha_2\text{FFN}(\tau_2 z_i^{attn} + \beta_2) + z_i^{attn} & i < N, \\ \gamma_2\alpha_2\text{FFN}(\tau_2 z_i^{attn} + \beta_2) + \text{FFN}(z_c) + z_i^{attn} & else. \end{cases} \tag{14}$$

C. Anatomical Fidelity (AF) Loss

Medical image generation tasks are characterized by complex and highly structured patterns, therefore preserving global shapes and anatomical patterns is crucial. Diffusion models

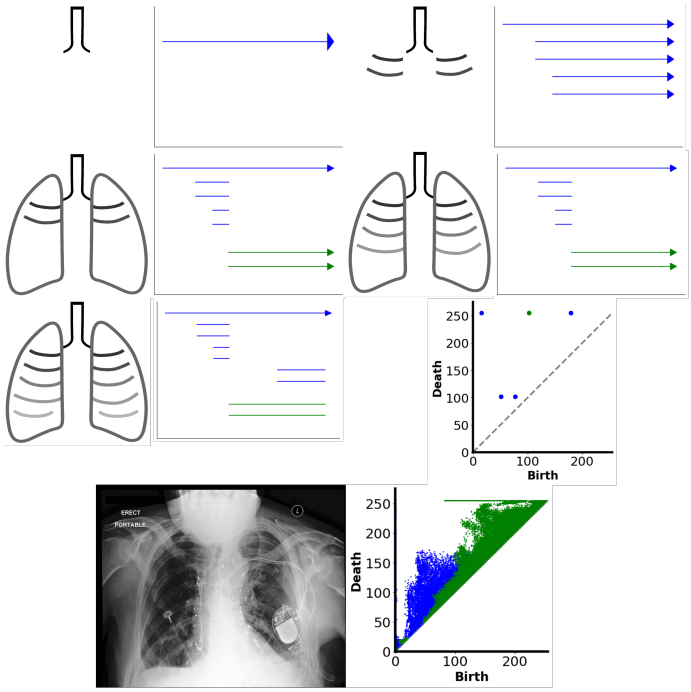

Fig. 3. Illustration of the creation of a Persistence Diagram (PD) from a simplified chest image using cubical complex. As the threshold $\boldsymbol{t}$ increases, more pixels are revealed, leading to the appearance and disappearance of topological structures. At low $\boldsymbol{t}$ values, only the upper portion of the chest is visible, forming a single connected component ($\boldsymbol{H_0}$), recorded as a blue bar in the persistence barcode. As $\boldsymbol{t}$ increases, additional components—representing parts of the ribs—emerge, adding more $\boldsymbol{H_0}$ features. At higher thresholds, the lung contours become visible, forming loops ($\boldsymbol{H_1}$), which appear as green bars in the barcode. Concurrently, the previously separated rib components merge into a unified structure, resulting in the death of all but one $\boldsymbol{H_0}$ component. The PD reflects this simplified chest structure in the following order of persistence: the overall chest (excluding lower ribs) as the top-left point in the PD, the two lungs (green points), the lower ribs (top-right points), and the remaining ribs, which gradually merge into the chest structure (points near the diagonal). Thie PD captures multiscale topological features, with higher persistence (i.e., greater significance) on the global chest shape and the lungs—both key anatomical components. The bottom row shows an example form a real CheXpert X-ray (left) and its corresponding PD (right).

traditionally utilize the MSE loss, which primarily focuses on pixel-level accuracy and local features. However, this approach does not adequately capture the broader, more abstract structural relationships within the data. As a result, synthesized images, while visually similar to target images at a local level, may lack correct anatomical structure or exhibit altered global features. Edge-based losses can potentially aid in preserving anatomical structure by emphasizing boundaries and contours. However, applying such losses directly in the latent space is challenging due to the abstract nature of latent representations, while implementing them in image space for diffusion models entails computationally expensive decoding at each training step. In contrast, we propose an AF loss, inspired by TDA, to enhance the model's understanding of anatomical structures. By comparing the topological structures between the model's output and target directly at latent level, the AF loss guarantees high anatomical and shape fidelity, without requiring a decoding step.

To extract the topological features from an image, a filtration process is performed. An example of the filtration process with a cubical complex is illustrated in Fig. 3. Consider an image represented as a function $f : \mathbb{Z}^2 \to \mathbb{R}$, where $\mathbb{Z}^2$ indexes the pixel grid and $\mathbb{R}$ represents the intensity values at each pixel. A filtration is constructed based on these intensity values using sublevel sets:

$$S_t = x \in \mathbb{Z}^2 : f(x) \leq t \tag{15}$$

where t is a threshold parameter. As t increases from the minimum (dark) to the maximum intensity (bright) value, more pixels are included in the sublevel sets, gradually revealing the structure of the image. Initially, only the pixels with low intensities are included. As t increases, these components grow and begin to merge, forming larger connected regions and eventually enclosing loops. This process can be tracked and visualized using a Persistence Barcode (Fig. 3 right panel). Each bar in the barcode represents a topological feature (connected component or loop) with its birth b and death d times along the threshold parameter t, depicted as:

$$\text{Birth-death pair} = (b, d) \tag{16}$$

Finally, the persistence barcode is converted into a persistence diagram (Fig. 3 second bottom panel), shown as a scatter plot where the x-axis is the birth time b and y-axis is the death time d. Each point in the persistence diagram corresponds to a bar in the barcode plot and is given by:

$$PD(\mathbb{Z}) = \{(b_i, d_i) \in \mathbb{R}^2 | b_i < d_i\} \tag{17}$$

Points near the diagonal $y = x$ (gray dotted line) represent features with short lifetimes, considered as noise. In contrast, points far from the diagonal represent features that persist across a large range of the threshold value t and are indicative of important topological features in the data.

During training, a one-step de-noised latent $\hat{z_0}$ is obtained by subtracting the model-predicted noise σ from the noised latent z. Filtration is then applied to both the one-step de-noised latent $\hat{z_0}$, and the un-noised latent (target) z_0, to obtain two persistence diagrams $PD_1(\hat{z_0})$ and $PD_2(z_0)$.

To compute the distance between two persistence diagrams, we apply the p-Wasserstein distance:

$$W_p(PD_1, PD_2) = \left(\inf_{\gamma \in \Gamma(PD_1, PD_2)} \sum_{(x,y) \in \gamma} \|x - y\|_p^p \right)^{\frac{1}{p}} \tag{18}$$

where $\Gamma(PD_1, PD_2)$ denotes all possible bijections γ that map points from PD_1 to points in PD_2. Features with higher persistence have a greater influence on the p-Wasserstein distance when misaligned, inherently guiding the loss function to prioritize the alignment of anatomically significant structures. Minimizing the p-Wasserstein distance thus effectively promotes the alignment of topological features between the ground truth and reconstructed latent representations. Anatomical patterns that exhibit high persistence—reflecting their structural stability and global spatial extent—are consequently emphasized throughout training.

Finally, the Wasserstein distance is summed with the MSE loss, weighted by a trade-off factor λ:

$$L = L_{MSE} + \lambda W_p(PD_1, PD_2) \tag{19}$$

In data-limited conditions, pixel-wise losses such as MSE treat all image regions equally, which can lead the model to overfit to dominant visual patterns (e.g., textures) while underrepresenting semantically critical structures (e.g., organ boundaries, lung contours). The AF loss addresses this limitation by prioritizing the alignment of high-persistence topological features that typically correspond to anatomically important regions. These features exert greater influence on the Wasserstein term, guiding the model to reconstruct spatial structures most relevant for clinical realism. In this way, the AF loss implicitly injects anatomical priors into the training process—without requiring explicit annotations—and improves generation fidelity when supervision is sparse. This property directly supports the goal of synthesizing anatomically faithful radiographs from limited training data.

To speed up the training process, the PD computation is performed directly at the latent level, by considering the latent representation with dimension 32×32 as an image, eliminating decoding overheads. Our experiments show that both H_0 and H_1 homology groups are necessary to achieve optimal results. To accelerate the PD and Wasserstein distance computations, we verify that performing PD computation by randomly selecting a channel within each latent, rather than using all channels, achieves similar performance. This approach results in 1.8× speed increase compared to using the full latent space, without compromising performance.

IV. Datasets, Settings and Metrics

A. Datasets

GOSH NEC [37] is a pediatric abdominal X-Ray (AXR) dataset of NEC cases collected from the Great Ormond Street Hospital for Children (GOSH)[1]. NEC is a rare but severe intestinal disease in premature neonates with high mortality [52]. The dataset comprises 1,398 AXRs from 380 patients, categorized into four classes: 497 Surgical NEC images from 86 patients, 346 Medical NEC images from 97 patients, 307 No Pathology images from 133 patients and 248 Other Pathology images from 64 patients. Due to the complexity of the neonatal abdominal structure and to minimize collapsing in GANs, we allocate patient-by-patient basis 80% (∼1,000 images) for training and 20% (∼280 images) for testing, repeated five times with different splits.

CheXpert [28] is a comprehensive CXR dataset comprising 224,316 images from 65,240 patients, including both frontal and lateral view radiographs. We randomly sample 1,500 images and split into 500 for training and 1,000 for testing, repeated five times with different samplings seeds.

MURA [36] is a bone X-ray dataset containing 40,561 images from 12,173 patients. We use the Humerus subcategory, sampling 500 images for training and the original test set of 288 images, repeated five times with different sampling seeds.

B. Metrics

Following [20], FID, IP and IR are used for evaluation.

[1]Available at: https://doi.org/10.5522/04/26042824.v1

FID compares the distribution of generated to real images using typically an Inception network, defined as:

$$\mathrm{FID} = \|\mu_r - \mu_g\|_2^2 + \mathrm{Tr}\Big(\Sigma_r + \Sigma_g - 2\,(\Sigma_r\,\Sigma_g)^{\frac{1}{2}}\Big),$$

where μ_r, Σ_r are mean and covariance of real-image features, and μ_g, Σ_g are mean and covariance of generated-image features. A lower FID score indicates better similarity between generated and real images, suggesting higher performance.

IP measures generated image quality as the proportion classified as real by a trained classifier, defined as:

$$\mathrm{IP}(\Phi_r, \Phi_g) = \frac{1}{|\Phi_g|} \sum_{\phi_g \in \Phi_g} f(\phi_g, \Phi_r),$$

where ϕ_g is the generated images feature vectors, Φ_r and Φ_g are a sets of real and generated images feature vectors, and $f(\phi_g, \Phi_r)$ is denotes the k-th nearest neighbor of ϕ_r in Φ_r. IP quantifies if each generated image is within the estimated manifold of real images. Higher precision indicates that more generated images are of high quality and resemble real data.

IR assesses diversity of generated images, indicating how well the model captures the variability in the real dataset, defined as:

$$\mathrm{IR}(\Phi_r, \Phi_g) = \frac{1}{|\Phi_r|} \sum_{\phi_r \in \Phi_r} f(\phi_r, \Phi_g).$$

where ϕ_r is the real images feature vectors. IR quantifies if each real image is within the estimated manifold of the generated images. Higher recall means the model generates a wider variety of images, reflecting the diversity of real data.

C. Settings

We repeat experiments five times and compute the average performance along with the standard deviation. Settings and hyperparameters on components derived from DiT and DiffFit are kept unaltered. For all datasets, we empirically set the binary threshold value in SPM to 150, the number of clusters to 25 and the last N blocks to 7. The trade-off factor λ is 2×10^{-6}. We use both H_0 and H_1 homology groups and the 1-Wasserstein distance. AnatoDiff is trained with a batch size of 16 on a single NVIDIA RTX 4090 GPU. All images are generated at a resolution of 256×256. During inference, we generate 1,000 images and compute metrics against the testing set. All models are implemented using Pytorch. For ablations and stability analysis, all experiments are carried out on the CheXpert dataset and results are averaged across five times.

V. Results and Discussions

A. Quantitative results

We compare AnatoDiff with SOTA data-limited and few-shot generation models, including the GAN-based models MoCA [43], CDC [42], and RICK [44], as well as the diffusion-based model FSDM [44]. Additionally, we benchmark the popular GAN-based medical image generation model DC-GAN [53], as a reference point to illustrate the limitations of commonly used GAN-based methods [54]–[56]. To compare with more recent medical-specific generative models, we

TABLE I
QUANTITATIVE RESULTS FROM ALL 3 DATASETS. IMPROVEMENTS ARE WITH RESPECT TO DIFFFIT.

	GOSH NEC			CheXpert			MURA		
	FID↓	IP↑	IR↑	FID↓	IP↑	IR↑	FID↓	IP↑	IR↑
DC-GAN [27]	449.0±100.9	0.00±0.00	0.00±0.00	359.9±123.4	0.01±0.00	0.00±0.00	349.2±48.0	0.06±0.10	0.00±0.00
MedGAN [49]	312.6±18.7	0.00±0.00	0.06±0.03	259.3±21.3	0.02±0.02	0.00±0.00	347.6±15.3	0.06±0.10	0.00±0.00
MoCA [43]	245.0±16.0	0.20±0.15	0.00±0.00	217.6±48.4	0.01±0.02	0.00±0.00	199.4±17.9	0.27±0.13	0.00±0.00
CDC [42]	169.4±17.0	0.01±0.02	0.06±0.03	107.2±18.2	0.08±0.05	0.04±0.03	242.5±6.9	0.01±0.01	0.01±0.01
RICK [44]	165.8±6.2	0.08±0.03	0.08±0.02	68.5±0.5	0.27±0.02	0.28±0.02	160.7±9.7	0.05±0.01	0.12±0.02
LDM (BitFit) [46]	148.3±3.2	0.13±0.03	0.10±0.06	115.0±3.0	0.02±0.01	0.53±0.03	127.8±3.3	0.23±0.03	0.35±0.02
MT-DDPM [50]	130.7±8.3	0.26±0.1	0.07±0.03	144.6±6.9	0.21±0.04	0.08±0.02	149.1±7.4	0.22±0.06	0.10±0.05
FSDM [38]	127.6±10.1	0.37±0.07	0.27±0.06	66.9±5.2	0.35±0.03	0.30±0.03	111.6±6.6	0.47±0.04	0.22±0.03
LDM (LoRA) [47]	100.9±6.3	0.28±0.02	0.31±0.03	89.4±4.5	0.23±0.03	0.40±0.05	105.2±2.9	0.40±0.03	0.22±0.04
LDM (Full FT) [4]	100.8±2.2	0.21±0.04	0.30±0.03	87.5±8.8	0.22±0.06	0.35±0.03	101.3±1.3	0.43±0.05	0.25±0.06
LDM (SeLoRA) [51]	89.5±4.3	0.34±0.03	0.51±0.04	69.6±4.0	0.22±0.02	0.58±0.04	98.5±5.9	0.41±0.07	0.52±0.02
DiT (Full FT) [35]	85.3±4.1	0.48±0.03	0.46±0.02	48.2±3.9	0.42±0.05	0.46±0.04	81.1±4.5	0.53±0.04	0.55±0.05
DiffFit [39]	85.5±5.4	0.41±0.05	0.49±0.05	48.0±3.2	0.34±0.04	0.60±0.04	82.9±3.6	0.47±0.02	0.58±0.02
AnatoDiff (Ours)	73.1±3.0	0.54±0.01	0.54±0.03	40.6±3.8	0.43±0.02	0.61±0.03	70.7±3.1	0.54±0.05	0.59±0.05
Δ	-14.5%	+13%	+5%	-15.4%	+9%	+1%	-14.7%	+7%	+1%

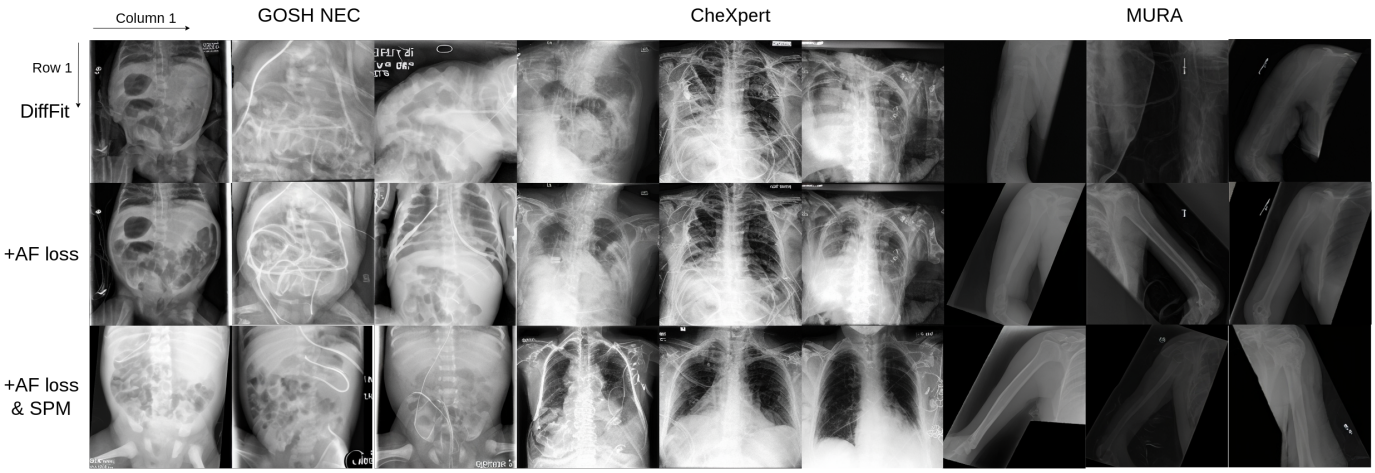

Fig. 4. Qualitative comparisons of the effects of AF loss and SPM on image generation. GOSH NEC (columns 1-3), CheXpert (columns 4-6), MURA (columns 7-9). (Top) Images using DiffFit. (Middle) Images with AF loss. (Bottom) Images with both AF loss and SPM. Using DiffFit alone results in suboptimal outcomes, with images often exhibiting distorted or unrecognizable shapes and mixing of anatomical structures. By integrating AF loss and SPM, images from the same noise achieve more realistic anatomical shapes and improved fidelity.

include MedGAN [49], MT-DDPM [50], and LDM [4] using both full fine-tuning (FT) and parameter-efficient approaches (BitFit [46], LoRA [47] and SeLoRA [51].

Table I presents the performance comparisons of the evaluated models across all three datasets. Diffusion-based model consistently outperformed GAN-based models. In particular, GAN-based models exhibited significantly lower IR, which can be attributed to model collapse under limited training samples. DiffFit demonstrated superior performance on all three datasets. Compared to its fully fine-tuned counterpart DiT, DiffFit demonstrates that updating only a small subset of parameters can achieve performance on par with or even better than full fine-tuning, aligning with findings in [57], [58].

Our proposed AnatoDiff outperforms fully fine-tuning DiT, along with medical image generation models including MT-DDPM and SeLoRA. Moreover, AnatoDiff showed noticeable improvements over DiffFit across all datasets, with average enhancements of 14.9% in FID, 9.7% in IP, and 2.3% in IR, indicating robust capability in data-limited scenarios.

B. Ablation study

Fig. 4 shows a modular comparison of generated images from the same noise under different ablation settings. In row 1, images generated by DiffFit exhibit defects in general object shape (column 1), a mixture of organs (columns 3, 4, 6), and incorrect structures (columns 2, 8). By including AF loss, images in row 2 show improvements in both shape and anatomy. Specifically, AF loss corrects shape defects in column 1, resolves the mixture of organs in columns 3, 4, and 6, and produces correct structures in columns 2 and 8. However, artifacts such as excessive medical lines/tubes in columns 2 and 5, and minor shape defects in columns 1, 4, and 6 still persist. SPM further supports accurate generation of object shapes by using clustered training images as prototypes. The bottom row shows that images generated with both AF loss and SPM exhibit the most anatomically accurate and structurally correct representations, with noticeably fewer artifacts.

Quantitative results is listed in Table II. The effectiveness of the AF loss is demonstrated by a 6.3% improvement in FID, 6.0% in IP, and 2.0% in IR. Incorporating the SPM leads to an additional 9.8% improvement in FID and 3.0% in IP.

TABLE II
ABLATION ON PROPOSED MODULES

	FID↓	IP↑	IR↑
DiffFit	48.0±3.2	0.34±0.04	0.60±0.04
+ Anatomical Fidelity Loss	45.0±3.7	0.40±0.03	0.62±0.03
+ Shape Prototype Module	40.6±3.8	0.43±0.02	0.61±0.03

TABLE III
ABLATION RESULTS ON SPM PARAMETERS. (TOP) LAST N BLOCKS. (BOTTOM) NUMBER OF PROTOTYPES K.

Last N Blocks	FID↓	IP↑	IR↑
2	42.8±2.0	0.39±0.05	**0.64±0.02**
5	42.7±2.0	0.39±0.04	0.63±0.05
7	**40.6±3.8**	**0.43±0.02**	0.61±0.03
9	43.0±3.9	0.40±0.01	0.62±0.03
Number of Prototypes K			
15	44.9±2.7	0.33±0.02	**0.68±0.04**
20	42.8±2.3	**0.43±0.02**	0.60±0.03
25	**40.6±3.8**	**0.43±0.02**	0.61±0.03
30	41.6±1.3	0.39±0.03	0.66±0.02
35	42.9±1.5	0.37±0.02	0.62±0.04
40	42.0±4.0	0.40±0.05	0.60±0.06

TABLE IV
ABLATION ON NUMBER OF CHANNELS CONTRIBUTING TO AF LOSS

No. of channel	FID↓	IP↑	IR↑	Train Steps/Sec↑
1	44.4±2.1	0.40±0.01	0.61±0.04	∼0.93
2	45.9±3.0	0.39±0.02	0.59±0.04	∼0.66
3	45.9±3.0	0.39±0.02	0.61±0.04	∼0.56
4	42.1±3.3	0.43±0.02	0.60±0.03	∼0.51
Average All	43.9±1.9	0.39±0.02	0.59±0.02	∼0.92
Random				
H_0 only	49.8±3.3	0.33±0.02	0.60±0.03	∼1.00
H_1 only	46.9±3.4	0.37±0.02	0.61±0.05	∼0.98
H_0 & H_1	40.6±3.8	0.43±0.02	0.61±0.03	∼0.92

TABLE V
SENSITIVITY ANALYSIS ON λ IN AF LOSS

λ	FID	IP	IR
2×10^{-5}	53.0±4.1	0.31±0.03	0.46±0.02
1×10^{-6}	42.0±2.9	0.42±0.04	0.63±0.05
2×10^{-6}	40.6±3.8	0.43±0.02	0.61±0.03
3×10^{-6}	41.2±2.8	0.42±0.03	0.63±0.03
2×10^{-7}	48.6±5.3	0.36±0.03	0.63±0.03

Table III details the hyperparameter ablation study on the last N blocks and number of prototypes K in the SPM. Increasing K initially improves generation quality. However, beyond $K = 25$, further increases yield diminishing returns, suggesting excessive variability does not necessarily enhance performance. To achieve high fidelity, we select 7 for N and 25 for K based on their FID and IP performance.

Table IV examines the effects of the number of latent channels used in the AF loss. Using fewer channels offers faster training times but degrades FID and IP. Although using all four channels yields the best results, it incurs high computational overhead. Instead, we randomly select a single channel for each latent to compute the AF loss. Interestingly, this approach not only accelerates training time but also improves FID.

Table V assesses the sensitivity of λ. We evaluate a range of values centered around the default setting (2×10^{-6}), including variations one order of magnitude smaller and larger. Generation performance remains stable for small perturbations around the default value (1×10^{-6} and 3×10^{-6}), reflected by consistent IP and IR scores, with only slight variations in FID. However, performance noticeably degrades for more extreme deviations. Specifically, smaller values by one order of magnitude (2×10^{-7}) lead to higher FID and lower IP, while larger values by an order of magnitude (2×10^{-5}) cause significant deterioration across all metrics, including IR. These results indicate that the model is robust to minor adjustments in λ, with a clear optimal operational range. However, more aggressive deviations can adversely affect generation quality.

C. Qualitative results

We present generated images from various methods alongside examples from the training set in Fig. 5. Notably, DC-GAN fails to generate meaningful images from the GOSH NEC dataset, indicating the complexity of neonatal abdomen structure under limited-data. Few-shot and limited-data GAN approaches such as MoCA, CDC, and RICK improve the quality relative to DC-GAN but still produce significantly flawed images. The generated abdomen and chest shapes are distorted, and bone shapes are unrecognizable, as seen in rows 2-4. Furthermore, the scarcity of training images and the complexity of radiographs cause GAN methods to produce a limited variety of outputs and frequently collapses. Diffusion models like LDM, FSDM and DiffFit generate higher quality images compared to GANs, but key challenges remain in achieving truthful shapes and correct anatomy. This issue is particularly noticeable in the MURA dataset (columns 8-10). Although DiffFit produces the highest quality images among diffusion models, they still contain anatomical inconsistencies — the intestinal and chest areas in columns 2 and 3 are inaccurately represented, and the humerus shape is incorrect in the last 2 columns. Our proposed AnatoDiff (2nd last row), consistently generates accurate images with anatomically correct structures and shapes across all three datasets.

D. Image Memorization Study

Memorization is a well-known issue in diffusion models trained on limited data, where the model tends to replicate its training data. To investigate this, we follow the methodology described in [59] to identify the closest training images.

The top two rows of Fig. 6 illustrate examples from directly fine-tuning DiT, where the generated images are nearly identical to the training image, differing only by being unmodified or horizontally flipped. With the addition of DiffFit, the generated images introduce novel elements and avoid the direct copying effect observed in fully fine-tuned DiT.

Building on this, we incorporated our novel AF loss and SPM to further improve image generation quality. These modules impose additional constraints during training (topological constraints from the AF loss and general shape constraints from the SPM), therefore, it is important to verify that they do

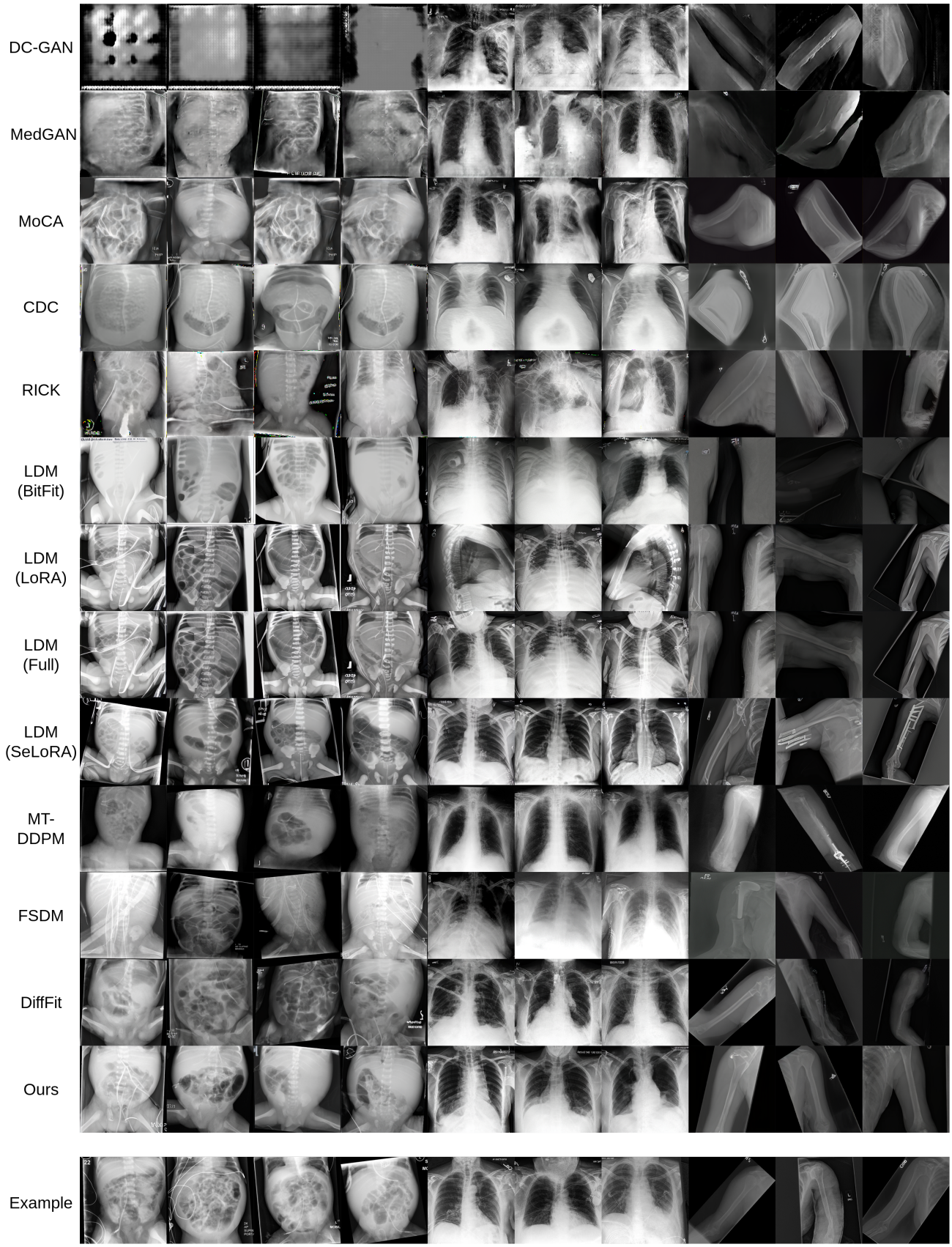

Fig. 5. Qualitative comparisons of images generated from different methods alongside samples from the training set. GOSH NEC (columns 1-4), CheXpert (columns 5-7), MURA (columns 8-10). GAN-based models (MoCA, CDC, RICK) generate images with unsatisfactory quality, often producing very similar images due to mode collapse. Diffusion-based methods (LDM, FSDM, DiffFit) achieve higher quality but still suffer from untruthful shapes and inaccurate anatomy. Our proposed AnatoDiff addresses these issues with AF loss and SPM, enabling the generation of high-quality images while maintaining truthful shapes and anatomical accuracy.

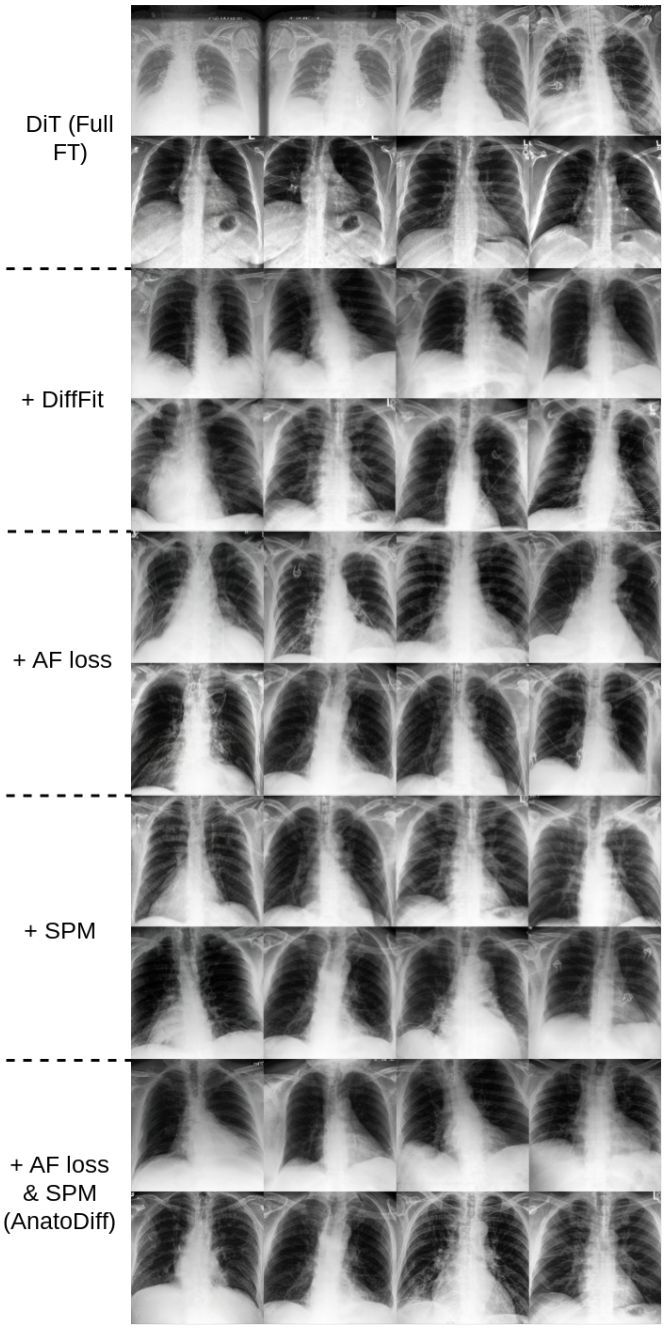

Fig. 6. Comparison of images generated by DiT and DiffFit with different combinations of the proposed modules (first column) and the three most similar training images from the CheXpert dataset (next columns).

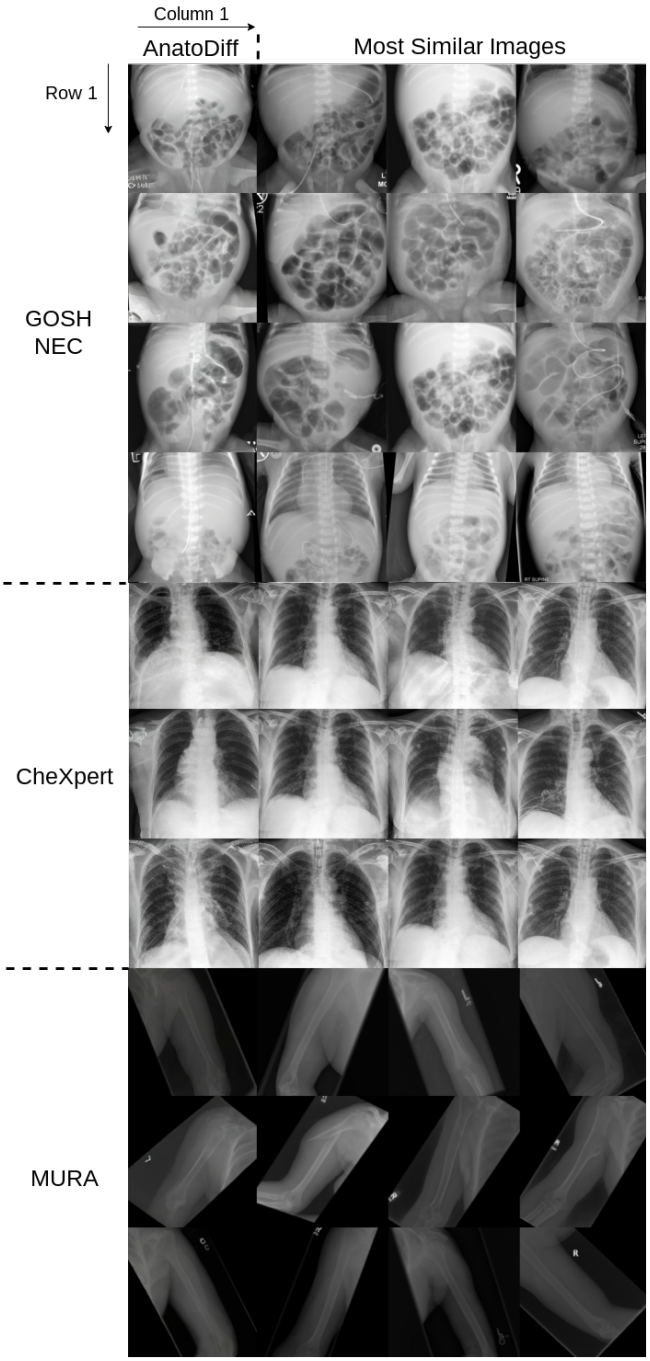

Fig. 7. Comparison of images generated by AnatoDiff (column 1) with the top three most similar images from the training set (columns 2-4). GOSH NEC (rows 1-4), CheXpert (rows 5-7), MURA (rows 8-10). AnatoDiff is able to synthesize realistic images from limited data without replicating training samples.

not reintroduce memorization effects. As shown in the bottom rows of Fig. 6, generations with either or both modules preserve novelty while enhancing structural fidelity. This analysis confirms that the proposed AF loss and SPM improve image quality without increasing memorization. Further examples from other datasets are shown in Fig. 7.

E. Prototype Visualization and Stability Analysis

Fig. 8 displays examples of prototypes from all three datasets, each representing a clustered common shape characteristic of the dataset (abdomen, chest, and arm). The prototypes serve as soft templates, encouraging the model to maintain anatomical structure without sacrificing generative diversity. Specifically, they guide the model to learn appropriate spatial relationships and morphological patterns without over-constraining fine-grained details. By ensuring prototypes only capture the coarse silhouette or structural cues of the anatomy without encoding fine details, the model maintains the balance between abstraction and informativeness, ensuring that generation remains flexible and not overly constrained.

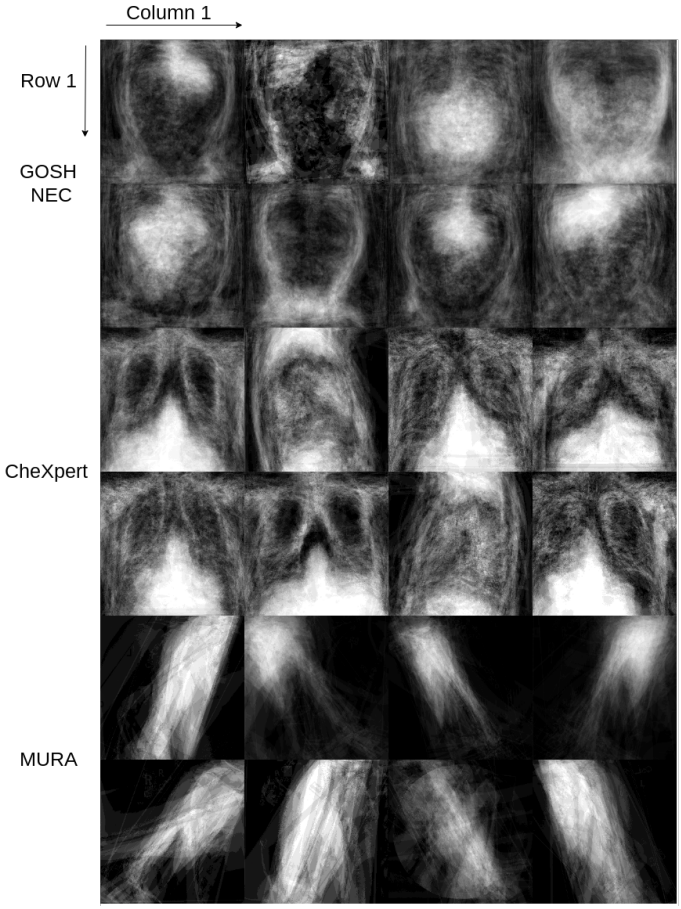

Fig. 8. Visualization of prototypes in SPM for GOSH NEC (rows 1-2), CheXpert (rows 3-4), MURA (rows 5-6). These prototypes capture the general outline shape of each anatomical structure, providing structural and anatomical guidance to model during image generation. Structures visibly defined include shape of thorax and abdomen, cardiac and liver silhouettes, outline of ribs, and lungs, and outline of humerus

For example, in GOSH NEC, the prototype captures rounded abdominal contours and approximate organ regions, which reduces pathological blending of thoracic and abdominal features. In CheXpert, the prototype encodes asymmetrical thoracic features (e.g., cardiac silhouette offset), which helps the model avoid symmetric artifacts like duplicated hearts. In MURA, the arm prototype promotes continuity and orientation, reducing the frequency of disjointed or mirrored limb artifacts. These prototypes provide a general anatomical structure, serving as foundational guides to ensure that the generated images adhere to realistic anatomical configurations while allowing for variability within those constraints.

To confirm the stability of the prototype clustering stage, we conduct three experiments: (1) re-running K-means with different random seeds while keeping model training and generation seeds fixed; (2) varying binarization thresholds, including a no-binarization setting; and (3) replacing K-means with an alternative method—Density Peak Clustering.

Results in Table VI show minor performance variations across different seeds (FID ranging from 40.2 to 41.0), consistently outperforming the DiffFit baseline (FID = 48.0). Similarly stable results are observed for IP and IR, confirming clustering process's robustness to initialization randomness.

TABLE VI
STABILITY ANALYSIS WITH DIFFERENT RANDOM SEEDS DURING K-MEANS CLUSTERING

Seed	FID↓	IP↑	IR↑
0 (default)	40.6±3.8	0.43±0.02	0.61±0.03
1	40.5±4.3	0.45±0.02	0.63±0.03
2	40.2±3.7	0.44±0.03	0.62±0.03
3	41.0±2.6	0.45±0.02	0.62±0.04

TABLE VII
STABILITY ANALYSIS ON BINARIZATION THRESHOLD VALUES

Threshold Value	FID↓	IP↑	IR↑
No binarization	44.1±2.5	0.42±0.04	0.62±0.04
180	46.0±2.5	0.39±0.06	0.61±0.06
160	42.8±3.4	0.42±0.05	0.62±0.04
150	40.6±3.8	0.43±0.02	0.61±0.03
140	41.6±2.6	0.42±0.04	0.63±0.06
120	42.6±2.6	0.42±0.04	0.63±0.06
100	43.1±4.9	0.40±0.04	0.64±0.05

TABLE VIII
COMPARISON WITH ALTERNATIVE CLUSTERING METHOD

Method	FID↓	IP↑	IR↑
K-means	40.6±3.8	0.43±0.02	0.61±0.03
Density Peak Clustering	40.7±3.7	0.44±0.04	0.63±0.03

Table VII shows a stability analysis of thresholding value during binarization. Performance remains stable within the threshold range of 120–160, with optimal performance at 150. Threshold values above 180 lead to degraded prototype quality and instability in the clustering outcome, excessively removing crucial anatomical outlines, as shown in Fig. 9.

The primary purpose of prototypes is to provide general structural outlines to be used by the model as a starting point, rather than detailed informative examples. As shown in Fig. 9, prototypes without binarization can lack clear structural delineation, resulting in lower overall generation performance. Moreover, we noticed the binarization does not greatly alter the overall structure of the clustered result, which can be seen in the comparison between column 1 and 2. The clustered outcome with the addition of binarization only results in a sparser image with more visible outlines, while the overall structure and shape in unchanged. Quantitatively, we found that performing binarization at thresholds below 180 consistently outperforms no binarization in terms of FID, binarization thus effectively removes low-activation noise, enforcing a sparse prior, as can be seen in Table VII and Fig. 9.

Performance comparison between K-means and Density Peak Clustering shows minimal differences, as shown in Table VIII, indicating our framework's robustness to clustering methodology. This validates that the generation approach is not critically dependent on the choice of clustering algorithm.

F. Synthetic images for classification

To assess the performance of generated images in a downstream task, we perform a comparison using a classification

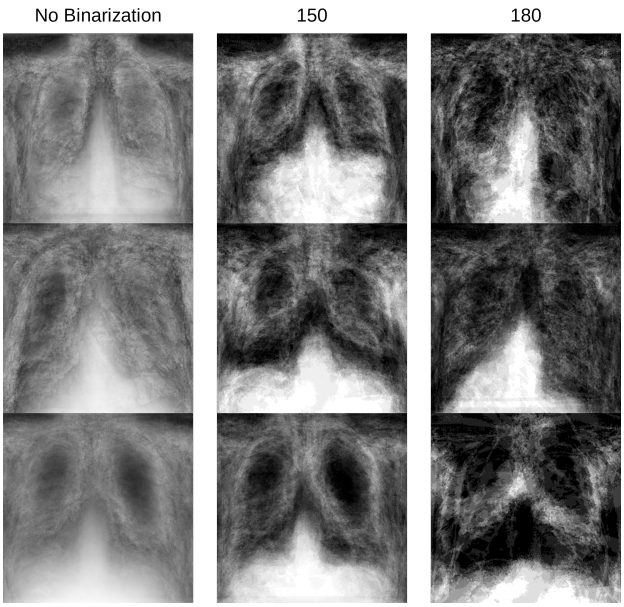

Fig. 9. Examples of prototypes generated with different binarization thresholds. Thresholding enhances structural outline clarity while preserving overall shape. However, excessively high thresholds can introduce clustering instability and lead to loss of anatomical delineation.

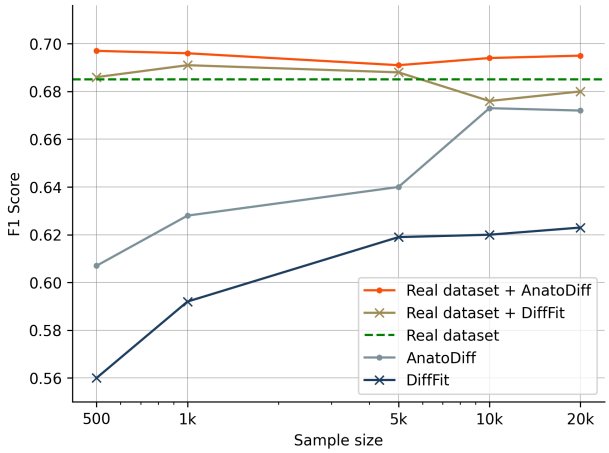

Fig. 10. Comparison of GOSH NEC classification performance using images generated by AnatoDiff and DiffFit. Compared to DiffFit, AnatoDiff generated images consistently improves F1 scores across all dataset sizes. Augmenting real data with AnatoDiff-generated images further enhances performance compared to using only real data and augmenting with DiffFit-generated images.

task with a ResNet-50 model. We generate 20,000 and 10,000 class-conditioned synthetic images using both AnatoDiff and DiffFit on the GOSH NEC dataset. From these, we sample subsets of 5,000, 1,000, and 500 images, and train the model five times per subset using different random initializations. For the 1,000 and 500 subsets, we ensure each sampling is performed with a different seed to avoid bias in the results. The trained models are evaluated on an isolated test set not used during the training of either the generation or classification models. Consequently, AnatoDiff is not trained to generate images from the test patients used in the classification stage.

We compare the classification performance trained on synthetic images, real images, and a combination of both. In all settings, the number of real training samples is fixed at 1,000, while only the number of generated samples varies. Importantly, both the diffusion model and the classifier are trained on the same set of real images, ensuring no data leakage and preventing any advantage that could arise if the diffusion model had access to additional unseen real samples. Thus, any improvement reflects AnatoDiff's ability to generate meaningful variations of a limited dataset. During classification training, the generated images are added to the real samples, thereby expanding the training set.

Fig. 10 illustrates this comparison. Model trained with AnatoDiff-generated images demonstrate superior performance over DiffFit across all subsets, with improvement of 5.3% F1 at 10,000 images and achieving performance comparable to models trained on real images, with a 1.2% F1 difference. When AnatoDiff-generated images are added to the real dataset, classification performance consistently surpasses that achieved with the DiffFit counterpart. Notably, adding more than 5,000 DiffFit-generated images leads to a decline in accuracy, which we hypothesize is due to low-quality synthetic samples overpowering the real, higher-quality data as their proportion increases. In contrast, AnatoDiff-generated images consistently improve performance, suggesting they are more informative and maintain greater anatomical coherence.

G. Expert Evaluation Study

As a preliminary investigation to assess the technical feasibility and visual plausibility of the generated images, a survey is performed on GOSH NEC images created using AnatoDiff. We generated 10 synthetic and randomly selected 10 real images from the GOSH NEC dataset. Images are shuffled and medical professionals are asked to decide if the shown image is real or generated. Eight consultant Paediatric Surgeons, one junior Paediatric Surgeon and one consultant Neonatologist participated in the survey. Results shows a Matthews correlation coefficient (MCC) of 0.277 and Fleiss' Kappa of 0.126 among participants, indicating responses are only slightly better than random guessing with a low level of consensus among raters. The low MCC score and Fleiss' Kappa implies AnatoDiff-generated images are difficult to differentiate from real images in the eyes of medical professionals. As part of future work, we plan to expand this into a larger-scale expert study involving a greater number of images and a broader panel of evaluators, enabling a more robust assessment of clinical relevance and informing further refinements—such as artifact-aware training objectives or post-generation filtering.

H. Limitations and Error Analysis

While AnatoDiff promotes anatomically plausible outputs and substantially reduces structural inconsistencies, in some instances, generated images may still appear slightly more regular or idealized than real X-rays. Fig. 11 presents such examples, where certain images appear smoother, with fewer of the subtle variations typically observed in clinical data. For example, the rib arcs in the chest X-ray are more uniform,

Fig. 11. Example of a failure case where the generated image looks overly smooth.

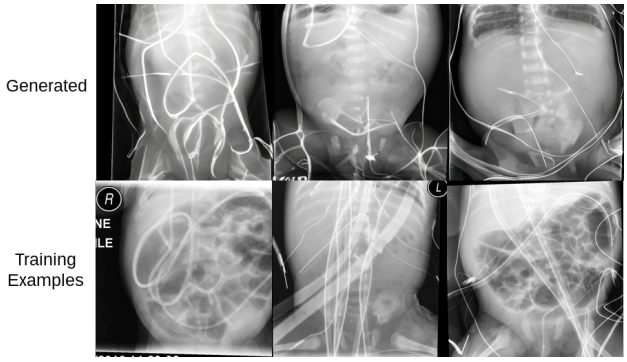

Fig. 12. A specific case of failure in the GOSH NEC dataset, where a generated image can occasionally be misled by the presence of excessive artifacts contained within some training images. (Top) Generated images. (Bottom) Training examples.

whereas in reality there are natural differences in the size, shape, and structure of some ribs. Similarly, the clavicles exhibit a slightly atypical structure and positioning, and in the humerus X-ray, the fine trabecular bone texture is less pronounced. The abdomen X-ray shows more homogeneous opacification compared to the expected radiolucencies from air in the gastrointestinal tract. These observations likely reflect the tendency of diffusion models trained on limited data to prioritise consistent global structure, while capturing fewer high-frequency details. Incorporating an edge-based loss, could further enrich local detail and visual realism. In future work, we aim to investigate edge-based objectives operating in the latent space that are compatible with our AF loss.

Another representative failure case, shown in Fig. 12, arises specifically from the GOSH NEC dataset. This dataset contains X-rays taken using portable scanners in the ICU setting, performed on neonates. As a result, the images frequently contain external objects such as tubes or monitoring devices affixed to the patient, which appear as linear artifacts. While these artifacts do not compromise the diagnostic validity of the X-rays, they may be inadvertently treated as meaningful anatomical features during training. Consequently, the model occasionally reproduces such artifacts in generated images, having learned them as part of the training distribution. For downstream clinical applications, such as generating de-identified training materials for medical education, manual screening can be employed to ensure that only high-quality samples are retained.

VI. Conclusion

We presented AnatoDiff, a diffusion method for generating high-quality X-ray images that accurately represent anatomical shapes and structures, even when trained on limited (500-1,000) images. AnatoDiff incorporates an SPM to guide the generation of anatomically accurate images, and an AF loss to regularize topological structures. AnatoDiff significantly outperforms the SOTA methods on three open-source X-ray datasets, achieving an average improvement of 14.9% in FID, 9.7% in IP, and 2.3% in IR, while consistently producing anatomically truthful images with accurate shape.

In downstream classification, a ResNet-50 trained on AnatoDiff images shows improved performance compared to training on SOTA DiffFit images, with F1-score increases of 2.1% to 5.3% across 500 to 10,000 samples. Furthermore, a survey involving medical professionals revealed that AnatoDiff-generated images are challenging to distinguish from real images, with an MCC of 0.277 and Fleiss' Kappa of 0.126. This underscores AnatoDiff's effectiveness in generating high-quality, anatomically accurate images representative of X-rays, highlighting its significant potential for practical applications.

As future work, we aim to investigate the potential synergy between the AF loss and edge-based losses, with the latter offering a means to recover fine local details that may not be fully captured by the former. Moreover, although this study focuses on 2D radiographs, the proposed framework could naturally extend to 3D modalities such as CT and MRI. This extension could be realized either by applying the AF loss across individual slices with aggregated results, or by generalizing both the AF loss and SPM prototypes to operate directly in 3D. These directions will be pursued in future work.

Acknowledgment

The authors thank survey participants, Dr Susan Shelmerdine and Dr Andrea Tomaselli for assisting survey design.

References

[1] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2020.
[2] P. Dhariwal and A. Q. Nichol, "Diffusion models beat gans on image synthesis," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, pp. 8780–8794, 2021.
[3] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.
[4] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 10674–10685, 2022.
[5] C. Saharia *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2022.
[6] A. Q. Nichol *et al.*, "GLIDE: towards photorealistic image generation and editing with text-guided diffusion models," in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 162, pp. 16784–16804, 2022.
[7] J. Ho *et al.*, "Imagen video: High definition video generation with diffusion models," *ArXiv*, vol. abs/2210.02303, 2022.
[8] J. Ho, T. Salimans, A. A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video diffusion models," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2022.
[9] A. Blattmann *et al.*, "Align your latents: High-resolution video synthesis with latent diffusion models," *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 22563–22575, 2023.
[10] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "Dreamfusion: Text-to-3d using 2d diffusion," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2023.
[11] H. Wang, X. Du, J. Li, R. A. Yeh, and G. Shakhnarovich, "Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation," *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 12619–12629, 2022.

[12] X. Zeng *et al.*, "LION: latent point diffusion models for 3d shape generation," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2022.

[13] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *ArXiv*, vol. abs/2204.06125, 2022.

[14] J. Yoon, L. N. Drumright, and M. van der Schaar, "Anonymization through data synthesis using generative adversarial networks (ads-gan)," *IEEE J. Biomed. Health. Inf.*, vol. 24, pp. 2378–2388, 2020.

[15] A. Bissoto, E. Valle, and S. Avila, "Gan-based data augmentation and anonymization for skin-lesion analysis: A critical review," *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshop (CVPRW)*, pp. 1847–1856, 2021.

[16] H. Shin *et al.*, "Medical image synthesis for data augmentation and anonymization using generative adversarial networks," in *Proc. Workshop at Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAIW)*, vol. 11037, pp. 1–11, 2018.

[17] I. E. Hamamci *et al.*, "Diffusion-based hierarchical multi-label object detection to analyze panoramic dental x-rays," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, vol. 14225, pp. 389–399, 2023.

[18] J. Rousseau, C. Alaka, E. Covili, H. Mayard, L. Misrachi, and W. Au, "Pre-training with diffusion models for dental radiography segmentation," in *Proc. Workshop at Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAIW)*, vol. 14533, pp. 174–182, 2023.

[19] T. Weber, M. Ingrisch, B. Bischl, and D. Rügamer, "Cascaded latent diffusion models for high-resolution chest x-ray synthesis," in *Proc. Pacific-Asia Conf. Knowl. Discov. Data Min. (PAKDD)*, vol. 13937, pp. 180–191, 2023.

[20] G. Müller-Franzes *et al.*, "Diffusion probabilistic models beat gans on medical images," *ArXiv*, vol. abs/2212.07501, 2022.

[21] S. Pan *et al.*, "2d medical image synthesis using transformer-based denoising diffusion probabilistic model," *Phys. Med. Biol.*, vol. 68, 2023.

[22] A. U. R. Hashmi *et al.*, "Xreal: Realistic anatomy and pathology-aware x-ray generation via controllable diffusion model," *ArXiv*, vol. abs/2403.09240, 2024.

[23] M. Aversa *et al.*, "Diffinfinite: Large mask-image synthesis via parallel random patch diffusion in histopathology," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2023.

[24] Z. Liu, Y. He, Y. Zhao, Y. Feng, and G. Zhang, "Overcoming pathology image data deficiency: Generating images from pathological transformation process," *ArXiv*, vol. abs/2311.12316, 2023.

[25] X. Xu, S. Kapse, R. Gupta, and P. Prasanna, "Vit-dae: Transformer-driven diffusion autoencoder for histopathology image analysis," in *Proc. Workshop at Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAIW)*, vol. 14533, pp. 66–76, 2023.

[26] S. Go, Y. Ji, S. J. Park, and S. Lee, "Generation of structurally realistic retinal fundus images with diffusion models," *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 2335–2344, 2024.

[27] D. Hu, Y. K. Tao, and I. Oguz, "Unsupervised denoising of retinal oct with diffusion probabilistic model," in *Med. Imaging 2022: Image Proc.*, 2022.

[28] J. A. Irvin *et al.*, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proc. AAAI Conf on Artif. Intell (AAAI)*, 2019.

[29] A. E. W. Johnson *et al.*, "Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports," *Sci. Data*, vol. 6, no. 317, 2019.

[30] P. Chambon *et al.*, "Roentgen: Vision-language foundation model for chest x-ray generation," *ArXiv*, vol. abs/2211.12737, 2022.

[31] K. Packhäuser, L. Folle, F. Thamm, and A. Maier, "Generation of anonymous chest radiographs using latent diffusion models for training thoracic abnormality classification systems," in *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5, IEEE, 2023.

[32] P. Chambon, C. Bluethgen, C. P. Langlotz, and A. Chaudhari, "Adapting pretrained vision-language foundational models to medical imaging domains," *arXiv preprint arXiv:2210.04133*, 2022.

[33] T. Weber, M. Ingrisch, B. Bischl, and D. Rügamer, "Cascaded latent diffusion models for high-resolution chest x-ray synthesis," in *Pacific-Asia conference on knowledge discovery and data mining*, pp. 180–191, Springer, 2023.

[34] D. Kyung, J. Kim, T. Kim, and E. Choi, "Towards predicting temporal changes in a patient's chest x-ray images based on electronic health records," *arXiv preprint arXiv:2409.07012*, 2024.

[35] W. S. Peebles and S. Xie, "Scalable diffusion models with transformers," *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pp. 4172–4182, 2022.

[36] P. Rajpurkar *et al.*, "Mura: Large dataset for abnormality detection in musculoskeletal radiographs.," *ArXiv*, vol. abs/1712.06957, 2017.

[37] K.-W. Yung, J. Sivaraj, P. D. Coppi, D. Stoyanov, S. Loukogeorgakis, and E. B. Mazomenos, "Diagnosing necrotising enterocolitis via fine-grained visual classification," *IEEE Trans. Biomed. Eng.*, 2024.

[38] G. Giannone, D. Nielsen, and O. Winther, "Few-shot diffusion models," *ArXiv*, vol. abs/2205.15463, 2022.

[39] E. Xie *et al.*, "Difffit: Unlocking transferability of large diffusion models via simple parameter-efficient fine-tuning," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pp. 4207–4216, 2023.

[40] K. Packhauser, L. Folle, F. Thamm, and A. Maier, "Generation of anonymous chest radiographs using latent diffusion models for training thoracic abnormality classification systems," *Proc. IEEE Int. Symp. Biomed. Imag. (ISBI)*, pp. 1–5, 2022.

[41] H. Ali, S. Murad, and Z. Shah, "Spot the fake lungs: Generating synthetic medical images using neural diffusion models," in *Irish Conference on Artificial Intelligence and Cognitive Science*, pp. 32–39, Springer, 2022.

[42] U. Ojha, Y. Li, J. Lu, *et al.*, "Few-shot image generation via cross-domain correspondence," *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 10738–10747, 2021.

[43] T. Li, Z. Li, A. Luo, A. B. Farimani, and T. S. Lee, "Prototype memory and attention mechanisms for few shot image generation," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022.

[44] Y. Zhao *et al.*, "Exploring incompatible knowledge transfer in few-shot image generation," *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 7380–7391, 2023.

[45] J. Zhu, H. Ma, J. Chen, and J. Yuan, "Few-shot image generation with diffusion models," *ArXiv*, vol. abs/2211.03264, 2022.

[46] E. B. Zaken, Y. Goldberg, and S. Ravfogel, "Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models," in *Proc. Assoc. Comput. Linguist. (Proc. Assoc. Comput. Linguist. (ACL))*, pp. 1–9, 2022.

[47] E. J. Hu *et al.*, "LoRA: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022.

[48] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," tech. rep., Stanford, 2006.

[49] K. Guo, J. Chen, T. Qiu, S. Guo, T. Luo, T. Chen, and S. Ren, "Medgan: An adaptive gan approach for medical image generation," *Computers in Biology and Medicine*, vol. 163, p. 107119, 2023.

[50] S. Pan, T. Wang, R. L. Qiu, M. Axente, C.-W. Chang, J. Peng, A. B. Patel, J. Shelton, S. A. Patel, J. Roper, *et al.*, "2d medical image synthesis using transformer-based denoising diffusion probabilistic model," *Physics in Medicine & Biology*, vol. 68, no. 10, p. 105004, 2023.

[51] Y. Mao, H. Li, W. Pang, G. Papanastasiou, G. Yang, and C. Wang, "Selora: Self-expanding low-rank adaptation of latent diffusion model for medical image synthesis," *arXiv preprint arXiv:2408.07196*, 2024.

[52] C. Battersby, T. Santhalingam, K. Costeloe, and N. Modi, "Incidence of neonatal necrotising enterocolitis in high-income countries: a systematic review," *Arch. Dis. Child Fetal Neonatal. Ed.*, vol. 103, no. 2, pp. F182–F189, 2018.

[53] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings* (Y. Bengio and Y. LeCun, eds.), 2016.

[54] H. Salehinejad, S. Valaee, T. Dowdell, E. Colak, and J. Barfett, "Generalization of deep neural networks for chest pathology classification in x-rays using generative adversarial networks," in *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 990–994, 2018.

[55] A. Madani, M. Moradi, A. Karargyris, and T. Syeda-Mahmood, "Semi-supervised learning with generative adversarial networks for chest x-ray classification with ability of data domain adaptation," in *Proc. IEEE Int. Symp. Biomed. Imag. (ISBI)*, pp. 1038–1042, 2018.

[56] M. J. M. Chuquicusma, S. Hussein, J. Burt, and U. Bagci, "How to fool radiologists with generative adversarial networks? a visual turing test for lung cancer diagnosis," in *Proc. IEEE Int. Symp. Biomed. Imag. (ISBI)*, pp. 240–244, 2018.

[57] E. Xie, L. Yao, H. Shi, Z. Liu, D. Zhou, Z. Liu, J. Li, and Z. Li, "Difffit: Unlocking transferability of large diffusion models via simple parameter-efficient fine-tuning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4230–4239, 2023.

[58] E. B. Zaken, S. Ravfogel, and Y. Goldberg, "Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models," *arXiv preprint arXiv:2106.10199*, 2021.

[59] G. Somepalli, V. Singla, M. Goldblum, J. Geiping, and T. Goldstein, "Diffusion art or digital forgery? investigating data replication in diffusion models," *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 6048–6058, 2022.