Physics in Medicine &
Biology



PAPER

# Guided synthesis of annotated lung CT images with pathologies using a multi-conditioned denoising diffusion probabilistic model (mDDPM)

To cite this article: Arjun Krishna *et al* 2025 *Phys. Med. Biol.* **70** 065007





IOP Publishing Phys. Med. Biol. 70 (2025) 065007 https://doi.org/10.1088/1361-6560/adb9b3

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

PAPER

RECEIVED
27 September 2024

REVISED
13 February 2025

ACCEPTED FOR PUBLICATION
24 February 2025

PUBLISHED
6 March 2025

# Guided synthesis of annotated lung CT images with pathologies using a multi-conditioned denoising diffusion probabilistic model (mDDPM)

**Arjun Krishna**[1,*], **Ge Wang**[2] **and Klaus Mueller**[1]

[1] Computer Science Department, Stony Brook University, Stony Brook, NY 11794, United States of America
[2] Biomedical Imaging Center, Rensselaer Polytechnic Institute, Troy, NY 12180, United States of America
[*] Author to whom any correspondence should be addressed.

**E-mail:** arjkrishna@cs.stonybrook.edu, wangg6@rpi.edu and mueller@cs.stonybrook.edu

**Keywords:** diffusion models, generative AI, synthetic pathology, lung-CT, annotated image synthesis, mammography

## Abstract

*Objective*. The training of AI models for medical image diagnostics requires highly accurate, diverse, and large training datasets with annotations and pathologies. Unfortunately, due to privacy and other constraints the amount of medical image data available for AI training remains limited, and this scarcity is exacerbated by the high overhead required for annotation. We address this challenge by introducing a new controlled framework for the generation of synthetic images complete with annotations, incorporating multiple conditional specifications as inputs. *Approach*. Using lung CT as a case study, we employ a denoising diffusion probabilistic model to train an unconditional large-scale generative model. We extend this with a classifier-free sampling strategy to develop a robust generation framework. This approach enables the generation of constrained and annotated lung CT images that accurately depict anatomy, successfully deceiving experts into perceiving them as real. Most notably, we demonstrate the generalizability of our multi-conditioned sampling approach by producing images with specific pathologies, such as lung nodules at designated locations, within the constrained anatomy. *Main results*. Our experiments reveal that our proposed approach can effectively produce constrained, annotated and diverse lung CT images that maintain anatomical consistency and fidelity, even for annotations not present in the training datasets. Moreover, our results highlight the superior performance of controlled generative frameworks of this nature compared to nearly every state-of-the-art image generative model when trained on comparable large medical datasets. Finally, we highlight how our approach can be extended to other medical imaging domains, further underscoring the versatility of our method. *Significance*. The significance of our work lies in its robust approach for generating synthetic images with annotations, facilitating the creation of highly accurate and diverse training datasets for AI applications and its wider applicability to other imaging modalities in medical domains. Our demonstrated capability to faithfully represent anatomy and pathology in generated medical images holds significant potential for various medical imaging applications, with high promise to lead to improved diagnostic accuracy and patient care.

## 1. Introduction

Data scarcity and imbalance present significant challenges to developing robust AI models for medical imaging. Unlike large-scale natural image collections such as ImageNet, medical imaging datasets are typically much smaller, and privacy regulations often limit data sharing or public release. These datasets also frequently suffer from severe imbalances, with rare pathologies underrepresented compared to common conditions or healthy cases. The lack of high-resolution images and detailed annotations further exacerbates

these challenges. All these shortcomings complicate the training and evaluation of AI models and increases the risk of systemic performance biases (Obermeyer *et al* 2019, Oakden-Rayner *et al* 2020),

To address these challenges, researchers have increasingly turned to generative AI for creating synthetic medical images (Khosravi *et al* 2023). A key advantage of synthetic data generation is the ability to incorporate pixel-level annotations from the outset, reducing the burden of manual labeling while maintaining or even improving model accuracy (Gao *et al* 2022, Krishna *et al* 2023b). Furthermore, synthetic data can help enrich underrepresented populations, addressing the scarcity of certain patient groups, rare diseases, or protected communities. For instance, inserting rare abnormal tumors into MRI images has been shown to improve model performance on real patient cases (Shin *et al* 2018), while synthetically generated CT images have been used to complement patient x-ray data for TB classification, enhancing the reliability of AI models by addressing dataset imbalances and supporting more accurate diagnoses, particularly in resource-constrained settings (Lewis *et al* 2021).

Yet, these approaches face notable limitations. They often struggle to generate high-resolution, diverse outputs with sufficient generalizability. Generating images with full resolution and precise anatomical detail remains a substantial challenge, particularly when annotations—such as those marking tumors or other pathologies—with intricate spatial features are to be modeled. (Krishna and Mueller 2019, Han *et al* 2023, Krishna *et al* 2023b). In this paper, we present a novel approach to conditional image generation that addresses this shortcoming.

Our method is based on a general-purpose, condition-free trained denoising diffusion probabilistic model (DDPM) (Ho *et al* 2020). While DDPMs generate images through a denoising process, existing methods have primarily focused on iteratively guiding sampling toward specific regions of the image distribution (Choi *et al* 2021, Ho and Salimans 2021, Chung *et al* 2023). Such single-conditioning approaches can produce unsatisfactory results, particularly in generating image features faithfully across multiple scales, such as detailed annotations within the broader anatomical context. Building on Choi *et al* (2021), we propose a significant extension that incorporates multiple conditions dynamically throughout the image generation process. By separately conditioning on anatomy and annotations, our approach achieves a substantially higher level of detail in the generated annotations, surpassing the limitations of single-conditioning methods.

Our approach generates high-quality, full-resolution annotated CT images that successfully pass the Visual Turing Test (Krishna *et al* 2023a). Moreover, it can dynamically produce new annotations with their corresponding images, even when such annotation types were not initially present in the training pipeline. We demonstrate our method in an application that synthesizes lung pathology within anatomy-constrained lung CT images, using additional inputs for pathology within an anatomical context. The new capability enhances the model's versatility, enabling it to handle unseen scenarios and generate meaningful annotations, broadening its applicability and potential use cases. To the best of our knowledge, this is the first work to achieve this degree of flexibility—an ability to produce full-resolution images with accurate annotations while maintaining anatomical consistency across all clinically relevant Hounsfield unit (HU) windows.

## 2. Related work

Image synthesis has a rich history, beginning with the breakthrough innovation of generative adversarial networks (GANs) which has found widespread applications. Aside from discussing prior work in image synthesis, we also focus on the modeling of pathologies, the main motivation for the development of our multi-conditioned DDPM.

### 2.1. GANs

GANs (Goodfellow *et al* 2014) have established themselves as a pivotal class of generative models. They have provided fast methods for synthesizing varied, high-quality medical images (Prezja *et al* 2022) and they currently represent the state-of-the-art in certain applications, such as brain image generation (Xing *et al* 2021). However, challenges persist in terms of controlling output diversity and ensuring stable training. Despite their success, classical GANs are unconditioned generative models that offer no direct control over the attributes of the generated data. Their reliance on an implicit prior learned through the adversarial training interplay can lead to training instabilities that negatively affect image fidelity. Conditional GANs (cGANs), such as P2PGAN (Isola *et al* 2017) and CycleGAN (Zhu *et al* 2017), incorporate conditioning signals to guide the generation process, and they have found several applications in medical imaging (Popescu *et al* 2021, Sun *et al* 2022). However, even cGANs have limitations. One key issue is that their conditioning codes are often too coarse, capturing only high-level class information but failing to adequately represent the full spectrum of morphological nuances within a class.

In another line of research, Atli *et al* (2024) explored GAN-based selective state space modeling combined with a novel attention mechanism to address the attention-related challenges in transformers and CNNs for multi-modality image translation. While such specialized frameworks represent valuable advances in tackling complex tasks, they do not match the background of our application, which focuses on a more general-purpose scenario. Specifically, we aim to enrich rare or missing annotations within a single modality, ensuring that the resulting synthetic data directly addresses imbalances in annotated and non-annotated medical imaging datasets.

### 2.2. cGANs in pathology

cGANs have been extensively used for pathology synthesis. Ghorbani *et al* (2020) utilized cGANs to enhance the diversity of skin lesion images, while Waheed *et al* (2020) and Jiang *et al* (2020) developed cGAN-based generators for chest x-ray and CT data augmentation. Other studies (Shin *et al* 2018, Thambawita *et al* 2021) employed cGANs to produce synthetic polyp images with masks. However, most of these approaches rely on the availability of well-labeled data, limiting their applicability when labels are scarce or incomplete.

In Shaham *et al* (2019), Shaham *et al* introduced SinGAN, a model architecture that enables multi-conditioning across different scales. While this approach bears some resemblance to our own, SinGAN is designed to learn from single-image statistics rather than addressing rare pathological cases. This reliance on internal statistics inherently limits semantic diversity, often causing the model to replicate common tissue patterns rather than capturing the subtle traits of pathology. In related work, Thambawita *et al* (2022) tackled these limitations for polyp images by training a separate model for each image, but this approach is highly resource-intensive.

We propose a more direct and versatile approach, employing multi-image guidance to manage the synthesis of both pathology and its associated anatomical context. Our method aims for simplicity and can leverage any pre-trained diffusion model.

### 2.3. Diffusion models and controlled generation

Recently, diffusion models such as DDPMs (Ho *et al* 2020, Nichol and Dhariwal 2021) have emerged as powerful alternatives to GAN architectures. Although they can be slower, their output is often superior (Dhariwal and Nichol 2021, Khader *et al* 2023) and offers improved image diversity and quality. DDPMs also allow for conditioning on medically relevant attributes (Rouzrokh *et al* 2023, Sizikova *et al* 2023), enabling the generation of images that meet specific clinical criteria.

To enable more precise and controllable medical image generation through image-based inputs, numerous studies have explored integrating explicit controls, such as segmentation masks, into text-to-image diffusion frameworks (Mou *et al* 2023, Qin *et al* 2023, Ye *et al* 2023, Zhang *et al* 2023). Alternatively, diffusion models can be trained from scratch with carefully designed conditions (Rombach *et al* 2022, Qin *et al* 2023); however, this approach is often limited by the scarcity of large public datasets. A persistent challenge lies in achieving both highly precise and finely grained control. Existing guided generation frameworks, such as ControlNet (Zhang *et al* 2023) and T2I-Adapter (Mou *et al* 2023), still struggle to produce images that fully align with the given image conditions. This limitation becomes particularly critical when generating rare or diverse pathological cases within medical datasets for balancing or augmentation purposes.

To address these limitations, recent research has begun modeling image-based controllable generation as an image translation task. Özbey *et al* (2023) introduced a diffusion-based equivalent of cycle-consistent adversarial translation, while Arslan *et al* (2024) proposed a supervised approach that combines the strengths of GANs and DDPMs. In contrast, our approach focuses exclusively on a single modality and targets specific conditions, such as pathology or demographic attributes. This focus ensures that the image generation process is optimized for addressing data imbalance and scarcity, particularly in scenarios where existing methods fall short in these tasks.

Another line of research (Khader *et al* 2023, Pan *et al* 2023, Tudosiu *et al* 2024) has explored the integration of vector-quantized latent spaces and transformer architectures with diffusion models. By leveraging transformer-based architectures, these methods aim to capture complex spatial dependencies in medical images more effectively than traditional CNN-based approaches. However, these transformer-based methods often face fidelity challenges in 2D images and typically require large annotated datasets spanning multiple categories to ensure consistent performance within the same category.

### 2.4. Decoupling pathology/labels, anatomy and background image

The scarcity of annotated data is particularly critical for rare and underrepresented pathologies. Studies examining pathology-specific performance (Khosravi *et al* 2024) have shown that the most substantial improvements from synthetic data augmentation occur in pathologies representing less than 5% of the

population. This highlights the potential of generative AI to address data imbalance by focusing on infrequent and underrepresented cases.

Generative approaches have been proposed to synthesize various types of lesions (Salem *et al* 2019, Kadia *et al* 2022, Liu *et al* 2023), aiming to integrate realistic pathological appearances into existing anatomical structures. However, many of these methods face limitations in terms of resolution, diversity, and the complexity of the datasets required. For instance, techniques like partial convolutions (Liu *et al* 2023) focus on refining local patches but struggle to generalize to larger contexts. Similarly, other approaches (Shin *et al* 2018, Salem *et al* 2019) often produce low-resolution images, rely heavily on extensive labeled datasets, or lack applicability across varied anatomical contexts. In some specialized scenarios, domain-specific simulation frameworks (Amirrajab *et al* 2022, Al Khalil *et al* 2023) have been developed to generate anatomically plausible cardiac MRI (CMR) labels. While effective for their specific applications, these frameworks are limited by their domain dependency.

Our approach addresses these challenges by decoupling pathology from anatomy through the use of compressed B-spline curve representations to model anatomical structures. This methodology provides greater flexibility and adaptability, as the curves are not tied to any specific medical domain. By enabling a modular and generalizable framework, our approach facilitates broader applications in generating anatomically accurate and diverse pathological images.

## 3. Contributions

Our approach builds on the prior work through a design that enables comprehensive control over all aspects of medical image generation. We demonstrate its capabilities and use cases using lung CT data, utilizing two datasets: a small annotated dataset and a large non-annotated dataset. Notably, our method requires only the large non-annotated dataset for effective performance. The top-left corner of figure 1 highlights these datasets (marked in red boxes).

The small dataset consists of 512 $\times$ 512 high-dose chest CT images with segmentation maps from 30 patients, publicly available via TCIA (Yang *et al* 2017). It includes annotations for lungs, heart, spinal cord, esophagus, surrounding tissue, and bones, as illustrated in the first two images of figure 2. Such small annotated datasets are commonly available from repositories like TCIA, NIH, and OpenNeuro. In contrast. the large dataset comprises non-annotated low-dose chest CT images of similar resolution from approximately 14 000 patients, collected using various scanners across multiple locations (provided by NIH but not publicly available).

Our contributions are as follows:

- We propose a novel set of pre-processing steps to generate annotations and anatomy maps for large medical datasets, facilitating structured and scalable data preparation.
- We enhance a diffusion model-based framework to generate diverse, multi-conditioned CT images with precise annotations spanning the full HU range, achieving a quality virtually indistinguishable from real images.
- We extend the model to handle multiple constraints, enabling the generation of medical images with high fidelity annotated pathology without requiring the annotations to be present in any of the training datasets.
- By leveraging a novel diffusion-based strategy that allows precise manipulation of generated imagery, we provide an effective means to augment datasets with rare or highly specific pathological and anatomical instances.

The last contribution ensures that synthetic data not only increase the training volume but also enrich it with nuanced examples critical for improving model robustness. We demonstrate the effectiveness of our approach by generating images with anatomy outlines and annotated pathology, integrated seamlessly within the generated anatomy, using lung CT and mammograms as test cases.

## 4. Methods

In the following sections we describe all of these contributions in detail. We begin with annotation modeling and then describe our multi-conditioned DDPM methodology.

### 4.1. Annotations modeling and augmentation using controlled diffusion

As shown in figure 1(a), we extract annotations from a small dataset of 12 patients' annotated lung CT scans. Subsequently, we generate new images corresponding to these annotations by leveraging the images from a larger, non-annotated dataset. This is achieved by training a DDPM on the large dataset.

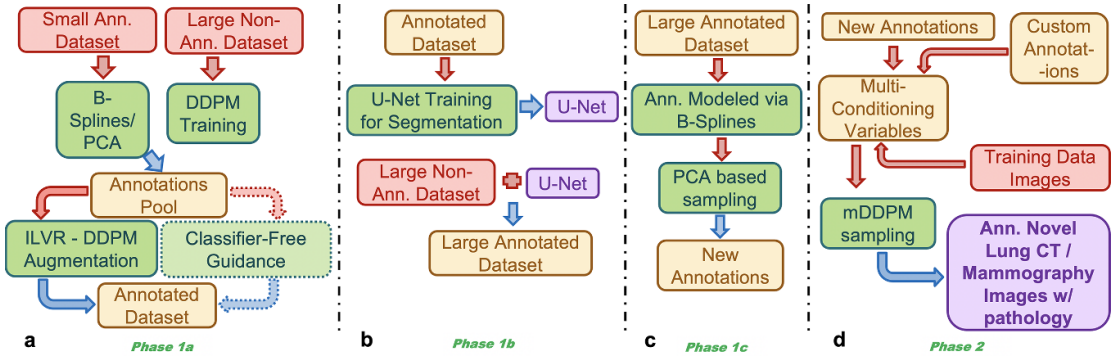


**Figure 1.** Our synthesis pipeline consists of two main phases. The first phase is divided into three sub-phases that transform a small annotated dataset into a large annotated dataset, with the capability to generate an infinite number of additional annotations using a large non-annotated dataset. These sub-phases use the strategies of modeling and sampling annotations through B-Spline/PCA-based sampling as suggested by Krishna *et al* (2023a). The second phase focuses on our primary multi-conditioning sampling step. Blue arrows indicate outputs such as images, data, or models, while red arrows represent the consumption of inputs or data.

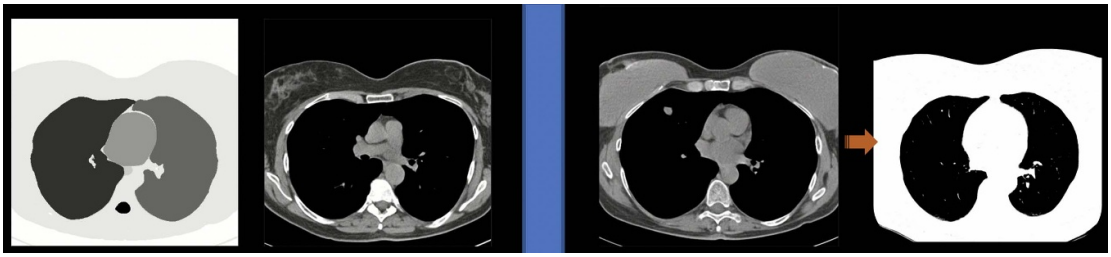

**Figure 2.** The first two images are from our small dataset. The last two images demonstrate that, in the absence of annotations, anatomy maps can be generated using a simple thresholding technique, which is applicable to most medical imaging domains.

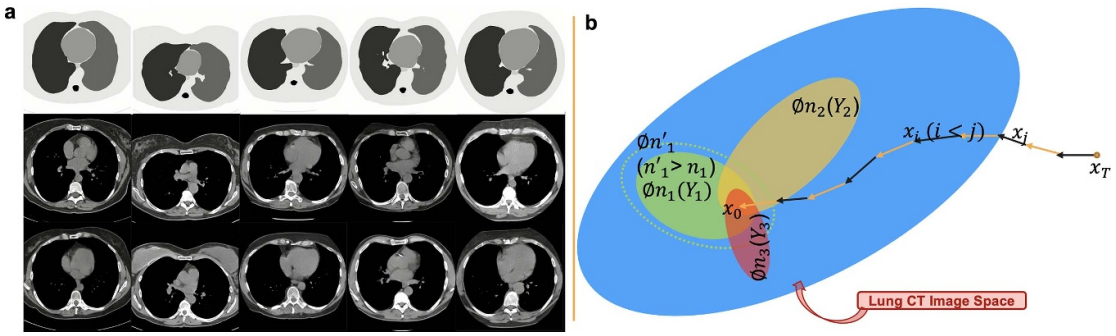


**Figure 3.** (a) Generated images ($512^2$) corresponding to the sampled anatomy maps (top row) defined by B-Spline curves (Krishna *et al* 2021) (b) Multi-Conditioned Guided Sampling. The blue area represents the image space for all CT lung images; the yellow, green and red circle represent the spaces closer to the three guidance images $y1$, $y2$ and $y3$, the size of the circles depends on the downsampling factors $n1$, $n2$, $n3$ of the filter used corresponding to these images.

1) Annotations sampling (figure 1(a)): Annotations or anatomy maps extracted from the small dataset are encoded using B-Spline curves, following the approach by Krishna *et al* (2021). As illustrated in figure 2, in cases where annotations are unavailable, simpler anatomy maps can be generated through thresholding. The control points of these B-Spline curves are projected into a variance-maximizing principal component analysis (PCA) space, which enables sampling to create new anatomy maps. This process is reused to sample diverse annotations (figure 1(c)) after annotating the large dataset of previously non-annotated images.
2) Diffusion models with guided sampling (figure 1(a)): We train an unconditional DDPM on the large dataset, comprising low-dose CT images and incorporating refinements suggested by Nichol and Dhariwal (2021). We use extracted annotations or anatomy maps as conditioning image inputs (Choi *et al* 2021) to the DDPM trained on the large dataset. Figure 3(a) highlights some of the results.

Alternative guidance architectures, such as classifier-free guidance (Ho and Salimans 2021) or ControlNet (Zhang and Agrawala 2023), can also be employed for this step. All generated images maintain a resolution of 512 × 512.

3) Generated annotations via U-net and PCA space (figures 1(b) and (c)): Combining the previously generated annotated images, we train a U-Net (Ronneberger *et al* 2015) to predict annotations from an image, enabling annotation generation for the entire large dataset as the augmented dataset's textures align with those in the large dataset. However, training a U-Net is not required if the small dataset lacks annotations and anatomy maps are instead extracted via thresholding. The resulting increased annotations / anatomy maps create a richer PCA space for B-spline control points, significantly enhancing the ability to generate diverse anatomy maps.

### 4.2. Multi-conditioned DDPM

We investigate the sampling strategies of our trained DDPM suggested by Choi *et al* (ILVR) (2021), and extend it to incorporate multiple conditional or guidance images. Our findings highlight the significance of this strategy for the purpose of synthesizing medical imaging datasets that are not only highly accurate but also annotated. Moreover, as these guidance techniques are not bound by annotations, they can be effectively employed to enhance annotated images featuring rare anatomies and pathology, thereby fostering the development of a more comprehensive and diversified dataset.

#### *4.2.1. Preliminaries*

Our DDPM iteratively transforms an isotropic Gaussian distribution into a full HU window lung CT image distribution. The Markov Chain model learns the reverse of the forward diffusion process which is

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

with $x_t$ as the latents with added noise and $\beta_t$ as a fixed variance schedule.

Equation (1) can be decomposed by the reparameterization trick and $x_t$ can be further derived in terms of the image $x_0$ as:

$$x_t = \sqrt{\overline{\alpha}_t}x_0 + \sqrt{1-\overline{\alpha}_t}\epsilon \tag{2}$$

with $\alpha_t := 1-\beta_t$ and $\overline{\alpha}_t := \prod_{i=1}^{t}\alpha_i$. The added noise $\epsilon \sim N(0, I)$ has the same dimensionality as the image and the sampled latents during training.

The reverse diffusion process is learned via a neural network $p_\theta$ and is expressed in terms of $\mu_\theta$ (Ho *et al* 2020):

$$p_\theta(x_{t-1}|x_t) = N\left(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 I\right) \tag{3}$$

$\mu_\theta$ is further decomposed (Ho *et al* 2020) in terms of noise approximator $\epsilon_\theta$:

$$\mu_\theta = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_\theta(x_t, t)\right). \tag{4}$$

By formulating the loss function (Ho *et al* 2020) as the log likelihood of $x_0$ and computing a variational lower bound (similar to the case of variational auto-encoders) as the KL divergence between $q$ and $p$, Ho *et al* (2020) framed the loss function as the L2 distance between the actual mean of the image ($\mu$) and $\mu_\theta$ which can be further simplified to as the L2 distance between the predicted noise $\epsilon_\theta$ and added noise $\epsilon$ at any given time $t$:

$$\text{Loss} = \|\epsilon - \epsilon_\theta(x_t, t)\|^2 = \|\epsilon - \epsilon_\theta\left(\sqrt{\overline{\alpha}_t}x_0 + \sqrt{1-\overline{\alpha}_t}\epsilon, t\right)\|^2. \tag{5}$$

Equations (2) and (5) are used to train our DDPM, incorporating refinements from Nichol and Dhariwal (2021). Our DDPM was trained on a large dataset of 5000 lung CT scans, with images extracted at full HU width of 2000. This ensures that the generated images span the entire HU width during sampling and can be visualized at other clinically relevant windows, including lung, bone, and soft-tissue. Utilizing equation (3) and the reparameterization trick, $x_{t-1}$ can be sampled as:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_\theta(x_t, t)\right) + \sqrt{\beta_t}\epsilon. \tag{6}$$

Using the above equation repeatedly, we can sample lung CT images starting from random noise after training a DDPM on our large dataset. Both training and sampling steps are outlined in prior works related

**Algorithm 1.** Sampling.

**1 Input:** Conditional / guidance images $y_1, ....y_M$
**2 Output:** Generated image x
**3 Filter-scales:** $\phi_{n_1}, ....\phi_{n_M}$
**4 Time-steps (T, a):** $a_1, ....a_M$
**5** $x_T \sim N(0, I)$
**6 for** $t = T$ **to** 1 **do**
**7** $z \sim N(0, I)$
**8** **if** t = 1 **then**
**9** z = 0
**10** $x_{t-1} = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_\theta(x_t, t)) + \sigma_t z$
**11** $X = 0$
**12** **for** $s = 1$ **to** $M$ **do**
**13** $y_{s_{t-1}} \sim q(y_{s_{t-1}}|y_s)$
**14** **if** $t \geqslant a_s$ **then**
**15** $X = X + \phi_{n_s}(y_{s_{t-1}}) - \phi_{n_s}(x_{t-1})$
**16** $x_{t-1} \leftarrow x_{t-1} + X$
**17 return** $x_0$

to DDPMs (Ho *et al* 2020). Next, we will focus on the sampling algorithm of our DDPM to facilitate multi-annotation guidance during our lung CT image generation.

*4.2.2. Multi-condition guidance (figure 1(d))*

As previously discussed, there are diverse approaches available to steer the sampling procedure of a trained DDPM. Here, we explore the guidance techniques introduced by Choi *et al* (2021) and utilize them to exert precise control over the generation of lung CT images within all HU windows. Choi *et al* propose the feasibility of directing the sampling process towards a subset of image distributions surrounding a reference image $y$, provided that we can establish similarity between the downsampled reference image $y$ and the downsampled generated image $x_0$.

To approximate this condition in each Markov transition during the sampling process, Choi *et al* consistently enhance the downsampled latent variable $x_t$ within steps $(T, a)$ to resemble the corresponding downsampled noisy version of the reference image $y_t$ ($y_t = \sqrt{\overline{\alpha}_t}y + \sqrt{1-\overline{\alpha}_t}\epsilon$). This ensures that both $x_t$ and $y_t$ exhibit shared low-frequency contents. Specifically:

$$p_\theta(x_{t-1}|x_t, c) \approx p_\theta(x_{t-1}|x_t, \phi_N(x_{t-1}) = \phi_N(y_{t-1})) \tag{7}$$

where $\phi_N(\dots)$ is a low-pass linear filter with N as the downsampling factor. The term is approximated by ensuring the latent $x_{t-1}$ captures the missing low-frequency contents of $y_{t-1}$ after sampling from the unconditional DDPM,

$$x_{t-1} = x_{t-1} + \phi_N(y_{t-1}) - \phi_N(x_{t-1}). \tag{8}$$

We argue that by manipulating the degree of downsampling and the number of steps a linear filter $\phi$ is used, for a given set of $M$ conditional images $y_1, y_2, \dots, y_M$, we can fine-tune the sampling using sets of integers $n_1, n_2, \dots, n_M$ and $a_1, a_2, \dots, a_M$. Here, $n_m$ denotes the downsampling extent for a linear filter corresponding to each conditional image during the steps [T, $a_m > 1$). This allows for valid image generation through a trained DDPM that shares low-level features (or similarity) with each of the conditional images. We modify equation (8) as:

$$x_{t-1} = x_{t-1} + \sum_{s=1}^{M} \begin{cases} \left(\phi_{n_s}\left(y_{s_{t-1}}\right) - \phi_{n_s}\left(x_{t-1}\right)\right), & \text{if } t \geqslant a_s \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

The integers $n_s$, $a_s$ for a conditional image will depend on the purpose and the nature of the conditional image in the generation of the final images. In practice, the above strategy may only work well for a maximum of three conditional images. Figure 3(b) visualizes our multi-conditional guidance and the steps in sampling where with each step the generated image gets closer to the desired super-subset of the image distribution. As is evident from the visualization; if above sets of integers are not chosen carefully, there may not be a significant overlap between the subset distributions of conditional images in which case the image samplings may start generating inaccuracies. Steps 11–16 in algorithm 1 illustrate the above process in the sampling of the synthetic lung CT images.

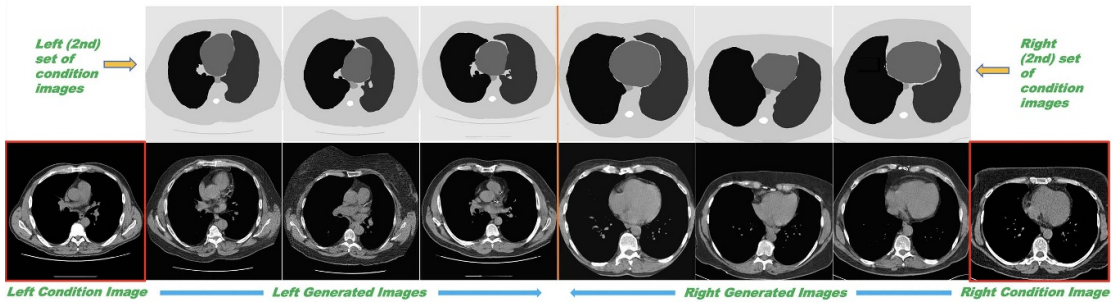


**Figure 4.** Three generated lung CT images, each with two conditional images (top and left). The images generated are for three different anatomy maps for the same conditional CT image (marked with red outline). The generated images follow the anatomy of the top row but texture features such as the heart follow the CT image. The results are displayed in the soft-tissue window to highlight the similarity and accuracy of the generated anatomy w.r.t. the guidance images.

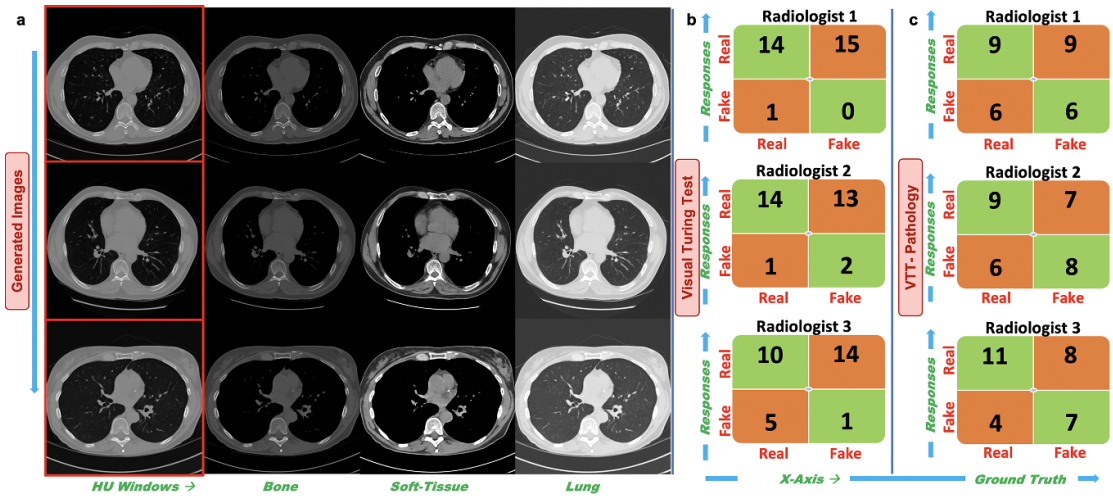


**Figure 5.** (a) The leftmost column of the figure highlights some of our sampled images in the full HU window range. The next three columns show the same sampled images in the Bone, Soft-Tissue and Lung windows. (b) Confusion matrices for the responses of the 3 radiologists. The overall accuracy of the responses is 45.56% which is close to 50%; a requirement for passing our Visual Turing Test. (c) Confusion matrices for the responses of the 3 radiologists (a different group) for the images with pathologies. The overall accuracy of the responses is 55.56% which is again close to 50%.

## 5. Experiments and results

As outlined in figures 1(a)–(c), aside from pre-processing steps that produce custom annotations, we train a DDPM (Nichol and Dhariwal 2021) on our large dataset which consists of low-dose lung CT-scans of 5000 patients. The images from the scans were extracted from the mid-abdomen regions, clearly showing the lungs along with the heart. The images are extracted in the entire relevant width of 2000 HU (−1000 HU to 1000 HU) for training our model; it enables the generation of images in the same HU range during the sampling process post training. That way, the images can be viewed at any HU window during their evaluation in the Visual Turing Test.

### 5.1. Visual and quantitative evaluation

Figures 4, 5(a) and 6 present images generated using our sampling strategy. Figure 4 illustrates sets of guidance images, each consisting of an anatomy map and a CT image, which are used as conditional inputs for generating the images shown (see the caption for more details). Figure 5(a) displays generated images in bone, soft-tissue, and lung windows, allowing for visual inspection of anatomical consistency across clinically relevant windows. Figure 6 provides additional examples where both an anatomy map and a real CT image serve as guidance inputs for each generated image.

#### *5.1.1. Visual turing test*

We began by verifying the anatomical accuracy and fidelity of the generated images through a Visual Turing Test (Krishna *et al* 2023a). With the assistance of three radiologists, we evaluated the realism of our generated

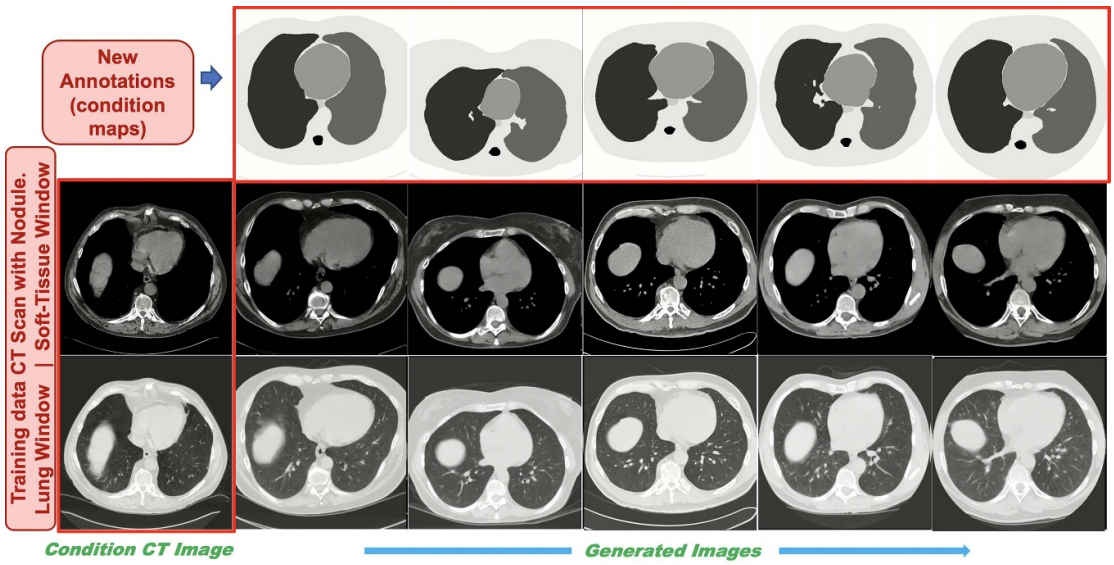


**Figure 6.** Another example of multi-conditioned sampling; The figure shows multiple generations of the same reference CT image (leftmost red column) in soft-tissue and lung windows based on different input reference anatomy maps (top red row). All the anatomy maps shown in the figures are new anatomy maps generated via PCA-sampled B-Splines (Krishna *et al* 2023b).

lung CT images. The radiologists were asked to label 30 randomly selected lung CT images as 'Real' or 'Fake', with the images chosen from bone, lung, and soft-tissue windows.

The purpose of this test is to assess whether our model can generate medically accurate images. If the experts are unable to correctly distinguish between real and fake images at least 50% of the time (chance baseline), the model is considered to have passed the Visual Turing Test. The results, compiled in figure 5(b) and table 1(a), demonstrate that our generative framework successfully passed the test. Expert radiologists were unable to identify most of the synthesized lung CT images, with many of their 'Fake' labels incorrectly assigned to real images. This indicates that our guidance-based DDPM sampling scheme produces images that are indistinguishable from real ones (see the caption for more details).

*5.1.2. Comparisons with unconditional generative models*

Table 2(a) presents a quantitative comparison of approximately 10 000 generated full HU window lung CT images against state-of-the-art generative models, including StyleGAN (Karras *et al* 2018), StyleGAN2 (Karras *et al* 2020), DDPM (Unguided Sampling) (Ho *et al* 2020), and diffusion-transformers (DiT) (Peebles and Xie 2023), all trained on the large dataset. We also evaluated images sampled without guidance from the same trained DDPM to assess the impact of our sampling strategy on image fidelity. As shown in the table, the FID and the MMD scores indicate that our model performs at least as well as the state-of-the-art unconditional models, if not better. Moreover, our approach uniquely focuses on generating both raw data and their corresponding annotations, setting it apart from these methods. Inception scores (IS) are less reliable for medical image synthesis as they rely on unrelated ImageNet classifications and do not compare against real medical image distributions. Nonetheless, our model performs comparably to state-of-the-art fidelity models.

In addition, we conducted a set-level comparison of structural similarity index (SSIM) scores with the training data, finding that our guided-sampling approach outperformed StyleGANs and DiT. While unconditional models generate accurate images, they often lack anatomical consistency due to the lack of an anatomy-controlled generation framework. Visual inspection confirms our method produces more anatomically coherent results. We also calculated the Diversity Score (DS), the average pairwise L2 distance among feature embeddings from a pre-trained InceptionV3, and found our model performs comparably to the state-of-the-art fidelity models.

*5.1.3. Comparisons with unconditional generative models in medical imaging*

In recent years, researchers have adapted models from the Computer Vision literature for synthetic medical image generation, modifying and integrating some for improved performance. For evaluation, we selected three state-of-the-art pipelines and compared their outputs against ours. The first pipeline (Khader *et al* 2023) employs a latent diffusion model (LDM) with VQ-GAN quantized latent space; the second (Tudosiu *et al* 2024) utilizes VQ-VAE's compressed latent space as input to a transformer for image generation; and the

**Table 1.** Summary of Three Radiologists' Feedback on Lung CT Images and Lung CT Images with Pathology.

| (a) Visual Turing Test | | | |
|---|---|---|---|
| | Accuracy | Real → Fake (FN) | Fake → Real (FP) |
| Radiologist 1 | 46.7% | 6.7% | 100% |
| Radiologist 2 | 53.3% | 6.7% | 86.7% |
| Radiologist 3 | 36.7% | 33.3% | 93.3% |
| Averages | 45.6% | 15.6% | 93.3% |
| (b) Visual Turing Test (Pathology) | | | |
| | Accuracy | Real → Fake (FN) | Fake → Real (FP) |
| Radiologist 1 | 50% | 40.0% | 60.0% |
| Radiologist 2 | 56.7% | 40.0% | 46.7% |
| Radiologist 3 | 60% | 26.7% | 53.3% |
| Averages | 55.6% | 35.6% | 53.3% |

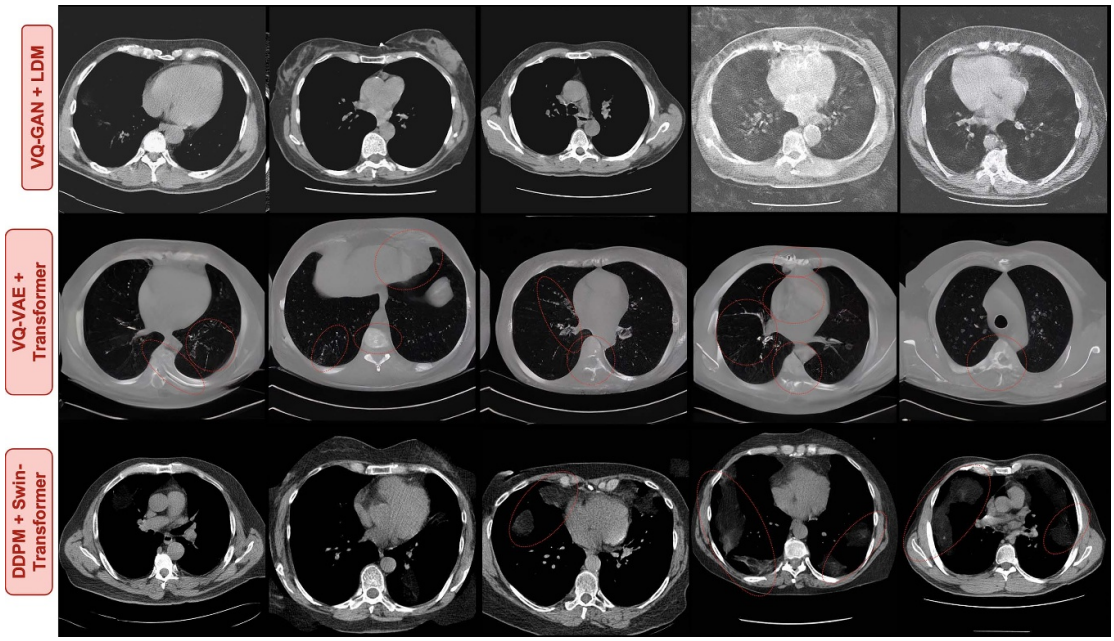


**Figure 7.** Images are generated at a resolution of 512 × 512, with each row showcasing images in either a soft-tissue window (first and third rows) or a full-HU window (third row), best suited to highlight anatomical inconsistencies and fidelity issues in different frameworks. Row 1: The framework by Khader *et al* (2023) demonstrates background noise in several cases, particularly visible in the last two images. Row 2: A latent-space-based framework (Tudosiu *et al* 2024) with a transformer proposed by another study displays inconsistencies across key areas, including bones, lungs, and hearts, as indicated by red dotted ovals. Row 3: A DDPM-based framework (Pan *et al* 2023) with intermediate swin-transformer layers generates excessive ground-glass opacity artifacts, particularly evident in the lung region.

third (Pan *et al* 2023) integrates swin-transformer layers with convolutional layers in a DDPM framework (MT-DDPM). While these models were primarily designed to generate lower-resolution images, some could create consecutive 2D slices. We modified them to generate higher-resolution 512 × 512 images for comparative analysis. Table 2(b) illustrates that our model performs equally well or outperforms these pipelines across various metrics, while figure 7 showcases sample outputs.

Enhancing the latent size of the VQ-GAN+LDM pipeline improved image quality but introduced residual noise in some outputs, visible in the last two columns of row 1. The second row displays results from the VQ-VAE+transformer, which showed inconsistencies in generating anatomical structures such as bones, tissue, hearts, lungs, spinal cords, and esophagi, especially in the full HU window. The MT-DDPM, combining transformer and convolutional layers, consistently produced clear, high-resolution images but occasionally exhibited ground-glass opacity-like artifacts. These artifacts, prominent in soft-tissue windows within the lung region, are highlighted with red circles in the third row of figure 7.

**Table 2.** Comparing generative models with multi-guided sampling (best scores are highlighted in bold).

| | FID | IS | MMD | DS | Set-level SSIM |
|---|---|---|---|---|---|
| (a) Unconditional Generative Models | | | | | |
| PGGAN | 146.09 | 2.21 | 0.141 | 15.71 | 0.33 |
| DiT | 82.83 | 1.87 | 0.156 | 14.49 | 0.38 |
| StyleGAN | 81.57 | 2.09 | 0.135 | 13.78 | 0.31 |
| StyleGAN2 | 72.31 | **2.28** | 0.135 | 14.88 | 0.30 |
| Unguided Sampling (DDPM) | 83.24 | 2.12 | 0.143 | 14.11 | 0.27 |
| Guided Sampling (Ours) | **69.85** | 2.22 | **0.129** | 14.40 | 0.45 |
| (b) Unconditional Generative Models in Med. Imaging | | | | | |
| VQ-GAN + LDM | 160.94 | 2.18 | 0.150 | **15.81** | 0.13 |
| VQ-VAE + Transformer | 196.34 | 2.00 | 0.160 | 13.64 | 0.18 |
| MT-DDPM (Swin-Trans.) | 85.33 | 2.14 | 0.143 | 14.57 | 0.32 |
| Guided Sampling (Ours) | **69.85** | 2.22 | **0.129** | 14.40 | 0.45 |
| (c) Conditional Generative Models | | | | | |
| P2P + CycleGAN | 81.96 | 2.12 | 0.133 | 14.69 | **0.46** |
| Cross-Att. DDPM | 79.02 | 2.26 | 0.131 | 14.83 | 0.26 |
| ControlNet (Multi-Cond.) | 106.74 | 2.01 | 0.140 | 14.03 | 0.35 |
| Guided Sampling (Ours) | **69.85** | 2.22 | **0.129** | 14.40 | 0.45 |

#### *5.1.4. Comparisons with conditional generative models*

We also compared our approach with some popular conditional generation methods. While quantitative metrics for these methods are presented in table 2(c), our primary focus was on evaluating how well the generated images adhered to conditional inputs while maintaining anatomical accuracy and fidelity-crucial for generating rare and diverse pathology and anatomy. Figure 8 illustrates this comparison.

Given our access to CT images with and without anatomy maps, and the ability to generate a large number of these maps by sampling our B-Spline space, we benchmarked our method against Image-Factory (Krishna *et al* 2023b). This state-of-the-art approach employs a combination of paired and unpaired training in a GAN setting for CT image generation. The results in the second and third columns of figure 8(top three rows) highlight that our method significantly outperforms Image-Factory in maintaining anatomical accuracy, particularly in the heart region, as seen in the soft-tissue window, while still adhering closely to the input anatomy maps, even at a high downsampling factor of 64. The fourth column presents the same generated images in the full-HU window, reinforcing this observation.

Next, we modified a multi-view diffusion framework to train a conditional-DDPM (Watson *et al* 2023) using cross-attention U-Nets and stochastic conditional sampling. However, as shown in the last three columns, conditionally trained DDPMs struggled to accurately follow conditional maps, performing less effectively compared to our method, even under high downsampling factors of 64 and 128.

Next, we trained a ControlNet (Zhang and Agrawala 2023), which enables multi-conditional sampling by concatenating conditional images as inputs during the sampling process. The results of conditional sampling via ControlNet are shown in the last two rows. While ControlNet performs well for single-conditional sampling using anatomy outlines, it tends to generate anatomically inconsistent results when handling multiple conditional images, as highlighted in the last row. One significant limitation of ControlNet is its inability to fine-tune the extent of conditioning for each conditional image. Unlike our approach, which allows nuanced control, ControlNet lacks mechanisms to resolve conflicts arising from multiple conditions. Consequently, it struggles to balance competing conditional inputs, leading to inaccuracies in anatomy generation under complex scenarios.

### 5.2. Anatomy features/pathology generation

For our next multi-conditional sampling experiments, we selected lung CT images from our training dataset that exhibit unique anatomical features and pathologies. For instance, figure 6 demonstrates multiple generations where the diaphragm of a reference CT image is distinctly visible, integrated into various lung CT anatomies through additional guidance provided by anatomy images used as inputs.

We extended our experiments to generate CT images featuring lung pathologies, including both benign and malignant tumors and nodules. With the assistance of our radiologists, we handpicked specific CT images from our training dataset that exhibited these pathologies and manually annotated them to serve as guidance image patches. Figure 9 displays generated images for five different pathologies across three different anatomies, guided by anatomy maps. We observed that in many cases, the generated lesions

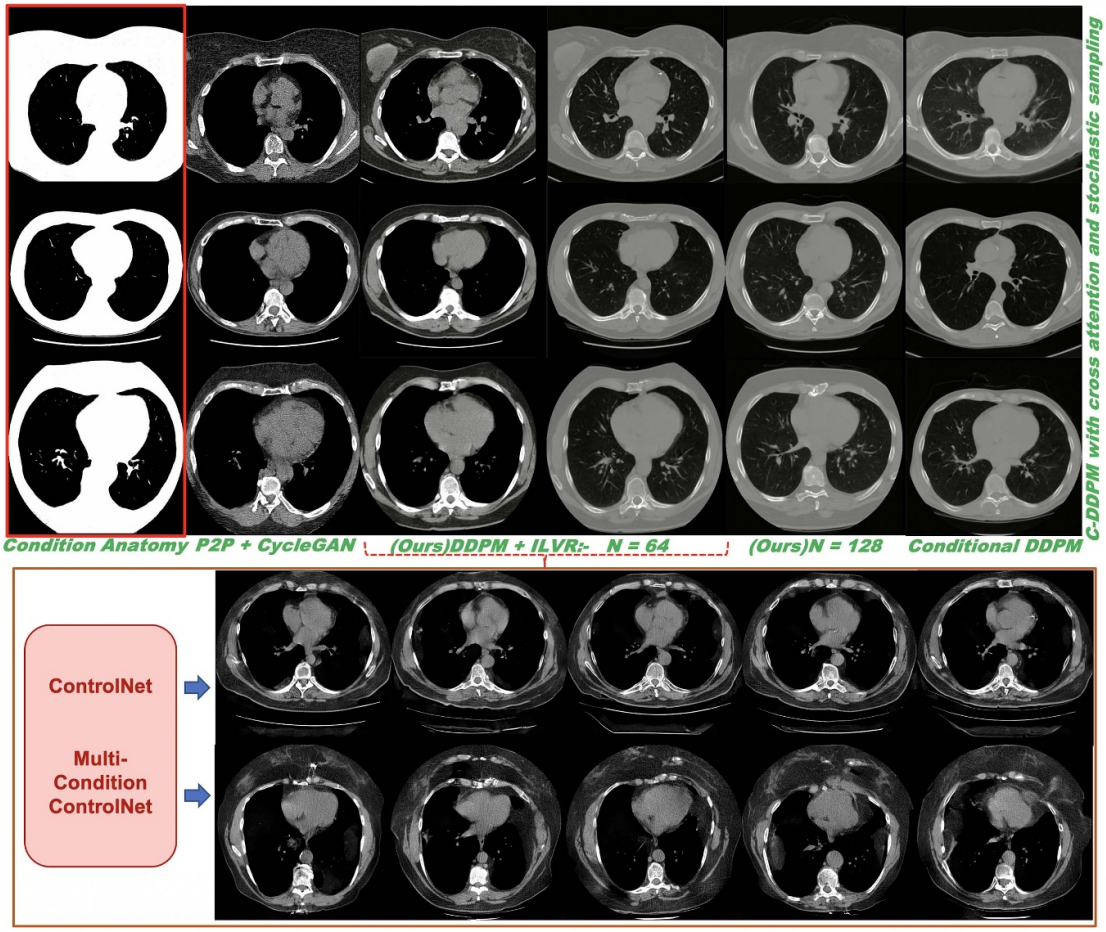


**Figure 8.** Top Three Rows: 2nd Column: Conditional generation via the P2PGAN + CycleGAN framework corresponds to the anatomy maps in the first column. 3rd Column Multi-guided sampling generation with a downsampling factor (N) of 64 on the anatomy maps; shown in soft-tissue window with accurate generation of the hearts unlike the 2nd column images. 4th column: Same images as the 3rd in the full-HU window. 5th column: Higher N but still adhere to the anatomy maps better than the conditionally trained DDPMs with cross-attention U-Nets whose generations are shown in the 6th Column. Bottom two rows: ControlNet generations with the top row conditioned on an anatomy map or a CT image and the bottom row conditioned on an anatomy map and a CT image. The results from the bottom row reveal that ControlNet struggles to consistently satisfy two conditional inputs, leading to inconsistent generations.

realistically adhered to lung walls or surrounding tissues when in close proximity, effectively mimicking the behavior of tumors in these regions.

In figure 10, we demonstrate that with a single pair of guidance images-a pathology-annotated CT image and an anatomy map-we can generate multiple CT images displaying the same pathology with slight variations in the spiculation patterns of the generated lesions. This variation can be achieved by adjusting the downsampling factor N of the linear filter or modifying the number of refining steps $[T, A]$ corresponding to the conditioning pathology shown in the leftmost column of the figure.

### *5.2.1. Visual turing test (pathology)*

We conducted a Visual Turing Test to evaluate the fidelity of pathology images generated via our multi-guided sampling approach. The goal was to assess whether the anatomical integration of pathology within the simultaneously generated anatomical context affected the overall fidelity of the images. The radiologists (a different group) were informed that half of the images were synthetic to encourage them to classify more images as fake. Each radiologist spent approximately thirty minutes on the test, averaging one minute per image.

The results, presented in figure 5(c) and table 1(b), show that while the radiologists performed slightly better than chance in distinguishing real from fake images, they still misclassified fake images as real more than half the time. This resulted in an average accuracy only slightly above 50%, further demonstrating that our generative framework effectively produces realistic images, even for rare pathological cases within the training dataset.

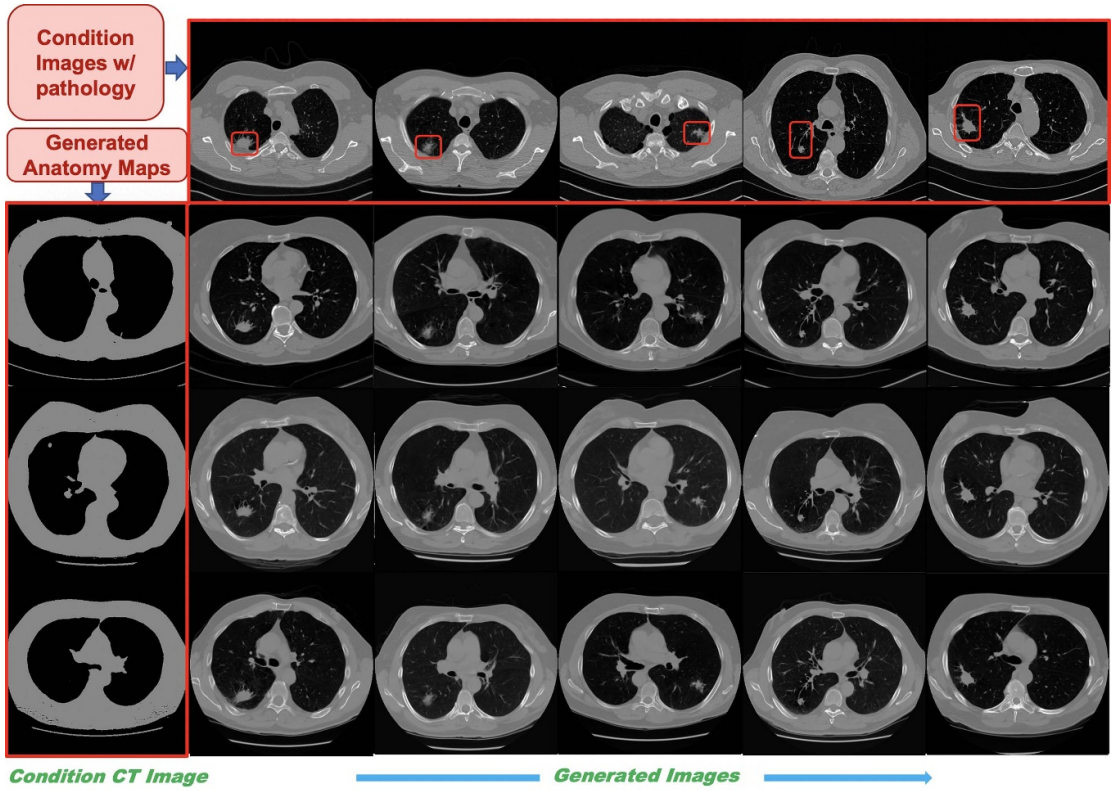


**Figure 9.** Examples of generated lung-CT images with pathology using multi-conditioned sampling. These images follow the anatomy of the conditional images in the left column (generated with B-Spline curves) and the pathology of the real CT images in the top row.

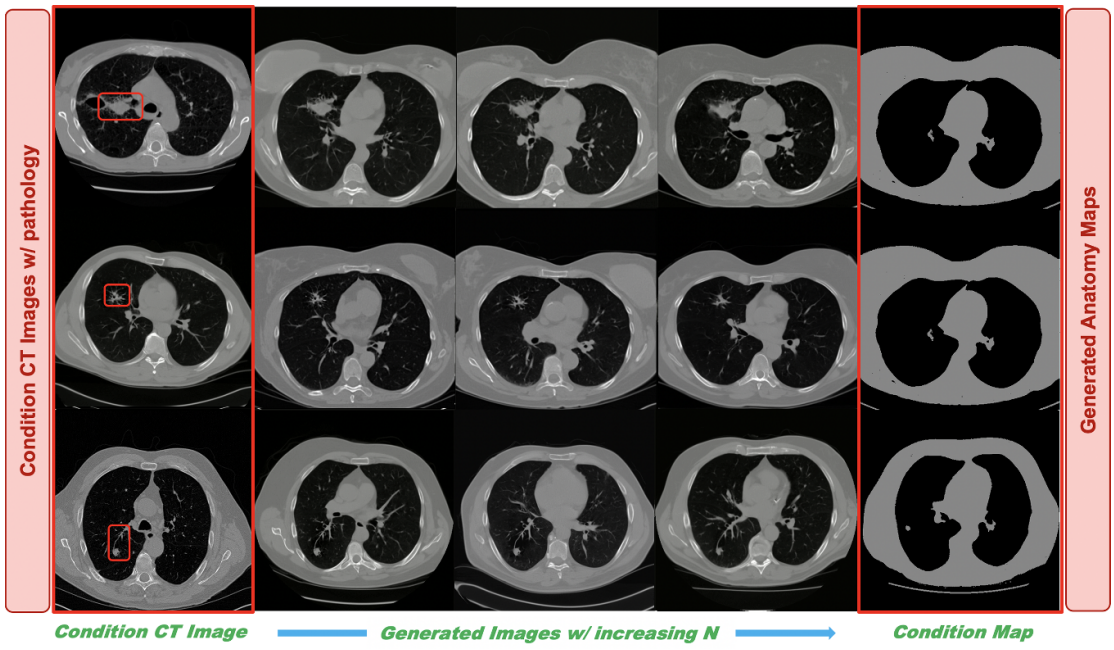


**Figure 10.** Examples of generated lung-CT images with pathology using multi-conditioned sampling; which follows the anatomy of the conditional images (generated using B-Spline curves) on the right and pathology of the real CT images on the left. By varying N for the conditional pathology images, we can generate different versions of the same tumor/nodule having varying spiculation patterns.

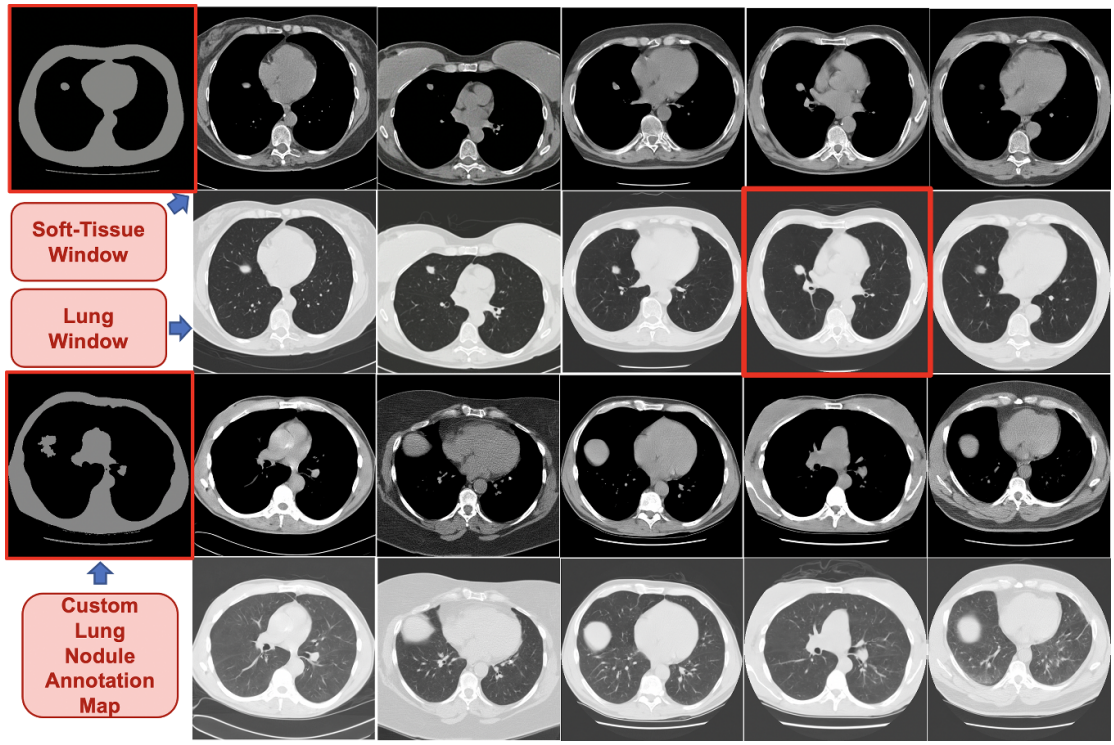


**Figure 11.** Examples of CT images with pathology generation using only one conditional anatomy map shown in the red boxes; for the first case (first two rows) we were partially successful in generating circular nodules along with diversified generation; for the second case (last two rows), we were unable to generate a spiculated tumor corresponding to the one shown in the anatomy map, when generating diverse images.

### 5.3. Ablation

To assess whether our multi-conditioned framework outperforms a conventional single-conditioned framework, such as that by Choi *et al* (2021), we conducted an ablation study. In this study, we reduced our condition image set to a single reference image containing both anatomy and pathology, as shown in figure 11 for two tumor cases. The red-boxed reference image in the top row (first case) depicts a reference anatomy with a simple, blob-shaped tumor. In contrast, the red-boxed reference image in the third row (second case) shows a reference anatomy with a more complex, spiculated tumor.

The remaining columns of figure 11 show the results of this experiment. The top two rows display the soft-tissue and lung windows, respectively, for the first case with a simple tumor. We observe that the generated tumors fairly closely resemble the reference tumor, with the intended variations. However, challenges arise when the nodules are positioned near surrounding organs (e.g. the heart), causing them to blend with the connecting tissues, as seen in the red-boxed result image (row 2, column 5).

While the generated tumors in the first case have a similar general appearance than the tumor in the reference image, for the second case with a complex tumor (bottom two rows), this general resemblance is not observed. In most generated images, the spiculated reference tumor degrades into a large, shapeless blob. This suggests that the ablated, single-conditioned approach struggles to capture more challenging pathologies, reducing them to simpler forms. Our multi-conditioned approach, on the other hand, is able to generate complex-shaped tumors and without unwanted tissue connectivity, as demonstrated in previous figures.

### 5.4. Showcasing versatility

We tested the versatility of our framework by generating cancerous tumors on mammography images. Using a trained DDPM on non-annotated mammograms, we applied our multi-conditioned sampling method, similar to the setup used for lung-CT pathology generation. Figure 12 presents preliminary results of generated mammograms containing tumors. To mitigate texture discrepancies between the two guidance images for a generated image, we pre-processed the tumor images using style transfer techniques. The preliminary results suggest that our framework can be effectively applied across multiple domains of medical imaging, enabling the generation of large, annotated, and balanced datasets from a minimal number of annotated images.

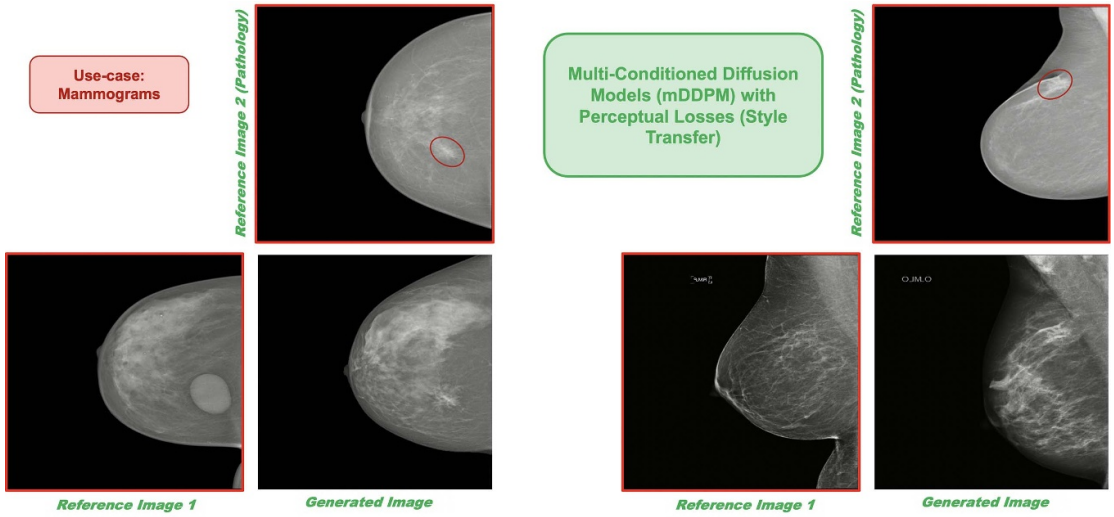


**Figure 12.** Multi-conditioned sampling in mammograms; the figure shows two generated images corresponding to their conditional images (left and top) highlighted in red. The images show the successful generation of the cancerous tumors of the top images embedded in the generated images along with its surrounding structures and textures similar to the left images.

## 6. Discussion

Despite significant advancements, generating high-resolution chest CTs with precise anatomical details, such as small nodules or subtle tissue boundaries, remains a major challenge. Existing diffusion and Vision Transformer-based models often fail the Visual Turing Test for chest CTs, even when successful in other domains (Khader *et al* 2023), highlighting their limitations in capturing fine-grained anatomical structures. While transformer-based guidance can enhance contextual coherence, it does not consistently ensure fine boundary accuracy or morphological fidelity-both of which are essential for reliable pathological assessment.

Our proposed method can be seen as complementary to other recent approaches, such as swin-transformer layers, ControlNet-augmented diffusion or classifier-guided strategies, which excel at preserving certain spatial or semantic properties. Specifically, ControlNet is well suited to tasks that depend on precise boundary delineations-like segmentation maps or contour constraints-while our ILVR-inspired (Choi *et al* 2021) guidance is more flexible for tissue-structure nuances and subtle morphological variation. Combining our multi-guidance with boundary-focused modules (e.g. ControlNet) may further improve the consistency of boundaries without sacrificing the natural variability of soft tissue. Similarly, techniques like classifier-free guidance may be integrated to further shape the sampling process toward clinically relevant features, possibly outperforming earlier approaches in terms of image realism and diversity balance.

A natural extension of our work involves integrating with text-based conditioning to improve control over pathological features. Text-guided approaches-such as those leveraging large language models-could enable more nuanced or multi-faceted constraints on the synthesis process (e.g. 'generate a lung CT with two small nodules in the upper lobe'). By combining text embeddings with image-based control, one might guide the synthesis toward specific pathologies or demographic profiles more effectively.

Despite these benefits, our method still operates entirely in image space rather than a latent space, making sampling relatively slower and more computationally expensive. This cost grows significantly in 3D, where the dimensionality is high and each diffusion step becomes more time-consuming. Such constraints can limit the practicality of generating large-scale synthetic datasets intended for broad data augmentation. For smaller datasets, or for specific subsets of rare pathologies requiring more diversity, our approach still offers important benefits in capturing subtle morphological nuances. However, future research should explore latent-space sampling methods tailored for 3D volumes-potentially by integrating our multi-conditioning approach with a latent diffusion framework. These enhancements could deliver comparable fidelity while alleviating the computational overhead. For now; although our model excels at capturing subtle pathologies, it is tested primarily on 2D slices; full 3D volumetric sampling remains computationally intensive and hence limit large-scale quantitative evaluations.

## 7. Conclusion and future work

Our results demonstrate that diverse and anatomically accurate lung-CT images, complete with annotated pathology or anatomy, can be generated even from small datasets. Remarkably, our approach was able to simulate pathology without relying on any pre-existing annotations for that pathology in the training datasets. Both visual inspections and quantitative assessments confirm the anatomical accuracy of the generated images across all clinically relevant HU windows, underscoring the effectiveness of DDPMs enhanced by guidance images in capturing the intricacies of anatomical structures in lung CT images. Beyond boundary-level precision, we emphasize multi-condition sampling that can scale up conditional diversity (e.g. multiple pathology types, style variations, or demographic attributes). Although conditional guidance can sometimes reduce variation in the generated samples, our results suggest that proper sampling strategies-particularly with an ILVR-like formulation-can maintain or even increase diversity while still improving fidelity and resolution.

Looking ahead, we plan to extend our framework to cover a broader range of medical domains and pathologies. Building on our success with mammograms, we aim to explore a combination of conditioning strategies beyond ILVR, such as ILVR paired with style-transfer techniques. Additionally, we are interested in expanding our work to 3D image generation, such as creating entire lung-CT scan slices with annotated pathology.

Finally, while the above discussion has focused largely on fidelity and morphological correctness, future investigations should examine the diagnostic utility of such synthetic data. For instance, one can systematically measure whether generated CT scans improve downstream pathology classification or segmentation across a broader range of diseases. Likewise, user studies with radiologists could assess realism and detectability of subtle lesions. In that vein, combining multiple generative paradigms-such as our ILVR-based multi-conditioning, boundary-aware ControlNet modules, or classifier guidance may yield more anatomically precise, high-resolution CT data for both research and clinical applications. Furthermore, we intend to evaluate whether models trained on our synthetic images can match or even surpass the performance of those trained on real data. This will help us assess the potential for synthetic images to completely replace real data in training deep learning models for various medical applications.

## Data availability statement

The data cannot be made publicly available upon publication because they are owned by a third party and the terms of use prevent public distribution. The data that support the findings of this study are available upon reasonable request from the authors.

## Acknowledgment

The authors thank the National Cancer Institute for access to NCI's data collected by the National Lung Screening Trial (NLST). The statements contained herein are solely those of the authors and do not represent or imply concurrence or endorsement by NCI. This work was funded in part by National Institutes of Health (NIH) Grant No. R01EB032716. The authors also thank Dr David Yankelevitz, Dr Yeqin Zhu, Dr Natela Paksashvili, Dr Lyu Lyu, Dr Lijing Zhang and others from the Mount Sinai Hospital Radiology Department for their help in conducting the Visual Turing Tests.

## ORCID iDs

Arjun Krishna https://orcid.org/0009-0004-4179-7068
Ge Wang https://orcid.org/0000-0002-2656-7705

## References

Al Khalil Y, Amirrajab S, Lorenz C, Weese J, Pluim J and Breeuwer M 2023 On the usability of synthetic data for improving the robustness of deep learning-based segmentation of cardiac magnetic resonance images *Med. Image Anal.* **84** 102688

Amirrajab S *et al* 2022 Label-informed cardiac magnetic resonance image synthesis through conditional generative adversarial networks *Comput. Med. Imaging Graph.* **101** 102123

Arslan F, Kabas B, Dalmaz O, Ozbey M and Çukur T 2024 Self-consistent recursive diffusion bridge for medical image translation (arXiv:2405.06789)

Atli O F, Kabas B, Arslan F, Yurt M, Dalmaz O and Çukur T 2024 I2I-Mamba: multi-modal medical image synthesis via selective state space modeling (arXiv:2405.14022)

Choi J, Kim S, Jeong Y, Gwon Y and Yoon S 2021 ILVR: conditioning method for denoising diffusion probabilistic models *Int. Conf. on Computer Vision*

Chung H et al 2023 Solving 3D inverse problems using pre-trained 2D diffusion models *Computer Vision and Pattern Recognition*
Dhariwal P and Nichol A 2021 Diffusion models beat GANs on image synthesis *NeurIPS*
Gao C et al 2022 Synthex: scaling up learning-based x-ray image analysis through in silico experiments (arXiv:2206.06127)
Ghorbani A, Natarajan V, Coz D and Liu Y 2020 DermGAN: synthetic generation of clinical skin images with pathology *Proc. Machine Learning for Health NeurIPS Workshop* vol 116 (PMLR) pp 155–70
Goodfellow I et al 2014 Generative adversarial nets *Advances in Neural Information Processing Systems (NeurIPS)*
Han K et al 2023 MedGen3D: a deep generative framework for paired 3D image and mask generation *Medical Image Computing and Computer Assisted Intervention Society*
Ho J, Jain A and Abbeel P 2020 Denoising diffusion probabilistic models *Neural Information Processing Systems*
Ho J and Salimans T 2021 Classifier-free diffusion guidance *Deep Generative Models and Downstream Applications Workshop at NeurIPS 2021* (available at: https://openreview.net/forum?id=qw8AKxfYbI)
Isola P, Zhu J Y, Zhou T and Efros A A 2017 Image-to-image translation with conditional adversarial networks *Proc. IEEE Conf. on Computer Vision and Pattern Recognition, (Honolulu, HI, USA)* pp 1125–34
Jiang Y, Chen H, Loew M and Ko H 2020 COVID-19 CT image synthesis with a conditional generative adversarial network *IEEE J. Biomed. Health Inf.* **25** 441–52
Kadia D, Nguyen T V and Asari V 2022 Lesion synthesis for robust segmentation of infected lung region on small-scale data *Soc. Sci. Res. Netw.* (https://doi.org/10.2139/ssrn.4029426)
Karras T, Laine S and Aila T 2018 A style-based generator architecture for generative adversarial networks *Computer Vision and Pattern Recognition* pp 4396–405
Karras T, Laine S and Aila T 2020 Analyzing and improving the image quality of stylegan *Computer Vision and Pattern Recognition* pp 8110–9
Khader F et al 2023 Medical diffusion: denoising diffusion probabilistic models for 3D medical image generation *Sci. Rep.* **13** 7436
Khosravi B et al 2023 Creating high-fidelity synthetic pelvis radiographs using generative adversarial networks: unlocking the potential of deep learning models without patient privacy concerns *J. Arthroplasty* **38** 2037–20343
Khosravi B et al 2024 Synthetically enhanced: unveiling synthetic data's potential in medical imaging research *EBioMedicine* **104** 105174
Krishna A, Bartake K, Niu C, Wang G, Lai Y, Jia X and Mueller K 2021 Image synthesis for data augmentation in medical CT using deep reinforcement learning *16th Int. Meeting on Fully Three-Dimensional Image Reconstruction in Radiology and Nuclear Medicine (Fully3D 2021), (Leuven, Belgium, 19 July–23 July)*
Krishna A and Mueller K 2019 Medical (CT) image generation with style *15th Int. Meeting on Fully Three-Dimensional Image Reconstruction in Radiology and Nuclear Medicine* p 1107234
Krishna A, Yenneti S, Wang G and Mueller K 2023 Image factory: a method for synthesizing novel CT images with anatomical guidance *Med. Phys.* **51** 3464–79
Krishna A, Yenneti S, Wang G and Mueller K 2023 Novel lung CT image synthesis at full hounsfield range with expert guided visual turing test *Fully 3D Image Reconstruction in Radiology and Nuclear Medicine*
Lewis A, Mahmoodi E, Zhou Y, Coffee M and Sizikova E 2021 Improving tuberculosis (TB) prediction using synthetically generated computed tomography (CT) images *Proc. IEEE Int. Conf. on Computer Vision (ICCV)* (IEEE) pp 3265–73
Liu Y, Yang F and Yang Y 2023 Free-form lesion synthesis using a partial convolution generative adversarial network for enhanced deep learning liver tumor segmentation *J. Appl. Clin. Med. Phys.* **24** e13927
Mou C, Wang X, Xie L, Wu Y, Zhang J, Qi Z, Shan Y and Qie X 2023 T2I-adapter: learning adapters to dig out more controllable ability for text-to-image diffusion models *Proc. AAAI Conf. on Artificial Intelligence*
Nichol A and Dhariwal P 2021 Improved denoising diffusion probabilistic models (available at: https://openreview.net/forum?id=-NEXDKk8gZ)
Oakden-Rayner L, Dunnmon J, Carneiro G and Ré C 2020 Hidden stratification causes clinically meaningful failures in machine learning for medical imaging *Proc. ACM Conf. on Health, Inference and Learning* pp 151–9
Obermeyer Z, Powers B, Vogeli C and Mullainathan S 2019 Dissecting racial bias in an algorithm used to manage the health of populations *Science* **366** 447–53
Özbey M, Dalmaz O, Dar S U H, Bedel H A, Özturk Ş, Güngör A and Çukur T 2023 Unsupervised medical image translation with adversarial diffusion models *IEEE Trans. Med. Imaging* **42** 3524–39
Pan S et al 2023 2D medical image synthesis using transformer-based denoising diffusion probabilistic model *Phys. Med. Biol.* **68** 105004
Park H Y et al 2021 Realistic high-resolution body computed tomography image synthesis by using progressive growing generative adversarial network: visual turing test *JMIR Med. Inf.* **9** e23328
Peebles W and Xie S 2023 Scalable diffusion models with transformers *Int. Conf. on Computer Vision (ICCV)*
Popescu D, Deaconu M, Ichim L and Stamatescu G 2021 Retinal blood vessel segmentation using Pix2Pix GAN *Proc. 29th Mediterranean Conf. on Control and Automation, (Puglia, Italy)* (IEEE) pp 1173–8
Prezja F, Paloneva J, Pölönen I, Niinimäki E and Äyrämö S 2022 DeepFake knee osteoarthritis x-rays from generative adversarial neural networks deceive medical experts and offer augmentation potential to automatic classification *Sci. Rep.* **12** 18573
Qin C et al 2023 UniControl: a unified diffusion model for controllable visual generation in the wild *Proc. NeurIPS*
Rombach R, Blattmann A, Lorenz D, Esser P and Ommer B 2022 High-resolution image synthesis with latent diffusion models *Proc. CVPR*
Ronneberger O, Fischer P and Brox T 2015 U-net: convolutional networks for biomedical image segmentation *Medical Image Computing and Computer Assisted Intervention Society* pp 234–41
Rouzrokh P, Khosravi B, Mickley J P, Erickson B J, Taunton M J and Wyles C C 2023 THA-net: a deep learning solution for next-generation templating and patient-specific surgical execution *J. Arthroplasty* **39** 727–33
Salem M et al 2019 Multiple sclerosis lesion synthesis in MRI using an encoder-decoder U-NET *IEEE Access* **7** 25171–84
Shaham T R, Dekel T and Michaeli T 2019 SinGAN: learning a generative model from a single natural image *Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV)* pp 4570–80
Shin H-C et al 2018 Medical image synthesis for data augmentation and anonymization using generative adversarial networks *Int. Workshop on Simulation and Synthesis in Medical Imaging* pp 1–11
Shin Y, Qadir H A and Balasingham I 2018 Abnormal colon polyp image synthesis using conditional adversarial networks for improved detection performance *IEEE Access* **6** 56007–17
Sizikova E, Saharkhiz N, Sharma D, Lago M, Sahiner B, Delfino J G and Badano A 2023 Knowledge-based in silico models and dataset for the comparative evaluation of mammography AI for a range of breast characteristics, lesion conspicuities and doses *NeurIPS 2023 Datasets and Benchmarks Poster*

Sun J, Du Y, Li C, Wu T-H, Yang B and Mok G S 2022 Pix2Pix generative adversarial network for low dose myocardial perfusion SPECT denoising *Quantum Imaging Med. Surg.* **12** 3539

Thambawita V L *et al* 2021 Data augmentation using generative adversarial networks for creating realistic artificial colon polyp images: Validation study by endoscopists *Gastrointest. Endosc.* **93** AB190

Thambawita V *et al* 2022 SinGAN-Seg: synthetic training data generation for medical image segmentation *PLoS One* **17** e0267976

Tudosiu P-D *et al* 2024 Realistic morphology-preserving generative modelling of the brain *Nat. Mach. Intell.* **6** 811–9

Waheed A, Goyal M, Gupta D, Khanna A, Al-Turjman F and Pinheiro P R 2020 COVIDGAN: data augmentation using auxiliary classifier GAN for improved COVID-19 detection *IEEE Access* **8** 91916–23

Watson D, Chan W, Martin-Brualla R, Ho J, Tagliasacchi A and Norouzi M 2023 Novel view synthesis with diffusion models *Int. Conf. on Learning Representations (ICLR)*

Xing S, Sinha H and Hwang S J 2021 Cycle consistent embedding of 3D brains with auto-encoding generative adversarial networks *Medical Imaging With Deep Learning*

Yang J *et al* 2017 Data from Lung CT segmentation challenge 2017 (LCTSC) *Cancer Imaging Arch.* (https://doi.org/10.7937/K9/TCIA.2017.3R3FVZ08)

Ye H, Zhang J, Liu S, Han X and Yang W 2023 IP-adapter: text compatible image prompt adapter for text-to-image diffusion models (arXiv:2308.06721)

Zhang L and Agrawala M 2023 Adding conditional control to text-to-image diffusion models *Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV)* pp 3813–24

Zhang L, Rao A and Agrawala M 2023 Adding conditional control to text-to-image diffusion models *Proc. ICCV*

Zhu J-Y *et al* 2017 Unpaired image-to-image translation using cycle-consistent adversarial networks *Proc. ICCV* pp 2223–32